Evaluation of the most recent Machine Learning approaches for the wastewater treatment

Pasquale MERLA*, Michele Dassisti**, Giambattista Stigliano***
*Ali6 Srl, Monopoli (BA), Italy
**Politecnico di Bari DMMM, Viale Japigia 182, 70046 Bari, Italy
*** InresLab Scarl, Monopoli (BA), Italy
(e-mail: *merla@ali6.it, **michele.dassisti@poliba.it, ***g.stigliano@inreslab.org)

***Abstract*— The main goal of this paper is to present a novel approach of the Machine Learning (ML) in predicting the trend of the water properties within wastewater treatment processes. Bibliographical analysis shows legacy approaches the most frequently adopted, such as the Feedforward Neural Network – FNN with few layers. The ML models proposed here uses, instead, Convolutional Neural Network – CNN with several layers and a number of parameters. The use of the embeddings is also proposed to manage the categorical features and thus reach an higher performance. A real case example of application is presented, by analysing real data in a given period to prove the quality of the ML algorithm architecture designed.**

I. INTRODUCTION

The wastewater treatment is a delicate and important activity for the health of humankind as well as for the environmental sustainability. The optimal control of the water treatment processes plays a critical role, provided an excess or defect on the estimation of chemical treatment components quantities can hurt the effectiveness of the water purification process. The goal of the present work is to benchmark the most modern approaches of Machine Learning (ML) in predicting the properties of the fluid subdue to a purification treatment process.

Compared to the past, low-cost information archiving techniques allows to accumulate it without necessity of filtering [1]. As a consequence, a growing availability of information results that increases the difficulties related to the abstraction of valid models for the analysis of the data themselves. The increased number of samples and the variables involved make it difficult to analyse data using purely statistical methods or rules-based programming. Rule-based programming needs the definition of the IF-THEN rules that link an action to an event: this implies the need to indicate a priori all the situations that may occur, to codify the rules necessary to manage them. It is thus necessary to create new rules to adapt the operation of the system to data variation. In a context in which data often change at a faster rate than the time needed to code new rules, this type of programming results ineffective [2] [3]. This complex computational scenario has favoured the increasing of the applications that use ML algorithms. In this type of algorithms, it is not necessary to specify exactly the expected behaviour, leaving the algorithm to deduce from a training set supplied.

Unfortunately, models for wastewater treatment existing in literature rarely adopt the most recent ML techniques even if, in recent years, strong advancements have been done in this field. To the best of our knowledge, there aren't literature work that apply temporal CNN [4] to predict parameters of an incoming fluids to a wastewater treatment plant, while this approach reached good results in similar field. In the model proposed here we tested the use of the embeddings [5] to take into account categorical features. Embeddings allows to treat categorical values as a point in a multidimensional hyperspace.

The paper is organized as follow: Section II describes the state of the art about the use of the ML to predict the properties of the fluid entering into a wastewater treatment plant. Section III describes the proposed approach and a real test case that use the CNN to predict COD and NO_3. Section IV shows the obtained results. Finally, the conclusions and the future research questions are drawn in Section V.

II. RELATED WORKS

This Section describes the state of the art about the use of the ML for the prediction of the properties of the fluid in waste-water treatment. The work proposed in [6] describes a model based on FNN [2], [7]–[9], to predict DO (dissolved oxygen) and BOD (biochemical oxygen demand), in Gomti river in India. The Dataset was created by the monthly monitoring of the Gomti water in eight different points, in two different periods of time January 1994 – December 1999, and January 2002 – December 2005. The Dataset includes 13 features (11 plus DO and BOD). Of the 960 available sample, 576 was used for training, 192 for the validation, and the last 192 for the test. Two different FNN was created, to predict DO and BOD. Both the FNN had one input layer with 11 neurons, one hidden layer with 23 and 11 neurons for DO and BOD prediction respectively, and one output layer with one neuron to represent the feature to predict (DO or BOD).

