BTO 2022.021 | April 2022

Deep Explorations: an explorative study for machine learning and deep learning applications in the water sector

Joint Research Programme

KWR

Bridging Science to Practice

Deep Explorations: an explorative study for machine learning and deep learning applications in the water sector

BTO 2022.021 | April 2022

This research is part of the Joint Research Programme of KWR, the water utilities and Vewin.

Project number

402045-228

Project manager
dr. G. J. (Geertje) Pronk

Client
BTO - Verkennend onderzoek

Author(s)
dr. X. (Xin) Tian, I. (Ina) Vertommen MSc., dr. F. (Frederic) Béen, dr. P. (Patrick) Bäuerlein

Quality Assurance
dr. P. (Peter) van Thienen

Sent to
This report is distributed to BTO-participants, and it is public after 1 year.

Keywords

deep learning, water sector, neural networks, natural language processing, spectra classification

Year of publishing
2022

More information
dr. Ina Vertommen
T +31 30 606 9739
E ina.vertommen@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

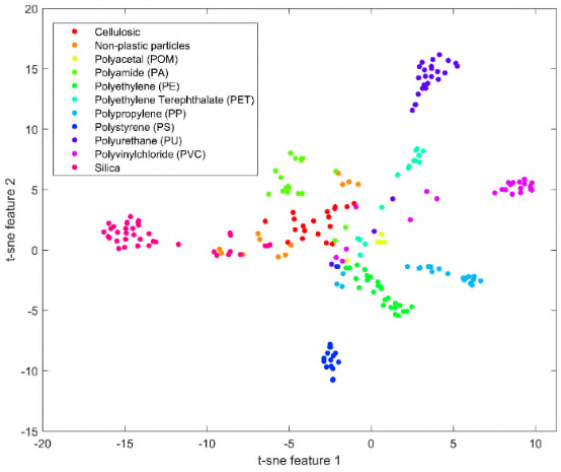
T +31 (0)30 60 69 511
F +31 (0)30 60 61 165
E info@kwrwater.nl
I www.kwrwater.nl

Management samenvatting

Verkennend onderzoek toont potentie aan van machine learning en deep learning voor de watersector

Auteurs: dr. Xin Tian, Ina Vertommen MSc., dr. Frederic Béen, dr. Patrick Bäuerlein

Voor de watersector kan machine learning (ML) en deep learning (DL) van grote waarde zijn om problemen op het gebied van classificatie, regressie of besluitvorming aan te pakken. Dit blijkt uit verkennende studie naar de mogelijkheid om ML/DL toe te voegen als instrument in de gereedschapskist voor onze onderzoekers. Cases zijn bestudeerd met het oog op de geautomatiseerde verwerking van klantmeldingen van drinkwaterbedrijven en de identificatie van microplastics in het milieu. Sinds 2016 is DL uitgegroeid tot een van de meest populaire onderwerpen in verschillende wetenschappen. DL, of breder, ML, is een krachtige methode voor het omgaan met grote reeksen gegevens en is dus in potentie ook interessant voor de watersector. Zo kan DL bijvoorbeeld worden toegepast om visuele data te analyseren voor identificatie van landgebruik en waterlekken, de beoordeling van de toestand van leidingen of de detectie van flocculatie bij waterbehandeling. Wat tekstanalyse betreft, kan DL worden gebruikt voor literatuurstudies, waarbij onderzoekers worden geholpen om wetenschappelijke publicaties efficiënter te doorzoeken. Een beperkende factor in het gebruik van ML en DL voor de watersector is een gebrek aan hoogwaardige gegevens. Daarom moet bij ontwikkelaars en gebruikers van de modellen hier prioriteit aan worden gegeven.



Classificatie van microplastics op basis van hun spectrale verdeling, ontgonnen via dit project.

Belang: deep learning is het snelst groeiende veld in kunstmatige intelligentie

Sinds 2016 is DL uitgegroeid tot een van de populairste onderwerpen in de wetenschap, in engineering en – zonder dat mensen het beseffen – het dagelijks leven. Het is nu het snelst groeiende veld binnen de kunstmatige intelligentie. Zelfrijdende auto's, gezichtsherkenning, live transcriptie en vertaling, actuele weersvoorspellingen en ga zo maar door; zonder DL zijn al deze nieuwe toepassingen onmogelijk. Ook de watersector heeft DL gevonden. In onderzoek en toepassingen worden op grote schaal visuele, tekst- en tijdreeksgegevens gebruikt om met behulp van DL problemen op te lossen. Voor meer zicht op de mogelijkheden heeft dit verkennend onderzoek antwoord gezocht op vragen zoals: wat betekent DL voor de watersector? Hoe kunnen we van deze nieuwe technologie profiteren? En waar is DL bruikbaar in ons onderzoek of zelfs in onze dagelijkse activiteiten?

Aanpak: wat is deep learning en wat betekent het voor de watersector?

DL draait om het toepassen van neurale netwerkalgoritmen in uiteenlopende omstandigheden. Op basis van een literatuurstudie hebben we veelgebruikte benaderingen bestudeerd, zoals ‘Convolutional Neural Networks’ en ‘Recurrent

Neural Networks', en hoe deze worden toegepast in geofysische wetenschappen, waaronder stedelijke waterstudies. Uit interviews met KWR-onderzoekers zijn we te weten gekomen hoe machine learning in de organisatie wordt gebruikt. En vooral hebben we gebieden besproken waar DL toegevoegde waarde zou kunnen hebben voor de watersector. We hebben twee voorbeelden nader onderzocht, door DL/ML toe te passen op (1) geautomatiseerde verwerking van klantmeldingen en (2) identificatie van microplastics.

Resultaten: kunstmatige intelligentie inzetten voor wateronderzoek en voor klachtenbehandeling van klanten

Uit de twee toepassingscases blijkt de waarde van DL en ML bij het automatiseren van meerdere activiteiten die momenteel handmatig worden uitgevoerd. We hebben aangetoond dat natuurlijke taalverwerking, aangedreven door DL, een effectieve tool is voor het automatiseren van tekstverwerking. Toegepast op een case study met klantmeldingen verzameld door een Nederlands drinkwaterbedrijf, waren de gebruikte algoritmes in staat om de emoties en verzoeken van klanten te begrijpen op basis van de tekstuele beschrijving van de melding. Ook hebben we laten zien dat machine learning, in combinatie met Laser Direct Infrarood-spectroscopie, kan worden gebruikt bij het identificeren van polymeren in watermonsters. In dit geval werd een beperkt aantal gelabelde microplastic monsters gebruikt om meer monsters te identificeren waarvan het type aanvankelijk onbekend was. Deze bevindingen suggereren dat kunstmatige intelligentie het analytisch-chemisch wateronderzoek sterk kan ondersteunen en bevorderen. Voor onderzoekers die polymeren willen classificeren (op basis van classificatiemodellen) of het biochemisch gedrag hiervan willen voorspellen (op basis van regressiemodellen), is dit van belang.

Toepassing: een chatbot voor waterbedrijven

Zoals uit deze studie blijkt kan de watersector naar verwachting in de toekomst veel baat hebben bij DL en mogelijke toepassingen daarvan. Door de snelle ontwikkeling van DL speelt het al snel een centrale rol in projecten met tekst-, audio- en beeldgegevens. We stellen voor dat waterbedrijven denken aan het gebruik van DL wanneer ze geconfronteerd worden met de noodzaak om dit soort gegevens te verwerken. Zo kan DL bijvoorbeeld worden gebruikt om drinkwaterbedrijven te helpen bij de ontwikkeling van een chatbot voor de automatische behandeling van consumentenklachten, wat een vervolgactie op deze studie zou kunnen zijn. Wat spectroscopie betreft, richten we ons in vervolgonderzoek op het minimaliseren van de inspanning van het labelen van monsters door gebruik te maken van actief leren. Ook willen we, wanneer in de toekomst meer gelabelde monsters beschikbaar komen, ons richten op het gebruik van diepe neurale netwerken in modellen met en zonder supervisie. Een beperkende factor in het toepassen van ML en DL voor de watersector is een gebrek aan hoogwaardige gegevens. Daarom moet bij ontwikkelaars en gebruikers van de modellen hier prioriteit aan worden gegeven.

Het Rapport

Dit onderzoek is gepubliceerd in het rapport: Deep Explorations: an explorative study for DL applications in the water sector (BTO 2022.021).

Contents

Meer informatie
dr. ir. Nikki van Bel
T 030-6069516
E Nikki.van.Bel@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

1 Introduction

1.1 Motivation and objectives

Deep learning (DL) is a class of techniques that turns out to be highly effective in applications like image and speech processing. In these applications, DL techniques beat more traditional artificial neural networks (ANN) relying on a limited number of hidden layers or other algorithms such as Random Forest or Support Vector Machine (SVM). DL is a family of machine learning (ML) algorithms based on (deep) artificial neural networks (i.e., multi-layer neural networks). DL is able to derive more complex relations between input and output than other ML techniques. To do this, DL requires more data, which can be a drawback. Because of its successful applications in image and speech processing (e.g., face recognition, auto-translating), DL has been widely applied to many other fields as well.

This study aims to conduct an explorative study to assess the value of deep learning (DL) for KWR and the water sector in general, by investigating different types of deep learning and their strengths and weaknesses for water-related applications. This fits very well within the trend at KWR and the water sector as a whole, where more and more often ML approaches show their potentials to be used to answer (research) questions. To do so, we illustrate the value of DL via two main case studies: (1) datamining in customer complaints received by drinking water utilities about pipe failure data and water quality measurements, and (2) datamining in infrared spectroscopy for microplastics analysis and polymer classification, which can later be extended to other areas, such as chromatography, UV adsorption and pattern recognition in data sets.

With the experience gained with these use cases, DL will become an essential tool in the researcher's toolbox, for solving practical problems of water utilities. As such, it will be recognized when it is a proper tool to address different problems across the water sector. As is also reflected by the two selected case studies.

1.2 Approach and reading guide

To exploit deep learning for the water sector, we first drafted an overview of different types of DL and their strengths and weaknesses. This will allow the selection of topics or problems in which DL has high potential to contribute. This overview is presented in Chapter 2. We provide an overview of deep learning and its applications to different engineering fields, with a particular focus on two commonly used algorithms (namely, convolutional neural networks and recurrent neural networks). Via an interview of five KWR colleagues, their insight into possible applications of DL within the research of KWR and the water sector is summarized in Chapter 3. Furthermore, two cases are selected: (i) One case regards the analysis of text data of customer communications with the water utilities while relating this to other data sources on pipe failures, network maintenance and water quality measurements in the distribution network. From previous research (BTO 2015.024) we learned that by using statistics it is possible to identify incidents in the water distribution network from customer notifications made to the water utility. The idea here is to go a step further, by automatically extracting the topic of the message to increase the number of available messages if necessary and understanding the requests from customers. (ii) The other case regards the analysis of infrared spectroscopy data in the field of microplastics identification. In fact, a large number of data are unknown for their types and only a small number of data have been labelled manually. For this particular type of data analysis, we used *ensemble learning* to solve the problem, i.e., using unknown labels to classify unlabeled samples. In Appendix IV, we also present a case study, carried out by a MSc student during his internship at KWR. The proof-of-concept study aims to test the value of DL models applied to predict river water levels. Finally, the conclusion and recommendations are given in Chapter 6.

1.3 Acknowledgement

We would like to thank Waterbedrijf Groningen (especially André van Toly), Brabant Water (especially Kristina Arsova), WMD (especially Ingrid Folkersma), and for providing data and suggestions for this research.

1.4 List of abbreviations

Abbreviations	Terms	Discussed/used in Chapter
AI	artificial intelligence	2, 3, 4, 5
CNN	convolutional neural network	2, 4
DL	deep learning	2, 3, 4
EL	ensemble learning	5
LSTM	long short-term memory	2, appendix IV
ML	machine learning	2, 3, 4, 5
NLP	natural language processing	2, 4
NLG	natural language generation	4
NLU	natural language understanding	4
RNN	recurrent neural network	2, 4, appendix IV

2 An overview of deep learning and its applications to water research

2.1 A general introduction of deep learning

Deep learning (DL) is an Artificial Intelligence (AI) branch that mimics the human brain's processing of data for application in object detection, speech recognition, language translation, and decision-making. DL is able to learn without the need for human intervention, using both unstructured and structured data (i.e., data that are defined with or without attributes). DL has not attracted much attention until recent years, thanks to the development of hardware and neural network theory. One of the most inspirational applications of DL is Google AlphaGo [1], which stunned the world at the start of the year 2017, by winning 60 online games of go in a row against human professional players.

In recent years, the explosive growth and availability of data and the remarkable advances in hardware technologies have led to the emergence of new studies in DL. DL, which has its roots in conventional **neural networks**, significantly outperforms its predecessors. It utilizes graph technologies with transformations among neurons to develop many-layered learning models. Many of the most cutting-edge deep learning approaches have been shown in a variety of applications, including Natural Language Processing (NLP), visual data processing, visual data processing, speech and audio processing, and many others [1].

Figure 1 shows the timeline of the development of DL from Machine learning (ML) and, even broader, AI. Traditionally, the efficiency of ML algorithms relied heavily on the goodness of **the representation of the input data**. A poor data representation often leads to lower performance compared to a good data representation. Therefore, **feature engineering** has been an important research direction in ML for a long time, which focuses on finding representative characteristics (features) from raw data and has led to lots of research studies. Furthermore, feature engineering is often very domain-specific and requires significant human efforts. Compared to classic ML algorithms, DL algorithms perform **feature extraction in an automated way**, which allows researchers to extract discriminative features with minimal domain knowledge and human effort. These algorithms include a layered architecture of data representation, where the high-level features can be extracted from the last layers of the networks while the low-level features are extracted from the lower layers [1].

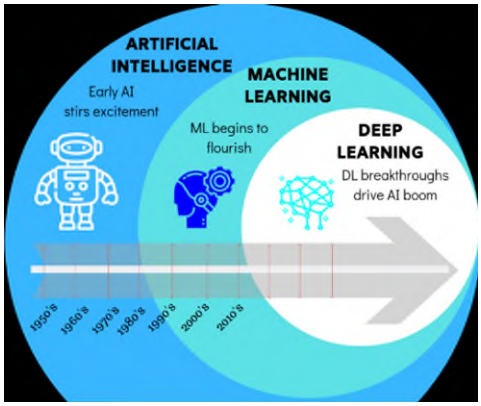


Figure 1. Deep learning is a branch of machine learning. The latter is also a branch of Artificial Intelligence. Source: [2]

The core of DL is a **(deep) neural network** used to link input(s) and output(s), as shown in Figure 2. ANNs, usually simply called neural networks (NNs), are computing systems vaguely inspired by biological neural networks that constitute animal brains [3]. An ANN is based on a collection of connected units or nodes called artificial neurons (nodes in Figure 2), which loosely model neurons in a biological brain. Each connection (lines in Figure 2), like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron receives a signal, processes it, and connects to other neurons. The "signal" at a connection is a real number, and the output of each neuron is computed by pre-defined functions. The connections are called edges. Neurons and edges have a weight that is optimized as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons may have a threshold such that a signal is sent only if the aggregate signal crosses that threshold. Typically, neurons are aggregated into layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times [3]. Numerous ANNs have been proposed in recent research with the purpose of studying a variety of challenges across multiple domains. Below we introduce two common ANNs: convolutional neural network (CNN) and recurrent neural network (RNN). Note that a good understanding of DL requires a strong foundation in mathematics and statistics. To read the following sections, readers can refer to Appendix I for background information. Also, note that we mainly introduce CNN and RNN because they are widely used and successful in practical water applications. Readers can refer to other materials to learn all other popular architectures of NN structures, e.g., autoencoders, multi-layer perceptron, self-organizing maps [4]. An overview of other ANN architectures, used in historical or modern machine learning applications, can be found via this link: https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464.

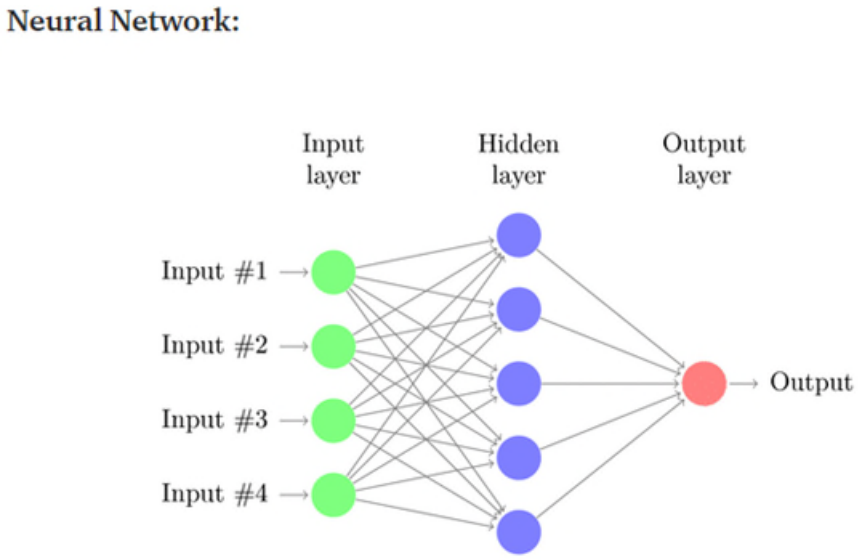


Figure 2. A schematic diagram of ANN. Source: [5]. Nodes stands for neurons and lines stands for connections.

2.2 A short history of DL

DL did not develop at the same rate as it does today until 2006. In the recent decade, as hardware performance has improved, more researchers have begun exploring DL. A brief history of DL is provided in [6], which is summarized as follows:

From Emergence to the AI winter

In 1943, the first mathematical model of a neuron was proposed, aimed at providing an abstract formulation for the functioning of a neuron without mimicking the biophysical mechanism of a real biological neuron. This model did not

consider learning yet. In 1949, the first idea about biologically motivated learning in neural networks was introduced by D.O. Hebb. Hebbian learning is a form of unsupervised learning of neural networks. In 1957, the Perceptron was introduced by F. Rosenblatt. The Perceptron is a single-layer neural network serving as a linear binary classifier. In the modern language of ANNs, a perceptron uses the Heaviside function as an activation function. In 1960s, the Delta Learning rule for learning a Perceptron was introduced by B. Widrow and M.E. Hoff. The Delta Learning rule, also known as Widrow & Hoff Learning rule or the Least Mean Square rule, is a gradient descent learning rule for updating the weights of the neurons. It is a special case of the backpropagation algorithm. In 1968, a method called Group Method of Data Handling (GMDH) for training neural networks was introduced by A.G. Ivakhnenko. These networks are widely considered the first deep learning networks of the Feedforward Multilayer Perceptron type. For instance, A.G. Ivakhnenko used a deep GMDH network with 8 layers. Interestingly, the number of layers and units per layer could be learned and were not fixed from the beginning. In 1969, an important paper by M. Minsky and S. Papert was published which showed that the XOR problem cannot be learned by a Perceptron because it is not linearly separable. This triggered a pause phase for neural networks called the “AI winter.”

Second Wave

In 1974, error backpropagation (BP) has been suggested by P. Werbos to use in neural networks for learning the weighted neurons in a supervised manner. In 1980, a hierarchical multilayered neural network for visual pattern recognition called Neocognitron was introduced by K. Fukushima. After the deep GMDH networks, the neocognitron is considered the second artificial NN that deserved the attribute deep. It introduced convolutional NNs (CNNs). In 1982, J. Hopfield introduced a content-addressable memory neural network, nowadays called Hopfield Network. Hopfield Networks are an example of recurrent neural networks. In 1987, T. Sejnowski introduced the NETtalk algorithm. The program learned how to pronounce English words and was able to improve this over time. In 1989, a Convolutional Neural Network was trained by Y. LeCun with the backpropagation algorithm to learn handwritten digits. A similar system was later used to read handwritten checks and zip codes, processing cashed checks in the United States in the late 90s and early 2000s.

