How can data science contribute to a greener world?

An exploration featuring machine learning and data mining for environmental facilities and energy end users

Dong Wang

Department of Chemistry
Umeå 2021
If you look on the ground in search of a sixpence, you don't look up, and so miss the moon.

– William Somerset Maugham
# Table of Contents

Abstract................................................................................................................................. ii

Sammanfattning på svenska ................................................................................................ iv

Abbreviations....................................................................................................................... vi

List of Publications ............................................................................................................. ix

Author’s Contributions ...................................................................................................... x

1. Introduction .................................................................................................................... 1
   1.1 Industry 4.0, big data, and data science ................................................................. 1
   1.2 Data science in environmental domains ................................................................. 2
   1.3 Gaps in literature and aims of study ..................................................................... 4

2. Material and methods ................................................................................................ 6
   2.1 Overview of case study subjects and data sources ............................................... 6
   2.2 Data preprocessing approaches ............................................................................ 9
   2.3 Methods for modeling the correlation between input and output .................... 11
   2.4 Methods for interpreting the models .................................................................. 13
   2.5 Methods for extracting the inherent patterns in data ....................................... 15
   2.6 Metrics for assessing model performance ......................................................... 18
   2.7 Workflows of the study ....................................................................................... 20

3. Model performances .................................................................................................... 25

4. Post-interpretation/-analysis based on the models .............................................. 28
   4.1 Identification of the most influential factors for WWTP effluent parameters 28
   4.2 Investigation into impact patterns of the most influential factors in WWTP 29
   4.3 Identification of different operational conditions and most responsible operational factors for boiler failures in WtE plant ............................... 35
   4.4 Discussion on the data-algorithm-interaction in energy anomaly detection for buildings .............................................................................................. 37

5. Conclusions and future perspectives ....................................................................... 40

Acknowledgment ............................................................................................................. 44

References ....................................................................................................................... 45
Abstract

Human society has taken many measures to address environmental issues. For example, deploying wastewater treatment plants (WWTPs) to alleviate water pollution and the shortage of usable water; using waste-to-energy (WtE) plants to recover energy from the waste and reduce its environmental impact. However, managing these facilities is taxing because the processes and operations are always complex and dynamic. These characteristics hinder the comprehensive and precise understanding of the processes through the conventional mechanistic models. On the other hand, with the development of the Fourth Industrial Revolution, large-volume and high-resolution data from automatic online monitoring have become increasingly obtainable. These data usually reflect abundant detailed information of process activities that can be utilized for optimizing process control. Similarly, data monitoring is also adopted by the resource end users. For example, energy consumption is usually recorded by commercial buildings for optimizing energy consumption behavior, eventually saving running costs and reducing carbon footprint. With the data recorded and retrieved, appropriate data science methods need to be employed to extract the desired information. Data science is a field incorporating formulating data-driven solutions, data preprocessing, analyzing data with particular algorithms, and employing results to support high-level decisions in various application scenarios.

The aim of this PhD project is to explore how data science can contribute to a more sustainable world from the perspectives of both improving the operation of environmental engineering processes and optimizing the activities of energy end users. The major work and corresponding results are as follows:

(1) **(Paper I)** An ML workflow consisting of Random Forest (RF) models, Deep Neural Network (DNN) models, Variable Importance Measure (VIM) analyses, and Partial Dependence Plot (PDP) analyses was developed and utilized to model WWTP processes and reveal how operational features impact on effluent quality. The case study was conducted on a full-scale WWTP in Sweden with large data (105,763 samples). This paper was the first ML application study investigating cause-and-effect relationships for full-scale WWTPs. Also, for the first time, time lags between process parameters were treated rigorously for accurate information uncovering. The cause-and-effect findings in this paper can contribute to more sophisticated process control that is more precise and cost-effective.

(2) **(Paper II)** An upgraded workflow was designed to enhance the WWTP cause-and-effect investigation to be more precise, reliable, and
comprehensive. Besides RF, two more typical tree-based models, XGBoost and LightGBM, were introduced. Also, two more metrics were adopted for a more comprehensive performance evaluation. A unified and more advanced interpretation method, SHapley Additive exPlanations (SHAP), was employed to aid model comparison and interpret the optimal models more profoundly. Along with the new local findings, this study delivered two significant general findings for cause-and-effect ML implementations in process industries. First, multi-perspective model comparison is vital for selecting a truly reliable model for interpretation. Second, adopting an accurate and granular interpretation method can profit both model comparison and interpretation.

(3) **(Paper III)** A novel workflow was proposed to identify the accountable operational factors for boiler failures at WtE plants. In addition to data preprocessing and domain knowledge integration, it mainly comprised feature space embedding and unsupervised clustering. Two methods, PCA + K-means and Deep Embedding Clustering (DEC), were carried out and compared. The workflow succeeded in fulfilling the objective of a case study on three datasets from a WtE plant in Sweden, and DEC outperformed PCA + K-means for all the three datasets. DEC was superior due to its unique mechanism in which the embedding module and K-means are trained simultaneously and iteratively with the bidirectional information pass.

(4) **(Paper IV)** A two-level (data structure level and algorithm mechanism level) workflow was put forward to detect imperceptible anomalies in energy consumption profiles of commercial buildings. The workflow achieved two objectives – it precisely detected the contextual energy anomalies hidden behind the time variation in the case study; it investigated the combined influence of data structures and algorithm mechanisms on unsupervised anomaly detection for building energy consumption. The overall conclusion was that the contextualization resulted in a less skewed estimation of correlations between instances, and the algorithms with more local perspectives benefited more from the contextualization.
Sammanfattning på svenska

Dagens samhälle har vidtagit många åtgärder för att lösa miljöproblem. Exempel är reningsverk för avloppsvatten för att minska effekter av vattenföroreningar och öka mängden av användbart vatten och använda avfall-till-energi (WtE) anläggningar för att återvinna energi från avfallet och minska dess miljöpåverkan. Det kan dock vara krävande att hantera dessa anläggningar eftersom processerna oftast är komplexa och dynamiska. Komplexiteten i processerna kan hindra en heltäckande och exakta förståelsen av processerna genom konventionella mekanistiska modellerna. Å andra sidan, med utvecklingen av s k Industri 4.0 har stora volym och högupplösta data från automatisk onlineövervakning blivit alltmer tillgängliga. Dessa data återspeglar vanligtvis detaljerad information om processaktiviteter som kan användas för att optimera processkontroll. På samma sätt antas dataövervakning också ha ett stort värde för slutanvändare. Till exempel registreras energiförbrukningen av kommersiella byggnader vanligtvis för att optimera energiförbrukning, för att i förlängningen spara driftskostnader och minska koldioxidavtryck. För att uppnå maximal kunskap från inhämtade data måste dessa bearbetas med lämpliga datavetenskapliga metoder för att extrahera ut den önskade informationen. Datavetenskap är ett område som innefattar formulering av datadrivna lösningar, dataförbehandling, analys av data med speciella algoritmer och användning av resultat för att stödja beslut på i olika tillämpningsscenarier.

Syftet med detta doktorandprojekt har varit att utforska hur datavetenskap kan bidra till en mer hållbar värld utifrån perspektiven att både förbättra driften av miljötekniska processer och optimera energianvändning i byggnader. Sammanfattning av resultaten är följande:


Paper II - Ett uppgraderat arbetsflöde utformades för att förbättra undersökningen av reningsverkets orsak-verkansamband. Förutom RF


Paper IV - Ett arbetsflöde på två nivåer (datastrukturvipp och algoritmmekanismnivå) presenterades för att upptäcka anomaler i energiförbrukningsprofiler för kommersiella byggnader. Arbetsflödet uppnådde två mål, dels identifierade de kontextuella energianomalier som gömdes bakom tidsvariationen i fallstudien och dels undersöktes den kombinerade inverkan av datastrukturer och algoritmmekanismer på oövervakad anomaldetektering för byggnaders energiförbrukning. Den övergripande slutsatsen var att kontextualiseringen resulterade i en mindre skev uppskattning av korrelationer mellan instanser, och algoritmerna med mer lokala perspektiv gynnades mer av kontextualiseringen.
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>AUC_PR</td>
<td>Area Under Precision-Recall Curve</td>
</tr>
<tr>
<td>BOD</td>
<td>Biological Oxygen Demand</td>
</tr>
<tr>
<td>CART</td>
<td>Classification and Regression Tree</td>
</tr>
<tr>
<td>CBLOF</td>
<td>Cluster-Based Local Outlier Factor</td>
</tr>
<tr>
<td>COD</td>
<td>Chemical Oxygen Demand</td>
</tr>
<tr>
<td>COF</td>
<td>Connectivity-Based Outlier Factor</td>
</tr>
<tr>
<td>DEC</td>
<td>Deep Embedded Clustering</td>
</tr>
<tr>
<td>DM</td>
<td>Data Mining</td>
</tr>
<tr>
<td>DNN</td>
<td>Deep Neural Network</td>
</tr>
<tr>
<td>DO</td>
<td>Dissolved Oxygen</td>
</tr>
<tr>
<td>DWT</td>
<td>Discrete Wavelet Transform</td>
</tr>
<tr>
<td>EFB</td>
<td>Exclusive Feature Bundling</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>GBDT</td>
<td>Gradient boosting decision tree</td>
</tr>
<tr>
<td>GOSS</td>
<td>Gradient-based One-Side Sampling</td>
</tr>
<tr>
<td>HF</td>
<td>High-pass Filter</td>
</tr>
<tr>
<td>ID</td>
<td>Induced Draft</td>
</tr>
<tr>
<td>IF</td>
<td>Isolation Forest</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>KL</td>
<td>Kullback–Leibler</td>
</tr>
<tr>
<td>LF</td>
<td>Low-pass Filter</td>
</tr>
<tr>
<td>LOF</td>
<td>Local Outlier Factor</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>MDI</td>
<td>Mean Decrease Impurity</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>OOB</td>
<td>Out-of-Bag</td>
</tr>
<tr>
<td>PAO</td>
<td>Polyphosphate-accumulating organism</td>
</tr>
<tr>
<td>PC</td>
<td>Principal Component</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PDP</td>
<td>Partial Dependence Plot</td>
</tr>
<tr>
<td>PI</td>
<td>Permutation Importance</td>
</tr>
<tr>
<td>PO4</td>
<td>Phosphate</td>
</tr>
<tr>
<td>PR</td>
<td>Precision-Recall</td>
</tr>
<tr>
<td>R²</td>
<td>Coefficient of Determination</td>
</tr>
<tr>
<td>RF</td>
<td>Random Forest</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Squared Error</td>
</tr>
<tr>
<td>SAE</td>
<td>Stacked Autoencoder</td>
</tr>
<tr>
<td>SHAP</td>
<td>SHapley Additive exPlanations</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>TSS</td>
<td>Total Suspended Solids</td>
</tr>
<tr>
<td>VIM</td>
<td>Variable Importance Measures</td>
</tr>
<tr>
<td>WtE</td>
<td>Waste-to-Energy</td>
</tr>
<tr>
<td>WWTP</td>
<td>Wastewater Treatment Plant</td>
</tr>
</tbody>
</table>
List of Publications

All the following papers are provided as appendices.


III. Dong Wang, Lili Jiang, Måns Kjellander, Eva Weidemann, Johan Trygg, Mats Tysklind*. Investigation into the causes of boiler failures in a waste-to-energy plant with a coupled engineering and data mining solution. *Manuscript*.

IV. Dong Wang, Therese Enlund, Amanda Fors, Johan Trygg, Mats Tysklind, Lili Jiang*. Towards delicate anomaly detection of energy consumption for buildings: enhance the performance from two levels. *Submitted* for publication.
Author’s Contributions

Paper I

The author was involved in formulating the research problem, processing data, building models, interpreting models, visualizing results, writing and revising the paper.

