Development of a Digital Twin to estimate flow in a water supply network

De Araújo, José * **Ravel, Pedro.** * **Silva, Diego.** * **Macedo, Euler** *
Villanueva, Juan * **Neto, Aristóteles** ** **Silva, Halan** **

* *Faculdade de Engenharia Elétrica, Universidade Federal da Paraíba, PB, (e-mail: jose.araujo@estudante.cear.ufpb.br).*
** *Fábrica de Vidros Planos, VIVIX, aristoteles.neto@vivix.com.br*

Abstract:
The digital twins have been emerging as an important solution for industrial processes for providing simulation models capable of imitating the physical system. However, there is not yet an unanimous approach for creating digital twins, considering the difficulties of using an accurate technique that have a rapid update to the new outputs caused by changes in the system as equipament efficiency, for example. In this sense, this work compares 3 machine learning techniques to modeling a water supply network (specifically, predict the flow in a point of the network), specifically a classic Artificial Neural Network and CatBoost, against the proposed algorithm, the KNN (K-Nearest Neighbors) which have incremental learning. Due the incremental learning approach it can be shown that KNN have superior performance than CatBoost and ANN (these techniques have transfer learning and need retraining, respectively), having better performance in both the first phase (training) and the last (the incremental learning), showing its potential for application in digital twins.

Keywords: digital twin; incremental learning; water system modelling; machine learning; continual learning; flow estimation.

1. INTRODUCTION

The digital twins has emerged as a great tool for optimization of industrial processes, through digital replication of the corresponding physical entity. This concept was developed first by NASA during the Apollo 13 mission in 1970 and came out after in 2010 in the final release of the NASA modeling, Simulation, Information Technology & Processing Roadmap (Shafto et al. (2010)). Since then, there were many attempts to establish an well defined conceptualization of the digital twins (Jones et al. (2020)), but in general, the concept should include three components Grieves (2014):

- A physical product
- A virtual representation of that product (included the modeling, testing, optimisation, etc of the physical process)
- The bi-directional data connections feeding data from the physical to the virtual representation and conversely information and processes from the virtual representation to the physical.

In these context, it is particularly important that the digital twin is a faithful copy of the physical counterpart, so that the systems modifications through the time in the physical twin be absorbed by the digital representation.

On the other hand, it is important to improve the conditions of water supply throughout the world, mainly because of high waste rates involved in the use of this essential and finite resource. The world is far away from the goals established in the 2015 United Nations Summit. According to the Sistema Nacional de Informações sobre Sanemento (SNIS), the waste rate of distribution in Brazil is around 40% SNIS (2022). This means for each 100 liters of water, 40 liters are lost due to leakings. This is an alarming rate and justifies the search for solutions.

A difficult of this kind of system is the pressure regulation in the network, since the demand varies throughout the day Anele et al. (2018). During high demand periods, the systems may work fine; however in the low demand periods, the lack of an adequate control leads to overpressurization in the network. When much more water is fed into the network than is consumed, burst pipe rupture may occur and therefore waste of water Jara-Arriagada and Stoianov (2024). In addition to that, pipe breaks causes are not fully understood Jara-Arriagada and Stoianov (2021). One possible way of solving this problem is make a pressure control through frequency inverters and pressure regulating valves. Moreover, it is fundamental to have knowledge of network behavior in different consumption situations in order to evaluate the optimized operating scenario. Therefore, modeling the distribution network, whether through mathematical equations or using artificial intelligence, is essential.

In this sense, this work aims to present a performance comparison of different machine learning techniques to enable the development of digital twin of an hydraulic network. Each approach has its own methods of model updating and intelligence building, so that its prediction accuracy and learning strategy can be evaluated.

1.1 State of Art

In specialized literature, some approaches for modeling systems for developing digital twins are presented. For some approaches, modeling through physical equations is preferable as in Zhou et al. (2022) and Zeng et al. (2023), in which the software Modelica was used for modeling thermoelectric systems. This software has packages for electric, thermal and mechanical modeling, etc. The main advantage of this method was the development of a tool for operation and failures simulation. Nonetheless, these works did not present a method for updating the model if it began to deviate from the behaviour of the physical system.

In this sense, modeling approaches based on machine learning techniques stand out because they are more flexible to be updating, not depending of physical equations to correct the deviations of its predictions from the real system. For example, in the work of Huang et al. (2022) a Combined Cooling, Heating, and Power-Cold Energy Recovery system was modeled, which included a series of turbines and heat exchangers, which would be very difficult to accomplish through mathematical modeling. So they chose to model the system using a Cascade-Forward Neural Network that had only 4 inputs and 3 outputs, simplifying plant modeling. Also in this work, they managed to simulate optimal operation situations through an optimization algorithm for each season of the year.