A new era

The year 2016 is commonly seen as a turning point because of a particular neural network called Deep Belief Networks. This particular neural network architecture can be efficiently trained based on a strategy called greedy layer-wise pre-training. This initiated the third wave of neural networks that made also the use of the term deep learning popular. In 2012, Alex Krizhevsky won the ImageNet Large Scale Visual Recognition Challenge by using AlexNet, a Convolutional Neural Network utilizing a GPU, and improved upon LeNet5. This success started a convolutional neural network renaissance in the deep learning community. In 2014, generative adversarial networks were introduced. The idea is that two neural networks can compete with each other in a game-like manner. Overall, this establishes a generative model that can produce new data. In 2019, Yoshua Bengio, Geoffrey Hinton, and Yann LeCun were awarded the Turing Award for conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing.

2.3 Commonly used neural networks of DL

2.3.1 Convolutional Neural Network (CNN)

CNN is a class of deep neural networks, most commonly applied to analyzing visual images. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics [7]. They are commonly applied in image and video recognition, recommender systems, image classification, medical image analysis, natural language processing, brain-computer interfaces, and financial time series analysis.

CNNs are regularized versions of multilayer perceptrons. Multilayer perceptrons usually mean fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer [7]. The fully-connected nature of these networks makes them prone to overfitting data. Typical ways of regularization (introducing a cost to overly complex, presumably overfitted outputs) include adding some form of magnitude measurement of weights to the loss function. CNNs use relatively little pre-processing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand-engineered. This independence from prior knowledge and human effort in feature design is a major advantage.

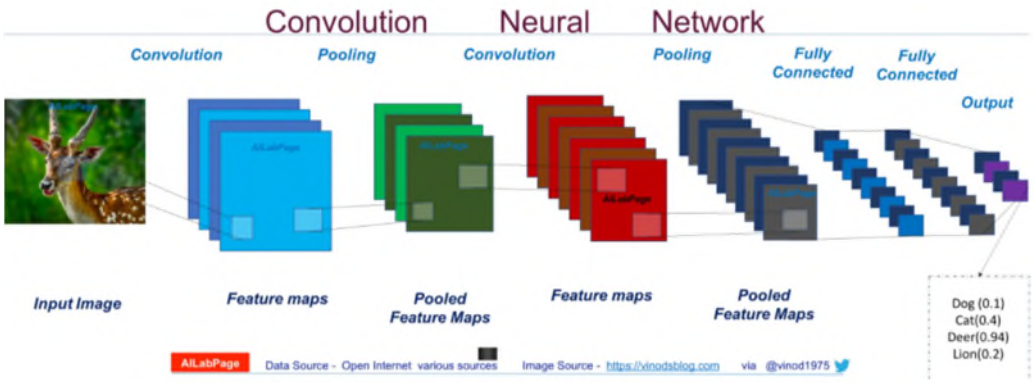


Figure 3: The use of CNN for object recognition in images [8].

In the above figure, on receiving a deer image as input, the network correctly assigns the highest probability for it (0.94) among a few pre-defined categories. The sum of all probabilities in the output layer should be one. There are 4 main operations in the ConvNet shown in the image above: Convolution, Pooling, and Classification (Fully Connected Layer). These operations are the basic building blocks of every Convolutional Neural Network. A more comprehensive example for illustrating CNN can be found via the link: https://poloclub.github.io/cnn-explainer/, where readers can observe how layers are connected from a clear and dynamic angle.

Typical CNN architectures stack a few convolutional layers (followed by a Rectified Linear Unit (ReLU) layer), then a pooling layer, then another few convolutional layers and a ReLU layer, then another pooling layer, and so on. For object recognition, the size of image convolution gets smaller and smaller, in terms of pixels, as it progresses through the network, but it also typically gets deeper and deeper (i.e., with more feature maps) thanks to the convolutional layers [4]. Classical CNN architectures include LeNet-5 (https://homl.info/lenet5), AlexNet (https://homl.info/80), GoogLeNet (https://homl.info/81), VGGNET (https://homl.info/83), and ResNet (https://homl.info/82).

A problem with CNNs is that the convolutional layers require a huge amount of (random access) memory (RAM), especially during training. This is because the parameters in a CNN problems is determined in a backpropagation way, i.e., from end to beginning. Due to this fact, the reverse pass of backpropagation all the intermediate values computed during the forward pass, occupying much RAM [4]. The readers should take this problem into account when applying CNN models.

2.3.2 Recurrent neural network (RNN)

A recurrent neural network (RNN) is a type of artificial neural network (ANN) that allows previous outputs to be used as inputs in subsequent connections. Derived from feedforward neural networks, RNNs can use their internal state (memory) to process variable length sequences of inputs [4]. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition. The term “recurrent neural network” is used indiscriminately to refer to two broad classes of networks with a similar general structure, where one is finite impulse and the other is infinite impulse. Both classes of networks exhibit temporal dynamic behavior. Both finite

impulse and infinite impulse recurrent networks can have additional stored states, and the storage can be under direct control by the neural network. The storage can also be replaced by another network or graph, if that incorporates time delays or has feedback loops. Such controlled states are referred to as gated state or gated memory, and are part of long short-term memory networks (LSTMs) and gated recurrent units. This is also called Feedback Neural Network (FNN). What makes RNNs unique is that the network contains a hidden state and loops. The looping structure allows the network to store past information in the hidden state and operate on sequences (Figure 4).

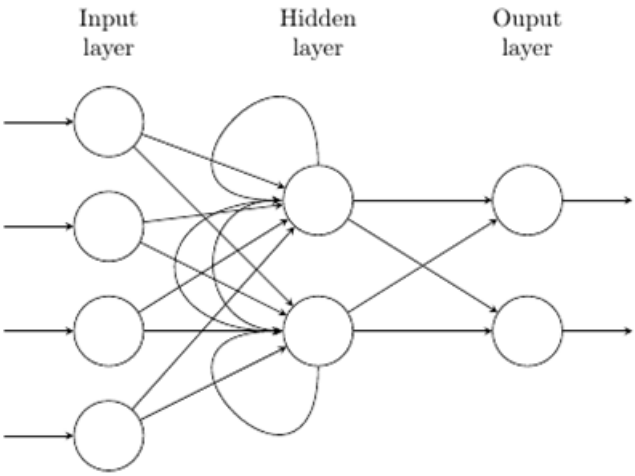


Figure 4: The architect of a simple RNN model [9].

The hidden layer contains a temporal loop that enables the algorithm not only to produce an output but to feed it back to itself. This means the neurons have a feature that can be compared to short-term memory. The presence of the sequence makes them "remember" the state (i.e., context) of the previous neuron and pass that information to themselves in the "future" to further analyze data. Overall, the RNN neural network operation can be one of the three types:

1. One input to multiple outputs - as in **image recognition**, image described with words;
2. Several contributions to one output - as in **sentiment analysis**, where the text is interpreted as positive or negative;
3. Many to many - as in **machine translation**, where the word of the text is translated according to the context they represent as a whole;

To conclude, the above-mentioned applications show the value of RNN in image recognition and text analysis. This relates to the present project, of which one the aims is to extract information from users' reports. We will discuss more practical methods and applications in Section 3.

2.4 Other neural networks

In this report, we focus CNN and RNN because these two could be more potentially important for applications in the water sector. But apart from CNN and RNN, there are also a few NN's that could be useful:

- Generative adversarial network (GAN): Given a training set, GAN learns to generate new data with the same statistics as the training set. For example, a GAN trained on photographs can generate new photographs that look at least superficially authentic to human observers, having many realistic characteristics.

- Autoencoder: Autoencoders are artificial neural networks capable of learning efficient representations of the input data, called codings, without any supervision (i.e., the training set is unlabeled). These codings typically have a much lower dimensionality than the input data, making autoencoders useful for dimensionality reduction. More importantly, autoencoders act as powerful feature detectors, and they can be used for unsupervised pretraining of deep neural networks. Lastly, they are capable of randomly generating new data that looks very similar to the training data, which is doen via the combination with GAN.
- Multilayer perceptron (MLP): An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

The above-mentioned NN's were either developed in the early stage of DL or not applicable yet to water-related problems. Therefore, we do not discuss them in detail in this report. Readers can refer to [10] for more details.

2.5 Data requirements of DL

Figure 5 shows how the classification error of a benchmark problem drops with the increased number of samples [6]. It can be seen that DL demands a huge training samples to achieve a model error of less than 5% (blue dashed line). More precisely, over 25, 000 training samples are typically required (with respect to the benchmark problem). Given the relative simplicity of the benchmark problem which was considered to make Figure 5, results show that a deep learning model cannot do miracles with a small dataset. If the sample size is too small, the method fails. As a result, the combination of a model and data is critical for task completion [3]. Readers can also refer to case study I and II in subsequent Chapters about the data requirement.

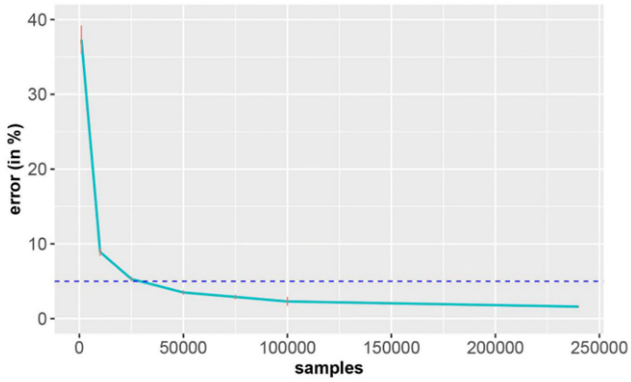


Figure 5. Classification error decreases with the increased number of training samples, implemented on a benchmark dataset. Source: [6].

2.6 Trendy applications of DL

The primary applications of DL are for visual data processing (e.g., images containing objects), audio data processing (e.g., natural languages), and text data processing (sentimental sentences).

I: Natural Language Processing (NLP)

NLP is a collection of methods and approaches that are primarily used to teach computers to intepret human languages. Classification of documents, translation, paraphrase identification, text similarity, summarization, and question answering are just a few of the NLP tasks. The richness and ambiguity of human language make the development of NLP difficult. Furthermore, natural language is largely context-dependent, with literal meanings

changing according to word forms and topic specializations. Recent research has demonstrated multiple successful attempts at reaching high accuracy in natural language processing tasks using deep learning approaches. The majority of natural language processing models begin with a similar preprocessing step: (1) the input text is tokenized, and then (2) these words are reproduced as vectors, or n-grams. It is critical to represent words in a low dimension to ensure that the similarities and contrasts between diverse words are perceived accurately. The difficulty arises when the length of the words contained in each n-gram must be determined. This approach is context-dependent and necessitates domain knowledge.

Sentiment Analysis is a subfield of NLP concerned with studying a text and categorizing the author's feelings or opinions. The majority of datasets used for sentiment analysis are classified as positive or negative. Similarity Analysis is another subfield of Natural Language Processing that analyzes two sentences and projects their similarity based on their underlying hidden semantics. It is a critical feature that is advantageous for a variety of NLP tasks, including plagiarism detection, question answering, context detection, summarization, and domain identification. Because this project involves text analysis of client messages, we go into detail about it in Section 3.

II: Visual Data Processing

DL approaches have become central aspects of numerous cutting-edge multimedia systems and computer vision systems. More precisely, CNNs have demonstrated remarkable performance in a variety of real-world tasks, including image processing, object detection, and video processing [1]. DL techniques play a major role in advancing object detection in recent years. Before that, the best object detection performance came from complex systems with several low-level features and high-level contexts. However, with the introduction of new deep learning approaches, object detection has advanced to a new level of maturity. These advances are driven by successful methods such as Region-based CNN (R-CNN). **R-CNN** bridges the gap between object detection and image classification by introducing **region-based object localization** methods using deep networks. However, training in R-CNN is extremely expensive in terms of computational time and memory, particularly for deep networks. In recent years, this technique has been extended to address the aforementioned issues by introducing two successful techniques: Fast R-CNN and Faster R-CNN [11]. The former leverages sharing computation to speed up the original R-CNN and train a very deep VGGNet, while the latter enables almost real-time object detection. Another commonly used real-time object detection is YOLO (You Only Look Once), which only contains a single CNN. The convolutional network performs bounding-box detection and class probability calculation for each box simultaneously. The benefits of YOLO include its fast training and testing (multiple frames per second) and its reasonable performance compared to previous real-time systems [11][12].

III: Speech and Audio Processing

Audio processing is a method of manipulating electrical or analog audio signals directly. It is required for voice recognition (or transcription of speech), speech augmentation, phone classification, and music classification. Speech processing is a hot research topic due to its critical role in achieving optimal human-computer interaction. Automatic Speech Recognition (ASR) technology has advanced to a new level in the twenty-first century. However, it is still a long way from imitating human behavior to interact with humans. In recent years, interest in speech recognition has expanded beyond the development of an acoustic model for use in ASR systems. A deep RNN architecture, called Deep Speech 2, maintains the robustness of the overall network in a noisy environment. Besides, the approach has shown the capability of quickly being applied to new languages with high-performing recognizers. Initially, CNNs were incorporated into ASR to address the computational challenge. They are, however, exceedingly difficult to train and slow to converge. The core module inside RCNN is the Recurrent Convolutional Layer (RCL), whose state evolves over discrete time steps. Besides speech recognition tasks, many research studies focus on Speech Emotion Recognition (SER), Speech Enhancement (SE), etc. [1]

IV: Prediction of sequential data

Apart from the three application types described previously, DL can also be used to create predictions on numeric time series data [6]. While CNNs are feedforward networks, connections in RNNs can form cycles. This allows the modeling of dynamic changes over time. LSTM is a type of RNN that addresses the drawbacks of RNNs, such as their inability to handle long-term dependencies. LSTMs have feedback connections, in contrast to Deep Feedforward

Neural Networks. Additionally, they are capable of processing not only single data points, such as vectors or arrays, but also sequences of data. As a result, LSTMs are especially well-suited for evaluating sequential data, such as time series [6].

To summarize, a few more practical examples are listed in

Table 1. Overall, DL, a relatively new and rapidly evolving method, presents a variety of obstacles as well as opportunities and solutions in a wide range of applications. While DL can well utilize enormous amounts of data and information for training models because of its architecture, it also presents a black-box solution for many applications. DL's interpretability should be researched further in the future. Especially deep learning requires a large amount of data (ideally labeled data) for training and predicting unknown data. When available datasets are small or when data must be analyzed in real-time, this situation becomes even more challenging. While the majority of existing DL implementations use supervised algorithms, machine learning is gradually transitioning toward unsupervised and semi-supervised learning to handle real-world data without the need for manual human labeling. Note that supervise, unsupervised and semi-supervised refer to a machine learning task aiming to map inputs to labeled, unlabeled, partially labeled outputs respectively. Readers can find these fundamental terms in a general machine learning textbook.

Table 1. Several applications of deep learning in disciplines other than water research [2].

Application domains	Categories	How is DL applied
Social Network Analysis	I + II	The popularity of numerous social media platforms such as Facebook and Twitter enables users to share a wealth of information, including photos, thoughts, and opinions. Due to the promising performance of deep learning on visual data and natural language, various deep learning algorithms have been applied for social network analysis, including semantic assessment, link prediction, and crisis response.
Information Retrieval	I + II	Deep learning has a significant impact on information retrieval; for example, a technique called deep-structured Semantic Modeling (DSSM) has been proposed for document retrieval and web search, in which a DNN performs latent semantic analysis and the queries, as well as click-through data, are used to determine the retrieval results.
Autonomous Driving	II + III	Numerous huge firms and unicorn startups are developing self-driving automobile technologies, including Google, Tesla, Uber, etc,. Recent research has classified autonomous driving systems into robotics approaches for recognizing driving-relevant objects and behavioral cloning approaches for learning direct mapping from sensory input to driving action.
Disaster Management Systems	I + II	Another example is disaster management systems, which have gained considerable attention in the machine/deep-learning community. A well-designed disaster information system can assist the general public and city managers in being well informed of the present threat situation and assisting in the decision-making process. At the moment, the primary obstacle to applying deep learning approaches to disaster information systems is that the systems must deal with time-sensitive data and deliver the most accurate assistance in near real-time. When a natural disaster strikes unexpectedly, a large volume of data must be collected and examined. While there is research demonstrating the use of deep learning in disaster information management, the technology is still in its infancy in terms of data, methodologies, and hardware processing capacities.

2.7 Applications of DL to geophysical sciences

This subsection summarizes the application of DL to geophysical studies, most of which is related to water research, as shown in Table 2.

Table 2. A few DL applications in geophysical research [13] [14].

Application domains	Categories	How is DL applied
Remote sensing	II	In domains that involve water, such as remote sensing, DL is becoming the favored method for extracting information from raw images. Due to CNNs' superior ability to extract information from geometric shapes, textures, and spatial patterns, they readily outperform previous methods that relied solely on spectral signatures or handcrafted features. The primary remote sensing applications include classifying or segmenting images (assigning classes to each pixel in an image based on its content, e.g., land use classes and crop types, and object recognition (finding targets in a series of images).
Climate	II	The number of uses of deep learning in climate science is rapidly increasing, with applications focused on the identification of extreme climate occurrences. Liu et al. (2016) trained a CNN with two convolutional layers to detect extreme events using hundreds of photos of tropical cyclones, weather fronts, and atmospheric rivers. This novel method detects extreme events with an accuracy of 89–99 % and is valuable for assessing climate models(Tropical Cyclones, Atmospheric Rivers and Weather Fronts).
Hydrology and water resources	II	In comparison to several other fields, hydrology has not seen the widespread use of DL. There are now two different types of DL applications in hydrology: (i) analyzing photos for hydrometeorological and hydrologic information, such as precipitation; (ii) dynamic modeling of hydrologic variables or sensor network data, e.g., hydrological variable prediction using LSTM models.

2.8 Applications of DL to urban water resources research

While DL has demonstrated its promise for resolving challenges linked to urban water management and drinking water distribution systems that are challenging using other conventional methods, the applications are far fewer than in other disciplines. To effectively apply deep learning into urban water resources research, we must first increase our knowledge of the underlying mathematical and computational methods/models used by DL algorithms and then re-formulate some conventional issues in a way that DL can solve. To begin, Table 3 summarizes several recent studies on urban water resources.