Paper II

The author was involved in formulating the research problem, processing data, building models, interpreting models, visualizing results, writing and revising the paper.

Paper III

The author was involved in formulating the research problem, processing data, building models, visualizing and interpreting results, writing and revising the paper.

Paper IV

The author was involved in formulating the research problem, processing data, building models, visualizing and interpreting results, writing and revising the paper.
1. Introduction

1.1 Industry 4.0, big data, and data science
We are currently living in the era of the Fourth Industrial Revolution (Industry 4.0). Leveraging the technologies like cyber-physical systems (CPS), Internet of Things (IoT), cloud computing, and machine-to-machine communication (M2M), Industry 4.0 was designed to improve the interconnection and communication between humans, machines, and sensors; to enable the systems to function and fulfill tasks autonomously with minimal human intervention; to offer people real-time, precise, and comprehensive information for operational optimization and decision-making (Hermann et al., 2016; Lasi et al., 2014; Lu, 2017; Xu et al., 2018). Having its roots in the manufacturing industry, Industry 4.0 has evolved to cover the digitization and smartization for numerous industries, and not only for the production in factories but also the whole value chain (Liao et al., 2017; Schroeder et al., 2019; Xu et al., 2018). Even though the degree of involvement of Industry 4.0 can vary across industries and enterprises, monitoring the behaviors of functional entities and collecting the corresponding data through the metering system, as an entry-level but fundamental practice, have been adopted ubiquitously across various industries (Chen, 2020; Hopmann and Schmitz, 2021; Kemal et al., 2020; Li and Pandey, 2017).

The data generated in the wave of Industry 4.0 are usually large and complex, and they are referred to as ‘big data’. Big data are characterized by 5V: high Volume, high Velocity, high Variety, low Veracity, and high Value. Volume refers to the amounts of data; Velocity refers to the speed of data generation and accumulation; Variety refers to the diverse sources of data and the various types of data (structured, semi-structured, or unstructured); Veracity refers to the quality of data (can be messy, inconsistent, or incomplete); Value refers to the benefits data can provide (Iqbal et al., 2020; Jin et al., 2015; Sen et al., 2016). Due to the nature of big data, it is always demanding to extract information or knowledge from them. For example, the calculation can be expensive, data are usually not in the form to be directly analyzed, not all the data records or attributes are informative, or the information desired for the task is too complicated to extract. To tackle all these challenges to meet requirements, optimize decision-making, or create new value opportunities, we need the employment of data science.

Data science is a field incorporating formulating data-driven solutions, data preprocessing, analyzing data with particular algorithms, and employing results to support high-level decisions in various application scenarios. Besides the knowledge from the application domains, data science requires expertise from
data integration, computer science, mathematics, information visualization, communication, etc. (Martínez et al., 2021; Molina-Solana et al., 2017; Van Der Aalst, 2016).

There are countless tasks from industries, but the types of data-driven solutions are limited. Hence, in the stage of formulating data-driven solutions, the data and the particular task will be examined to judge whether the task can be fulfilled by data-driven solutions. If so, the type of data-driven solution suitable for the particular task will be decided, which sets the base for the whole analysis. In the stage of data preprocessing, the missing data will be interpolated, the non-informative records and attributes will be removed, and the form/structure of the data will be reorganized to adapt to the specific data-driven solution set in the previous stage. After the data are preprocessed, the algorithms corresponding to the solution will be employed to either model the correlation between input and output data or discover the undefined patterns within the data. The ones for correlation modeling are usually supervised Machine Learning (ML) algorithms in which the correlation learning is supervised by the labels/output data. On the contrary, the ones for intrinsic pattern discovery are usually unsupervised ML/Data Mining (DM) algorithms, and most of them do not have the supervision mechanism because there are no labels/output data. For the rest, supervision exists, but the learning is supervised by the input data itself. Depending on the specific demands in different scenarios, some modeling results can be used directly to serve the purpose of the task, whereas in other cases the post-analysis of the modeling is needed, for example, interpreting the models/modeling results in the particular domain contexts. Along the whole pipeline of data science application, the domain knowledge and mathematics/computing techniques are always intertwined and therefore influence each other. Thus, even though ‘Data science’ as a broad high-level term conveys a few concepts shared by various application scenarios, only within a specific application context the profile of data science can be identified.

1.2 Data science in environmental domains
Throughout the history of human society’s civilization and development, the activities of human beings have caused and are still causing enormous adverse impacts on the natural environment in various aspects. Examples include air, soil, and water pollution, ozone layer destruction, acidic deposition, global warming, biodiversity loss, waste disposal, and deforestation (Owen et al., 2006). In turn, environmental degradation affects the wellness of human society in terms of both health and living conditions. For example, air pollution can cause ischaemic heart disease and stroke; water pollution can cause cholera, dysentery, and organ damage; ozone layer depletion can cause skin cancer; global warming
can cause more weather-related disasters and rising sea level (Hasan et al., 2019; Liu et al., 2021; Pandey et al., 2021; Parker, 2021).

Realizing the consequences of environmental degradation, human society has taken many measures to prevent the situation from worsening and keep the development sustainable. Those measures include both hard and soft measures. The hard measures are usually the deployment of environmental engineering facilities. For instance, wastewater treatment plants (WWTPs) are deployed to alleviate both the pollution in water and the shortage of usable water; waste-to-energy (WtE) plants are used to both recover energy from the waste and reduce the environmental impact of the waste; wind farms, hydropower stations, and photovoltaic power stations are adopted to reduce the dependence on fossil fuel and reduce the emissions of greenhouse gases and other pollutants (Maka et al., 2021; Munir et al., 2021; Perveen et al., 2014; Salgot and Folch, 2018). The examples of soft measures are monitoring energy consumption and optimizing the corresponding activities, use of life-cycle assessment (LCA) to guide the design of environmentally friendly products, and adoption of environmental impact assessment and strategic environmental assessment to ensure the planned projects and policies consider environmental impact (Ilbeigi et al., 2020; Morgan, 2012; Okada et al., 2021; Therivel et al., 2013).

In the background of Industry 4.0, data science plays a significant role in both hard and soft measures of tackling environmental degradation. For the applications to the hard measures, the contribution of data science to solving environmental issues can be direct or indirect, from promoting the control of end products or emissions of the environmental facilities to optimizing the facilities’ function for a more efficient and cost-effective operation. Some examples are as follows. An intelligent decision-making system, comprised of multiple data science methods and several sensors, was developed to decrease the occurrence of membrane fouling in WWTPs, as membrane fouling can lead to effluent quality degradation, waste of energy, and collapse of the process (Han et al., 2020). A data-driven approach consisting of a set of techniques was implemented to a WWTP to both optimize the coagulant dosage and enhance Cu removal for industrial wastewater (Wang et al., 2021). In order to optimize the combustion process control of WtE plants, a set of ML methods was employed to fulfill a quick and accurate estimation of the high heating value of municipal solid waste (Bagheri et al., 2019). In order to improve the control of carbon monoxide (CO) cost-effectively in a clinical waste incineration plant, a method integrating both neural networks and fuzzy logic principles was used to predict the CO emission (Norhayati and Rashid, 2018).

In the applications to soft measures, data science is usually closely associated with the ultimate goal – energy reduction or environmental performance
improvement. For example, a clustering-based approach and a genetic algorithm were applied to optimizing the energy consumption in buildings (Sonta et al., 2021). In a study of improving energy efficiency for energy-intensive manufacturing industries, cube-based models were used for data preprocessing and long short-term memory models were used for forecasting energy consumption (Ma et al., 2020). A big-data-oriented data science framework was proposed to deal with the challenges associated with big data and grasp the opportunities big data offer in the scenario of environmental performance evaluation (Song et al., 2017). Big data analytics and artificial intelligence techniques were employed to help supply chain managers to develop sustainable supply processes and improve environmental performance (Benzidia et al., 2021).

1.3 Gaps in literature and aims of study

Despite the various data science applications to environmental domains mentioned above, there is still plenty of room for improvements, or there are simply gaps in particular areas. Two examples in hard measure scenarios are WWTPs and WtE plants. The vast majority of published studies on data science applications to WWTPs solely focus on prediction using ML models without interpreting the models to obtain knowledge about how the studied targets are influenced. Also, the models usually treat time lags between process steps in WWTP processes with insufficient rigor or neglect it entirely (Cao and Yang, 2020; Guo et al., 2015; Shi and Xu, 2018; Verma et al., 2013). This neglect could lead to incorrect interpretation and analysis of modeling results, which in turn could cause problems when attempting to control processes based on cause-and-effect relationships identified through such analyses and interpretations. Besides, there have been no systematic data science applications examining boiler failures in WtE plants or ordinary thermal power stations. However, given the great complexity and variability of the mechanisms of boiler failure (Ding et al., 2017; Haghhighat-Shishavan et al., 2019; Jones, 2004; Paz et al., 2017), a data-driven solution is needed for a more effective failure investigation. The data-driven solution can help identify the linkage between a failure and the specific operational parameters and guide the operation without digging deeply into the details of the failure mechanism.

In soft measure scenarios, one example is that there have been no studies investigating the combined influence of data structure and algorithms’ mechanisms on the performance of unsupervised anomaly detection for building energy data (Fan et al., 2018; Mariano-Hernández et al., 2021). Data and algorithms are the two key factors, so their individual influences and interactive influences on the detection performance should be researched for establishing a better understanding of what data form should be prepared and what algorithm
should be adopted for a more accurate and robust anomaly detection result in practice.

The aim of this thesis is to bridge the gaps mentioned above and achieve improvement in process control for WWTPs, boiler failure investigation for WtE plants, and detection of anomalous energy consumption for buildings. Figure 1 shows the structure of this thesis and illustrates how the different equipped papers are connected in terms of the overall objective, and how they are differentiated as far as the application scenarios are concerned. In Paper I, a workflow based on multiple ML algorithms and model interpretation methods was developed and used to model WWTP processes and investigate how operational factors influence effluent quality. The time lag along the process line was rigorously considered and estimated. In Paper II, we designed an upgraded workflow to improve the reliability and functionality of its counterpart in Paper I. More ML algorithms were explored and a more advanced interpretation method was adopted to expect a more informative and comprehensive cause-and-effect investigation for WWTPs. In Paper III, a novel workflow, in which engineering knowledge and multiple data science methods were coupled, was put forward as a means of investigating the causes of boiler failures in WtE plants. In Paper IV, we proposed a workflow to evaluate the difference in anomaly detection effect between the original data (with only behavioral attribute) and contextualized data (with both behavioral and contextual attributes), and between the algorithms from global, local, and global-local-hybrid perspectives.