The work in [10], aims to apply a generic FNN for regression (GRNN) [11], to predict BOD in a waste-water treatment plant in Algeria. The proposed GRNN had four layers: input, pattern, summation, output. The Dataset was composed by 691 samples (one per day) with 6 features, for two years, from August 1st 2009 to July 31th 2011. The 80% of the 691 samples was used for training, the last 20% for validation. The number of neurons in the input layer corresponds to the number of input features. The number of neurons in the pattern layer corresponds to the number of training pairs, identified as each possible pair of input – output units. The summation layer is characterized by the presence of only two neurons, while one only neuron forms the output layer.

In [12] a FNN to predict DO is proposed. The case study refers to the estimation of DO concentrations

downstream of the city of Mathura, in India, which rises along the banks of the Yamuna river. The dataset is made of monthly samples with 5 features, carried out on three survey stations which were located respectively upstream, in the centre and downstream of the city of Mathura. Three different FNNs were created, which corresponded to different inputs, but which shared the output represented by the DO concentration downstream of Mathura. In the first case the neurons of the input layer was represented by the samples deriving from all three stations except for the output, for a total of 14 neurons; in the second case, the neurons of the input layer correspond to the samples deriving from the station in the centre and from those upstream of the city were input (10 neurons), while in the last case the neurons of the input layer were represented by the measurements of the Mathura upstream station alone (only 5 neurons). In the hidden layer there was an unspecified number of neurons that could not exceed $2 \times n + 1$ (with n = number of inputs). The output layer had only one neuron for the unique feature to predict (downstream DO of the city). The reference dataset is composed of 72 surveys of samples for each station made monthly from 1990 to 1996. For the training they use 48 of the 72 total samples, while for the validation the remaining 24.

In [13] the authors apply the Wavelet approach to predict the DO concentration in water, comparing the results with other architectures. Each sample included 6 features: solar radiation, water temperature, DO, pH, humidity and wind speed. Data were collected every 60 minutes from 21 to 27 July 2010 for a total of 168 samples. Of these 144 they formed the training set and the remaining 24 (which referred to the last sampling day), were used as a test set. According with some of the other works presented so far, the proposed architecture provides a three-layer FNN: one input, one hidden and one of output. Similarly, in this case the number of neurons for the input and output layers depends on the number of features entering and leaving the FNN. The number of neurons in the input layer will therefore be equal to 6, while in output there is only one neuron. In the middle there is a hidden layer consisting of 4 neurons. The input will then be characterized by all six features sampled in the first half hour of detection, which will output the DO forecast in the next half hour. For the training phase a maximum number of 500 epochs was envisaged, with a desired error of 0.005 and a learning rate of 0.3.

In [14] authors propose to predict the water quality of the Büyük Menderes river in Turkey, using an algorithm that combines an ARIMA model and a FNN model. The hybrid model used considering that the ARIMA models are not able to approximate the non-linear characteristics of the data, while the FNN models, approximate the non-linearity better than they are able to do with the linear type characteristics. The dataset includes monthly samples of three parameters: water temperature, DO and boron B concentration, for a period of nine years (1996-2004). Of the 108 total samples, the first 72 were used for the training phase, while the remaining 36 were used for the test phase. The FNN used in this work includes one input layer, one output layer and a single hidden layer. The data are processed in a first step by the ARIMA, which captures its linear characteristics. The result of this phase is the production of residues characterized exclusively by non-linear characteristics, ready to be processed by the FNN model.

In [15] the authors developed a set of ANNs mediated by an ensemble method, for the prediction of pH, DO and turbidity of the waters of the Nakdong river, in South Korea. The choice of using ensemble method is justified by the authors' attempt to eliminate the influence on the performance of the model of the initial choice of weights, mediating between them the results obtained by different ANNs. In the specific application, there are three architectures mediated:

- only one hidden FNN layer;
- a multilayer FNN;
- a Recurrent Neural Network (RNN).