Table 3. DL applications in urban water resources research [13], [14]

Application domains	Categories	How is DL applied
Recognition of water bodies in urban remote sensing images	II	In [15], a novel deep-learning architecture is proposed for the extraction of urban water bodies from high-resolution remote sensing imagery. First, an adaptive simple linear iterative clustering algorithm is applied for segmentation of the remote-sensing image into high-quality super-pixels. Then, a new convolutional neural network (CNN) architecture is designed to extract features of water bodies from input data in a complex urban background and mark the super-pixel as one of two classes: an including water or no-water pixel. Finally, a high-resolution image of water-extracted super-pixels is generated, with an overall accuracy of 99.14%.
Disaster analyses from crowdsourcing data and citizen science	I + IV	While DL has been rapidly developing, citizen science and crowdsourcing data provide new opportunities for data-hungry DL models [16]. In [17], a study was conducted to reveal the occurrence of urban pluvial flooding based on citizen observations. Using a ten-year dataset of radar rainfall maps and 70,000 citizen flood reports for the city of Rotterdam, authors derived critical thresholds beyond which urban pluvial flooding is likely to occur. The encouraging results suggest that citizen observatories, although prone to larger errors and uncertainties, constitute a valuable alternative/additional source of information for gaining insight into urban pluvial flooding. Although only machine learning models were applied because of the size of the database, we can foresee that citizen observation can greatly enlarge the database and realize the possibility of applying DL models for analyzing disasters.
Leakage detections	IV	In [18], the authors proposed a novel leakage detection model based on density-based spatial clustering of applications with noise and multiscale fully convolutional networks to detect water loss. The leakage detection model is built based on the proposed method to detect the leakage area. Compared with the support vector machine, Naive Bayes Classifier, and k-Nearest Neighbor, the accuracy of the proposed method is improved by 78%, 72%, and 28%, respectively. Therefore, the proposed method can contribute to solving the problem of leakage area detection, improve leakage detection efficiency and reduce water loss.
Water demand prediction	IV	In [19], the authors investigated the potential of deep learning in short-term water demand forecasting, developing a gated recurrent unit network (GRUN) model to forecast water demand 15 min and 24 h into the future with a 15-min time step. The performance of GRUN was compared with a conventional artificial neural network (ANN) model and seasonal autoregressive integrated moving average (SARIMA) model. The results show that the deep learning method improves the performance of water demand prediction. In general, deep neural network models like GRUN outperform the ANN and SARIMA models for both 15-min and 24-h forecasts. These findings can provide more flexible and effective solutions for water demand forecasting. In [20], the authors applied a novel methodology that includes data pre-processing and an Artificial Neural Network optimized with the Backtracking Search Algorithm (BSA-ANN) to estimate monthly water demand concerning previous water consumption. The BSA-ANN model yielded the best result.
Flood prediction	IV	In [21], the authors used a deep belief network (DBN) based on an extreme learning machine (ELM) that is structured by back propagation and optimized by particle swarm optimization algorithm, named DEBP, for flood susceptibility mapping in the Vu Gia-Thu Bon watershed, central Vietnam. Comparing with several well-known machine learning algorithms, including artificial neural network-based radial base function (ANNRBF), logistic regression (LR), logistic model tree (LMTree), functional tree (FTree), and alternating decision tree (ADTree), the new proposed model, DEBP, has the highest goodness-of-fit and prediction accuracy of all of the tested models and thus shows promise as a tool for flood susceptibility modeling.

To summarize, the majority of deep learning applications for urban water research are focused on (1) information retrieval from remote sensing images, which enables us to identify and classify land-use types and provide this information to downstream research, and (2) time series analysis and prediction, which can be applied to water demand forecasting, flood forecasting, and leakage detection, among other applications. It should be highlighted that for most applications, the limited availability of data remains a challenge. Future studies should look into crowdsourcing data to further expand the database. Additionally, many urban water studies do not use audio data and only a few do so with text data. It is currently unclear how these two categories can aid our research. We will examine text data (i.e., citizen reports about water nuisances) and assess the value of text data in this project.

3 Application of machine and deep learning at KWR and their potential for the water industry

As part of this project, we interviewed five KWR colleagues, to gain insight into possible applications of DL within the research of KWR and the water sector. This will serve as input to develop a strategy regarding the use of DL within KWR's research.

3.1 Interviews

The interviewed colleagues, who specialize in different research fields, including water treatment, water quality, water resources, drinking water networks, citizen science, and the interdisciplinary domain – Hydroinformatics. Table 4 summarizes this information.

Table 4. Interviewees from KWR and their corresponding research fields.

Name	Fields of specialization
Dr. Bas Wols	Water treatment; statistical models
Dr. Dirk Vries	Dynamic systems modeling and control; data fusion; data science
Prof. Dr. Dragan Savic	Hydroinformatics; environmental engineering; urban water systems; optimization
Dr. Peter van Thienen	Hydroinformatics; drinking water distribution network modelling; optimization
Dr. Stijn Brouwer	Citizen science; strategic innovation processes; the role of policy entrepreneurs

Although a few questions were prepared in advance (see Table 5), the interview was conducted in a flexible manner, this is, we did not get through each question individually with each interviewee, but rather focused on the subject on which the interviewee desired to impart experience and perspective.

Table 5. List of questions that served as a guide for the interviews.

Topic	Question
Good lessons from past projects which involve ML/DL	1. Did you participate in any project that used DL? If yes, could you give a brief introduction about them? 2. Why did you adopt DL in your projects?
Insight about opportunities and challenges of applying DL in the water industry	3. While DL has been applied in many other domains, do you think it will bring new opportunities to the water industry? In what aspects? 4. Do you think it is also a challenge to couple deep learning with conventional water-related research? What are the bottlenecks? 5. Some researchers hold negative attitudes towards deep learning as it is a black box, which cannot reveal the actual physical process. Do you agree with them or see it from a different perspective?
Implementation	6. In recent years, DL has been greatly developed to understand images and human languages. Do you think that will bring new opportunities? 7. What do you think of citizen science, smart sensors, and remote sensing, which will bring new data for deep learning? Is it something KWR aims to develop a research line in parallel to deep learning?

3.2 Outcomes

3.2.1 Good lessons from past projects which involved ML

Due to the fact that DL did not gain popularity until 2016 (let alone its application in water research), there have been no research projects at KWR that have used DL to solve classification, regression, or clustering problems until now. While DL was explored in certain project ideas, a lack of appropriate data samples frequently hampered its deployment in a variety of real-world situations. To say the least, multiple participants discussed their experiences with machine learning in research projects. The applications span from unsupervised learning (e.g., anomaly detection) to supervised learning (e.g., prediction of flows, water demand, or failures; recognition of vegetation from remote sensing images). Some interviewees stated a desire to include DL in their future research and projects, once sufficient data becomes available, .

Apart from the amount of data, we notice that nearly all (previous) machine learning projects described by interviewees involved numeric data. Given that DL techniques are often built and utilized to handle image/text/audio data or sequence data (e.g., time series), they may not be the best answer for projects using quantitative data. However, in any future project that incorporates images and long-term time-series data, DL will surely demonstrate its strength in classification and forecasting.

3.2.2 Opportunities for the water sector

Machine learning, or artificial intelligence in general, is believed to offer a huge opportunity for the water sector, specifically, to gain a deeper knowledge of various systems and result in more effective and efficient actions. Colleagues also shared their perspectives on how they want to use deep learning in future research and projects, ranging from picture to text data and language processing. Table 6 provides a summary of possible applications and opportunities for the water sector.

Concerning image data, some promising applications are identification of land-use, vegetation, droughts, and leakages, among others, detection of the flocculation in water treatment, and assessment of pipe conditions in drinking water supply networks. To apply DL in such applications, thousands of images, with respect to different conditions, are required to train a good model. Meanwhile, we also need inspection instruments to capture these images, for instance, HD cameras and inspection robots.

Concerning text data, several interviewees mentioned an interesting idea of literature mining the documents and publications stored at KWR's library or content server, or more broadly to mine scientific publications available online (this latter idea coming from a remark that Google scholar is not smart enough) and automatically create libraries on specific topics. In this case, DL can be used to find and list documents according to their relevance to a certain topic or a collection of keywords. To realize it, natural language processing (NLP) is needed to "vectorize" the words used in the literature and measure the distance, as an indicator of similarities, between found materials (papers, conference proceedings, reports, books, etc.). Literature mining could also make our body of work more accessible to the outside world (the idea being that KWR provides a smart literature mining tool and creates comprehensive overviews of available documents). A new two-year project, aiming to implement text mining for highly concerning compounds in water (Project: Impact van zeer zorgwekkende stoffen in het milieu, Project number: 402545-233) was started in early 2021. This gives us the opportunity to explore the added value of DL in this context.

Another interesting NLP application is to improve the customer service of water utilities, by personalizing it. For instance, we expect that DL can help better understand the concerns and needs of customers when they get in touch with the water utility (for instance by phone, email, or chat) by automatically analyzing the spoken words and the possible sentiment behind it, and combine it with information about the client (including historical records from previous contacts and information based on customer profiles), help the customer service employees in addressing the client in a more customized manner. This could be extended even further to automatically reply to clients through

a chatbot, which would allow for 24h service. This last option is not as simple as training a binary classification model. It involves lots of expert knowledge to improve and validate the model. One of the application cases in this project regards the analysis of client messages to the water utilities (see Chapter 4). This personalized approach could also be used in a message regarding water-saving tips. For instance, if customers would use apps or use smart water meters, depending on how water is used and on the customer profile, the water utility could give personalized water-saving tips. Another possibility to improve customer service would be to allow customers to upload a picture of their water meter (tested by some Dutch water companies, e.g., Evides), instead of asking customers to report the meter readings to the water utility, as done now.

Other applications include the evaluation of data quality data gathered by customers in crowdsourcing projects, a time-consuming step that is now done manually; audio processing, for instance, for leakage detection; the classification/recognition of different substances based on their spectrograph (Chapter 0 of this report).

A final example is the possibility to combine reinforcement learning (a particular machine learning approach to making decisions) and real-time control, given that reinforcement learning shares some similarities with Model Predictive Control. The latter is a typical real-time control technique used to operate water systems. Nevertheless, deep reinforcement carries a very large computational burden and it can be challenging to apply to real-time applications, in which systems require prompt decisions.

3.2.3 Concerns

DL will certainly become one of commonly used tools in the researcher's toolbox. However, the application of DL (and again, of ML and AI in general) also comes with challenges. When implementing DL in future applications, the following concerns have been raised by the interviewees:

1. Data availability and quality: DL requires a large amount of data. Moreover, as Andrew Ng suggested, we should move from model-centric ML to data-centric ML [22]. In other words, rather than spending much effort in building DL models, the quality of data is usually the key to influence the model performance.
2. Keeping the results explainable: there is a general need for explainable AI, so we should make the effort to build DL models as explainable as possible. This benefits the understanding of outcomes for both researchers and water utilities. Especially in the field where physically-based models are less developed, DL might help to mimic the process but we still need to find a reasonable way to explain the DL model based on physical processes. Some researchers believe that since DL has several layers of abstraction, it might be a bridge able to reconcile the 'black box' feeling to explainable AI.
3. Ethical concerns: artificial intelligence, including DL, can lead to some ethical issues (Neelke Doorn, 2021), such as the responsibilities of AI models and the protection of privacy. Therefore, we have to consider that '*AI techniques for the water sector should not be left to data scientists alone, but requires a concerted effort by water professionals and data scientists working together, complemented with expertise from the social sciences and humanities*.'
4. Trend breaks: DL models largely rely on historical trends to make predictions for the future, assuming these trends will hold and continue in the future. This assumption might not stand in some cases, for instance, climate change. In these circumstances, we need to be careful about the implementation of DL and be aware of its limitations.
5. Combination of DL expertise and domain knowledge: a good implementation of DL models requires domain experts and their knowledge.
6. Keep in touch with the end-users: it is important to know and understand the needs of end-users, in order to validate the assumptions made for training a DL model and testify the practical applicability of the launched model(s).

Table 6. Overview of possible applications of DL in KWR's research.

Application	Aim	Type of data	Required techniques	Status
Flocculation	To examine the structure of flocs based on their images	Images	Computer vision; Supervised learning	conceptual stage
Pipe condition assessment	To assess pipe conditions based on the images captured by inspection robots (e.g., AIR)	Images	Computer vision; Supervised learning and unsupervised learning	conceptual stage
Literature mining	To build a smart KWR library that can automatically detect similar studies to a selected study	Text	Natural language processing; Unsupervised learning	Ongoing research in the BTO-project " Impact van zeer zorgwekkende stoffen in het milieu" (402545-233)
Client-oriented service	To build a real-time client-oriented service, which can provide 24/7, tailor-made service to each client and detect their emotions based on tones, words, and profiles	Text	Natural language processing	Primary research undertaken in case study 1 of this project/report.
Classification of substances	Classification of different substances based on their spectrograph and machine learning	Numeric (signals)	Machine learning	Case study 2 of this project/report.
Decision making	To combine reinforcement learning with real-time control (e.g., Model Predictive Control) for making a decision-support system for water systems	Numeric	Real-time control; Reinforcement learning	conceptual stage

3.2.4 Overview of ML/DL project at KWR

Main machine learning (including deep learning) applications:
A. Classification B. Regression C. Clustering D. Reinforcement
Data Types:
I. Text (language) II. Image (or Video) III. Numeric (excluding time series) IV. Time series

Projects/ideas	Team	Category	Data type	Use deep NN	Remark
Customer complaint processing	HI	A	I	Yes	based on NLP models
Microplastic identification	HI + CWG	A + C	III	No	Ensemble learning
Prediction of groundwater levels	HI + ECO	B	IV	Yes	Time series analysis
Idea: Reinforcement for water system planning and management	HI	D	III	Maybe	
QSAR	DWB	A/B	III		
Idea: Dimensionality reduction	Several	B	III + IV	Depends	PCA, Autoencoder
Fiware4Water/AI-based Data Validation	WTR/ECS	B	IV	Yes	Simple statistical methods for anomaly detection, LSTM deep autoencoders for reconciliation of wastewater treatment process parameters
Information Extraction task on pathogen characteristics	WKG	A	I+III	yes	based on custom-NER (Spacy) model + unsupervised techniques (regex+ rule-based matching)
Question-answering model on pathogen contamination events	WKG	A	I+III	yes	based on BERT+fine tuning (Squad format) with domain-specific corpora
Digital Twin for district heating system of WarmteStad	ECS	B, maybe D as well	IV	maybe	forecasting time series data, maybe Reinf.L for control/optimization

4 Case Study I: Promoting Automated Complaint Processing for Water Utilities Based on Natural Language Processing

4.1 Introduction

Water utilities often consider customer complaints to be a valuable source of information for identifying system malfunctions and improving their services. Traditionally, complaints are handled by phone operators. The essential information abstracted from verbal communication can only be concisely recorded. As a result, it can occasionally result in misunderstandings and also requires water utilities to maintain an adequate number of telephone operators to handle incoming calls, particularly during rush hours or in the event of a malfunction affecting a larger area. In recent years, an increasing number of (Dutch) water utilities have put an online system in place for customers to easily submit complaints. It enables information to be stored in a more organized and efficient manner. This is a critical step in advancing the digitalization of the water sector since the information provided by customers is commonly more detailed than short-handed transcriptions made by phone operators. Additionally, accuracy significantly improves, particularly for names, addresses, and zip codes. While complaints are now being collected digitally, the content still needs to be processed manually. The main complaint processing tasks include extracting critical information, classifying the major issue, and responding to the customer with a solution.

The primary goal of NLP is to instruct machines to understand human languages and then to assist humans in processing text messages. We demonstrate in this section, using a case study of customer complaints, that an NLP model can understand the syntactic meanings of words, classify a customer complaint correctly, identify the customer's emotion in the complaint, and recognize the intent and request mentioned in the complaint. NLP includes basic processing (e.g., Language processing, lemmatization, lexical and syntactic analysis of words), Natural Language Understanding (NLU, e.g., sentiment analysis, sentence classification, and topic modeling), and Natural Language Generation (NLG, e.g., automatic generation of responses to human languages). Our study is primarily concerned with the task of implementing basic text processing and NLU to investigate and resolve customer complaints about water related nuisance.

4.2 Materials and Methodologies

4.2.1 Customer complaints about drinking water

Our study focuses on the Province of Groningen in the northern Netherlands (Figure 6), where drinking water is managed by the Water Utility Groningen (in Dutch: Waterbedrijf Groningen, referred to as WBG hereafter). WBG is one of only a few Dutch water utilities that have begun collecting customer complaints digitally. Data collection and management have noticeably improved as a result of the deployment of a database and a well-designed website (Figure 7). When submitting complaints, customers need to select a category for the problem, describe the issue in detail, and include their address or customer number so that the water utility can quickly track the reported problem.

Figure 6. The ten water utilities in the Netherlands. The study area of this study is the Water Utility Groningen (in Dutch: Waterbedrijf Groningen) in the north of the Netherlands. Source: https://www.vewin.nl/sector-in-beeld.

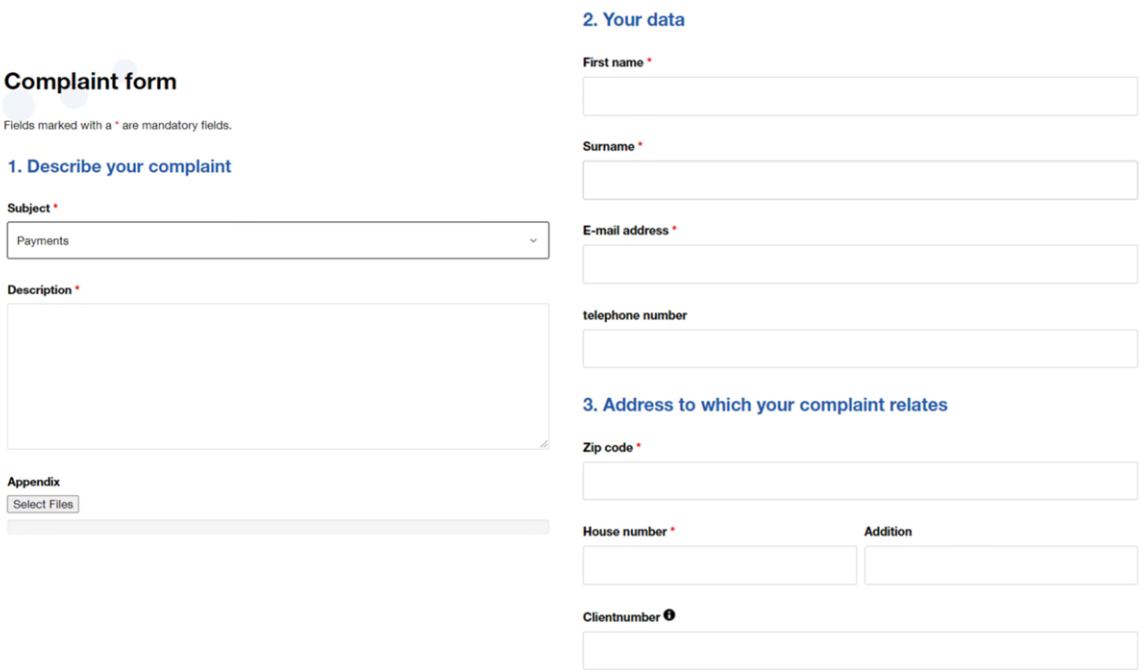


Figure 7. An example of a complaint form designed by the Water Utility Groningen (WBG) for collecting customer complaints. (Adapted from the website of WBG: https://waterbedrijfgroningen.nl/). Subject categories include Payment, Water meter reading, Water meter replacement, Malfunction, Construction of main lines, Moving houses, Maintenance, Connections, Planned projects.

With 597,000 inhabitants spread over an area of 2,960 km^2, WBG produces 44 Mm^3 drinking water per year and distributes it via its 5,000 km-long water distribution network. We retrieved a database of 4,730 customer

complaints, collected by WBG via telephone, email, and webpage between February 2013 and January 2021 (i.e., ≈ 590 complaints per year).

Figure 8 shows the geographical distribution of customer complaints in the province. The majority of complaints came from densely populated cities of the province, such as the city of Groningen (c.a., 231,000 inhabitants) and Veendam (c.a., 28,000 inhabitants).