![Diagram of thesis structure and paper characterization](image)

Figure 1. Diagram of thesis structure and paper characterization
2. Material and methods

2.1 Overview of case study subjects and data sources
As is shown in Figure 2, the case study subjects of these two papers were the Umeå WWTP, the largest WWTP in the Umeå municipality, Sweden. Umeå WWTP consists of three phases: primary treatment, secondary treatment, and tertiary treatment. It was studied at the full scale but without the involvement of the processing system of waste activated sludge. Besides a few manual samplers, there were seven kinds of online probes through the whole treatment process – flow rate (FT), total suspended solids (TSS), pH, phosphate (PO4), temperature, dissolved oxygen (DO), and total solids (TS) probes. The detailed properties and names of the online probes are shown in Table S1 of Paper I and II. In both papers, only the data from online probes were used. The online probes monitored the water parameters every millisecond, which guaranteed the large volume and high resolution of data essential for accurate and dependable analysis results. However, the data were averaged to enlarge the interval between two adjacent instances from 1 millisecond to 10 minutes, decreasing the number of instances from 63,516,600,000 to 105,861. The averaging was adopted because wastewater treatment is usually a slow process without prominent rapid changes. Also, it alleviated the adverse influence of noise on the probe signals. Two effluent parameters, TSSe and PO4e, were used as the labels for the ML modeling, whereas the other parameters monitored by the online probes (the ones in Paper I and II (Table S1) excluding TSSe and PO4e) were used as the features. The difference in data between Paper I and II was that the data used for PO4e modeling were further refined in Paper II. Many uninformative instances were removed to ensure a more accurate information extraction regarding the relationships between PO4e and the operational factors. These uninformative instances had the fixed PO4 concentration, 0.0500000007450581 mg/L, which was the lower limit of the PO4 probe's measurement. This collection of fixed values cannot reflect the actual dynamics of the process, and it can infuse noise into the data, eventually impairing the precision of models and the subsequent interpretation.
Figure 2. Schematic diagram of process units and monitoring probes in the Umeå WWTP
In the case study of **Paper III**, the subject was a WtE plant named Dåva 1, which belongs to Umeå Energi AB, Sweden. As a 65 MW Combined Heat and Power (CHP) plant, Dåva 1 has the capacity of incinerating approximately 20 t/h of waste and takes around half municipal solid waste and half industrial waste as fuel. The operation time for the plant is about 8,000 hours per year, with one stoppage scheduled for maintenance. The boiler-related layout of Dåva 1 is shown in Figure 3. A boiler is a heat exchanger for increasing the temperature of water/steam while decreasing the temperature of flue gas. There were over 1,000 online sensors monitoring the whole process line, resulting in more than 1,000 features in the original data. Nevertheless, only a small portion of them was closely related to the boiler system. Hence, in collaboration with the engineers from Dåva 1, features unrelated to the boiler operational conditions were eliminated, and only 66 features remained (as shown in Table S1 of **Paper III**). Five unplanned stoppages (caused by boiler failures) were examined in the study. Stoppages happening in quick succession were analyzed together in the same dataset, so eventually, only three datasets were obtained and studied (as shown in Table 1 of **Paper III**). In order to guarantee enough data for analyzing the differences between the normal and abnormal operational conditions regarding the stoppage causes, instances of three to five months (subject to the availability of data) were included before the first stoppage. Even though the measurement frequencies of the probes were extremely high, the datasets were retrieved at a resolution of 30 minutes by averaging. In addition to the purpose of reducing noise in the data, the averaging helped weaken the time-lag impact exerted by the flow of water, steam, and flue gas.

![Figure 3. Boiler-related layout of the waste-to-energy plant Dåva 1](image)
In **Paper IV**, three datasets of commercial buildings’ energy consumption were studied. The datasets were provided by an energy management company, Mestro AB, in Sweden. They were the electricity consumption records of Mestro’s clients with the monitoring frequency of one record per hour. Three datasets represented two types of commercial buildings in three different cities – Karlskoga, Göteborg, and Jönköping. We selected the data in 2018 for all the three buildings as the datasets studied in **Paper IV**. The original datasets only contained one feature, which was the behavioral feature ‘energy consumption.’ Dataset A was retrieved from the main electricity meter serving a property in Karlskoga of 1622 m² heated area. The property is used for retail business. Dataset B was retrieved from the main electricity meter serving a property in Göteborg of 47,166 m² heated area. The property is a university building, with most of the rooms being offices. Dataset C was retrieved from the main electricity meter serving a property in Jönköping of 28,046 m² heated area. The function of this property is retail. Initially, there were no labels in the datasets indicating normal or abnormal instances. However, labels were needed for evaluating anomaly detection performances. Thus, manual labeling was conducted by leveraging both empirical domain knowledge and statistical methods.

### 2.2 Data preprocessing approaches

Data preprocessing is a crucial stage in data science implementations, and it usually takes considerable amounts of time and effort. This is because the data from the real world are usually far from perfect or in a form not suitable for DM or ML algorithms (García et al., 2015; García et al., 2016). For example, in this study, the original data were in pieces (from multiple files), with missing values, bad data cells, noise, and unnecessary information. Leaving these problems untreated will hinder yielding accurate results from algorithm employment and subsequent analysis. Sometimes it even can prevent the DM and ML algorithms from functioning. In this study, some preprocessing tasks were basic, such as dropping irrelevant data or information, interpolating for missing or bad data, and standardizing data. To fulfill these tasks, we used both domain knowledge and basic programming techniques from Python modules Numpy (Oliphant, 2006) and Pandas (McKinney, 2010).

However, other preprocessing tasks were more complicated and challenging, requiring more profound domain knowledge and advanced data science techniques. One example is the time lag calculation and data transformation for WWTP data in **Paper I** and **Paper II**. Wastewater treatment processes include various flows of both water and sludge, and it takes time (usually significant) for them to travel through the process line. Thus, there are time lags between the times when the water/sludge reaches different probes. In order to ensure the ML models explainable from the perspective of WWTP processes, the original time
series data must be transformed into batch series data. The transformation was done by calculating the time lags using the integration approach and aligning the time windows in a new structure with the time lags being considered. In Paper III, since the DM results turned out to be very sensitive to noise, an extra denoising technique was adopted in addition to averaging data points that was adopted in Paper I and Paper II. This technique is discrete wavelet transformation (DWT), a tool that can effectively denoise signal data (Burrus et al., 1998). Generally, there are three steps for DWT-based denoising. First, work out the DWT coefficients from the decomposition expansion of the signal with noise. Second, retain the coefficients corresponding to the signal, but substitute the ones associated with noise by zeros. Third, rebuild the signal from the adjusted coefficients. In Paper IV, the contextualized data contained many binary features after one-hot encoding (also a preprocessing technique) (Yu et al., 2020), but these features were sparse and not information-intensive. Moreover, the high dimensionality of the data could hamper the good performance of the anomaly detection algorithms. Hence, the data with those binary features needed to be compacted. Stacked autoencoder (SAE) was utilized to fulfill this task since it can keep much more informative variance compared to the conventional linear reduction technique, for example, principal component analysis (PCA) (Wold et al., 1987). SAE is an unsupervised technique utilizing an artificial neural network to map the data to a latent space with lower dimensionality representing the original data. The representation is learned through the iteration of rebuilding the original data from the representation, and it is validated by the difference between the rebuilt data and the original data (Hinton and Salakhutdinov, 2006; Le, 2013).

The approaches used for data preprocessing in Paper I-IV are listed below in Table 1.
Table 1. Summary of data preprocessing approaches adopted in **Paper I-IV**

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Function</th>
<th>Adopted for</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merging, dropping, interpolating, standardizing, etc.</td>
<td>Data cleansing</td>
<td>Paper I-IV</td>
</tr>
<tr>
<td>Averaging</td>
<td>Denoising</td>
<td>Paper I-III</td>
</tr>
<tr>
<td>Transforming time series data to batch series data</td>
<td>Ensuring ML models explainable from the perspective of WWTP process</td>
<td>Paper I and II</td>
</tr>
<tr>
<td>DWT</td>
<td>Denoising</td>
<td>Paper III</td>
</tr>
<tr>
<td>Adding contextual features</td>
<td>Constructing the context surrounding the instances</td>
<td>Paper IV</td>
</tr>
<tr>
<td>One-hot encoding</td>
<td>Transforming each categorical feature into multiple binary features</td>
<td>Paper IV</td>
</tr>
<tr>
<td>SAE</td>
<td>Mapping data to a low-dimensional latent space</td>
<td>Paper IV</td>
</tr>
</tbody>
</table>

### 2.3 Methods for modeling the correlation between input and output

The goals in **Paper I** and **II** were obtaining the cause-and-effect information between operational factors and effluent quality parameters, so **Paper I** and **II** were the cases where supervised learning should be employed. Supervised learning is a subcategory of ML in which models are trained using labeled data (with ground truth). Labeled data consist of features (input) and labels (output), and they are in pairs for every data point (instance). Supervised learning aims to learn the mapping function from the input to the output so that the learned model is capable of predicting the output for the new data fed into it. The supervised learning model uses the feedback from the output of training data to check whether it maps the input-output relation accurately. This is similar to a student learning things under the supervision of a teacher. Supervised learning problems can be generally grouped into two types – regression and classification problems. In **Paper I** and **II**, we tackled regression problems.

Two supervised learning models, Random Forest (RF) and Deep Neural Networks (DNN), were used in **Paper I**. RF is an interpretable model, whereas
DNN is not interpretable (or not suitable to be interpreted) (Guidotti et al., 2018; Rudin, 2019). DNN was used to validate whether RF caught sufficient variance for the subsequent RF model interpretation due to DNN’s superb ability to learn the complex relationship between input and output to yield precise predictions (Oliveira et al., 2019; Ozoegwu, 2019; Parisi et al., 2019; Shabanpour et al., 2017). RF was selected as the representative of tree-based models capable of modeling complex information and readily interpretable (Breiman, 2001; Chen and Guestrin, 2016; Ke et al., 2017). RF was chosen because of its several advantages. For example, only a few hyperparameters needed to be tuned in the training procedure; RF’s performance is relatively stable regardless of the change of hyperparameters; RF is not prone to overfitting. (Breiman, 2001; Breiman, 2002; Fawagreh et al., 2014).

The selection of RF as the representative of tree-based models in Paper I has been justified, and RF has been proved to be appropriate for cause-and-effect interpretation. However, exploring other tree-based models were still perceived to be needful for a more unbiased and informative interpretation. Thus, two more typical tree-based models, XGBoost (Chen and Guestrin, 2016) and LighGBM (Ke et al., 2017), were used and evaluated with RF in Paper II for comparison. XGBoost and LightGBM are two efficient engineering implementations of Gradient boosting decision tree (GBDT) (Friedman, 2001). For XGBoost, one of the major improvements over GDBT is utilizing Newton’s method instead of gradient descent. For LightGBM, the main improvement over GDBT is the reduction of computing time in both row-wise and column-wise dimensions.

The methods used for modeling input-output correlations in Paper I and II are listed below in Table 2.
Table 2. Summary of methods modeling input-output correlations adopted in Paper I and II

<table>
<thead>
<tr>
<th>Methods</th>
<th>Function</th>
<th>Adopted for</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>Modeling complex correlations between operational factors and effluent quality (Interpretable)</td>
<td>Paper I and II</td>
</tr>
<tr>
<td>DNN</td>
<td>Modeling complex correlations between operational factors and effluent quality (Non-interpretable); Validate RF models’ performances</td>
<td>Paper I</td>
</tr>
<tr>
<td>XGBoost</td>
<td>Modeling complex correlations between operational factors and effluent quality (Interpretable)</td>
<td>Paper II</td>
</tr>
<tr>
<td>LighGBM</td>
<td>Modeling complex correlations between operational factors and effluent quality (Interpretable)</td>
<td>Paper II</td>
</tr>
</tbody>
</table>

2.4 Methods for interpreting the models

The ultimate goals of Paper I and Paper II were to understand the correlations between operational factors and effluent quality parameters, i.e., the relationship between model input and output. Thus, there was a need to interpret the trained and tested models to extract the relationship they learned from the training data, rather than merely using the trained and tested models for predicting the output for the future new input data. This is also one of the major highlights of Paper I and II that distinguish them from the other ML application studies on WWTPs.