Clustering algorithms were also adopted for the data related to water turbidity, so that they were divided into classes, before using the ensemble method. As for the dataset, the daily data for a period ranging from 2009 to 2012, of PH, DO and turbidity were observed, for a total of 785 samples. Given the data collected on day t and t-1, the goal was to predict the water quality per day t+1. A number of tests were performed that varied with respect to the samples considered during the training phase and with respect to the number of neurons in the hidden layer.

To the best of our knowledge, there are no works in the literature that refer to the particular context of this work (tertiary wastewater treatment) and contemporarily use the most recent ML approaches.. Many of the proposed architectures, consider only FNN consisting of a single hidden layer and characterized by an extremely limited number of neurons. In none of the cases proposed the temporal CNN have been used, provided they are extremely effective in the analysis of time series: their effectiveness lies in the reduced demand for computational resources undergoing training and running. Likewise, under no circumstances were categorical variables using the definition of embeddings [5], whose use has proven to bring many benefits in terms of performance, as demonstrated in Google and GloVe's Word2vec [16], [17]. The papers in the literature use only legacy ML techniques, so in this paper, we have applied two of the most recent ML approaches (i.e. temporal CNN and embedding for categorical features) to make predictions of COD and NO_3 for the incoming fluid in a wastewater treatment plant.

III. PROPOSED APPROACH

A. ML algorithm

In this paper we propose two different types of CNN to analyse wastewater treatment sustainability, that differ in the use of embeddings to consider categorical values. In particular, the use of temporal CNNs has been devised as the optimal one. For each type, we have created two different CNNs to predict NO_3 or COD, as it will be shown in the subsequent case.

The goal of the proposed ML algorithm was to predict two critical wastewater parameter (say NO_3 and COD) for the a given period of observation (10 hours, corresponding to the mean time for waste-water treatment) starting from a set of samples (last 24 hours as explained in the next paragraph).

For all the CNN configurations (see example in Figure 1), there are 4 different branches, one for each feature in