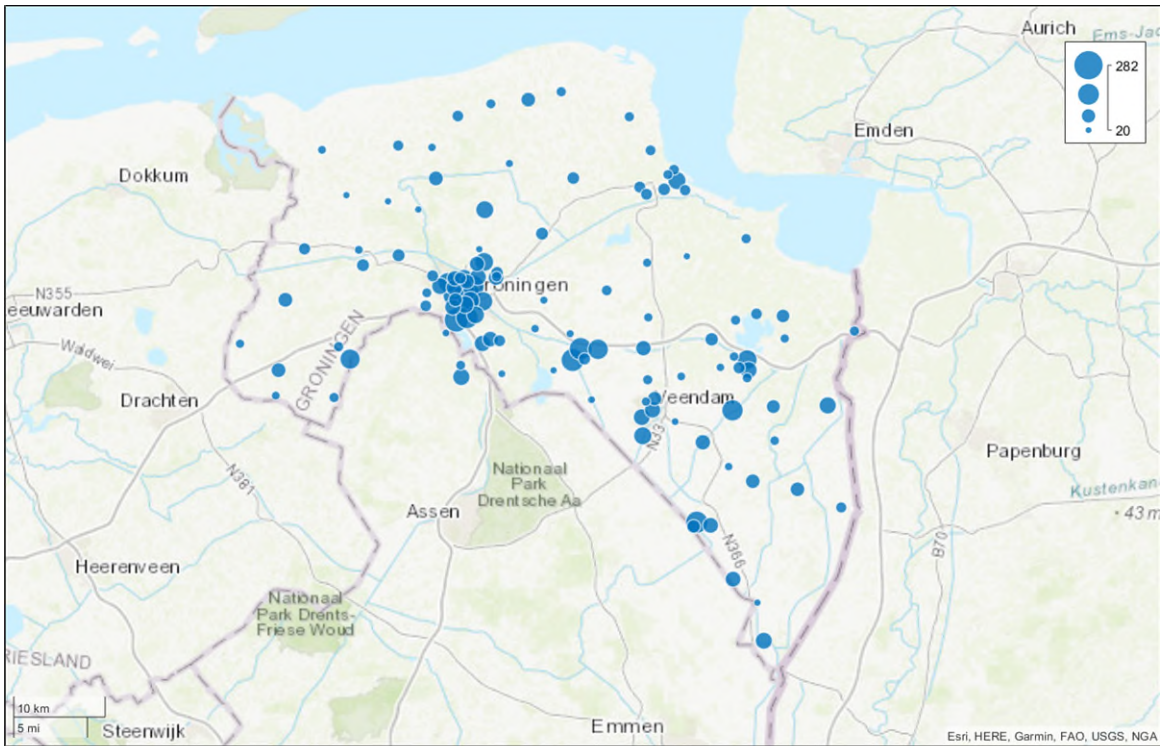


Figure 8. The spatial distribution of customer complaints in the province of Groningen between February 2013 and January 2021.

Figure 9 illustrates some basic statistics about customer complaints: (a) The number of reports received per year reflects the possible range of complaints over years, ranging from 300 to 800. Moreover, it implies that water utilities therefore need to consider flexibility in allocating an adequate number of employees to handle customer complaints. (b) Mondays are commonly the busiest day of the week, with complaints frequently doubling or tripling the volume received on Fridays. This is because non-emergency complaints submitted on weekends are only processed (and thus registered) on the following business day, i.e., Monday or Tuesday (if Monday happens to be a national holiday). (c) December to February received significantly more complaints (50+ per month) than the rest of the year (20-50 per month). Last but not least, we note a peak of approximately 150 complaints (roughly five per day) in August 2018, which was the result of multiple payment issues that month.

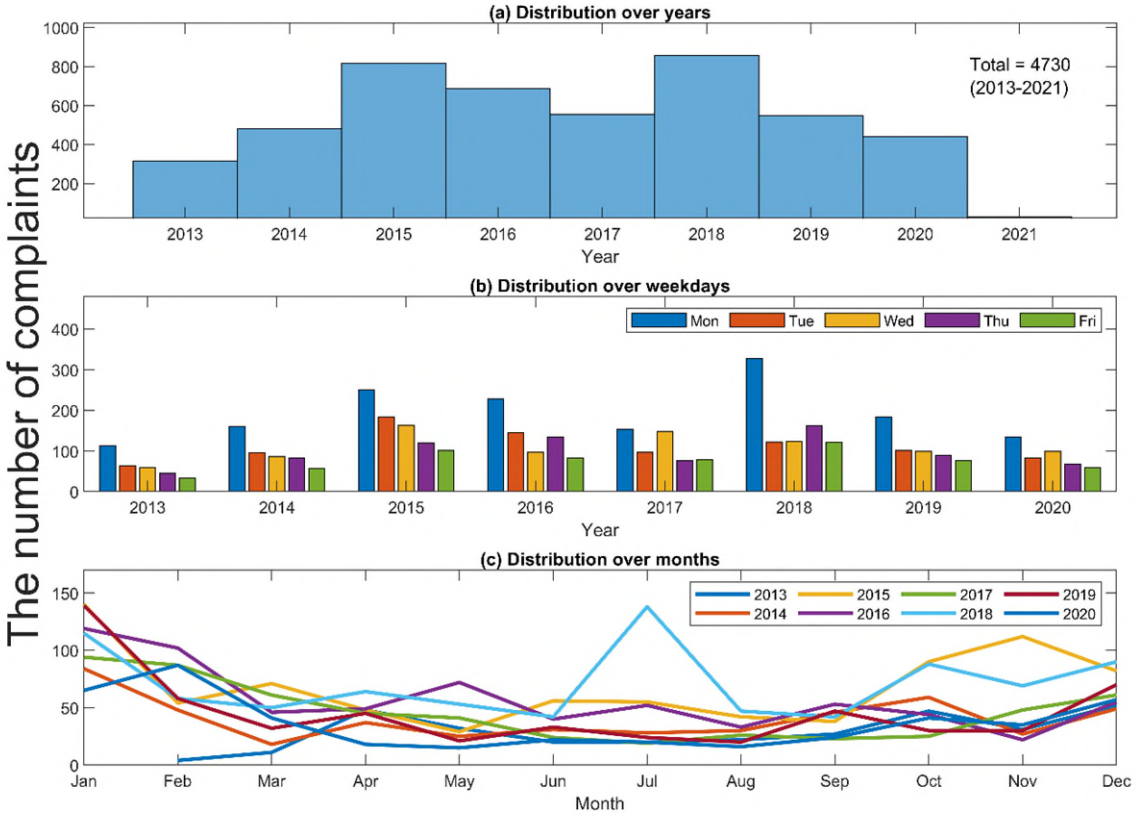


Figure 9. Basic statistics about customer complaints. Data were retrieved between February 2013 and January 2021 so the year 2021 shows a small number of complaints.

4.2.2 Natural language processing

Complaints from customers are a type of natural language data. Regardless of the language in which the complaint is written (in our case, Dutch), linguists and computer scientists have been looking for an interactive way of teaching computers how to understand the content of a conversation and then assist us back in processing and analyzing (colossal) amounts of natural language data in a variety of domains. Natural Language Processing (NLP) emerged naturally as a result of these efforts. Going through symbolic NLP (1950s-1990s) and statistical NLP (1990s-2010s), the state-of-the-art (neural) NLP techniques have entered a new era, benefiting greatly from the rapid development of deep learning and the information explosion. In this study, we only focus on the latest development of NLP techniques and, more importantly, how they can be used to help the water industry automate and simplify the processing of consumer complaints.

Customers frequently include the following information in their complaints: greetings, a description of the situation, and a request for help and support from the water utility. Regularly, the latter two are semantically and contextually related. For instance, a customer reported: *There is a leakage from my pipe. Please fix it asap*. We should be able to infer the pronoun *it* in the second sentence refers to *leakage* from the first sentence and the situation is rather urgent based on the word *asap*. As a result, the following analyses are frequently conducted to apply NLP to comprehend customer complaints:

(1) Lexical analysis

Lexical analysis is the first and most important step in parsing a sentence by first interpreting words. To begin, we need to tokenize words, which involves breaking a sentence into smaller units (commonly referred

to as tokens). Following that, words are lemmatized. In other words, they are transformed into their corresponding word stems. For instance, *am*, *are*, *is*, *was*, *were*, and *will be* all derive from the same root verb *to be*. While these words have distinct spellings and grammatical functions in a conversation, the distinctions introduce complexity when processing texts with identical meanings. As a result, regardless of the grammatical aspect of the word, they are commonly characterized as their stem words. Finally, we also need to remove stop words, which are a collection of frequently used words that have little meaning for textual data (e.g., the, in, a). To process texts in a variety of languages, we need to adopt suitable stop words. In this study, we use an open-sourced stop word list, including about 300 stop words in Dutch [23].

(2) Syntactic analysis

The syntactic analysis aims to parse the functions of words in a sentence based on the lemmatized tokens. It is critical to distinguish between, for instance, *book* as in *book a room* and *book* as in *read a book*. By carrying out the so-called part of the speech (POS) analysis, a trained algorithm typically looks for the type/function of the word. The POS is a linguistic term that refers to the classification of words, which includes nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, interjections, numerals, articles, determiners, and symbols. By doing so, machines ought to be able to determine *book* is a (transitive) verb in the first example (because of the following nominal object) and a noun in the second one (because of the preceding article and verb). After implementing the POS analysis, it is also necessary to parse the dependency of words in a sentence, which is accomplished by examining the word association and the sentence structure. The dependency parsing process enables the algorithm to discern essential information like humans do, such as the root of the sentence, the principal (nominal) subject, and or occasionally, an open clausal complement (i.e., the verb following the modal verb), as well as the descriptive information. In this study, we adopt a word-based dependency grammatic analysis, rather than a constituent-based one. As the names imply, the word-based method makes it easier to express the relationship between each word in a sentence. When conducting a dependency analysis, we typically regard the verb, particularly transitive verbs and their associated objects, as the central component of the sentence (or called headwords). The dependent word, or the so-called child word, functions as an modifier for the headword. Additionally, prepositions and their associated objects also need to be considered, which can also convey important information about time and space. For example, a user may specify that the leakage occurs *in the basement* or *beneath the water sink*.

(3) Vectorization of words

Whereas humans comprehend a language via words, computers comprehend it via numbers. Especially in NLP, word vectors are used to enable computers to understand the language. Word vectors are a series of real numbers representing words as a vector in a high-dimensional geometric space. Word vectors facilitate the arithmetic operations of words. In other words, using the l_2-norm of the subtraction of two word vectors, namely $\left\| \boldsymbol{wv}_1 - \boldsymbol{wv}_2 \right\|_2$, we can easily calculate the geometric distance between two words. The smaller the l_2-norm value, the more semantically similar the words are. For instance, we would expect that *problem* and *issue* have a small l_2-norm value while *problem* and *meter* have a relatively larger value. The word vectorization provides a way to map all the words into a high-dimensional space, suitable for machine learning algorithms to classify or cluster texts. In other words, word vectors are the actual connector between linguistics and AI.

To obtain the word vectors, we can either train a computationally expensive model based on a large-sized self-defined corpus or use a pre-trained model that performs well on a particular corpus, e.g., Wikipedia. With the advancement of DL and NLP models, there are multiple accurate pre-trained models accessible for general use, for example:

- An open-sourced English corpus for general use, which suits many cases for general texts in English [24].
- An open-sourced Dutch corpus for general use, which is adopted in this project, as we aim to process non-scientific and conversational texts [25].
- Biomedical and clinical text [26], which is used in the BTO-project Zeer Zorgwekkende Stoffen (402045/233), in which we aim to process scientific and biochemical texts. Note this specific corpus treats domain words differently from a general-purpose model.

At this moment, few corpora span nearly every scientific topic. Typically, it demands the collaboration of a large number of domain experts to define all the corpus's necessary words. However, given the wide application of NLP, we can anticipate that more corpora for new scientific fields will be available soon.

(4) Discourse analysis

The vectorized words enable us to infer the emotions of customers from their texts, which is also called sentiment analysis. This is a useful resource for water utility staff who are responsible for responding to customers. Specifically, the water utility can also track the total sentiment score year after year to assess the service satisfaction level. There are two ways to implement sentiment analysis. We can manually mark a few examples with a score between -1 (absolutely dissatisfied) and 1 (absolutely satisfied). Then, using the input (vectorized words in a sentence) and output (scores), a regression model can be trained. After fine-tuning the regression model until it is acceptable, it can be used to predict the sentiment score of new complaints. The second way is to employ a pre-trained model, in which words have been annotated with sentimental labels by using a large corpus. For instance, we adopt the open-sourced model and corpus provided by TextBlob, which supports multiply languages, including English and Dutch [27]. The use of a pre-trained model is a preferable way because it lowers the need for skilled staff to label sentiments of thousands of sample complaints.

In addition to sentiment analysis, another key component of discourse analysis is intent recognition. Although not every NLP project involves intent recognition, it is vital when we need to interactively respond to a conversation. Due to the domain-specific nature of intent recognition, we have to specify the categories of training samples manually. With the defined categories, the intent recognition problem is effectively transformed into a typical machine learning classification problem, i.e., predicting which category that a new complaint belongs to.

In this study, we utilize two Python libraries, Spacy and Rasa, for conducting NLP tasks. Spacy is an open-source software library for advanced NLP implemented in the programming languages Python and Cython. Spacy makes use of a CNN model (whose details can be found in Section 2), based on the library thinc [28] and a transition-based approach [29]. Spacy is employed in this work to perform the majority of NLP tasks, except intent recognition. The latter is dealt with by the Rasa. Rasa is an open-sourced contextual AI built on Spacy. Rasa provides flexibility for creating customized and automated interactions between humans and machines, which aligns with our goal of customizing intent recognitions. Rasa (version 3.0+) uses a new state-of-the-art lightweight, multitask transformer architecture for NLU: Dual Intent and Entity Transformer (DIET). Particularly, DIET is a multi-task transformer architecture that handles both intent classification and entity recognition together. Readers can refer to the architecture of the RASA model for more details [30].

4.3 Results

In this section, we present the findings of applying NLP to process customer complaints, aimed at understanding the extent to which machines can automate textual message processing. Note that we use *Italic font* to denote texts derived directly from the original or translated complaints.

4.3.1 Lexical and syntactic analysis of customer complaints

The following example complaint is used in subsequent sections to explain how to process the complaint using NLP techniques. Since the complaints were written in Dutch, we added an English translation where necessary to make the text more comprehensible. However, note that all analyses were implemented on Dutch texts, rather than English ones.

Textbox 1. An example complaint received by the water utitility WBG.

Original customer complaint in Dutch:
Geachte heer/mevrouw, Sinds het wisselen van de watermeter is de druk vrij laag. Vooral met douchen is dit hinderlijk. Hierbij een vriendelijk verzoek om dit probleem te verhelpen.

Translated complaint in English:
Dear sir or madam, since the change of water meter, the water pressure is rather low, especially when taking a shower. Hereby we kindly request you to fix this problem.

We began by conducting a lexical analysis to identify the lemma of each word (see Table 7). The distinction between original words and lemmas frequently occurs with verbs, which have multiple grammartical aspects. For example, the word *is* corresponds to the lemma *zijn* (*to be*). Next, we marked stop words in sentences, denoted as 'TRUE' in the column 'is_stop' of Table 7. Stop words are less important and routinely excluded in subsequent text analyses. Additionally, we also labeled punctuation marks in sentences, such as commas and full stops. This action enables the examination of the sentence structure. In other words, the punctuation marks determine the completeness of a sentence, as shown in the column 'is_punct' of Table 7.

After the lexical processing of words, we processed their dependency and the part of speech (POS) to which the words belong. The POS is a linguistic term that refers to the classification of words (see Section 4.2.2), as shown in the column 'pos' of Table 7. Generally, the POS can be deduced mostly from the context. Note that every algorithm or model has errors. In this example text, 2 out of 32 words were labeled with a wrong POS, namely *wisselen* and *vrij*. *Wisselen* (*change*) can work both as noun (as in our example) or verb while *vrij* (*rather*) functions as an adverb to describe the extent to which the pressure is low. It can also mean *free*, as an adjective, to describe time or a product. The algorithm failed to assess the POS for vrij and wisselen, mainly due to the coverage of similar samples in the training dataset. Although errors are sometime inevitable, they can be further analyzed together with word dependencies, which is discussed in the next paragraph.

Table 7. Original text and tokenized text with lemma, part of speech, stop words and punctuation. Note that the column 'translation' is only added to facilitate reading to non-Dutch readers, the translation is not needed for the lexical and syntactic analysis of the complaints.

Number	Orginal text	Translation	lemma	is_stop	is_punct	pos
1	Geachte	Dear	geacht	FALSE	FALSE	ADJ
2	heer/mevrouw	sir / madam	heer/mevrouw	FALSE	FALSE	NOUN
3	,	,	,	FALSE	TRUE	SYM
4	Sinds	Since	sinds	TRUE	FALSE	ADP
5	het	the	het	TRUE	FALSE	DET
6	wisselen	change	wisselen	FALSE	FALSE	VERB
7	van	of	van	TRUE	FALSE	ADP
8	de	the	de	TRUE	FALSE	DET
9	watermeter	water meter	watermeter	FALSE	FALSE	NOUN
10	is	is	zijn	TRUE	FALSE	VERB
11	de	the	de	TRUE	FALSE	DET
12	druk	pressure	druk	FALSE	FALSE	NOUN
13	vrij	rather	vrij	TRUE	FALSE	ADJ
14	laag	low	laag	FALSE	FALSE	ADJ
15	.	.	.	FALSE	TRUE	SYM
16	Vooral	Mostly	vooral	TRUE	FALSE	ADV
17	met	with	met	TRUE	FALSE	ADP
18	douchen	showering	douchen	FALSE	FALSE	AUX
19	is	is	zijn	TRUE	FALSE	VERB
20	dit	this	dit	TRUE	FALSE	PRON
21	hinderlijk	annoying	hinderlijk	FALSE	FALSE	ADJ
22	.	.	.	FALSE	TRUE	SYM
23	Hierbij	Hereby	hierbij	FALSE	FALSE	ADV
24	een	a	een	TRUE	FALSE	DET
25	vriendelijk	friendly	vriendelijk	FALSE	FALSE	ADJ
26	verzoek	request	verzoek	FALSE	FALSE	NOUN
27	om	for	om	TRUE	FALSE	ADP
28	dit	this	dit	TRUE	FALSE	DET
29	probleem	problem	probleem	FALSE	FALSE	NOUN
30	te	to	te	TRUE	FALSE	ADP
31	verhelpen	resolve	verhelpen	FALSE	FALSE	VERB
32	.	.	.	FALSE	TRUE	SYM

Finally, we applied dependency parsing to extract the grammatical structure of a sentence, i.e., the grammatical relationship between head words and child words. According to the dependency tree shown in Figure 10, the customer began the sentence with a signal adverb *hierbij* (*hereby*), followed by his/her request (*verzoek*), which was described by a determiner and an adjective modifier. Next, the request was followed by detailed information about resolving (*verhelpen*) the problem (*probleem*). By observing the orientation of arcs, the connection of words is presented. Furthermore, we extracted critical information from the sentence by examining nouns and verbs, and sometimes adverbs, and their dependencies. In this instance, the primary structure of the sentence recognized by the model is *hereby* (signal adverb) -> *request* (noun) -> *solve* (verb) -> *problem* (noun). This is a key step to extract the underlying meaning of a sentence by machines, which is also necessary for subsequent analyses, such as comparing sentence similarities or evaluating sentence sentiments. A detailed list of dependency labels can be found in the following reference [31].

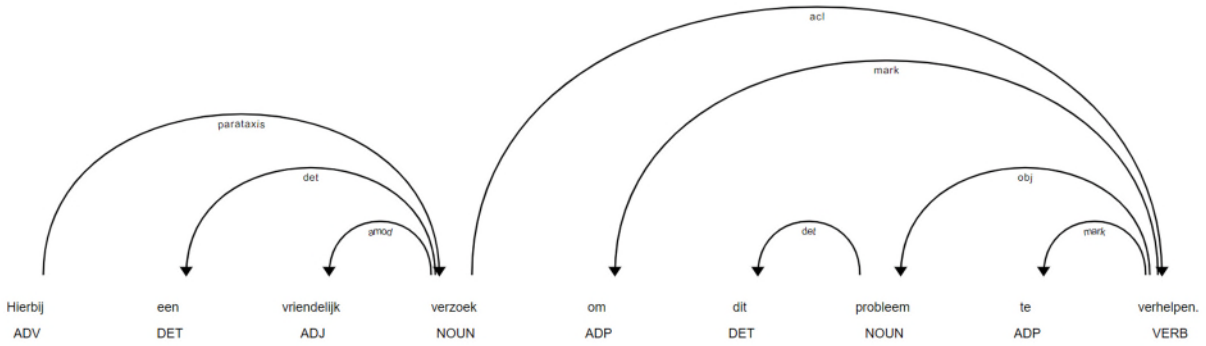


Figure 10. A dependency tree of a customer complaint, which kindly requested the water utility WBG to resolve the low water pressure issue. The dependency tree depicts the grammatical relationship between words in the sentence. English translation of this sentence: Hereby a friendly request to solve this problem.