In Paper I, the interpretation approach was Variable Importance Measure (VIM) + Partial Dependence Plots (PDPs). VIM was used to identify the most influential operational factors, and PDP was used to look into how the influential factors influenced the effluent quality. The two most typical methods of Variable Importance Measure (VIM) are the Permutation Importance (PI) and Mean Decrease Impurity (MDI) measures (Breiman, 2001; Breiman, 2002). We selected the MDI measure for identifying the RF models’ most important features. The adoption was justified by the reasons as follows: MDI is more robust in performance and less computationally expensive compared to PI (Calle and Urrea, 2011; Li et al., 2019); the data values for all the features were continuous values, which means using MDI will not lead to the cardinality bias problem reported in the literature (Boulesteix et al., 2012; Strobl et al., 2007). PDP shows the dependence of the target output labels on the input features of interest (the most influential ones identified by VIM). PDP marginalizes all the values of all other input features with respect to the current feature of interest. Thus, it can be
explained as the expected target output as a function of the input feature of interest. With this function, the correlation between the feature of interest and the predicted label can be intuitively illustrated and extracted (Friedman, 2001).

In Paper II, the interpretation approach was SHapley Additive exPlanations (SHAP). SHAP belongs to the group additive feature attribution methods, and it was developed from the combination of game theory and local explanations (Lundberg and Lee, 2017b). As of the time of writing this thesis, SHAP is the only method that possesses the properties of local accuracy, missingness, and consistency (Lundberg et al., 2018). SHAP maps the original feature space to a simplified space of binary features. Based on the simplified space, the original model is approximated by a simple linear function of the binary features -- as shown in Paper II (Equation (3)). The feature attribution value in this equation is defined in Paper II (Equation (4)), and the feature attribution value is named SHAP value. Calculating $E[f(x)|x_o]$ shown in Paper II (Equation (5)) is essential for obtaining SHAP values. However, computing $E[f(x)|x_o]$ accurately is rather demanding and expensive. Thus, multiple approximation approaches have been put forward, including the one adopted in Paper II, Tree SHAP. Tree SHAP is characterized by utilizing the structure of decision tree to estimate the $E[f(x)|x_o]$ value in an efficient fashion (Lundberg and Lee, 2017a). In comparison with the conventional VIM + PDP, SHAP is a more granular technique that can produce feature attributions for each instance. This feature is advantageous for a more thorough and precise understanding of the models’ behavior. SHAP is capable of uncovering more details about model fitting and revealing the combined impact of different operational factors. This is an additional function to discovering influential operational factors and investigating causality between them and effluent quality -- which the VIM + PDP approach presents in Paper I. To the best of the authors’ knowledge, Paper II was the first study applying SHAP analysis to WWTP research as of the time when the paper was written.

The methods used for interpreting models in Paper I and II are listed below in Table 3.
Table 3. Summary of model interpretation methods adopted in **Paper I** and **II**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Function</th>
<th>Adopted for</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIM + PDP</td>
<td>Interpreting RF models; Identifying operational factors influencing effluent quality and their independent and overall effects</td>
<td><strong>Paper I</strong></td>
</tr>
<tr>
<td></td>
<td>Interpreting RF, XGBoost, and LightGBM models; Serving as an auxiliary evaluation on model performance; Identifying operational factors influencing effluent quality and their detailed, independent, and joint effects</td>
<td><strong>Paper II</strong></td>
</tr>
</tbody>
</table>

### 2.5 Methods for extracting the inherent patterns in data

In contrast to the employment of supervised learning in **Paper I** and **II**, **Paper III** and **IV** adopted unsupervised learning, which is another subcategory of ML and is commonly used for DM tasks. Unlike modeling input-output relationships in supervised learning, unsupervised learning works with unlabeled data (no output). The goal of unsupervised learning is to model the patterns resulting from the intrinsic nature of the data, for example, the underlying structure of data or distribution of instances. The patterns usually can convey the desired imperceptible information of the data for a particular task. As mentioned in Section 2.3, supervised learning is like a student learns things under the supervision of a teacher. Suppose all the unsupervised algorithms are the students in the class of unsupervised learning. In that case, the vast majority of them do not get feedback or guidance while learning things, and the others get feedback or guidance from themselves instead of teachers – because there is no teacher in this class.

In **Paper III**, two unsupervised approaches were used to investigate the operational-factor-related boiler failure causes - Principal Component Analysis (PCA) + K-means and Deep Embedded Clustering (DEC). The case-based unsupervised clustering was the only approach suitable for this paper’s objective for the following two reasons. The monitored data in **Paper III** were not labeled, and the criteria for abnormal operation conditions could vary among different boiler components and at different times. PCA is a commonly used technique for dimensionality reduction. It maps the original data onto a lower-dimensional space formed by the first few principal components (PCs) while retaining as much data variance as possible. K-means is a clustering algorithm, and its mechanism is intuitive; all the instances in a particular dataset are divided into $k$ clusters according to their distances to each other. This is achieved through the process of
minimizing the distances among instances in every cluster, and meanwhile maximizing the distances among all the clusters (Jain, 2010). DEC is an algorithm learning feature embedding and instance clustering at the same time by leveraging deep neural networks (DNN) and K-means. DEC does not perform clustering on the original data $X$, instead, DEC first projects $X$ nonlinearly onto a low-dimensional latent space $Z$ through the encoder part (DNN structure) of SAE that is characterized by the parameters $\theta$. Subsequently, K-means is carried out on $Z$. In the reverse direction, both the encoder parameters and the $k$ centroids of K-means are trained simultaneously through the backpropagation. DEC comprises two major procedures: 1) Set up the encoder of SAE and K-means with the initial parameters $\theta$ and $k$ cluster centroids, respectively; 2) Iteratively optimize $\theta$ and cluster centroids by minimizing the Kullback–Leibler (KL) divergence between the soft assignment and the generated auxiliary target distribution (Xie et al., 2016). PCA + K-means was designed as the baseline for assessing the performance of DEC. For PCA + K-means, K-means is performed based on the result of PCA but performing PCA does not require any input from K-means’s results, i.e., the information is only passed from PCA to K-means. For DEC, the encoder of SAE and K-means are combined and trained simultaneously and iteratively. The information pass in DEC is bidirectional: from encoder to K-means to KL divergence, and also from KL divergence to encoder and K-means.

In Paper IV, four unsupervised anomaly detection algorithms were examined to understand how different algorithm mechanisms interact with different data structures and the interaction’s influence on detection performance. Unsupervised anomaly detection was the focus because it is the most adaptable and versatile approach in real application scenarios compared with supervised and semi-supervised anomaly detections (Fan et al., 2018; Goldstein and Uchida, 2016). The four algorithms were Local Outlier Factor (LOF), Connectivity-Based Outlier Factor (COF), Cluster-based Local Outlier Factor (CBLOF), and Isolation Forest (IF), and they are the representatives of different mechanisms in the sense of the perspectives from which the detection is carried out. LOF and COF are the representatives of local perspectives, CBLOF is the one of global-local-hybrid perspectives, and IF is the one of global perspectives. Through the employment of each of them, the instance more outlying will be assigned a higher score.

LOF is density-based, and it detects outliers by quantifying how much an instance deviates from its neighbors within a given neighborhood scope. The densities between the neighborhoods of the instance and its neighbors are compared, and the difference is quantified to be the LOF score indicating the instance’s degree of outlier-ness (Breunig et al., 2000). COF can be considered to be a variation of LOF, which also calculates the instance’s outlier-ness based on the difference between it and its neighbors. Nevertheless, it outperforms LOF in detecting the
outliers that diverge from a neighborhood of low density. The idea behind COF is that not all outliers are in a less dense neighborhood compared to their neighbors’ – some can be just isolated from the neighbors that are more connected with each other (Tang et al., 2002). The concept of CBLOF is that the instances not belonging to the large clusters should be considered as outliers. With CBLOF, the instances are grouped into clusters by employing a clustering algorithm, and it is followed by the procedure of computing the CBLOF score of an instance is calculated based on several measurements. The measurements are the capacity of the instance’s cluster, and the distance between the instance and its nearest large cluster (if this instance is in a small cluster), or the distance between the instance and its cluster (if this instance is in a large cluster) (He et al., 2003). IF is developed based on the concept that, compared to the normal instances, the anomalous instances are prone to be isolated faster by the random partitioning of the data. IF is an ensemble approach that is composed of multiple isolation trees (iTrees). In every iTree, the recursive division is applied to dataset \( X = \{x_1, x_2, \ldots, x_n\} \). Each division is based on the randomly chosen feature and the split value for the feature that is within the feature’s value range. After the iTrees are built, the path length \( h(x) \) of every instance \( x \) is computed. \( h(x) \) is defined as the number of partitions required for isolating the instance \( x \) in an iTree, from the root to the terminating node of the iTree. In order to normalize the average \( h(x) \) from all iTrees, the average path length of unsuccessful search in Binary Search Tree (BST) is used. Finally, the IF score is computed based on the normalized average \( h(x) \) (Liu et al., 2008).

The methods used for extracting data inherent patterns in Paper III and IV are listed below in Table 4.
Table 4. Summary of methods for extracting data inherent patterns in Paper III and IV

<table>
<thead>
<tr>
<th>Methods</th>
<th>Function</th>
<th>Adopted for</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA + K-means</td>
<td>(Linearly) Embedding data to a low-dimensional latent space and cluster instances in this space (Unidirectional information pass)</td>
<td>Paper III</td>
</tr>
<tr>
<td>DEC</td>
<td>(Nonlinearly) Embedding data to a low-dimensional latent space and cluster instances in this space (Bidirectional information pass)</td>
<td>Paper III</td>
</tr>
<tr>
<td>LOF</td>
<td>Detecting anomalies (Density-based; Local perspective)</td>
<td>Paper IV</td>
</tr>
<tr>
<td>COF</td>
<td>Detecting anomalies (Chaining-distance-based; Local perspective)</td>
<td>Paper IV</td>
</tr>
<tr>
<td>CBLOF</td>
<td>Detecting anomalies (Clustering-based; Local-global hybrid perspective)</td>
<td>Paper IV</td>
</tr>
<tr>
<td>IF</td>
<td>Detecting anomalies (Isolation-tree-based; Global perspective)</td>
<td>Paper IV</td>
</tr>
</tbody>
</table>

2.6 Metrics for assessing model performance

For both supervised and unsupervised learning studies, we need to evaluate the model performance to know whether the modeling is successful in fulfilling the requirements of the task. In Paper I, the evaluation metric we used was Coefficient of Determination ($R^2$). $R^2$ indicates how much percentage of variance in the label can be modeled/learned by the model. The value of $R^2$ ranges from 0 to 1, corresponding to the percentage of captured variance from 0 to 100%. In Paper II, to ensure a more comprehensive evaluation and comparison between the three tree-based models, two more metrics, Root mean squared error (RMSE) and Mean absolute error (MAE), were adopted besides $R^2$ (Bui et al., 2019). $R^2$ was used because it straightforwardly indicates the amount of general data variance that the model captures. RMSE is used in conjunction with MAE to reveal detailed fitting features for the models. Both of them are in the form of average error, but RMSE magnifies the average error compared to MAE, especially for the data with a large volume of instances. This effect is more apparent for large individual errors that the average error is calculated from. The large individual errors’ contribution to the RMSE value is much more prominent than their contribution to the MAE value. Hence, the comparison between the models’ RMSE and MAE values can help examine whether a model has the issue of polarized fitting, which provides another perspective for the evaluation and comparison between models.
In Paper III, the critical point was to distinguish normal and abnormal conditions from each other accurately. However, there were only labels of repair/stoppage period but no labels of the normal- or abnormal-conditions period. Thus, to evaluate whether the clustering models yielded a good separation between normal or abnormal conditions, the WtE plant domain knowledge was utilized. To be specific, the major time span of the abnormal conditions should be immediately before the time when the repair/stoppage starts to happen. As opposed to the abnormal conditions, there would be a normal-conditions period that should be earlier than the abnormal-conditions period. The clustering performance would be regarded as good if the clustering results matched this general pattern of operational conditions in time series.