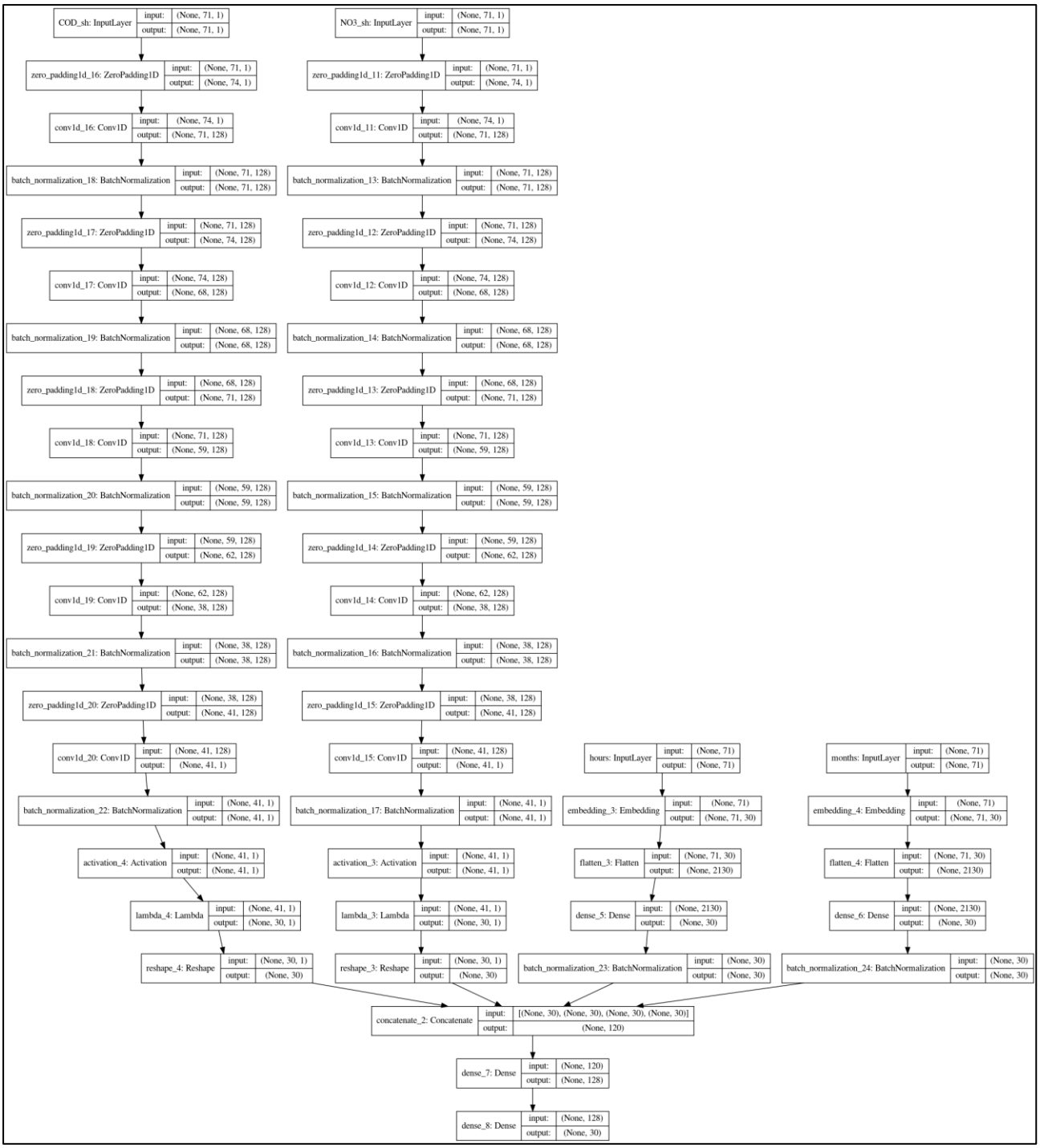

Figure 1. CNN architecture with embeddings for categorical features

input. First two branches (left of Figure 1) refer to two features characterized by real values. They are both composed by a set of convolutional layers (Conv1D) with Zero-Padding and Batch normalization, with different size. The other two branches refer to two categorical features (these features are categorical because they take only a limited number of integer values (from n to m)). Firstly, this type of feature have been taken into account with word embeddings, that allows to represent a word as a vector [16], [17]. In this case (example in Figure 1) the two branches are composed by an input layer, an embedding layer, a flatten layer, and a dense layer with Batch normalization. Secondly, to compare the accuracy of the prediction with and without the use of the embeddings, the two categorical features have been taken into account with only dense layers.

A concatenation layer has been added, to concatenate the four branches and a set of dense layer to allow reaching the desired output (30 values corresponding to the following 10 hours).

To benchmark the proposed approach we use the Mean Absolute Percentual Error – MAPE as the difference from predicted and actual data.

B. *A real test case*

The ML algorithm described in the previous section was tested in a real case that refers to an artificial basin for tertiary treatment of waste water in southern Italy..

1) Dataset and pre-processing

A real Dataset has been chosen to test the proposed algorithm characterized by measurements taken in a period from July 1st, 2014 to June 30th, 2016 every 4 minutes, in an artificial reservoir of wastewater treatment, for about 263000 samples. Each sample was characterized by the following features (Figure 2):

- **COD** (Chemical Oxygen Demand);
- **NO_3** (Nitrates);
- **Temp** (Temperature).

The sampling period equal to 4 minutes was considered too low due to excessive noise in the data and because, considering the phenomenon under examination, it is not possible to appreciate any significant variation in such a short time. For this reason, the dataset has been resampled considering the need to reduce noise, to be able to consider significant variations, and continue to have a sufficiently high number of data. The result was the determination of the new sampling period of 20 minutes, going from about 263000 samples to around 53500.