4.3.2 Similarity and sentiment analyses of customer complaints

As introduced in 4.2.2, word vectors are a critical component to enable machines to interpret natural languages and help humans with text processing automation. We vectorized words into 300-long vectors using a pre-trained NLP model for Dutch (also see 4.2.2).

The word vectors first enable the calculation of the complaint similarity, which is mathematically defined as the geometric distance of two vectors mapped in the 300-dimensional space. Based on the word vector, one can easily find similar complaints in the complaint database, one of the similar examples being shown below in Textbox 2. This similar complaint has been previously addressed by a WBG employee so the response to this complaint can be used to reply to the example complaint (which assumes to be a new complaint needing WBG staff to process). In other words, when responding to new complaints, we can relate back to formerly enclosed cases that were similar to the new cases. In doing so, it significantly increases the efficiency of complaint handling to a semi-automated level. When a significant number of processed instances accumulate, Natural Language Generation (NLG) models are expected to produce answer emails automatically. Although NLG is not included in this study, we explore it as a possible future research direction in 4.4.

Textbox 2. A complaint from the database, which is similar to the example complaint shown in Section 3.2.

Translated complaint in English:
Dear Sir / Madam,
We have been living at xxx since October 2020.
What strikes us about the drinking water supply compared to our previous home (xxx) is that the water pressure is much lower and that the capacity also leaves something to be desired. When two or more users used simultaneously, the pressure dropped sharply. For example, taking a shower and at the same time a toilet being flushed or the washing machine is turned on. Can you indicate whether this can also be remedied?
I would like to hear from you.
Yours sincerely,

Figure 11 shows the distribution of sentiment scores, which is between -1 and 1 based on the Textblob model, for all complaints. It is roughly in a Gaussian distribution shape with a slightly longer tail in the negative part (score < 0) and a high peak at the value 0.1. It shows that the majority of customers used a neutral tone when reporting water nuisances in complaints. By comparing positive and negative tones, we discover that the number of negative voices only slightly outnumbers the number of positive voices. However, a small number of customers expressed great dissatisfaction in their complaints (e.g., sore < 0.75). This is frequently the result of an unresolved earlier issue or an unacceptable bill of a large amount. Water utilities may need to pay close attention to these reports to assess whether the case is well resolved and can be enclosed or whether any of their services can be improved. When we

have a longer history of complaints (e.g., > 10 years), we can also study the interannual trend of sentiment scores and whether the trend meets the expectation of the water utility for their service improvement.

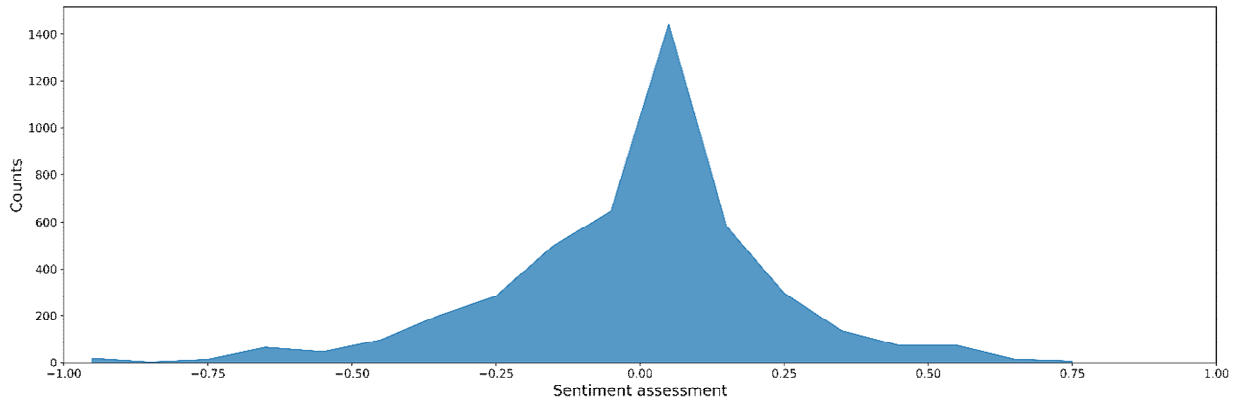


Figure 11. Sentiment score distribution of collected complaints.

4.3.3 Intent recognition

For intent recognition, a few examples for intent categories need to be manually marked prior to model training. Textbox 3 lists selected examples of the labeled categories we used to recognize intents in this study. Appendix I gives more details for each category (8-24 examples for each category). We included five categories, namely, (I) repair and replacement, (II) payment, (III) leakage, (IV) low pressure, and (IV) water quality. Consider the example complaint (Textbox 1), it falls on the fourth category with a probability greater than 99%. Therefore, we can infer that the example complaint intends to report and fix a water pressure issue. In real-world application, we also need to set an acceptance threshold (e.g., 80%). Once the maximum probability for prediction is lower than the acceptance threshold, we should reject the prediction and regard it as an outliner (not belonging to any pre-defined category). We used 5-folder cross validation to test the model performance. The test accuracy, precision, and F1-score were 0.825, 0.802, and 0.79 respectively.

The majority of collected complaints can be classified for their intents by NLP. Uncategorizable complaints often concern a minor topic, which is not listed above and lacks a sufficient number of examples to support it (e.g., a customer submitted a letter with an apology for using an aggressive tone in an earlier complaint). On the other hand, we did not consider processing a message involving multiple topics in this study. This is because our model does not aim to solve a multi-classification problem which could deteriorate the performance for predicting a single-topic problem.

Textbox 3. Some selected examples for intent recognition.

We only show English translations below while more original Dutch texts can be found in Appendix I. Note that the problem statement may sound not natural to a native English speaker due to the language difference. However, we attempt to keep the original word order and show their direct translation below. Also note that these examples can be complete sentences or just a semantic component (e.g., noun phrases and verb phrases).

Training dataset I – request for repair or replacement

1) Please replace the meter.
2) With this letter, I want to make an appointment for the installation of a new water meter.
3) would like to have a new meter installed

Training dataset II – request for investigation about payment

1) I would like to see this refunded to my account
2) This is a repair of the water company itself, so these costs are not on me.
3) For the above reasons, I hereby object to the imposed administration costs and do not agree to payment.

Training dataset III – request for investigation of water leakage/noise

1) Our water meter beeps very annoyingly.
2) I have a lot of water at the meter behind the front door.
3) There is also a leakage from the main water tap.

Training dataset IV, request for investigation of water pressure issue

1) a fairly low water pressure
2) reduced water pressure
3) We currently have no water.

Training dataset IV, request for investigation of water quality issue

1) The water from the tap was brown and sandy.
2) As of today, tap water smells and tastes strange.
3) The water has a strange metallic (copper) smell.

The recognized intent can significantly assist water utilities in processing and tracking the reported issue. According to the intent categories listed in Textbox 3, the issue with a recognized intent can proceed to a particular department within the water utility, which is in charge of this type of issue, significantly reducing the amount of time spent by staff who must perform this task manually.

We do not attempt to list the contents of all complaints in this report because intent recognition is a practical task in NLP for which each project needs to be handled particularly (e.g., about the definition of intents). In addition, there are no generic solutions that can be applied in all circumstances. Instead, algorithm and model developers must customize models practically based on the actual requirements of their water utilities. For instance, many water utilities are also responsible for flood risk management and need to address complaints about flooded/damaged properties. In this instance, a category pertaining to flooding and a sufficient number of associated examples must be added in the training dataset for an intent recognition model. In contrast to the previous tasks in this study (sentiment analysis), for which a pre-trained model can be used, intent recognition requires building a new NLP model from scratch. To enhance the model performance and adapt it to new topics, new texts need to be iteratively added to retrain the model until it achieves satisfactory performance for each defined category.

4.4 Discussion

4.4.1 How can water utilities benefit from the latest NLP techniques

Natural Language Processing is a field of research that largely relies on DL to enable computers to understand natural languages. This study demonstrates the use of two important components of NLP, namely language processing and understanding. Using the latest NLP approaches, we can extract the grammatical structure of a phrase, identify the most meaningful parts of speech (e.g., transitive verbs and their objects), and even utilize deep learning models to distinguish intents derived from texts. A critical component that enables this is the vectorization of words or the conversion of words into machine-readable numeric arrays. The feasibility of the words to be well-mapped into a high-dimensional space is dependent upon the selected corpus. Because we aimed to process conversational texts rather than scientific terms, this research study adopted a large general-purpose corpus trained on Wikipedia. Within a particular field of research, readers must either create their corpus or seek publicly accessible corpora.

NLP can potentially reduce the time for manually processing complaints. For instance, using the machine-aided information, a water utility employee who is responsible for email replies can quickly grasp the main topic of the complaint by only checking summarized key information, understand the main request of the complaint by reviewing the recognized intent, and respond by referring to a similar complaint that has been earlier processed. However, we should also notice that it also demands considerable efforts in collecting samples and feedback from users during the process of training the NLP model. Water utilities may need to consider the tradeoff between the effort in training a model and the benefit of using a trained model.

At this stage, we can only partially automate complaint processing because we still require extensive feedback from water utility staff on how the NLP model performs in practice. However, this study demonstrates a beginning point for considering the use of NLP model to replace unnecessary repeated tasks conducted by humans.

4.4.2 How will water utilities benefit more from NLP in the near future

We can anticipate that Natural Language Generation (NLG) will complete the cycle of NLP for facilitating interactive communication between customers and machines in the near future. Although NLG is not investigated in this study and is still being developed in NLP research, we can expect that it is a beneficial tool in assisting us in truly understanding different situations. The real understanding of a problem also demands information from customers, often more than once. Therefore, a chatbot, equipped with Natural Language Understanding and NLP, is more effective at gathering detailed information. Textbox 4 shows an example of the use of NLU and NLG to help water utilities automate customer complaint collecting, processing, and responding. In this example, we demonstrate that NLU can comprehend key information such as the time and the address, while NLG can generate messages confirming information, inquiring about the nature of the problem from the customer, or declining the proposed time. This is a way to standardize the complaint description by including precise time, location, problem narrative, and other relevant activities. Additionally, we can see from this example that lots of (repeated) work can be performed by machines, which substantially eliminates the need for human labor. However, for the entire system to function properly, we also need the support of other databases and models. For instance, we need a database that stores user profiles (with personal information securely maintained by water utilities) to quickly identify recent maintenances that can lead to this reported issue. Besides, we also need real-time information about ongoing or planned projects in the management area so that we can track whether the problem is related to the project and the number of customers affected in the same area. Additionally, if a frequently asked questions (FAQ) database has been established on historical complaints and responses, the NLP should be able to search for a query that is similar to a previous one and reply referring to the earlier response. As a return, the FAQ database can also be updated automatically with new questions that often appear. Finally, a model should determine the nature of the problem and find the appropriate technician from the internal employee staff database who is responsible for maintaining the specific type of work for the reported particular area. Textbox 4. An example of using NLU and NLG to help water utilities to process customers.

Customer

Hello there, I would like to report an issue about the low water pressure.

[NLU identifies the main object of the sentence - water pressure and the intent for reporting issues.]

Smart Assistant

Hello Mr./Ms. xxx, thanks for contacting us about the water pressure issue.

We would like to help you solve the problem as efficiently as possible. When did you start encountering this problem?

[The model finds the customer profile from the system and NLU generates a question to obtain detail about time.]

Customer

Since yesterday.

[NLU identifies the time.]

Smart Assistant

As you have logged in our system, we find your address is xxx, xxx, xxx. Is it correct?

[The model finds the customer address from the system and NLU generates a question to confirm.]

Customer

Yes.

[NLU identifies the intent of confirmation.]

Smart Assistant

There are no ongoing constructions in your district to cause the problem.

From our record, we see that a new water meter was installed in your kitchen last week. Is it correct?

[The model looks for surrounding projects and previous maintenance and NLU generates a question to confirm.]

Customer

Yes, on 1 July.

[NLU identifies the time.]

Smart Assistant

Have you tried to fully open the valve of the new meter?

[NLU looks for a solution from the pre-defined questions.]

Customer

Yes, I did but not helping at all.

[NLU identifies the intent of denial.]

Smart Assistant

We would like to send a technician to your address to investigate the problem.

Could you tell us the preferred date and time, please?

[NLU looks for a solution from the pre-defined questions.]

Customer

Any Friday, in the morning.

[NLU identifies the date and time.]

Smart Assistant

An appointment has been generated for 10:00 am on 9 July. Could you confirm it, please?

[The time management system looks for a suitable time based on the agenda of the technician and NLU generates a confirmation.]

Customer

OK.

[NLU identifies the intent of acceptance.]

Smart Assistant

The appointment is confirmed. You will receive an email in about 5 minutes with details of your appointment.

System messages: Thank you for contacting us! Have your problems been solved?

Yes and close the conversation window. | No. Contact an assistant.

[If yes, the case is to be recorded in the database and used as a reference for similar issues or the case with the same customer.]

Nonetheless, we also need to accept that NLP is not a perfect solution for all problems. Due to the complexity of natural languages, expressions may be difficult to comprehend even for native speakers, let alone machines. As a result, we continue to require assistance from experienced staff capable of resolving complex difficulties. As shown in Textbox 4 and Figure 12, we also need to include a part inquiring about the satisfaction of a customer with the automatic response. If not satisfied, it implies the NLP model encountered a case with fewer examples in the past to enable the model to cope with new cases. Therefore, a human assistant can excel at this task. This process is also referred to as human-in-the-loop or active (machine) learning. Theoretically, the more instances we use to train the model, the less expert assistance is required. But it is evident not all future situations are comparable to previous ones. Thus, it is always advisable to have a human assistant available for 'unexpected' cases during the early stage of implementing NLP models, but they only need to deal with a decreasing percentage of cases as the model matures. Human-addressed instances can be stored in the FAQ database, becoming new samples for training the model. Last but not least, this system can also be used to instruct new water utility employees to familiarize business issues and standardize responses to customers.

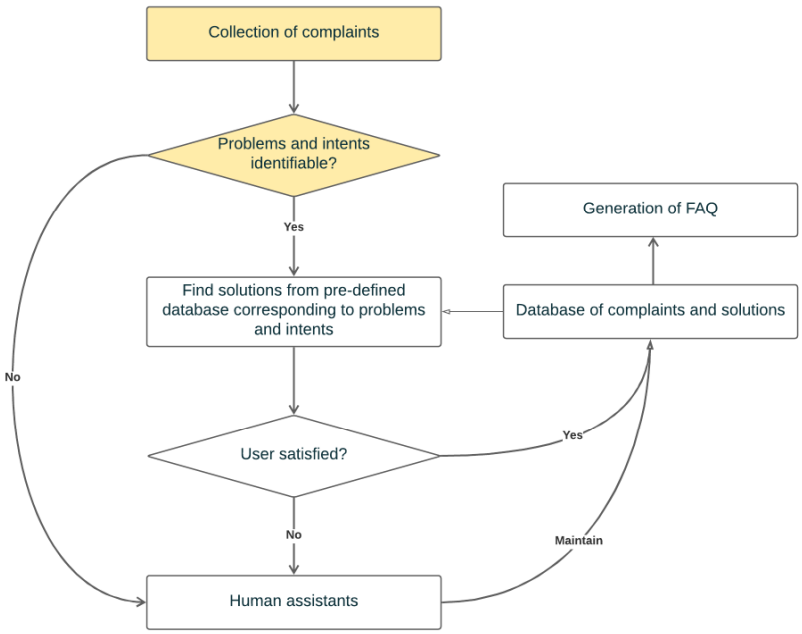


Figure 12. The big picture of using NLP for automating the complaint processing for water utilities. The blocks in yellow indicate what is studied in this research while others are part of future research.

4.4.3 Bottlenecks and limitations

We demonstrate in this report that when AI and DL are integrated with linguistics, the field of machine-aided NLP is formed, which can be used to cope with tasks involving natural languages. NLP can significantly reduce the amount of work that must be repeated by humans. However, NLP or DL is not a panacea. The critical aspect of applying NLP is to ensure that the model is clear and transparent to end-users (e.g., water utility staff).

Each sector has its lexicon and terminologies, which implies that model developers should therefore modify their NLP models for each situation. This stage, however, demands considerable time and effort for domain experts to define these vocabularies. Too few terminology instances may degrade the NLP model performance, resulting in a barrier for NLP applications in water research. Finally, it is important to collect feedback from customers regarding occasions when the NLP does not function properly. These particular circumstances require additional analysis by (experienced) water utility staff, which accordingly takes time and effort too. Given that machine learning improves as a result of massive and quality examples, data-centric NLP models, rather than model-centric models, are the key to making NLP perform successfully in real-world applications [22].

4.5 Concluding remarks

We demonstrated that natural language processing, as an interdisciplinary field combining linguistics and deep learning, is an effective tool for automating text processing. When applied to a case study involving customer complaints about urban water-related nuisances collected by the water utility WBG, NLP models were able to carry out many tasks that must be repeated by humans. The analyses yielded the following conclusions:

(1) Via a lexical analysis, NLP is capable of removing unnecessary words and symbols and determining the lemmas of essential words. This implies that current NLP techniques are capable of analyzing words in sentences. Lexical analysis is the most effective method for tasks that require extracting essential words from a sentence.

(2) Via a syntactic analysis, NLP addresses the relationship between words as well as the functionality of words. Therefore, syntactic analysis is more useful when attempting to extract the dependency of words. In other words, it enables us to extract grammatic constituents from the sentence, which is useful, for example, for identifying the detailed information of a problem or a request.

(3) Both the lexical and the syntactic analyses can be undertaken manually by humans. However, recent advances in deep learning enable machines to perform these activities in place of humans. The key component is word vectorization, which represents words as numerical vectors in a high-dimensional space. Using machine-readable vectors, similar information can be detected, emotions can be analyzed, and more importantly, intents can be recognized.

(4) This study presents a fundamental investigation of applying NLP to automate text processing for the water sector. With the anticipated research outcomes of NLP in the future, NLP will be able to further and deeper automate text processing, including text generation, interaction with humans, connection to digital databases, and so on.

(5) As with other deep learning models, the main constraint on implementing NLP is still data availability. To maximize the potential value of NLP, model developers and users should prioritize the collection of high-quality data and the use of active learning to progressively train the NLP model with new data.

5 Case study II: Implementation of LDIR and machine learning for the identification of environmentally exposed polymers

5.1 Introduction

Polymers, which in include microplastics, have been detected in oceans, rivers, lakes, bottled water and at considerably lower concentrations also in drinking water [32]–[36]. Research about the occurrence of polymers in water, especially drinking water, is highly relevant given that polymers are explicitly mentioned in regulations such as the new European Drinking Water Directive (DWD), which was adopted in December 2020 and came into force on January 12, 2021 [37], [38]. The EU mandates that Member States have to implement the provisions of the revised DWD into national laws and regulations by 12 January 2024. The directive aims to address growing public and scientific concern about the effects of polymers on human health, in particular if exposure occurs through ingestion of water intended for human consumption and to address the risks involved. This means that polymers will have to be monitored and risk assessments made. Fast and reliable analysis methods and knowledge about the occurrence of polymers are hence crucial.

Several optical techniques exist to detect and characterize polymers, such as Fourier Transform Infrared (FTIR) andRaman spectroscopy [39], [40]. These techniques rely on databases containing reference spectra to which acquired particle spectra are compared. Reference spectra are generally collected from pristine polymers. However, particles detected in environmental samples are seldom pristine and have often undergone more or less extensive weathering due to environmental conditions (e.g., UV-light, oxidation). Weathering will affect their physicochemical properties, and hence will influence their interaction with infrared light. Ultimately, this will translate in modified spectral profiles, which can potentially lead to misidentifications [41]. Compiling a database with spectra of polymers with different degrees of weathering would be extremely time consuming and is not a viable option. Therefore, alternative approaches combining more rapid analytical techniques, as well as more advanced data analysis approaches, which take advantage of existing data from both pristine and weathered samples, are necessary.