In paper IV, we used Precision-recall (PR) curve and the area under the PR curve (AUC_PR) to evaluate the anomaly detection models’ performances. PR curve is a plot that intuitively displays the performance of classification models under different possible thresholds of classification, and it is a variation of the conventional receiver operating characteristic (ROC) curve. PR curve is more precise than ROC curve for the classification on imbalanced data, for which ROC curve usually produces over-positive evaluation results. In a PR curve, the horizontal axis is recall ranging between 0 and 1, and the vertical axis is precision ranging between 0 and 1. PR curve is plotted by connecting all the ‘recall, precision’ dots resulting from the various given thresholds, and it displays the trade-off between precision and recall for different classification thresholds. AUC_PR is computed for PR curve to indicate the corresponding model’s prediction capacity. AUC_PR is a metric that generally indicates the model’s prediction capacity over all possible classification thresholds. AUC_PR can be interpreted as the likelihood that the model will rank a random positive instance higher than a random negative instance. An AUC_PR value can range from 0 to 1, and a larger AUC_PR value means a better class prediction capacity. AUC_PR was used to assess and compare models’ performances due to its two merits: 1) Scale-invariant. It assesses the ranking of predictions instead of the specific values predicted by the models. 2) Classification-threshold-invariant. It evaluates the prediction capacity of models regardless of what classification threshold is adopted.

The metrics used for evaluating model performances in Paper I-IV are listed below in Table 5.
### Table 5. Summary of metrics for evaluating model performances in Paper I-IV

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Function</th>
<th>Adopted for</th>
</tr>
</thead>
<tbody>
<tr>
<td>R²</td>
<td>Evaluating model performance regarding the percentage of overall variance in the label modeled/learned by the model</td>
<td>Paper I and II</td>
</tr>
<tr>
<td>RMSE</td>
<td>Evaluating model performance regarding average prediction error (Used in pair with MAE to reveal the issue of polarized fitting; Tend to amplify large individual errors)</td>
<td>Paper II</td>
</tr>
<tr>
<td>MAE</td>
<td>Evaluating model performance regarding average prediction error (Used in pair with RMSE to reveal the issue of polarized fitting)</td>
<td>Paper II</td>
</tr>
<tr>
<td>WtE engineering knowledge</td>
<td>Evaluating whether the clustering pattern is rational from the perspective of WtE operation</td>
<td>Paper III</td>
</tr>
<tr>
<td>AUC_PR</td>
<td>Evaluating the detection capacity of the model</td>
<td>Paper IV</td>
</tr>
</tbody>
</table>

#### 2.7 Workflows of the study

Figure 4, Figure 5, Figure 6, and Figure 7 depict the workflows of **Paper I**, **Paper II**, **Paper III**, and **Paper IV**, respectively. With the methods mentioned incorporated, the workflows elaborate on how the complete studies were carried out, from the data preprocessing to model training to model/result interpretation.
Figure 4. Flow chart of the study in Paper I
As is shown in Figure 4 and Figure 5, the workflows of **Paper I** and **II** were in the same vein. In general, both of them consisted of data preprocessing, model fitting, model comparison, and model interpretation. This is because the general objectives of **Paper I** and **II** were similar, and the datasets used in the case studies were also similar (except for the volumes of PO4 data). However, they differed significantly in the procedures of model comparison and model interpretation. In **Paper I**, the model comparison was carried out between the DNN and RF to check whether RF captured enough variance to support the subsequent model interpretation, and the evaluation metric was only $R^2$. In **Paper II**, the comparison was carried out between three tree-based models, RF, XGBoost, and LightGBM, to select the optimal one for interpretation. Two more metrics, MAE and RMSE, were used along with $R^2$ to reveal more details of the model fitting for a more comprehensive and unbiased comparison. In **Paper I**, the interpretation approach was Variable Importance Measure (VIM) + Partial Dependence Plots (PDPs). VIM was used to identify the most influential operational factors, and PDP was used to investigate how those influential factors can affect effluent quality. In **Paper II**, these two objectives achieved by VIM and PDP were fulfilled by SHAP analysis based on the SHAP values. SHAP was used.
to reveal more details about model fitting and uncover joint effects of multiple operational factors, besides the results VIM+PDP could yield.

Figure 6. Flow chart of the study in **Paper III**

![Flow chart of the study in Paper III](image1)

As is shown in Figure 6 and Figure 7, **Paper III** and IV did not have data split for model validation and testing since unsupervised learning was applied in these