After resampling, we deleted all wrong data (for example NO_3 equals to zero because it is impossible to have this value for NO_3), we introduced new features that are strongly related to NO_3 or COD, and after a preliminary test we decided to ignore Temp because poorly correlated with NO_3 and COD. The features used as input of the proposed ML algorithms were:

- **NO_3_sh**: difference between the logarithm of two following NO_3 measurements;
- **COD_sh**: difference between the logarithm of two following COD measurements;
- **Hours of the day**;
- **Month of the year**.

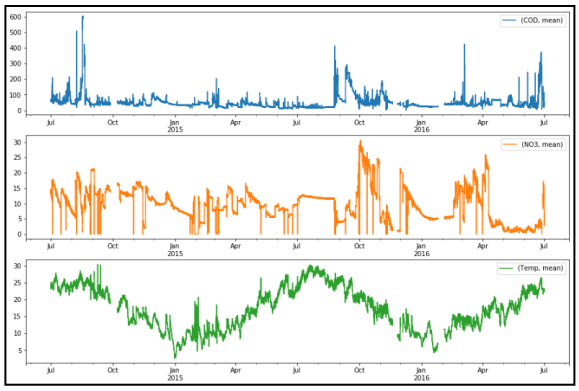

Figure 2. Data trend

2) Training, validation, and test sets

After resampling, we obtained 53500 samples (3 samples/hour). 34 consecutive hours, correspond to a total of 102 consecutive samples. There were approximately 520 surveys conducted, each survey contained 102 consecutive samples, for a total of 53500 samples measured. The test set was composed of approximately 100 surveys. Approximately 43300 samples remained after the test surveys selection which can only fulfil approximately 420 surveys. 420 surveys is not sufficient to create a CNN training set. In order to increase the number of surveys, and so, to satisfy the CNN training set sample size requirement, we created p+1 surveys from a set of 102+p samples. For example, 105 samples satisfy 4 surveys by a single shift in sample (0-101, 1-102, 2-103, 3-104). By adopting this procedure it was possible to obtain about 30100 distinct surveys, a sufficient number for the training set of the CNNs.

IV. RESULTS

In previous Section we described the ML architectures used in proposed ML algorithms. In this section we discuss the obtained results. For each data surveys in test set we predict the last 30 points (10 hours), starting from the first 72 (24 hours), as depicted in Figure 3.

To evaluate the prediction accuracy, we compare the predicted points (red in Figure 3) with the last 30 points of the surveys in test set (they were composed by 72 values used as x and 30 as y).

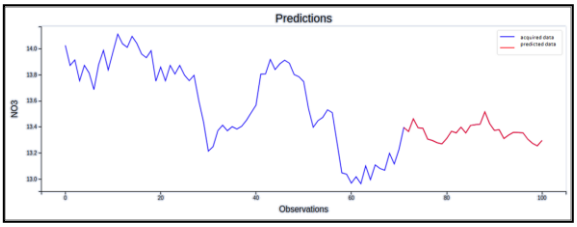

Figure 3. Predicted values (in red) of NO_3, starting from actual samples (in blue)

To compare the predicted value with the true ones we use the Main Absolute Percentage Error (MAPE). Figure 4 shows MAPE for each ML model.

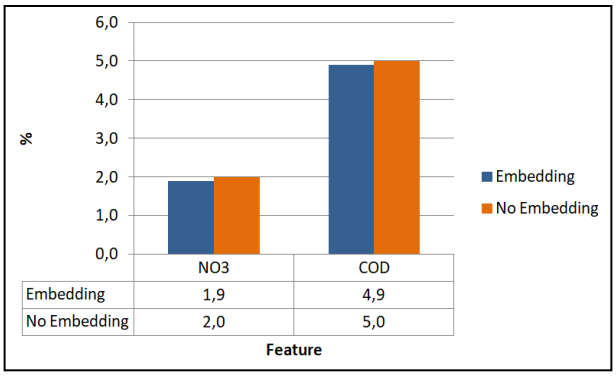

Figure 4. MAPE

The MAPE is less than 2% NO_3 prediction and less than in 5% for the COD prediction. The use of embedding has not produced the desired effects. Their use has in fact resulted in a performance improvement that is too small to consider the use to be advantageous.