LDIR has recently been introduced as an alternative technique for the analysis of polymers in environmental samples. Compared to approaches such as FITR, it has the advantage of scanning the surface area, that contains the sample, with at a single frequency prior to initiate a full frequency scan. As a full frequency sweep is time-consuming, this will safe time as now only area's with actual particles will undergo a full frequency scan.[42]. This has obvious advantages in terms of analysis time, as spectra will be acquired when particles have been detected during the preliminary scan. Samples that are almost void of particles are therefore measured considerably faster with LDIR. However, the infrared band recorded with LDIR instruments is narrower (ca. 1800 to 900 cm^{-1}) [43]. As less information is collected of a particular particle, LDIR is more prone to misidentification when analyzing weathered particles compared to FTIR [43]. Therefore, more effective classification methods are required to minimize the risk of misidentification. Machine learning (ML) can be a way to more reliable classification approaches; yet few examples of their implementation in the field of polymer identification exist.

ML, particularly its sub-branches deep learning and reinforcement learning, has greatly gained popularity, becoming an essential tool for various scientific fields and industrial sectors, including biomedical engineering and water research [13], [44], [45]. In a nutshell, ML aims to build a model architecture to learn a pattern from input to output.

By optimizing model parameters using a training dataset and evaluating model performance using a validation dataset, trained models are obtained that can be applied to new datasets [4]. Model inputs and outputs can be either numeric or textual. In practice, ML is classified based on the type of supervision utilized during the learning process. When the output of a dataset has been manually labeled prior to training, supervised learning is often used to determine which inputs are the major factors (also known as features) that contribute to the labeled output. For example, if multiple types of polymers and their related features are gathered, a supervised model seeks to discover a pattern in which type A is determined by one collection of features whereas type B is determined by another collection with different features or feature values. However, supervised learning demands considerable effort in labeling samples. When no labels are available, unsupervised learning is utilized to cluster samples. In practice, one often encounters a mixed dataset with a small number of labeled samples and a large number of unlabeled samples, prompting the use of active learning [4]. Active learning first seeks to develop a supervised model using labeled data and then apply it to classify unlabeled data. Next, the samples for which the model prediction is most uncertain are sorted out and to be handled by experts. The labeling effort iterates until the model achieves a desired level of model performance. In doing so, it is possible to use a relatively small number of labeled samples to find the type of polymers from the unlabeled samples.

In this case study, we developed an effective approach to identifying environmentally exposed polymers using a combination of LDIR and ML. Available databases of spectra of polymers and other plastics were collated to a large dataset (81k samples) of spectra of environmentally exposed polymers collected from surface and drinking water across the Netherlands. Using the available labeled examples as input, we trained a supervised ML model to build a classifier to identify polymer types based on their spectra, which was then applied to unlabeled samples. The latter were accepted and labeled with their appropriate types, only when the classifier determined their types with a probability larger than 70%. Finally, an unsupervised ML algorithm was used to cluster unaccepted samples and domain experts could determine the type of cluster by examining a small number of representative samples from each cluster. For now the model uses spectra of 10 different types of polymers.

5.2 Methodology

5.2.1 Machine learning models

Prior to introducing the ML models, it is important to specify the input and output. In this study, 1,650 features are used as input, which corresponds to absorption rates on wave numbers ranging from 975 to 1800 cm^{-1} with a resolution of 0.5 cm^{-1}. Moreover, due to the uneven distribution of labeled and unlabeled samples, we combined several supervised and unsupervised models to predict polymer types (Error! Reference source not found.). Note that we use the words labeled and classified to represent polymer types that are manually annotated and determined by the algorithms, respectively. Figure 13 shows the ML models adopted in this study, which comprises of the following components:

- Data pre-processing: We pre-processed the data for better training the models. First, samples whose maximum absorption value across all wave numbers was smaller than 0.01 cm-1 were removed as they likely involved measurement of non-polymeric particles (e.g., air bubbles). Next, the values were standardized to their mean value and unit variance of all observations of each wave number. In other words, normalized features were used to train the models.

- Training a classification model: Specifically, a supervised model (ensemble learning) was developed, aiming to classify 223 labeled samples into 11 pre-defined polymer types. Two commonly used classification models were trained and compared, including (random) subspace k-nearest neighbor [46], and boosted decision trees [47], i.e., Model #1 and #2 in Table 8. To achieve the optimal model performance, their hyperparameters were tuned by implementing a grid search for (i) the learning rates (i.e., the step that an optimization algorithm takes to minimize its loss function) and (ii) the number of weak learners (i.e., a

predictor that performs relatively poor and is aggregated with other predictors to generate a well-performing predictor). To evaluate the performance of the supervised models, the classification accuracy (= 1 - re-substitution error) is used as the performance indicator. The accuracy of a classifier is defined as the ratio of correctly identified samples to the total number of samples. Besides, to prevent overfitting the model, 10-fold cross validation was also performed, which was conducted by randomly partitioning the training dataset into 10 equal-sized subsets and computing the mean value of the accuracies of all subsets. In other words, total performance was measured as the accuracy of the cross-validated model based on ten subsets. Note that cross validation provides a rough estimate of the accuracy when applying the trained ML model to new data.

- Application of the classification model: The trained classification model was applied to all unlabeled samples. Particularly, the classification model computed the probability that a sample fell into a given category (i.e., polymer type). As a result, a minimum acceptance criterion (e.g., 70% or 85%) was specified, over which samples were considered classifiable and below which they were considered (model) outliers not belonging to any of the 11 pre-defined classes. In other words, outliers are samples whose classification model cannot detect their types.

- Training a clustering model: Outlier samples are still unlabeled and require additional analysis using unsupervised learning, for which, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) model was used [48]. The DBSCAN algorithm (model #3 shown in Table 8) is an efficient method for grouping data with similar features and meanwhile recognizing noises, i.e., independent sample points not belonging to any cluster. In doing so, samples with similar spectra patterns can be identified in the same cluster.

- Labeling the clusters: Finally, the central points of each cluster can be manually assessed by an expert operator to determine the type of polymer. Samples within the same cluster were then assumed to be of the same type and are labeled as such. As a result, unlabeled samples were spilt into two sets: identifiable samples from clusters and unidentifiable samples as outliers.

- Additionally, given the number of features is 1,650 (much higher than 2), we adopted an unsupervised learning approach to visualize the distance between samples in a 2D plane [49], namely t-distributed Stochastic Neighbor Embedding (t-sne, model #4 shown in Table 8).

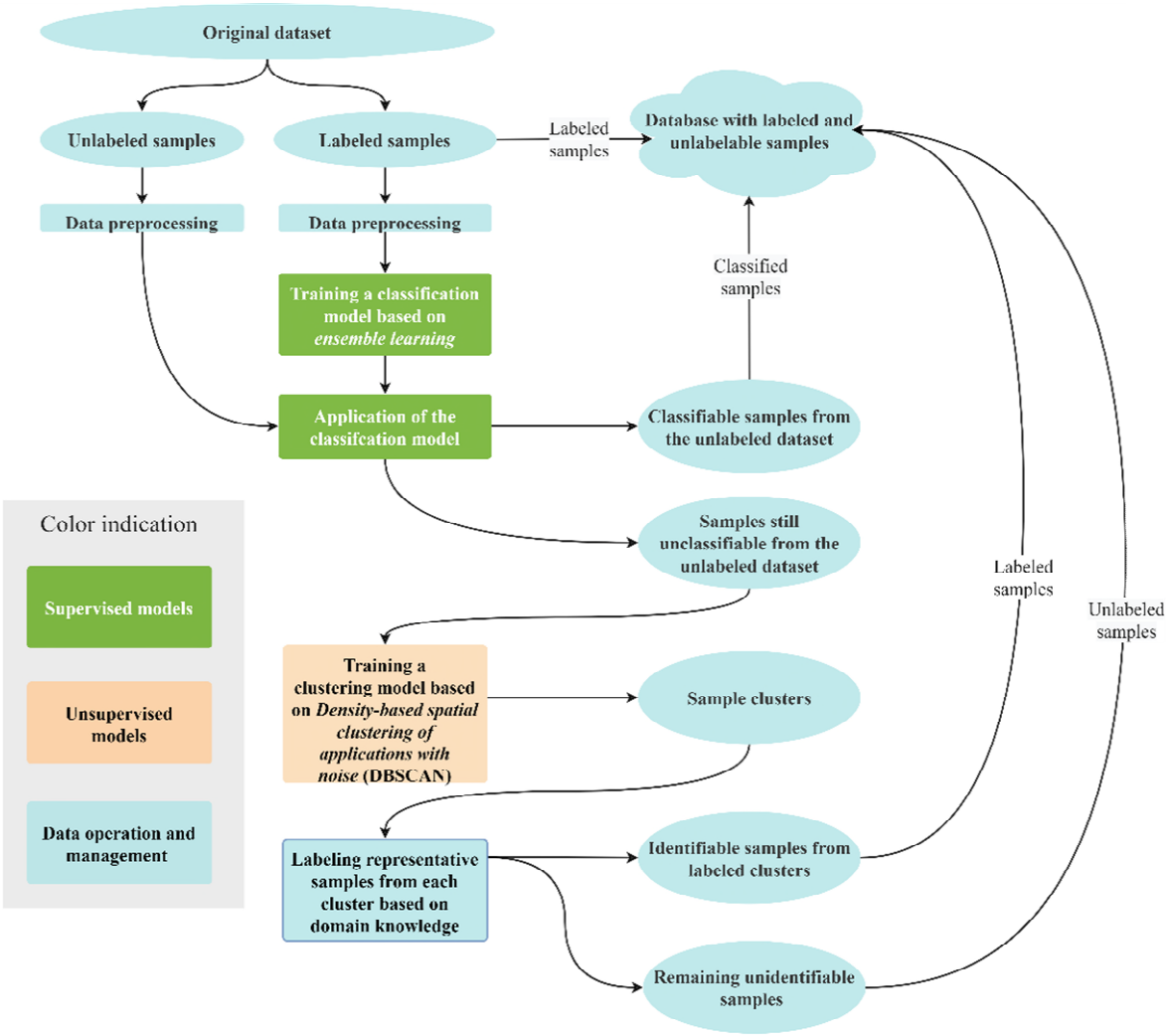


Figure 13. The framework to train combined ML models. In the diagram, the shapes of oval, rectangular, and cloud indicate input/output, processes/models, and database respectively.

Table 8. List of main models used in this study.

	Name	Type	Brief introduction
1	subspace k-nearest neighbor	ensemble supervised learning	The random subspace k-nearest neighbor method is an ensemble ML method, which often achieves high accuracy when used to solve classification problems.
2	boosted decision trees	ensemble supervised learning	By fitting the residuals of preceding decision trees, the boosting algorithm iterates to learn from data.
3	density-based spatial clustering of applications with noise	unsupervised learning	DBSCAN is a non-parametric clustering algorithm based on calculations of densities.
4	t-distributed stochastic neighbor embedding	unsupervised learning	The t-sne algorithm is a ML technique for reducing the dimensionality of data space.

5.3 Materials and Methods

5.3.1 Chemicals

The following chemicals were used in this study: sodium dodecylsulfonate and KOH were purchased from Merck (Darmstadt, Germany); H_2O_2 (30%), $ZnCl_2$ and Ethanol were purchased from Boom (Meppel, the Netherlands) and fluorescent green polyethylene microspheres were purchased from Cospheric (Santa Barbara, USA). Ultrapure water was made using an ELGA system (18.2 MΩ/cm, LabWater, Lane End, UK) . All solvents used, including the cleaning solvents, were filtered prior to usage over 5 µm sieves and stored in closed bottles. All solvents were exclusively used for polymer analysis to minimize the risk of cross-contaminations. All equipment was at all times covered when not used to prevent air contamination, except when the samples were taken on location.

5.3.2 Particle analysis

Samples were analyzed using the quantum cascade laser (QCL) based Agilent chemical imaging laser direct infrared (LDIR) system. Wavenumbers from 975 to 1800 cm-1 with a resolution of 0.5 cm-1 are covered. Particles ranging from 20 - 500 µm were measured and counted. The smallest particles which can be analyzed with the available instrument would be 3 µm, but as 10 µm was the smallest size-fraction collected, at no point particles smaller than 10 µm were scanned for.

5.3.3 Sampling

Two sampling methods were used here, namely a cascade of metal sieves and a metal candle filter cascade. The first one was used for samples with an expected high number of particles, the latter for samples with an expected low number of particles (additional information is provided in the Supporting Information). Samples were taken from different surface waters (river Rhine, river Meus, river Overijsselse Vecht), groundwater and at different stages of drinking water treatment. At each sampling event, between 1000 and 10,000 L of water were filtered through either the cascade of two metal sieves (Gilson, USA) of 500 µm and 100 µm, with a 10 µm plankton net (Hydro-Bios, Germany) at the end or the candle filter cascade (5, 10, 100 and 500 µm). The 500 µm sieve or filter were used to remove particles and prevent clogging of the smaller sieves/filters. Particles of 500µm or larger were not analyzed. Residues from other sieves/filters were transferred into separate glass bottles using Milli-Q ultrapure water (Millipore Sigma, USA) and stored at 4˚C.

5.3.4 Quality assurance

Potential contamination occurring during sample handling was minimized by cleaning all laboratory surfaces with ethanol. Equipment was rinsed with ethanol and covered immediately with aluminum foil, and a cotton laboratory coat was worn at all times by the analysts involved in sample preparation and analysis. In addition, all solvents and solutions were filtered prior to use. Used materials were not made from plastic wherever possible (e.g., a metal filter setup with a Teflon tube, the separation funnels made of glass). Negative controls were treated in parallel to each batch of actual samples to determine the degree of contamination. For the positive control, a known number of plastic particles (green fluorescent PE, average diameter 100 µm) were added to a water sample. The percentage of particle number before and after work-up was used as estimate of the recovery rates. These control particles were counted visually under a microscope.

5.3.5 Sample processing and analysis

Particle analyses were based on previously described methods [32]. Particle analysis focused on 10 µm and 100 µm residues of the sieves. The suspensions from these two fractions were combined and filtered over a 10 µm metal mesh. The latter was then transferred into a beaker with a 10% sodium dodecylsulfonate solution. After 24h, the suspension was filtered again over a 10 µm metal mesh and transferred into a beaker with 75 ml 12.5% potassium hydroxide solution and left standing for 5 days at 35°C. Subsequently, the suspension was filtered again through the 10 µm metal filter. The residues were then transferred into a beaker with 50 ml 30% hydrogen peroxide solution and left standing for 24h at 35°C. The sample was filtered again through the same metal filter, and the residues were

transferred into a separation funnel using a 100 ml zinc chloride solution (1.6 g/cm^3). The funnels were shaken and left standing to enable settling of denser materials. The settled material was discarded by continuously turning the valve of the funnel to prevent clogging, re-suspension and loss of plastics. About 10 mL liquid was allowed to remain in the funnel. These 10 mL were filtered again over a 10 µm metal filter. Using four mL ethanol, the retained materials were removed from the filter and transferred into a glass vial. A vortex was created in this suspension to distribute the particles evenly. An aliquot (2 x 50 µm) was taken and transferred onto a microscope plates for analysis. Aliquots were used to avoid having a too high particle density on the plates. Total number of particles per sample were extrapolated from these aliquots. Samples were analyzed using the LDIR. Between four to six samples were analyzed in parallel in combination with a negative control sample to detect (cross)-contamination during work-up and a sample spiked with a known amount of clearly visible fluorescent polyethylene microbeads to calculate the recovery rate (positive control. After each set, the whole equipment was cleaned using ultrapure water and ethanol. The equipment was then covered with aluminum foil to prevent contamination from the air.

5.3.6 Types of particles

In total, 81,291 particles from various Dutch aqueous matrices (drinking water, surface water, effluent form wastewater treatment plants) were used in this study, of which 223 were already labeled (≈0.3%) and 81,068 being unlabeled (≈99.7%). Ten different types of polymers are among the labeled particles, including cellulosic, polyacetal (POM), polyamide (PA), polyethylene (PE), polyethylene terephthalate (PET), polypropylene (PP), polystyrene (PS), polyurethane (PU), polyvinylchloride (PVC), and silica. An additional category called non-plastic particle was also added. Other types of polymers were not the main focus of the study and were regarded as outliers to these 11 predefined categories. These polymers were chosen as they are the most commonly found in the environment according to recent literature [32], [50].

5.4 Results and Discussion

5.4.1 Classification of labeled polymer samples

Two classification models (Model 1 and Model 2 shown in Table 8. List of main models used in this study.) were trained on 223 labeled samples spanning 11 different polymer types. The results are presented in figure 14. In general, the subspace KNN model outperformed the boosted decision tree model, with the former achieving 89.7% accuracy (12.6% higher than that of the latter). Figure 14 shows the confusion matrix for the two models, namely, a subspace KNN model and a boosted decision tree model. Rows represent true classes (predefined labels) whereas columns represent predicted classes (labels marked by the classification models). The subspace KNN model can predict the class of all polymer samples with no more than 3 mislabeled samples for each type. The best prediction was obtained for PE, as all samples were correctly labeled. The worst prediction was obtained for POM, where 2 out 5 samples were mislabeled. Besides, 5 out of 15 non-plastic samples were mislabeled (≈33.3%). This high inaccuracy is due to the fact that non-plastic particles do not have a distinct spectrum and only a limited number of spectra were available for training the classifiers. On the other hand, the boosted decision tree performed worse than the subspace KNN model, in particular because 22 particles were mislabeled as silica. This was because the boosted decision tree algorithm underfits the decision region for silica. Apart from the mislabeled silica, the boosted decision tree performs generally well, though it is slightly worse than the subspace KNN model across all polymer types. Based on these results, the subspace KNN model was deemed the most appropriate for classifying polymers in our case study.

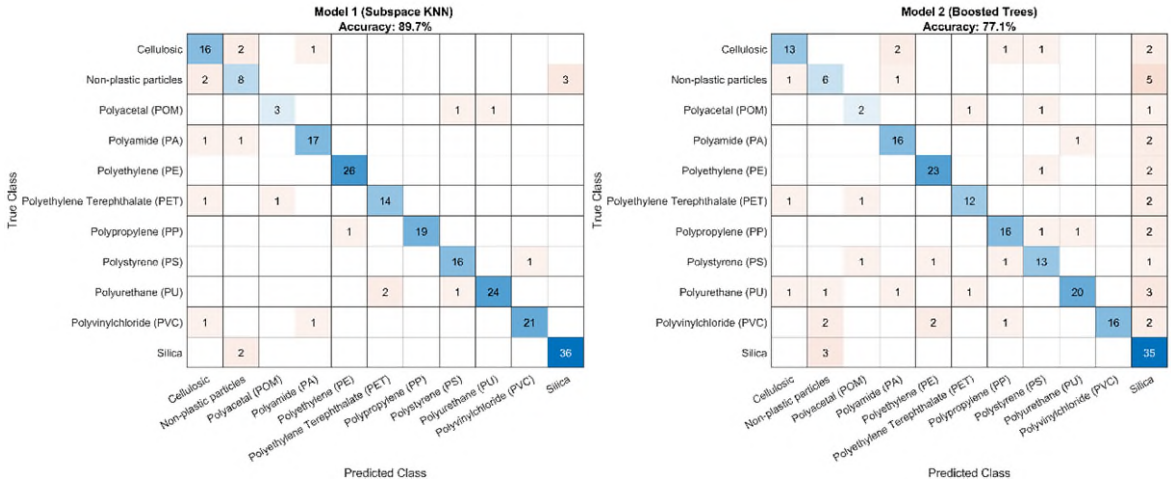


Figure 14. Results of classification model with 10-fold cross validation. Two models, subspace KNN and boosted decision tree, were developed and compared. Subspace KNN (accuracy=89.7%) outperforms boosted decision tree (accuracy=77.1%) for our classification problem.