Figure 7. Flow chart of the study in **Paper IV**

![Flow chart of the study in Paper IV](image2)

```
two studies. This is the most significant difference in data preprocessing compared to Paper I and II. Additionally, the model performance evaluation in Paper III and IV were in a totally different vein compared to the counterparts in Paper I and II. In Paper III, the domain knowledge in WtE was leveraged to assess the performance rather than any numerical metrics. In Paper IV, the scale-invariant and classification-threshold-invariant AUC_PR was adopted for performance evaluation and model comparison. It meant the absolute anomaly scores of the instances and the absolute thresholds of being anomalous were not decisive for the evaluation. Moreover, the interpretation of models in Paper I and II were the quantitative interpretations that were more systematic and based on a series of equations. However, the analyses of model performances in Paper III and IV were qualitative ones with a focus on the mechanisms of the algorithms. In Paper I and II, interpreting the models was for achieving the ultimate objective of the case studies – understanding the relationships between the operational factors and the effluent quality parameters for WWTPs. Nevertheless, in Paper III and IV, analyzing the model performances was for extracting the knowledge from the results of case studies. This knowledge was expected to help guide what type of algorithms should be considered for different application cases in the future.

One interesting difference between Paper III and Paper IV’s data preprocessing was that we dropped features for Paper III but added extra features for Paper IV. In Paper III, many irrelevant features were filtered out with the guidance of domain knowledge. This was carried out to remove the noise that might degrade the performances of the models. In Paper IV, adding contextual features was to include the information that was neglected in the original data form but was crucial for detecting the anomalies hidden in the time variation. These two approaches were opposite, but they served the same purpose – rearranging the data to make them contain essential and clean data, which was to guarantee the maximal relevant information or knowledge would be learned by the models.
3. Model performances

In Paper I, for the RF model on TSS$_e$ (TSS$_e$-RF) and the RF model on PO$_4$$_e$ (PO$_4$$_e$-RF), they both captured more than 90% of the variance in the training sets. Besides, according to the performance on the test sets, both of them could predict a large proportion of the variance in the unknown dataset (92% for TSS$_e$-RF and 88.6% for PO$_4$$_e$-RF). Both models also demonstrated superb generalization performances — they predicted the unknown dataset (test dataset) almost as well as the datasets used to train them. Both DNN models, TSS$_e$-DNN and PO$_4$$_e$-DNN, captured more than 90% of the variance in the training sets and could predict a large proportion of the variance (92% for TSS$_e$-DNN and 87.2% for PO$_4$$_e$-DNN) in the unknown dataset. The performances on training sets and test sets being close also suggested that both models had superb generalization performance. The results showed that TSS$_e$-RF and TSS$_e$-DNN had similarly good performances, so did PO$_4$$_e$-RF and PO$_4$$_e$-DNN. To summarize, the trained RF models learned sufficient and precise correlations between the operational factors and effluent parameters. Thus, they were qualified to be used for further (VIM and PDP) analysis.

In Paper II, as was indicated by $R^2$ values, every TSS$_e$ model captured a considerable amount (over 90%) of general variance from both training and test sets. Also, the RMSE values were in the same pattern as the $R^2$ values. Nevertheless, RF’s MAE value on the test set was substantially higher than the one on the training set, which suggested the problem of overfitting (i.e., RF did not have good generalization performance). Another point worth noting was that RF’s MAE values were considerably lower than those of XGBoost and LightGBM, despite their RMSE and $R^2$ values being very close. The reason for this contrast was that, for RF, the variation among individual errors (the difference between true labels and predicted labels) of the instances was considerable. Specifically, many instances possessed quite small errors that resulted in the low MAE values. In contrast, many instances had very large errors, leading to high RMSE values — since the large errors tend to be magnified in the calculation of RMSE, compared to the calculation of MAE. Unlike RF, the fitting performances of LightGBM, especially XGBoost, were superior in a more balanced fashion. Hence, despite the fact that RF’s MAE values were the lowest and RF possessed nearly the same RMSE and $R^2$ values as the other two models’, RF was the least optimal model among these three due to the overfitting and polarized model prediction. As opposed to RF, XGBoost was optimal because the difference between its RMSE and MAE values was the smallest, and overfitting was not observed with any metric. Hence, it was concluded that the SHAP analysis of the XGBoost model on the training set of TSS$_e$ would uncover comprehensive and precise information for the entire data, and the information would be applicable to future unknown data as well. The overall pattern for TSS$_e$ models mentioned above was also
identified for the PO4<sub>e</sub> models. XGBoost was the optimal PO4<sub>e</sub> model since the disparity between its RMSE and MAE values was the minimum, and there was no overfitting according to all the metrics. Hence, it was believed that the SHAP analysis of the XGBoost model on the training set of PO4<sub>e</sub> would deliver comprehensive and precise information for the entire data, even for future unknown data.

In Paper III, for DEC on dataset A, the optimum number of neurons in the hidden layer (nn<sub>hl</sub>) was 80. With the optimum nn<sub>hl</sub>, the model yielded the optimum clustering result (Fig. 4 in Paper III) when the number of neurons in the embedded layer (nn<sub>el</sub>) was 2. For PCA + K-means, no value in the predefined set of number of principal components (npc) yielded the demarcation between the normal-conditions instances and the abnormal-conditions instances. The full results from both DEC and PCA + K-means were shown in Paper III (Section S2). For DEC on dataset B, the optimum nn<sub>hl</sub> was 128. With the optimum nn<sub>hl</sub>, the model yielded the optimum clustering result (Fig. 6 in Paper III) when nn<sub>el</sub> was 8. For PCA + K-means, no value in the predefined set of npc produced a superior clustering result to the one from DEC. The full results from both DEC and PCA + K-means were shown in Paper III (Section S3). For DEC on dataset C, the optimum nn<sub>hl</sub> was 200. With the optimum nn<sub>hl</sub>, the model yielded the optimum clustering result (Fig. 8 in Paper III) when nn<sub>el</sub> was 5. For PCA + K-means, no value in the predefined set of npc yielded the demarcation between the normal-conditions instances and the abnormal-conditions instances. The full results from both DEC and PCA + K-means were shown in Paper III (Section S4).

As is shown in Paper III (Section S2, S3, and S4), for nearly every dimension, the clustering result from DEC was superior to the one from PCA+K-means. This is due to the difference in mechanisms between these two methods. PCA + K-means maps the original feature space into a latent space through the linear dimensionality reduction method PCA, and employs K-means to cluster the instances in the latent space. In contrast, DEC takes the encoder part of a pre-trained SAE as the embedding module (in a DNN structure) and trains the embedding module and K-means bi-directionally and iteratively. The embedding module passes the embedded information to K-means, and the embedding module and K-means also receive feedback for optimization through the process of minimizing KL divergence. Moreover, possessing multiple tunable core hyperparameters (such as nhl and nn<sub>hl</sub>) may also be one of the factors resulting in the better performance of DEC, whereas PCA+K-means merely has npc.

An example of the PR curves is shown in Figure 8, and the complete PR curves for all the three datasets are shown in Paper IV (Fig. 2, Fig. 3, and Fig. 4). The overall results for dataset A were: 1) All the detectors on the contextualized data
marginally outperformed their counterparts on the original data; 2) LOF > COF > CBLOF > IF with regard to anomaly detecting capacity, and the PR_AUC value of LOF was 0.989 on the contextualized data. The overall results for dataset B were: 1) Nearly all the detectors on the contextualized data considerably outperformed their counterparts on the original data – the exclusion was IF with a relatively slight gap of 0.109 of PR_AUC; 2) LOF > COF > CBLOF > IF concerning anomaly detecting capacity on the contextualized data, and the PR_AUC value of LOF was 0.941. The overall results for dataset C were: 1) All the detectors on the contextualized data considerably outperformed their counterparts on the original data; 2) LOF > CBLOF > COF > IF regarding anomaly detecting capacity on the contextualized data, and the PR_AUC value of LOF was 0.957.

Figure 8. An example of PR curves of detectors and the corresponding PR_AUC values (for dataset A in Paper IV) (CT stands for the contextualized data, OG stands for the original data)
4. Post-interpretation/-analysis based on the models

After the models were acquired and their performances were evaluated, further interpretations/analyses were carried out to either fulfill the ultimate objectives of the tasks or analyze the origins of the difference in models’ performances. In Paper I and II, model interpretation methods were used to interpret the input-output relationships learned by the models for obtaining the cause-and-effect findings associated with the WWTP process. In Paper III, histograms of normal-conditions instances and abnormal-conditions instances were plotted for every feature to discover the accountable feature for the boiler failures. In Paper IV, the reasons on two levels, data structure and algorithm mechanism, were identified for the performance differences between all the detectors.

4.1 Identification of the most influential factors for WWTP effluent parameters

Once the trained models were obtained and validated, VIM analysis on RF models (Paper I) and SHAP analysis on XGBoost models (Paper II) were employed to evaluate the importance of every feature in the models. The VIM values and feature rankings were shown in Paper I (Fig. 4). Every feature’s mean absolute SHAP value over all instances and their rankings were shown in Paper II (subfigure A of Fig. 3 (TSSₜₑ) and Fig. 5 (PO₄ₑ)). Both VIM value and mean absolute SHAP value are the indicators of impact -- a higher VIM value or mean absolute SHAP value means a higher impact on TSSₑ or PO₄ₑ.

We focused on the top three influential features in Paper I and the top four influential features in Paper II, and subsequently analyzed their detailed impact patterns. In Paper I, the top three features for TSSₑ are TTₑᵢᵣ, FTₛᵣₑ, and TSSₐ₃ₑ. For PO₄ₑ, they are TTₑᵢᵣ, TSSₐ₄ₑ, and TSSₐ₄ₑ. In Paper II, the top four influential features for XGBoost on TSSₑ are TTₑᵢᵣ, TSSₐ₅ₑ, FTₛᵣₑ, and TSSₐ₃ₑ, while for XGBoost on PO₄ₑ are TTₑᵢᵣ, DOₐ₄ₑ, FTₛᵣₑ, and TSₛₑ. It can be observed that Paper I and II share the same top three features, only with the positions of FTₛᵣₑ and TSSₐ₃ₑ switched. Even though different data (due to random split for training, validation, and test sets), different modeling methods, and different interpretation methods were used, the overall rankings of the top three features for TSSₑ are the same for Paper I and II. This phenomenon reflects the effectiveness and robustness of both workflows in Paper I and II to some extent. This still holds when the RF model on TSSₑ in Paper II is included in the comparison – the top three features for RF were exactly the same as XGBoost’s. This sameness suggests that the general importance measurements for the most influential features, mean absolute SHAP value, cannot reflect the fine difference in model fitting indicated by MAE difference. Instead, the difference might be shown in the detailed impact
patterns of these influential features, or the difference is associated with some less influential features. In contrast, for PO4, **Paper I** and **II** only share the first feature TT in, among the top three features. The root cause of this difference is the significant difference in data. In **Paper II**, for PO4, many uninformative instances were removed to ensure a more accurate information extraction regarding the relationships between PO4 and the operational factors. Thus, the overall correlation between PO4 and operational factors was considerably changed in the data of **Paper II** compared to the PO4 data in **Paper I**.

Since **Paper II** is a further refined study over **Paper I**, the subsequent investigation into the impact patterns of the most influential features in this thesis is for the top four features identified by XGBoost models in **Paper II**. They are TT in, TSS lr, FT sr, and TSS a3 for TSS, and TT in, DO a4, FT sr, and TS sr for PO4. Also, the approach applied to the model interpretation is SHAP.

### 4.2 Investigation into impact patterns of the most influential factors in WWTP

After identifying the four most influential features for both TSS and PO4, SHAP dependence plots were drawn to illustrate how these features specifically influence the two effluent quality parameters TSS and PO4. Figure 9 and Figure 10 are the collections of SHAP dependence plots for TSS’s and PO4’s top four influential features, respectively. For every feature, the scatter points were colored to reflect the other three features’ value variations, respectively. In this way, the current feature’s joint influences with the other three features can be intuitively observed. As is shown in **Paper II** (Equation (3)), the definition of SHAP value, positive SHAP values contribute positively to TSS, and negative SHAP values contribute negatively to TSS.
Subfigures A1, A2, and A3 in Figure 9 are the SHAP dependence plots of TTI\textsubscript{in}. There is no apparent variation of the SHAP value when TTI\textsubscript{in} varies from 6 °C to 10 °C, but the SHAP value decreases as TTI\textsubscript{in} increases from 10 °C. This decline in the SHAP value suggests that a rise in temperature between 10 °C and 16 °C can enhance the removal of TSS\textsubscript{e}, probably through promoting the microorganisms’ physiological properties (Adams et al., 2010; Garcia-Rios et al., 2016), boosting microbial reproduction (Rajeshwari et al., 2000; Spellman, 2013), enhancing microbial activities (Rajeshwari et al., 2000; Young et al., 2017), improving the variety of species and the structure of microbial community (Chen et al., 2017), promoting the settleability of sludge (Wilén et al., 2008; Yang and Li, 2009), and enhancing the process of coagulation and flocculation (Dayarathne et al., 2020). Moreover, the variation line based on the scatter points is thin without dispersion.
in the vertical direction, which suggests that the SHAP value of TT\textsubscript{in} is mainly subject to TT\textsubscript{in} values without the involvement of other features.

Subfigures B1, B2, and B3 in Figure 9 are the SHAP dependence plots for TSS\textsubscript{lr}. There is a noticeable increase in the SHAP value when TSS\textsubscript{lr} rises from 0 to 0.75 mg/L. Above 0.75 mg/L, there is no apparent variation of the SHAP value as TSS\textsubscript{lr} rises. TSS\textsubscript{lr} directly comes from the settled solids in the final sedimentation basin, and TSS\textsubscript{e} directly comes from the treated water from the same basin. The sedimentation is a stable process, which means the ratio of TSS settled to the TSS left in the water is approximately constant. Hence, TSS\textsubscript{e} rises as TSS\textsubscript{lr} rises, which is demonstrated by the pattern when TSS\textsubscript{lr} varies below 0.75 mg/L. Nevertheless, the SHAP value levels off when TSS\textsubscript{lr} changes above 0.75 mg/L. A likely explanation is that WWTP operators who observe TSS\textsubscript{e} rising beyond a threshold will take immediate action to lower it (e.g., by dosing more coagulant into the second coagulant mixing basin), leading to a significantly larger portion of TSS settling in the final sedimentation basin without the increase of TSS\textsubscript{e}. Besides, there is a modest dispersion of the SHAP value in the vertical direction when TSS\textsubscript{lr} is above 0.75 mg/L. As is indicated in subfigure B1, TT\textsubscript{in} is accountable for this dispersion, and the larger TT\textsubscript{in} values result in smaller SHAP values of TSS\textsubscript{lr} and vice versa. A likely explanation is that high temperatures improve the coagulation-flocculation process and settleability of sludge/solids (Wilén et al., 2008; Yang and Li, 2009), which leads to more sludge staying settled in the last return sludge pipe regardless of the agitating effect of the pumping. Hence, even with TSS\textsubscript{lr} unchanged, the solids settling in the last sedimentation basin at a higher temperature could be more, contributing to a lower TSS\textsubscript{e}.