V. DISCUSSION AND FUTURE WORK

This article describes how, through the use of temporal CNNs, it is possible to predict the properties of a fluid

entering a wastewater treatment plant with a MAPE less than 5%. The main purpose was to show how the use of this approach is promising with respect to existing approaches. Future research will try to improve the results obtained through the use of new features deriving from data not directly related to the surveys used (e.g. weather data) and through ad-hoc fine-tuning process. It will be thus necessary to compare the results obtained with those resulting from ML approaches commonly adopted in the state of the art.

ACKNOWLEDGMENT

The scientific contents described in this paper are disclosed with the permission of Società Chimica Mediterranea (SCM) Srl, which committed research project titled *"SPOTT - System for the prediction of the physico-chemical characteristics of the incoming water, necessary for the optimization of the polyelectrolyte administration in the purification process of the waste water"* to InResLab scarl.

REFERENCES

[1] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2016.

[2] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerg. Artif. Intell. Appl. Comput. Eng.*, vol. 160, pp. 3–24, 2007.

[3] K. Guruswamy, "TeradataVoice: Data Science: Machine Learning Vs. Rules Based Systems," *Forbes*, 15-Dec-2015. [Online]. Available: https://www.forbes.com/sites/teradata/2015/12/15/data-science-machine-learning-vs-rules-based-systems/.

[4] A. van den Oord *et al.*, "WaveNet: A Generative Model for Raw Audio," *ArXiv160903499 Cs*, Sep. 2016.

[5] C. Guo and F. Berkhahn, "Entity Embeddings of Categorical Variables," *ArXiv160406737 Cs*, Apr. 2016.

[6] K. P. Singh, A. Basant, A. Malik, and G. Jain, "Artificial neural network modeling of the river water quality—A case study," *Ecol. Model.*, vol. 220, no. 6, pp. 888–895, Mar. 2009.

[7] K. P. Murphy, *Machine learning: a probabilistic perspective*. Cambridge, MA: MIT Press, 2012.

[8] G. Bontempi and S. Ben Taieb, "Statistical foundations of machine learning," *Univ. Libre Brux.*, 2017.

[9] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[10] S. Heddam, H. Lamda, and S. Filali, "Predicting Effluent Biochemical Oxygen Demand in a Wastewater Treatment Plant Using Generalized Regression Neural Network Based Approach: A Comparative Study," *Environ. Process.*, vol. 3, no. 1, pp. 153–165, Mar. 2016.

[11] D. F. Specht, "A general regression neural network," *IEEE Trans. Neural Netw.*, vol. 2, no. 6, pp. 568–576, Nov. 1991.

[12] A. Sarkar and P. Pandey, "River Water Quality Modelling Using Artificial Neural Network Technique," *Aquat. Procedia*, vol. 4, pp. 1070–1077, 2015.

[13] L. Xu and S. Liu, "Study of short-term water quality prediction model based on wavelet neural network," *Math. Comput. Model.*, vol. 58, no. 3–4, pp. 807–813, Aug. 2013.

[14] D. Ömer Faruk, "A hybrid neural network and ARIMA model for water quality time series prediction," *Eng. Appl. Artif. Intell.*, vol. 23, no. 4, pp. 586–594, Jun. 2010.

[15] S. E. Kim and I. W. Seo, "Artificial Neural Network ensemble modeling with conjunctive data clustering for water quality prediction in rivers," *J. Hydro-Environ. Res.*, vol. 9, no. 3, pp. 325–339, Sep. 2015.

[16] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543.

[17] R. Řehůřek, "Deep learning with word2vec and gensim | RARE Technologies," 17-May-2013. [Online]. Available: https://rare-technologies.com/deep-learning-with-word2vec-and-gensim/.