5.4.2 Classification of unlabeled samples

The previously trained classification model was then applied to unlabeled samples (real samples). The model produced 11 probabilities describing how likely an unknown polymer belongs to one of the 11 polymer types. Therefore, a threshold to accept the prediction probability had to be defined. A range of possible threshold values ranging from 67.5% to 100% was evaluated and the corresponding number of unaccepted samples was computed. The greater the threshold value, the larger the number of unaccepted predicted labels, as can be seen in **Error! Reference source not found.**. For the purpose of the applications considered in this study, 10% of unclassifiable samples were deemed by the domain expert as acceptable. Hence, in Figure 15, a threshold of 0.7 was selected. Obviously, this threshold was defined subjectively, and different values should be selected depending on what is considered acceptable for the specific application. Furthermore, if more information about the occurrence of different types of polymers in samples becomes available, the threshold can be adjusted accordingly.

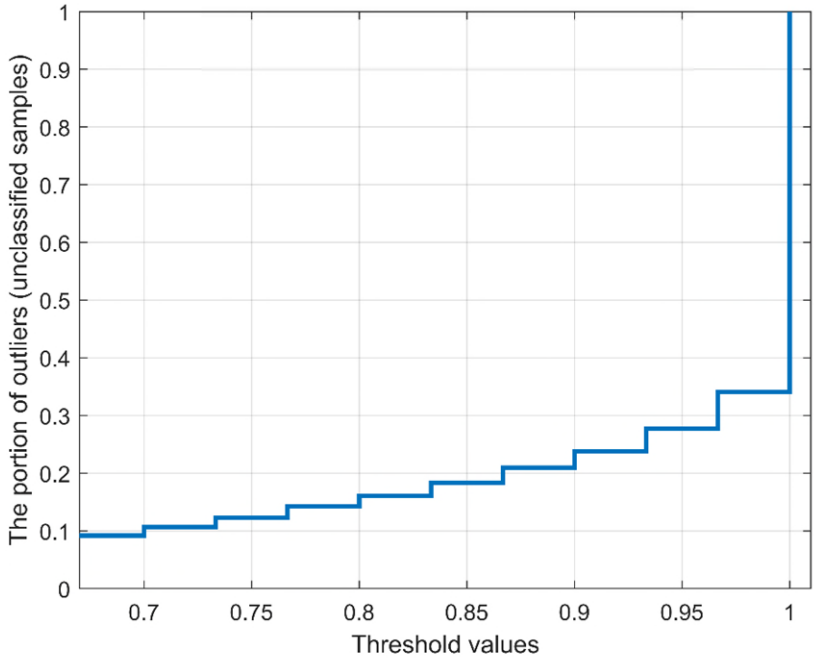


Figure 15. Percentage of unclassifiable samples as a function of the threshold for accepting the prediction of polymer types. The higher the threshold value is, the more samples are not accepted as labeled samples. This study uses the threshold of 0.7, leading to approximately 10% of unclassifiable samples (c.a., 8,100 samples).

5.4.3 Clustering unlabeled samples

The third part of the developed approach focused on the remaining 10% of particles which could not be classified using the 0.7 probability and required extra processing. As described in Figure 13, the unsupervised model DBSCAN (model #3 shown in Table 8) was used to cluster samples. Two parameters need to be set before training the model, namely the minimum number of points of a cluster (denoted as *minpts*) and the neighborhood searching radius for a cluster (denoted as *eps*). Different combinations of these two parameter can lead to the number of clustered types and the number of outliers (unclusterable samples). When the values for eps and minpts are fixed, one or more central points can be determined. These points represent those that contain at least minpts points inside the radius eps. It is, however, a pragmatic choice to determine these two values. Figure 16 illustrates three possible combinations of eps and minpts values that provide different clustering results. When the value of eps is increased, the number of outliers decreases as more samples can be clustered within a broader searching radius. When the value minpts is decreased, more clusters and fewer outliers appear, as more clusters of lower sizes can emerge. However, it is not a straightforward linear relationship. Assuming that one should have 30 polymer types, with at least 30 samples per type, and roughly 40% outliers that are minor types and cannot be clustered, one can find a compromise using the mentioned diagram (in this instance, eps = 2.75 and minpts = 30). By clustering these data, domain experts need to label a smaller number of samples with far less work. Due to the fact that this is a manual check performed by a domain expert, in which the domain expert must screen a large number of spectra and determine the type of polymers, we do not demonstrate the manual process in this study.

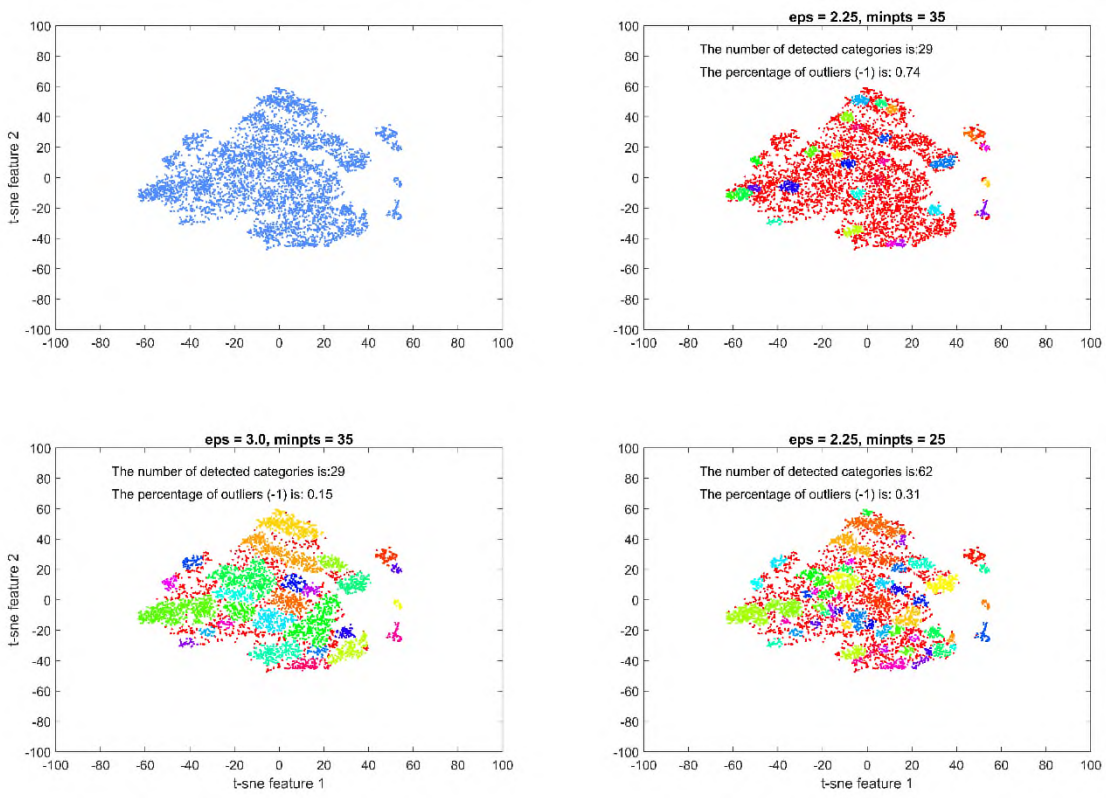


Figure 16. Clustering unlabeled samples based on two parameters (the distance and the minimum number of points contained in one cluster).

5.4.4 The implications of machine learning models

5.4.4.1 Selection of ML models

In Section 5.4.1, two ML models, namely, subspace KNN and boosted decision tree, were trained to classify 223 samples into 11 pre-defined categories, whose accuracy of predictions were 89.7% and 77.1% respectively. Besides these two models, we also trained other ensemble ML models for comparison, including bagged decision tree

(accuracy = 66.1%), ensemble subspace discriminant (accuracy = 61%), weighted KNN (accuracy = 67%), and RUSBoost decision tree (accuracy = 8%), which all performed worse than the two reported models. As a result, we only showed detailed results in Section 5.4.1 and considered the subspace KNN model as the best classifier for our polymer samples. Typically, it is difficult to know which model has the best performance before training and comparing candidate models, as the no free lunch theorem implies [51]. We suggest that the above-mentioned ensemble ML models be trained and compared before deploying an ML model for a new application.
Besides the ML models that are used in this study, deep neural networks can be utilized to classify polymers. For example, an autoencoder that can encode and decode the high-dimensional feature space and then predict polymer types may be applicable in our case. However, training such a model requires a significant number of samples, which were unavailable at the time of conducting this research. In future study, we intend to collect more samples and employ deep neural networks to address this issue.

5.4.4.2 Selection of features

In this study, absorption rates on wave numbers ranging from 975 to 1,800 cm^{-1} were used as features to train the classification model. Alternatively, the 1^{st}-order derivative with respect to the absorption rate is often used, too. To compare with the original absorption rate, a subspace KNN model and a boosted decision tree model were trained based on the 1^{st}-order derivative of the absorption rate, resulting in the prediction accuracy of 92% and 17% respectively. The boosted decision model is much worse than the one trained on the original absorption rate while the subspace KNN is slightly better. As a result, we did not adopt the 1^{st}-order derivative for the classification and clustering analyses because (i) the performance only improves insignificantly for the subspace KNN model; (ii) to apply the classification model to new samples, it takes extra computation time to calculate the 1^{st}-order derivative; (iii) it is more intuitive to interpret the original absorption rate than the 1^{st} order derivative.

5.4.4.3 The size of dataset for training an ML model

ML is a data-driven technique requiring a large number of observations, particularly in the training dataset, in order to develop a good model. In this work, initially we started with a dataset of 60 labeled samples and only achieved an accuracy of approximately 70%, which is 19.7% lower than the model trained on the 223 labeled samples. Ideally, adding extra samples would improve the model performance. There is, however, the concern of finding a balance between model performance and manual labeling effort. In practice, one can iterate the process of increasingly adding samples and analyze the model's performance as it evolves. When performance exceeds a predetermined level or when adding new samples improves performance little, no additional samples are required. Additionally, if a large number of samples are misclassified for a particular class, extra labelled samples for that class should be added. Consider the case of non-plastic polymers encountered in this study: it remains challenging for the classification model to accurately classify non-plastic samples type because (1) non-plastic samples, unlike other plastic samples, lack substantial group characteristics and (2) certain non-plastic samples may present characteristics that are very similar to those of other groups (**Error! Reference source not found.**). As a result, the model developer and user should consider the accuracy of categorizing a non-plastic polymer when applying the classification model or exclude non-plastic samples entirely and regard them as outliers.

5.4.4.4 Implementation of polymer screening

To conduct the polymer screening, one should first train a classification model, then detect unclassifiable samples (outliers), and finally build a clustering model to group outliers. Note that 90% of the unlabeled samples (≈73,100) in this study can be successfully identified based on the 223 labeled examples. Thus, considerable effort for the domain expert to label samples can be saved. The remaining 10% of the outlier samples (≈8,100) can be clustered and subsequently labeled by domain experts. The amount of effort required to identify polymer types is dependent on the number of clusters and representative samples within each cluster. One who aims to adopt this

method must weigh the trade-off between the overall performance and the time required to process individual samples. If the classifiable samples (≈90 percent) already match the criteria for conducting polymer screening, training a clustering model and manual labeling can also be omitted. In addition, the methodology proposed in this study can also be applied to the case with a large number of polymer types, as long as there are quality samples of each polymer type.

5.5 Concluding remarks

In this study, we investigated the applicability of LDIR data in combination with ML to identify polymers in water samples. This novel approach is necessary as the number of plastics at various stage of degradation in the environment and other aqueous samples will be unlimited. A database with a certain amount of reference spectra alone and comparison using Pearson correlation coefficient will not suffice. The major finding was that ML was able to further identify the type of polymers, based on the absorption measured by LDIR. This research shows a suitable way to identify polymers that are not pristine and have been weathered. Creating realistically weathered particles (each type of polymer at different stages of weathering) is not a viable option. For our research we used a limited number of weathered particles from the environment to train a supervised ML model. This model builds a classifier to identify polymer types based on their spectra, which was then applied to unlabeled samples. These findings suggest that in general artificial-intelligence-aided techniques can greatly support and promote analytical chemistry and water research, which is of interest to researchers who aim to classify polymers (based on classification models) or predict their bio-chemical behaviors (based on regression models). Nevertheless, the application of the methodology presented in this study could also be limited to the number of quality data, e.g., labeled samples. Efforts from domain experts are still needed when applying this method to a new application. However, we propose an approach which minimizes the labeling effort through active learning. In future work, we also aim to employ deep neural networks in both supervised and unsupervised models as more labeled samples become available in the future. The introduced model is in principal applicable to any kind of spectra, not just infrared spectra.

6 Conclusions

The following situations are often encountered in our research: (1) classification (e.g., of hydrological events or chemical compounds), (2) predictions (e.g., of weather, water levels or concentration), (3) recognition of objects (e.g., in images, videos, or audio records), (4) identification of key information in texts, (5) decision-making. All these applications can benefit from machine learning (ML) or deep learning (DL).

6.1 Lessons from case studies

In this BTO project, we demonstrated its value via two case studies, i.e., using ML/DL to deal with text, spectrum, or time series data in real-world water-related applications. This is starting point to illustrate that ML/DL can play an significant role in projects where such methods are needed. To summarize, the following findings have been made:

(1) Case study I shows how natural language processing (NLP) powered by DL can assist water companies in handling consumer complaints. Although NLP cannot deal with all kinds of complaints, it is incredibly beneficial for automating repetitive tasks such as complaint categorization, analysis of customers' emotions, and recognition of their demands and intents. We conducted this exploratory research to determine the feasibility of performing these tasks using NLP approaches. Given that complaint processing requires interaction between customers and computers, it is also worthwhile to consider the steps following up this project. For example, developing a chatbot might significantly assist water utilities in receiving instant feedback during communication (more efficiently than emails), which could become an interesting topic for BTO projects aimed at digitalizing urban water management.

(2) Case study II shows how ensemble learning can classify microplastic types. Although only a small number of samples have been identified for their types, they can be used to determine how the spectrum distribution leads to the microplastic type and to further identify hundreds of unknown microplastic samples. In this case, we must additionally evaluate the model's design with the assistance of experts. A typical classifier is capable of calculating the likelihood of a sample being classified into a specified category. When the probability is too low, experts must intervene to establish a probability threshold for accepting the classifier's results. In this case, we must additionally evaluate the model's design with the assistance of experts. A typical classifier is capable of calculating the likelihood of a sample being classified into a specified category. When the probability is too low, experts must intervene to establish a probability threshold for accepting the classifier's results. Via this case study, we see the potential of ML to classify polymers, rather than the conventional manual check.

6.2 Implications for the water sector

First of all, we suggest that readers need to understand that ML/DL is not a panacea to solve all problems. When applying ML/DL, readers should bear the following points in mind:

(1) Which type of problem is being addressed? Is this an application that goes beyond the capabilities of ML/DL?

(2) Are there sufficient and good-quality data to train a model?

(3) How is the quality of data? As [22] suggests, quality data are critical for training a good model.

(4) When applying ML/DL models to new instances, do they have similar characteristics with old instances used to train the model?

Generally speaking, ML/DL is a fast growing field which has great promise for future water-related applications. We suggest that water utilities think about the use of it in several potential applications as follows:

(1) Inspection of pipe conditions: as a typical classification problem to tell whether the condition is good or not, we can consider the use of inspection cameras and sensors to collect pipe conditions under different circumstances. Then a classifier should be able to assess the condition of new measurement. An follow-up idea or project could be based on this point. For instance, several water utilities in the Netherlands have been collecting image about inside conditions of pipes. DL will be useful to identify different conditions of pipes.

(2) Prediction of demand: as a typical regression problem, we can consider the use of household water consumption data to predict the water use, which is important for other research based on consumption. This is in fact a typical time series prediction, which DL is suitable for. The commonly used RNN models can learn the patterns of time series from this historical data and apply it to predict trends.

(3) Customer service chatbot: because the Dutch water utilities serve thousands of households, the need for customer complaints is also enormous. A chatbot powered by NLP can significantly reduce labor costs for water utilities and provide an effective and efficient method of communicating with customers. This is one of the most representative applications for DL techniques. The use chatbot can greatly reduce the workload from phone operators dealing with customers' complaints and enhance the communication efficiency between the water utility and customers. KWR will follow this point up as a research line for future projects.

(4) The use of AI/ML/DL can also be expanded to the handling and, in particular, interpretation of the large amounts of chemical data which is being generated by drinking water companies in their monitoring and regulatory activities. With respect to water quality monitoring, large amounts of chemical and biological data are being generated and the implementation of these tools can help detect events (e.g., calamities) and trends at an earlier stage, just to name a few.

Appendix I Fundamentals of deep learning

Readers who want to learn more about deep learning can refer to lots of open-sourced materials, for instance, https://github.com/L1aoXingyu/Roadmap-of-DL-and-ML, Figure 17 below also shows the mathematical and statistical fundamentals of deep learning. Generally speaking, DL is a natural development of matrix theory, calculus, probability, and optimization.

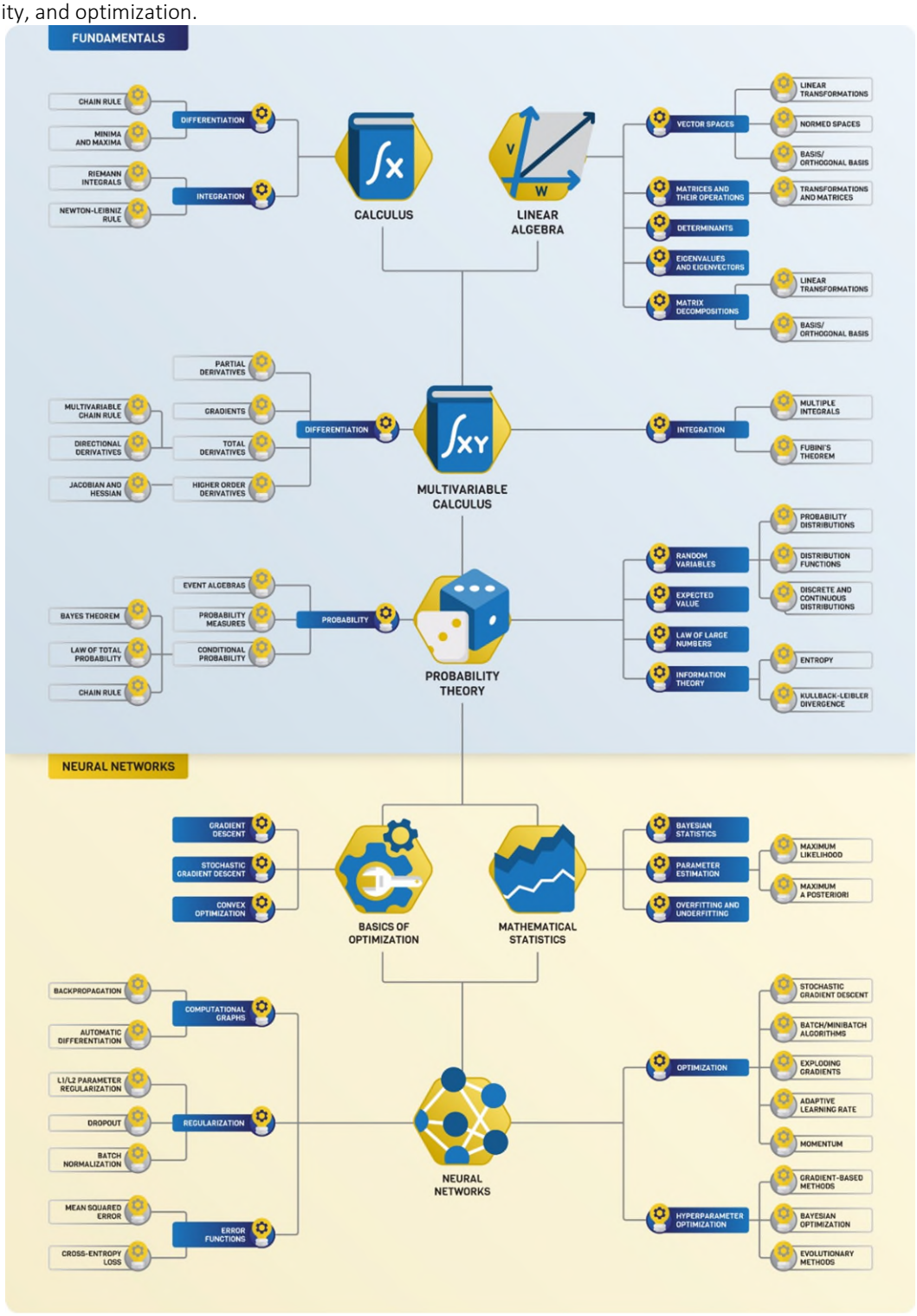


Figure 17. Fundamentals of deep learning. Source: [3].