Subfigures C1, C2, and C3 in Figure 9 are the SHAP dependence plots for FT\textsubscript{sr}. The main trend of FT\textsubscript{sr}'s SHAP values is level when FT\textsubscript{sr} varies below 28 m\textsuperscript{3}/h, despite several dispersions occurring because of the influence from other features not among the top four. Above 28 m\textsuperscript{3}/h, there is a dramatic rise overall. The likely reason lies in the layout of the process units. Both the second return sludge pipe (No. 14 to No. 10 in Fig. 1 of Paper II) and the first return sludge pipe (No. 14 to No. 12 in Fig. 1 of Paper II) transport settled sludge from the secondary sedimentation basin. Thus, there is a trade-off between the volume of sludge returned through the second return sludge pipe and the volume of sludge returned through the first. Once FT\textsubscript{sr} exceeds a threshold (about 28 m\textsuperscript{3}/h), the volume of activated sludge returned to the aeration basin becomes inadequate for biodegradation, contributing to the rise of untreated solids in the following units and the effluent. Furthermore, when FT\textsubscript{sr} exceeds this threshold, the first coagulant mixing basin and the primary sedimentation basin will expect an overload of suspended solids. The portion of suspended solids that cannot settle in the primary sedimentation basin will thus enter the aeration basin. Nonetheless, this part of sludge will lack biodegradation capacity since large
amounts of microorganisms will not survive the anoxic and anaerobic units before the aeration basin (Hussain and Bhattacharya, 2019), which will cause a larger solids load in the aeration basin. Besides the general pattern along the horizontal axis, as is presented in subfigure C2, TSSlr shows a correlation to the dispersion of the FTsr’s SHAP values when FTsr varies between 0 and 28 m³/h. A larger TSSlr value basically results in a smaller SHAP value of FTsr and vice versa. This indicates that enhancing TSS settlement in the final sedimentation basin will help decrease TSSe -- when the overload of TSS is not introduced by the excess return of sludge through the second return sludge pipe.

Subfigures D1, D2, and D3 in Figure 9 are the SHAP dependence plots for TSSa3. The SHAP value basically declines while TSSa3 increases from 0 to 2000 mg/L, but it levels out when TSSa3 varies beyond 2000 mg/L. The line formed by the scatters is generally thin, but some dispersion related to TTin occurs between 0 and 700 mg/L of TSSa3. Basically, the smaller TTin results in a larger SHAP value of TSSa3 and vice versa. TSS in the aeration basin (e.g., TSSa3) is the reflection of the microorganisms’ concentration to a large extent. When the microorganisms' concentration is low, the food and dissolved oxygen in the aeration basin are adequate for the microorganisms. Under this condition, increasing the temperature will boost the microorganisms’ metabolism and therefore promote the bio-degradation of the organic matter, which will finally result in the enhanced removal of TSSe. In contrast, the likely reason for the SHAP value being level when TSSa3 is beyond 2000 mg/L is as follows. When the food and oxygen are adequate, more microorganisms lead to better bio-degradation since all the microorganisms can perform well. Nonetheless, when the concentration of microorganisms keeps increasing to a certain level, the amount of food and oxygen will not be adequate any more for all the microorganisms to function, which will hamper any further improvement in the removal of TSSe.
Subfigures $A_1$, $A_2$, and $A_3$ in Figure 10 are the SHAP dependence plots for $T_{T_{in}}$. The variation of $T_{T_{in}}$ between 6 °C and 14.4 °C does not impact the SHAP value noticeably and coherently, except for the several considerable fluctuations. When $T_{T_{in}}$ rises beyond 14.4 °C, it causes an overall rise in the SHAP value, regardless of the apparent fluctuations. The likely reason lies in the rivalry for substrates between glycogen-accumulating organisms (GAOs) and polyphosphate-accumulating organisms (PAOs). PAOs contribute positively to PO$_4$ removal, whereas GAOs do not (Seviour et al., 2003). PAOs are psychrophilic, while low temperatures hamper the metabolism of GAOs. Accordingly, PAOs dominate at lower temperatures, but GAOs gain the upper hand at higher temperatures (Erdal et al., 2003; Lopez-Vazquez et al., 2009). As for the dispersions, there are a few apparent dispersions, but it is uncertain whether the other three top features are accountable for them.
Subfigures B1, B2, and B3 in Figure 10 are the SHAP dependence plots for DOn. It can be observed in the three subfigures that there is no evident variation of the SHAP value when DOn varies below 5 mg/L. Nonetheless, a jump of the SHAP value is seen at 5 mg/L of DOn, and the SHAP value generally increases when DOn increases beyond 5 mg/L. There are considerable dispersions of the SHAP value when DOn varies beyond 5 mg/L, whereas none of the other three top features are associated with them. The pattern beyond 5 mg/L is unusual – high SHAP values, high temperatures, low second return sludge flows, and high dissolved oxygen appear together. This pattern might reflect the situation where the wastewater load is at a rather low level without adequate food in the water for the microorganisms (PAOs included) to flourish and perform well. Therefore, a considerable proportion of the dissolved oxygen in the aeration basin cannot be consumed, and the functional PAOs that can ingest PO4 are insufficient.

Subfigures C1, C2, and C3 in Figure 10 are the SHAP dependence plots for FTsr. The SHAP value rises modestly as FTsr rises from 0 to 12 m³/h, and declines modestly when FTsr increases beyond 12 m³/h. Additionally, there is a dramatic decrease in the SHAP value when FTsr reaches 12 m³/h. When FTsr value is small (smaller than 12 m³/h), it can be deduced that the volume of sedimented sludge in the biosedimentation basin is minimal, which is caused by the poor biodegradation activity in the aeration basin. This deduction can be validated by the large volume of DOn in this range, as presented in subfigure C2. Therefore, the aeration basin requires a substantial volume of sludge to be returned to the aeration basin (first return sludge) to increase the concentration of microorganisms. Nonetheless, since the first and second return sludges share the direct sludge source, the rise in FTsr means more sludge is returned to the coagulant mixing basin, and gets sedimented and removed from the primary sedimentation basin instead of being returned to the aeration basin. This exerts a negative impact on the biological removal of PO4. In contrast, high FTsr (above 12 m³/h) and low DOn indicate effectual biodegradation. In this situation less first return sludge and more second return sludge can benefit the removal of the PO4 ingested by PAOs and help sustain the functional metabolism and multiplication of PAOs. The significant difference in SHAP values between these two levels of biodegradation intensity implies that biodegradation is crucial for PO4 removal: weak biodegradation induces low concentration of PAOs, causing large portion of PO4 unconsumed in the water. Besides the overall pattern along the horizontal axis, there is an evident dispersion shown in subfigure C3. It happens in the FTsr range of 19 m³/h to 33 m³/h, and it is associated with TSsr. A lower TSsr results in a lower SHAP value of FTsr and vice versa. The likely reason for the dispersion is that more total solids returned leads to more salts and suspended solids competing with PO4 for the coagulant FeCl₃, which negatively affects the removal of PO4e.
Subfigures $D_1$, $D_2$, and $D_3$ in Figure 10 are the SHAP dependence plots for $T_{SR}$. Generally, there is a rise in the SHAP value as $T_{SR}$ increases, but the trend line becomes much more level from 2.5% of $T_{SR}$. The likely cause of the general increase pattern is that more returned total solids introduce more salts and suspended solids to compete with $PO_4$ for the coagulant $FeCl_3$, which exerts an adverse impact on the removal of $PO_4$. Moreover, there are evident dispersions both below and above 2.5% of $T_{SR}$. For the one below 2.5%, no feature in the top four list is accountable for it. For the one above 2.5%, $TT_{in}$ and $DO_{a4}$ are accountable for it. The larger $TT_{in}$ and $DO_{a4}$ values result in smaller SHAP values of $T_{SR}$ and vice versa. The likely reason behind $TT_{in}$’s influence is that higher temperatures can boost the processes of coagulation-flocculation and chemical precipitation. The $DO_{a4}$ related dispersion is consistent with the unusual pattern shown in subfigures $B_1$, $B_2$, and $B_3$. Therefore, the explanation of that pattern can be used to interpret this dispersion.

4.3 Identification of different operational conditions and most responsible operational factors for boiler failures in WtE plant

After identifying the optimal clustering model and the associated clustering results in time series in Paper III, the operational conditions corresponding to each cluster were identified. As is shown in Paper III (Fig. 4, Fig. 6, and Fig. 8), there are three clusters – clusters 0, 1, and 2. Cluster 2 was identified as the repair periods/stoppages induced by the failures since all three cluster 2 periods matched the repair timelines recorded in the log information. Subsequently, according to the criterion mentioned in Section 2.6 -- ‘the major time span of the abnormal conditions should be immediately before the time when the repair/stoppage starts to happen,’ cluster 1 was identified as the abnormal conditions that induced the failures. Finally, based on the already identified cluster 1 and the criterion mentioned in Section 2.6– ‘As opposed to the abnormal conditions, there would be a normal-conditions period that should be earlier than the abnormal-conditions period,’ cluster 0 was identified as the normal conditions.

In Paper III (Fig. 4 (for dataset A) and Fig.6 (for dataset B)), we can observe that there are several short periods of normal conditions amid the broad range of abnormal conditions. This is explainable because the operational conditions in WtE plants are dynamic and sometimes fluctuate in a wide range. Despite this, the cluster lines in those two figures are relatively smooth. In Paper III (Fig. 6), an extremely short cluster 2 can be spotted amid the cluster 0 period. Nevertheless, the engineers at Dåva 1 established that this short cluster 2 was not associated with boiler failures. Instead, it was very likely induced by some transitory malfunction of the monitoring system. In Paper III (Fig.8 (for dataset
C)), unlike the situation for datasets A and B, the fluctuations between clusters 0 and 1 are more frequent. In order to obtain more refined and smooth clustering results for clusters 0 and 1, a higher degree of denoising of the original data should have been employed. Nonetheless, the overall demarcation between clusters 0 and 1 is not influenced by these noises. Besides, it is evident that there are a few transient periods of cluster 2 in addition to the two officially documented repair periods (the last two) for fixing the boiler failures. As stated by the engineers at Dāva 1, these non-boiler-related stoppages either are the reflection of transitory malfunction of the monitoring system (corresponding to the extremely short periods) or were induced by other problems outside the boiler system (corresponding to the modestly short periods).

Based on the separation of operational conditions shown in Paper III (Fig. 4, Fig. 6, and Fig. 8), histograms of each feature for normal and abnormal conditions were drawn to inspect the distributions of instances for the normal-conditions and abnormal-conditions clusters. An example of the histograms is shown in Figure 11. The features that display an apparent distribution shift between these two clusters are considered accountable for the failure. Therefore a closer look needs to be taken at these features to avert the same failure in the future. Histograms of the features with considerable shifts between normal and abnormal conditions were shown in Paper III (Fig. 5 (dataset A), Fig. 7 (dataset B), and Fig. 9 (dataset C)). The full results (histograms for all the features) were presented in Supplementary Material of Paper III (Section S5, S6, and S7). Both cluster 0 and cluster 1 have a roughly normal distribution for most of the features, so the difference between their most frequent values ($\Delta m$) was utilized to indicate the difference between the distribution of cluster 0 and the distribution of cluster 1. For the other features, their distributions are characterized by two peaks, especially for cluster 1. Of those two peaks in cluster 1, the one further from the peak of cluster 0 was utilized to approximate the $\Delta m$ between these two clusters.

For dataset A, the features with considerable $\Delta m$ are T-BbSH, T-BSH3rm, T-BSH3r, T-BSH2rm, T-Fun, FL-G. The accurate location information of the failures is missing in the log file, but it was recorded that both failures occurred in the tube grid close to the roof. This information is in line with the results of analyzing the histograms since the sensors of the identified accountable features are located around the tube-grid area. These features imply that the (overly) intense heat between the first flue-gas pass and the superheater 2 is the major cause for these two boiler failures. Besides, the leftward distribution shift of FL-G from normal conditions to abnormal conditions suggests that the decline in the workload of the ID fan might have boosted the intense heat of the flue-gas. For dataset B, the features with considerable $\Delta m$ are T-BSH2l, T-BSH2r, T-BSH1r, T-SbSH3. It should be noted that the shift in flue-gas temperature around superheater 2 (T-BSH2l and T-BSH2r) resulted in the shift in steam temperature before superheater 3 (T-SbSH3). The information about the accurate locations of
the failures is missing in the log file, but it was recorded that the failure occurred in the superheater area, which is in accordance with the result. For dataset C, the accurate location information of the failures is missing in the log file, but it was recorded that the failures occurred in the tube grid close to the roof. With the location information being considered, the accountable features were eventually identified as P-F and T-SbSH3. P-F being accountable suggests that the pressure decline in the furnace might have caused the rise in the pressure gap between the outside and the inside of the tubes, which might have contributed to the eventual tube bursting. Also, as is indicated by the shift in T-SbSH3, the rise in steam temperature and pressure might have been a critical factor for the tube bursting.

Figure 11. Example of histograms of features with evident shifts between normal and abnormal conditions (for dataset C in Paper III). Subfigures are independent; in each subfigure, feature values can be read from the horizontal axis, and the corresponding count/frequency of values can be read from the vertical axis.

4.4 Discussion on the data-algorithm-interaction in energy anomaly detection for buildings

In Paper IV, there are two sets of overall results shared by all the three datasets. The first one is that every detector on the contextualized data outperformed their counterparts on the original data. It can be deduced that the reason lies in the data structure. The contextualization introduces the contextual features (month, day class, and hour) to the original data, setting up the context of the behavioral feature (energy consumption) for each instance. The extra context information results in the redefinition of the instances’ correlations to each other. By contrast, for the original data, the information that can be used to compute the correlations among the instances is only the behavioral feature. Without the crucial temporal information being incorporated, bias can develop in the process of
identifying/calculating the neighborhoods and the neighborhoods’ densities for LOF, neighborhoods and the chaining distances for COF, the path length for IF, and the cluster centroids for CBLOF.