Appendix II. An example about word vectorization.

When building and training NLP models, words are mapped to numeric vectors that represent the semantic similarity. In our case study and many other NLP applications, word vectors are typically in a high-dimensional space (e.g., 300 dimensions). However, a simple example, adapted from [52], is used here to visualize how words can be mapped to vectors in a 2D space.

Assume we have four words: Italy, Greece, Rome, and Athens. Two of them are countries names while two are capital cities. By vectorizing these words, we aim to create a calculative distance such that the distance between Italy and Rome is comparable to that between Greece and Athens. In other words, Rome to Italy is the same as Athens to Greece. By doing so, we are able to make a simple calculation such as Italy – Rome + Athens = Greece. Mathematically, we need to build the following word vector space by using four coordinates:

	Country	Capital	Italian	Greek	Vector
Italy	1	0	1	0	[1, 0, 1, 0]
Rome	0	1	1	0	[0, 1, 1, 0]
Greece	1	0	0	1	[1, 0, 0, 1]
Athens	0	1	0	1	[0, 1, 1, 0]

In real-world applications, we need to adopt a word vector space of more than just 4 coordinates. For instance, readers can check the following links to see how words are mapped to a 300-dimensioanl space: https://fasttext.cc/docs/en/english-vectors.html. In [52], a 2D projection is presented.

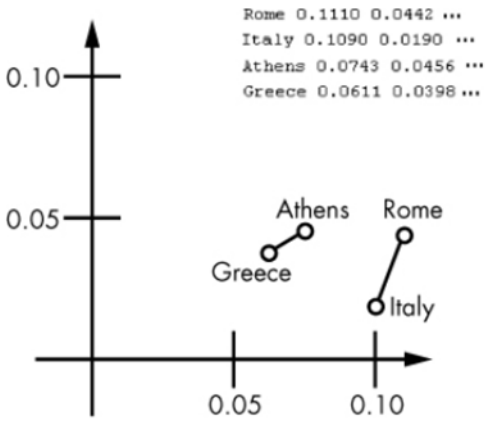


Figure 18. 2D projection of word vectors. Credit: [52].

Appendix III. Texts used to categorize intents in the NLU model.

Training dataset for IR-category I, problem statement about water meter or water network

1) Onze watermeter piept heel irritant.
2) Hierbij is door de werkzaamheden continue blijvende lekkage veroorzaakt achter de meter.
3) Lekkage van de waterleiding voor de meter
4) Ik heb veel water staan bij de meter achter de voordeur
5) We kunnen de hoofdwaterkraan niet goed open en dicht draaien.
6) Tevens is er sprake van lekkage uit de hoofdwaterkraan.
7) watermeter maakt nu wat lawaai
8) De watermeter maakt een tikkende geluid bij watergebruik, hoorbaar buiten de meterkast.
9) mijn hoofdkraan voor de meter is stuk.
10) Hoofdkraan sluit niet af.
11) De afsluitkraan onder de vloer valt zeer moeilijk dicht te draaien.
12) ik heb kapotte waterkraan van hoofdleiding.
13) Na het plaatsen van een watermeter is er lekkage bij de meter.
14) De hoofdkraan die voor de watermeter zit sluit niet meer goed af.
15) Als ik de kraan voor de watermeter dicht zet blijft de meter doorlopen en blijft er water doorlopen.
16) Slecht af te sluiten en te openen hoofdkraan plus piepende watermeter
17) Echter de nieuwe watermeter maakt veel meer lawaai.
18) De waterkraan lekt
19) een lek in de leiding
20) lekkage waterleiding
21) was er een waterleiding gesprongen vlak voor mijn huis
22) Leidingbreuk
23) Dikke lekkage vlak
24) hoofdleiding lek

Training dataset for IR-category II, problem statement about no water

1) een vrij lage waterdruk
2) de waterdruk is een stuk lager
3) Sinds maart 2020 is de waterdruk gehalveerd met af en toe weer normale waterdruk.
4) een verminderde waterdruk
5) Wij hebben op het moment geen water
6) Er komt geen water uit de kraan.
7) maar nu is onze waterdruk weer zeer laag te noemen
8) Sinds enkele dagen is de druk weg bij het openen van de kranen.
9) Minder druk op de leiding
10) Geen waterdruk meer sinds juli 2020
11) We hebben weinig druk
12) zeer geringe waterdruk

Training dataset for IR-category III, problem statement about water quality

1) bruin kraan water
2) het water uit de kraan was bruin en zanderig.
3) Echt donker van kleur

4) troebel water
5) last van bruine deeltjes in onze leidingen.
6) Sinds vandaag stinkt en leidingwater en smaakt het vreemd.
7) Het water heeft een vreemde metaalachtige (koper)geur.
8) Er zweven kleine witte (kalk)deeltjes in het water.
9) er zit rommel in het water

Training dataset for IR-category IV, problem statement about website and account

1) Helaas kan ik niet meer inloggen met mijn gegevens.
2) ik kan nog steeds niet inloggen.
3) Onze gegevens kloppen niet, graag dat het even goed gaat doorspitten.
4) Als ik met mijn account wil inloggen op jullie en mijn wachtwoord wil herstellen krijg ik niks in mail (geen spam, geen ongewenst).
5) Online is aanmelding als nieuwe klant niet mogelijk
6) echter uw app geeft aan mijn emailadres niet als klant te herkennen
7) Ik kan mijn e-mail adres niet wijzigen
8) Het werkt niet als ik een foto wil uploaden.
9) Ik heb vanavond voor de 3e keer een mail ontvangen om mijn waterstand door te geven.
10) Het lukt niet om online de meterstand door te geven.
11) Het lukt niet om op de goede pagina te komen.
12) de site werkt niet.
13) De website werkt niet naar behoren.
14) In tegenstelling tot vorige jaren werkt het online doorgeven van de meterstand niet.
15) het lukt niet om in te loggen

Training dataset for IR-category V, request for repair or replacement

1) deze graag vervangen
2) gaarne uw oordeel
3) Hoe gaat u dit oplossen?
4) Met dit schrijven wens ik een afspraak te maken voor het instaleren van een nieuwe watermeter
5) Zou u hier verbetering aan kunnen doen?
6) ik wil graag snel mogelijk vervangen
7) Graag zouden we dit gerepareerd zien
8) Graag hebben we hier spoedig mogelijk contact over omdat er nu water verspild wordt.
9) Gezien het zeer hoge waterverbruik, wil ik dat er zo spoedig mogelijk een onderzoek gestart wordt.
10) Wil graag een nieuwe meter laten plaatsen.
11) Kunt u aangeven of dit ook te verhelpen is?
12) Hierbij een vriendelijk verzoek om dit probleem te verhelpen.
13) Heel graag willen wij van bovenstaande problemen af, heeft u hier een oplossing voor?
14) Kunt u dit verklaren en verhelpen?
15) Graag zien wij dit opgelost
16) Graag met spoed de aansluiting weer herstellen.
17) Zouden jullie voor mij kunnen checken of alles klopt en goed doorgekomen is.
18) Graag zou ik willen dat mijn adres gecorrigeerd wordt

Training dataset for IR-category VI, request for reclaiming payment

1) ik zou dit graag teruggestort op mijn rekening zien
2) Nu heb ik een factuur gekregen van 377,- en de meterstanden zijn al doorgegeven, maar volgens mij klopt het niet in verband met de lekkage.
3) Het betreft hier een reparatie van het waterbedrijf zelf dus deze kosten zijn niet voor mij.
4) Onder voorbehoud van alle rechten en weren verblijf ik.
5) Dus ben het niet eens met deze nota

6) Dit is een reden waarom ik deze factuur wil declareren bij uw bedrijf als schade geval.
7) er zijn in de afgelopen jaren door het Waterbedrijf Groningen onterecht bedragen voor waterverbruik bij mij in rekening gebracht en afgeschreven.
8) Om bovengenoemde redenen maak ik hierbij bezwaar tegen de opgelegde administratiekosten en ga niet akkoord met betaling hiervan

References

[1] S. Pouyanfar *et al.*, "A Survey on Deep Learning," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–36, Jan. 2019, doi: 10.1145/3234150.

[2] H. Cenan, "What's the deal with AI, anyway?," 2019. https://medium.com/zasti/whats-the-deal-with-ai-anyway-56a30177f438.

[3] Y.-Y. Chen, Y.-H. Lin, C.-C. Kung, M.-H. Chung, and I.-H. Yen, "Design and Implementation of Cloud Analytics-Assisted Smart Power Meters Considering Advanced Artificial Intelligence as Edge Analytics in Demand-Side Management for Smart Homes," *Sensors*, vol. 19, no. 9, p. 2047, May 2019, doi: 10.3390/s19092047.

[4] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. O'Reilly Media, Inc., 2019.

[5] A. Kumar, "Complete Deep Learning Roadmap," 2020. https://medium.com/analytics-vidhya/complete-deep-learning-roadmap-8748c0475dc1.

[6] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, and M. Dehmer, "An Introductory Review of Deep Learning for Prediction Models With Big Data," *Front. Artif. Intell.*, vol. 3, no. February, pp. 1–23, Feb. 2020, doi: 10.3389/frai.2020.00004.

[7] M. V. Valueva, N. N. Nagornov, P. A. Lyakhov, G. V. Valuev, and N. I. Chervyakov, "Application of the residue number system to reduce hardware costs of the convolutional neural network implementation," *Math. Comput. Simul.*, vol. 177, pp. 232–243, Nov. 2020, doi: 10.1016/j.matcom.2020.04.031.

[8] "Deep Learning – Introduction to Convolutional Neural Networks." https://vinodsblog.com/2018/10/15/everything-you-need-to-know-about-convolutional-neural-networks/.

[9] "RECURRENT NEURAL NETWORKS: MAJOR APPLICATIONS." https://theappsolutions.com/blog/development/recurrent-neural-networks/.

[10] A. Géron, *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. 2017.

[11] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, vol. 2017-Janua, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.

[12] X. Yin, Y. Chen, A. Bouferguene, H. Zaman, M. Al-Hussein, and L. Kurach, "A deep learning-based framework for an automated defect detection system for sewer pipes," *Autom. Constr.*, vol. 109, no. August 2019, p. 102967, 2020, doi: 10.1016/j.autcon.2019.102967.

[13] C. Shen *et al.*, "HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community," *Hydrol. Earth Syst. Sci.*, vol. 22, no. 11, pp. 5639–5656, Nov. 2018, doi: 10.5194/hess-22-5639-2018.

[14] C. Shen, "A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists," *Water Resour. Res.*, vol. 54, no. 11, pp. 8558–8593, Nov. 2018, doi: 10.1029/2018WR022643.

[15] Y. Chen, R. Fan, X. Yang, J. Wang, and A. Latif, "Extraction of urban water bodies from high-resolution remote-sensing imagery using deep learning," *Water (Switzerland)*, vol. 10, no. 5, 2018, doi: 10.3390/w10050585.

[16] W. Buytaert *et al.*, "Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development," *Front. Earth Sci.*, vol. 2, no. October, pp. 1–21, 2014, doi: 10.3389/feart.2014.00026.

[17] X. Tian, M. ten Veldhuis, M. Schleiss, C. Bouwens, and N. van de Giesen, "PPT-Critical rainfall thresholds for urban pluvial flooding inferred from citizen observations," *Sci. Total Environ.*, vol. 689, no. 1, pp. 258–268, Nov. 2019, doi: 10.1016/j.scitotenv.2019.06.355.

[18] X. Hu, Y. Han, B. Yu, Z. Geng, and J. Fan, "Novel leakage detection and water loss management of urban water supply network using multiscale neural networks," *J. Clean. Prod.*, vol. 278, p. 123611, 2021, doi: 10.1016/j.jclepro.2020.123611.

[19] G. Guo, S. Liu, Y. Wu, J. Li, R. Zhou, and X. Zhu, "Short-Term Water Demand Forecast Based on Deep Learning Method," *J. Water Resour. Plan. Manag.*, vol. 144, no. 12, p. 04018076, 2018, doi: 10.1061/(asce)wr.1943-5452.0000992.

[20] S. L. Zubaidi *et al.*, "Urban water demand prediction for a city that suffers from climate change and population growth: Gauteng province case study," *Water (Switzerland)*, vol. 12, no. 7, pp. 1–17, 2020, doi: 10.3390/W12071885.

[21] B. T. Pham *et al.*, "Can deep learning algorithms outperform benchmark machine learning algorithms in flood susceptibility modeling?," *J. Hydrol.*, vol. 592, no. October 2020, p. 125615, 2020, doi: 10.1016/j.jhydrol.2020.125615.

[22] A. Ng, "MLOps: From Model-centric to Data-centric AI," 2021. [Online]. Available: https://www.deeplearning.ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centric-AI.pdf.

[23] Github resources, "Stop words in Dutch." 2021, [Online]. Available: https://raw.githubusercontent.com/stopwords-iso/stopwords-nl/master/stopwords-nl.txt.

[24] Spacy, "Spacy LG model for English." 2021, [Online]. Available: https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.1.0.

[25] Spacy, "Spacy LG model for Dutch." 2021, doi: https://github.com/explosion/spacy-models/releases/tag/nl_core_news_lg-3.1.0.

[26] Scispacy, "Scispacy model for biomedical texts." 2021, doi: : https://allenai.github.io/scispacy/.

[27] Github resources, "Textblob-nl sentiment model." 2021, doi: https://github.com/gvisniuc/textblob-nl.

[28] Github resources, "Thinc: A refreshing functional take on deep learning, compatible with your favorite libraries." 2021, doi: https://github.com/explosion/thinc.

[29] C. D. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, "Neural Architectures for Named Entity Recognition," 2016, doi: arXiv:1603.01360.

[30] Rasa, "Rasa model architecture," 2021. https://blog.rasa.com/introducing-dual-intent-and-entity-transformer-diet-state-of-the-art-performance-on-a-lightweight-architecture/.

[31] Spacy, "Dependency tree," 2021. https://github.com/clir/clearnlp-guidelines/blob/master/md/specifications/dependency_labels.md.

[32] S. M. Mintenig *et al.*, "A systems approach to understand microplastic occurrence and variability in Dutch riverine surface waters," *Water Res.*, vol. 176, p. 115723, Jun. 2020, doi: 10.1016/j.watres.2020.115723.

[33] B. E. Oßmann, G. Sarau, H. Holtmannspötter, M. Pischetsrieder, S. H. Christiansen, and W. Dicke, "Small-

sized microplastics and pigmented particles in bottled mineral water," *Water Res.*, vol. 141, pp. 307–316, Sep. 2018, doi: 10.1016/j.watres.2018.05.027.

[34] X. Xu, Y. Jian, Y. Xue, Q. Hou, and L. Wang, "Microplastics in the wastewater treatment plants (WWTPs): Occurrence and removal," *Chemosphere*, vol. 235, pp. 1089–1096, Nov. 2019, doi: 10.1016/j.chemosphere.2019.06.197.

[35] C. Lorenz *et al.*, "Spatial distribution of microplastics in sediments and surface waters of the southern North Sea," *Environ. Pollut.*, vol. 252, pp. 1719–1729, Sep. 2019, doi: 10.1016/j.envpol.2019.06.093.

[36] B. E. Oßmann, "Microplastics in drinking water? Present state of knowledge and open questions," *Curr. Opin. Food Sci.*, vol. 41, pp. 44–51, Oct. 2021, doi: 10.1016/j.cofs.2021.02.011.

[37] The European parliament and the council of the european union, "The quality of water intended for human consumption," 2020. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32020L2184&from=ES.

[38] D. Schymanski *et al.*, "Analysis of microplastics in drinking water and other clean water samples with micro-Raman and micro-infrared spectroscopy: minimum requirements and best practice guidelines," *Anal. Bioanal. Chem.*, vol. 413, no. 24, pp. 5969–5994, Oct. 2021, doi: 10.1007/s00216-021-03498-y.

[39] A. B. Silva, A. S. Bastos, C. I. L. Justino, J. P. da Costa, A. C. Duarte, and T. A. P. Rocha-Santos, "Microplastics in the environment: Challenges in analytical chemistry - A review," *Anal. Chim. Acta*, vol. 1017, pp. 1–19, Aug. 2018, doi: 10.1016/j.aca.2018.02.043.

[40] S. Primpke, M. Godejohann, J. Rowlette, and G. Gerdts, "High-throughput environmental microplastic identification and quantification using a wide-field QCL-IR based microscope," in *Optical Fibers and Sensors for Medical Diagnostics, Treatment and Environmental Applications XXI*, Mar. 2021, p. 33, doi: 10.1117/12.2578414.

[41] P. Liu, X. Zhan, X. Wu, J. Li, H. Wang, and S. Gao, "Effect of weathering on environmental behavior of microplastics: Properties, sorption and potential risks," *Chemosphere*, vol. 242, p. 125193, Mar. 2020, doi: 10.1016/j.chemosphere.2019.125193.

[42] M. Dong, Z. She, X. Xiong, and Z. Luo, "Automated analysis of microplastics based on vibrational spectroscopy: Are we measuring the same metrics?," *Anal. Chem.*, 2021, doi: 10.33774/chemrxiv-2021-hqr34.

[43] S. Primpke, M. Godejohann, and G. Gerdts, "Rapid Identification and Quantification of Microplastics in the Environment by Quantum Cascade Laser-Based Hyperspectral Infrared Chemical Imaging," *Environ. Sci. Technol.*, vol. 54, no. 24, pp. 15893–15903, Dec. 2020, doi: 10.1021/acs.est.0c05722.

[44] M. Mowbray *et al.*, "Machine learning for biochemical engineering: A review," *Biochem. Eng. J.*, vol. 172, p. 108054, Aug. 2021, doi: 10.1016/j.bej.2021.108054.

[45] W. Sha *et al.*, "Machine learning in polymer informatics," *InfoMat*, vol. 3, no. 4, pp. 353–361, Apr. 2021, doi: 10.1002/inf2.12167.

[46] L. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009, doi: 10.4249/scholarpedia.1883.

[47] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*. Routledge, 2017.

[48] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN Revisited, Revisited," *ACM Trans. Database Syst.*, vol. 42, no. 3, pp. 1–21, Aug. 2017, doi: 10.1145/3068335.

[49] A. C. Belkina, C. O. Ciccolella, R. Anno, R. Halpert, J. Spidlen, and J. E. Snyder-Cappione, "Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets," *Nat. Commun.*, vol. 10, no. 1, p. 5415, Dec. 2019, doi: 10.1038/s41467-019-13055-y.

[50] C. Scherer *et al.*, "Comparative assessment of microplastics in water and sediment of a large European river," *Sci. Total Environ.*, vol. 738, p. 139866, Oct. 2020, doi: 10.1016/j.scitotenv.2020.139866.

[51] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997, doi: 10.1109/4235.585893.

[52] Y. Vasiliev, *Natural Language Processing with Python and Spacy*. William Pollock, 2020.