The second set of results common to all the three datasets is the consistent sequence of detectors with regard to their detecting capacity on the contextualized data. For datasets A and B, it is LOF > COF > CBLOF > IF. For dataset C, it is almost the same except for the swapped positions of COF and CBLOF: LOF > CBLOF > COF > IF. This pattern may be associated with the perspectives these algorithms calculate the anomalous scores from – either local or global perspective. LOF computes the anomalousness of instances from a local view because the anomaly score of the instance of interest is calculated based upon the densities of the neighborhood of it and the neighborhoods of its neighbors. COF possesses a very similar general mechanism with the exception of using chaining distance instead of density. Different from the utter local view of LOF and COF, CBLOF calculates the anomaly score in a hybrid manner with both global and local perspectives involved. Clustering instances and the subsequent identification of large and small clusters are on the global track. Every instance is examined to produce the cluster centroids for the whole data, and all the clusters’ sizes are used to calculate the standards of being ‘large’ and ‘small.’ Nonetheless, the final procedure of computing CBLOF values hinges on the size of the cluster that the instance belongs to and the distance between the instance and its nearest large cluster. For IF, the whole process of isolating instances is based on a sheer global perspective. For example, the random sub-sampling in every iTree is carried out over the whole input data, and the feature value used to split an iTree is randomly selected in the range between the feature’s minimum and maximum values. With the performance sequence and the mechanism details presented above, it can be concluded that local approaches have better detecting capacity than global approaches in the scenarios where the objective is to detect the anomalous instances with regard to those instances’ contextual neighbors rather than the rest of the whole data.

In summary, the better detecting capacity of LOF, COF, and CBLOF on the contextualized data can be ascribed to the elements at two levels – data structure level and algorithm mechanism level. At the data structure level, the instances’ contextual information is established through introducing the contextual features. The contextualization transforms the original mono-dimensional data space to a multi-dimensional one with the redefinition of every instance’s coordinates, which results in a more comprehensive and less skewed computation of the correlations between the instances. At the algorithm mechanism level, the algorithms with local perspectives can make the most of the contextual information introduced by the contextual features because defining the contextual neighborhood is crucial for their computation of anomaly scores.
Moreover, LOF shows the best performance for all the contextualized datasets, indicating that the density comparison mechanism is the most appropriate for the datasets in Paper IV. Because these three datasets are representative of the commercial buildings’ energy consumption profiles, it can be expected that LOF will be applicable to other similar datasets and yield good detecting performance. Nevertheless, it is always worthwhile to carry out COF as a comparison since ‘density difference’ is not the only manifestation of an instance being anomalous – there are some variants, for example, ‘pattern distinctness.’
5. Conclusions and future perspectives

This thesis investigated how data science can boost both hard and soft measures to tackle environmental issues nowadays. Through the case studies in three scenarios, WWTP effluent quality control, WtE plant boiler failure investigation, and commercial buildings' anomalous energy consumption detection, the details of the implementation of data science were demonstrated and illustrated – from data preprocessing to model building to model/result interpretation. The thesis yielded benefits for the case study subjects and, on a higher level, the knowledge that can be applied in a broader context of the related domains.

Paper I and II provided information on the operational factors that influence TSS, and PO4, and their effects, which can help operators understand the complex processes occurring in Umeå WWTP and similarly configured WWTPs. Several initial findings regarding the influence of the most significant factors were delivered by Paper I, then Paper II provided further and updated findings due to the implementation of new methods and refined PO4 data. In conjunction with traditional (i.e., chemical, biochemical, and hydromechanical) analyses, these results could be helpful for more effective and reliable decision-making about whether current operational parameters are appropriate or whether pre-emptive action is required to prevent deterioration in effluent quality.

More generally, instead of yielding predictions and developing soft sensors, Paper I inspected the suitability of utilizing ML to obtain a comprehensive understanding of WWTP processes. A corresponding workflow was proposed and described in detail, from data preprocessing to model building to model explanation. During the data preprocessing phase, the original data in time series were transformed into batch series data to assure that all the features of the same instance were the properties of the same batch of water. DNN and RF models were built in the model-building phase. DNN models were used to validate whether the RF models captured adequate variance that is essential for a convincing and representative interpretation result. The interpretation phase was based on both VIM and PDP analyses that were for identifying the influential features and those features' impact patterns, respectively. By and large, this workflow can also be employed to other process parameters (and even for the processes in other industries) if adequate and high-resolution data of them are obtainable. Paper II yielded two significant general findings concerning leveraging ML to investigate the cause-and-effect relationship in WWTPs (or even other process industries). First, comparing model performances in a multi-perspective manner is vital for uncovering comprehensive details required for
selecting a truly reliable and robust model for the subsequent interpretation. For both TSS and PO4, the RF models showed very similar performances to those of XGBoost and LightGBM if the MAE values were not considered. Only with MAE values being considered and compared with RMSE values were the problems of overfitting and polarized fitting of the RF models uncovered. Second, adopting an accurate, robust, and granular interpretation method can benefit both model comparison and model interpretation. For example, because of SHAP’s granular nature, the SHAP dependence plots demonstrated the dispersions reflecting the combined effects of multiple features. The information of combined effects is usually desired by the process industries where compound kinetics exist, and different underlying interactions occur between operational factors.

The conditions in and outside boilers are usually intricate and dynamic, making it nearly impossible to discover the precise physical or chemical mechanisms behind those failures. Therefore, in Paper III, the failure investigation for the boiler system was carried out in an operational-parameter-centric manner. The result can help sustain the operational conditions in a safe range to delay the potential failures and make the boiler system function longer until the scheduled annual maintenance. Two approaches, PCA + K-means and DEC, were employed to cluster the instances into different states of operational conditions, and their performances were compared. For all three cases, DEC was superior to PCA + K-means. This is because the embedding module and K-means in DEC are trained simultaneously and iteratively with the bidirectional information pass: from the embedding module to K-means to KL divergence, and from KL divergence to the embedding module and K-means. Finally, the features showing apparent distribution shifts between normal and abnormal conditions were identified as accountable for the boiler failures. The results can be easily understood by the operators since they are directly associated with the operational factors. Based on the results, it will be very viable for the operators to work out solutions to manipulate the process and keep the operational factors within a safe range. Besides waste-to-energy plants, this workflow is applicable to any other production line contingent on various operational factors while lacking operational-conditions labels in the data, which comprises a significant percentage of failure analysis cases.

Paper IV achieved two objectives – it precisely detected the contextual anomalies hidden behind the time variation of the energy consumption data from real buildings; it investigated the combined influence of data structure and algorithm mechanisms on the performance of unsupervised anomaly detection for building energy consumption data. Energy consumption for buildings, especially commercial buildings, is characterized by the high dependence on multi-dimensional (e.g., the dimension of hour, day, month, season, and year)
time variation. The anomalies within a specific temporal space (context) can usually appear to be normal in another context, so it is rather challenging to identify those anomalies without considering the specific context surrounding them. In this paper, the reconstruction of the original data with the introduction of three contextualized features (i.e., hour, day class, and month) significantly improved the anomaly detection performances on all the three given datasets regardless of the specific algorithms applied. The contextualization reconstructed the original data space to a new data space with three more temporal dimensions, which led to the alteration of the relative locations of the instances. The alteration contributed to a much less biased estimation of the correlations between instances, which were the key to defining the context of a particular instance. Since defining the context is a crucial procedure for the algorithms with local perspectives involved, the algorithms with more local perspectives benefited more from the contextualization and yielded better results.

All the studies in this thesis possess the potential of being improved in the future. One major improvement lies in data enrichment. For example, in WWTPs, if a new situation falls out of the range of the historical data, then the trained model obtained in Paper I and II may have to be re-tuned. This is the universal limitation for ML, especially when the models are nonlinear. Although nonlinear models are usually much more powerful and accurate than linear models, they have a significant disadvantage compared to linear models -- nonlinear models cannot extrapolate. In order to make nonlinear models perform well on unknown data, the unknown data should be within the range defined by the data used to train the models. Thus, if more diverse data are collected and used to train the model in the future, both the model performance and the interpretation outcome will be more robust and applicable. Besides, the number of studied effluent parameters left much space for increase. There were only two effluent parameters studied in Paper I and II, but there are some other important parameters for effluent as well. These two parameters were studied because only their data were in large scale and high resolution. Others’ data were too sparse and messy to yield convincing results to guide the process control. The bottleneck is the water determination technology. We expect that in the future, Umeå WWTP will deploy reliable online sensors for other effluent parameters, then all the parameters can be determined in a real-time manner. This will contribute to a more comprehensive study. Furthermore, for Paper III, more detailed log information about failures is needed in the future to better guide the selection of the responsible features. Also, more failure cases need to be analyzed to build the fingerprints of failures at various locations. The fingerprint library will help promote the operation and management in Dåva 1 WtE plant to a much more sophisticated level.
Another main improvement lies in the evolution of data science methods. With the fast evolution of computing power and the increasing availability of big data in recent decades, data science methods keep springing up either in the form of original methods or the optimization of existing methods. This trend will continue in the future and even at a more rapid pace since more industries and practitioners are getting involved, especially when driven by the demands in Industry 4.0 context. This will provide ample opportunities for refining the presented studies in this thesis. As shown in Paper I and II, the newer algorithm XGBoost outperformed RF, and the newer explanation system SHAP delivered more detailed and more informative influence patterns than the traditional interpretation system VIM + PDP. Moreover, in Paper III, the newer clustering algorithm DEC outperformed its basic counterpart PCA + K-means. Of course, there is no guarantee that newer methods will always outperform old ones. The success or improvement of data science applications is dependent on the perspectives of the particular tasks, available data, and algorithms/methods. A comprehensive and deep understanding of these three aspects is essential for the practitioner to know whether an improvement is possible, which part can be improved, and what methods should be explored. Sometimes, the improvement can even be brought by an old method, which has been overlooked because it was not designed for the task of interest. In this thesis, we explored and adopted the methods based on our best knowledge as of the time when we were writing the papers. However, there might be more appropriate methods existent beyond our knowledge. Thus, with more time and effort devoted, we should try to think out of the box and expand the radius of method exploration.
Acknowledgment

First and foremost, I am grateful to my main supervisor, Mats Tysklind, for providing Green Technology and Environmental Economics platform for my PhD study. This thesis would not be possible without his contribution. His consistent support, patient guidance, and inspiring advice are beyond price. He is also a role model in terms of personality and behavior. He is such a forbearing and witty person whom I always feel comfortable talking with and happy to work with.

I would like to extend my thanks to my other two supervisors, Johan Trygg and Lili Jiang. Johan always inspires me by offering insightful comments and constructive advice on my work. Lili’s input from the computing science perspective has served as valuable validation to the data science implementations. Her guidance has broadened my understanding of the data science domain.

I would like to thank my reference persons, Erik Björn and Solomon Tesfalidet, for their time and energy spent evaluating my PhD study and ensuring I was on the right track. Solomon is the first professor who bought me a beer, and that is one of the cool memories about my PhD life.

I would like to express my appreciation to Anna Linusson Jonsson. Without her help, the PhD study would have been uncertain and tougher in multiple facets. Her strong sense of responsibility, genuine care for PhD students, and great effort in working out solutions are truly impressive.

Great thanks also go to our collaborators in the companies -- Daniel Fredlander, Sven Thunell, and Ulrika Lindberg at Vakin AB; Måns Kjellander and Eva Weidemann at Umeå Energi AB; Therese Enlund and Amanda Fors at Mestro AB. Their contributions are significant to this multi-disciplinary thesis. The data, domain knowledge, validation, and feedback provided by them are invaluable.

I am thankful to have friendly and helpful colleagues at Umeå University, especially Sereilakhena Phal and Xuan-Son Vu. Sereilakhena is my former officemate who helped me adapt to the working environment. The office was always full of laughter when she was there. Xuan-Son’s perceptive comments gave me profound inspiration for Paper III.

Last but not least, I would like to offer special gratitude to my parents and wife for being my backbone and cornerstone. Although they do not contribute scientifically to this thesis, their understanding and support are the source of my courage, determination, and persistence that are essential for completing the thesis.
References


Benzidia S, Makaoui N, Bentahar O. The impact of big data analytics and artificial intelligence on green supply chain process integration and hospital environmental performance. Technological Forecasting and Social Change 2021; 165: 120557.


Erdal U, Erdal Z, Randall C. The competition between PAOs (phosphorus accumulating organisms) and GAOs (glycogen accumulating organisms) in EBPR (enhanced biological phosphorus removal) systems at different temperatures and the effects on system performance. Water Science and Technology 2003; 47: 1-8.


Hussain A, Bhattacharya A. Advanced Design of Wastewater Treatment Plants: Emerging Research and Opportunities: Emerging Research and Opportunities: IGI Global, 2019.


Li B, Pandey MD. An advanced statistical method to analyze condition monitoring data collected from nuclear plant systems. Nuclear Engineering and Design 2017; 323: 133-141.