In some other digital twin solutions, ANNs have also been used in the context of combining the attributes of machine learning models with physical equations. In these approaches, when possible, equations representing the physical process to be modeled are used to improve the ANN's estimate, as was done in Sun and Shi (2022) and Yang et al. (2024). However, as is obvious, this approach requires advanced knowledge of the process.

But one of the approaches that meets the concept of continuous learning of digital twins and is present in a Python implementation is the River package Montiel et al. (2021a). This module is specialized in algorithms based in incremental learning, which has implementations of algorithms of different types, such as dicision trees, linear regressions and KNN (K-Nearest Neighbors). KNN is a excelent technique for modeling systems, as it is entirely based on data storage, which in the River's implementation has a FIFO (First in, First out) structure, erasing older data of the memory.

On the other hand, another machine learning technique that has gained notoriety is the CatBoost, an algorithm based on Gradient Boostings Prokhorenkova et al. (2018). This family of algorithm has been applied in many competitive machine learning problems, as it can be find out in the Kaggle platform. In this context, two approaches that will be compared for the digital twins updating arise: through online learning and through transfer learning.

In the online learning, the learning process is continuous: the learning-model update its parameters from continuous data entry, that is, adjust its intern parameters one at a time Hoi et al. (2018). In this context, the online learning can be mede as presented in the Figure 1. In this case, as suggestion of learning metric, the online learning model can adjust its own parameters only when the relative error is higher than a defined target accuracy.

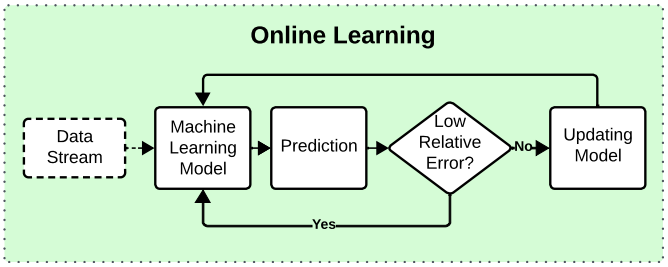

Figure 1. Online machine learning code diagram.

On the other hand, approaches based in transfer learning have a characteristic of having two learning stages: In the first (the preparation stage), the data are divided into two parts (train and test) which are used for training and model calibration so that it become the most accurate possible relating to the output values; after that the machine learning model begin the operation phase in real-time, which naturally as the time goes start to became inacurate due to context changing Weiss et al. (2016). This way, the model is retrained again using new data which was not used in the first training because they was collected during the operating time. An ilustration of this aproach of model updating is shown in the Figure 2.

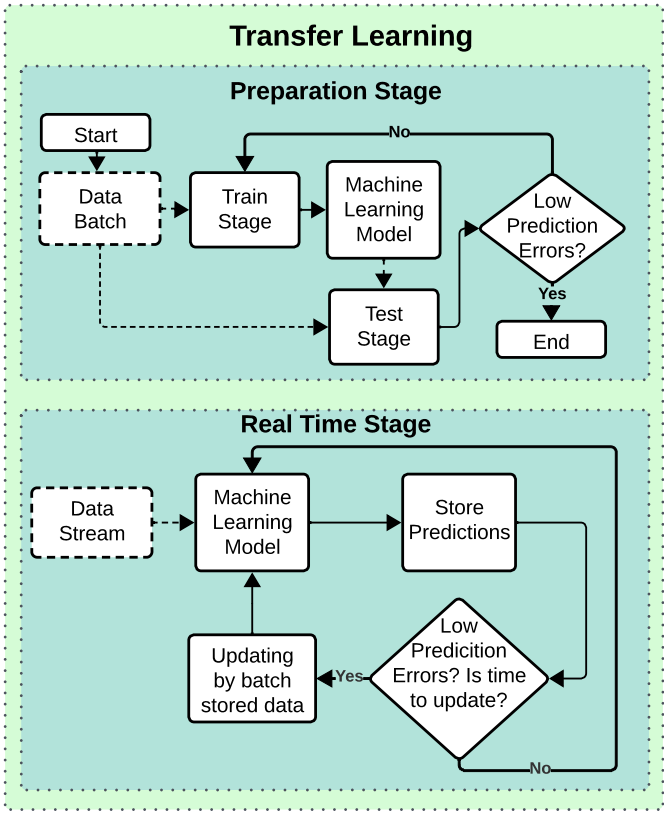

Figure 2. Transfer learning code diagram.

In this case, one way to achieve better quality in the model updating is to keep fixed some of the parameters callibrated in the first stage and update only a small part of them, as validated in Kumar et al. (2022). For example,

if the model is an artificial neural network, the weights can be callibrated in the first stage and in the seccond stage change only the weights of the last layer, as a way to retain knowledge gained in the hidden layers. This way, the knowledge of the model remains and only a small adjust is done to make predictions closer to the actual scenario. An ilustration of this example is shown in the Figure 3.

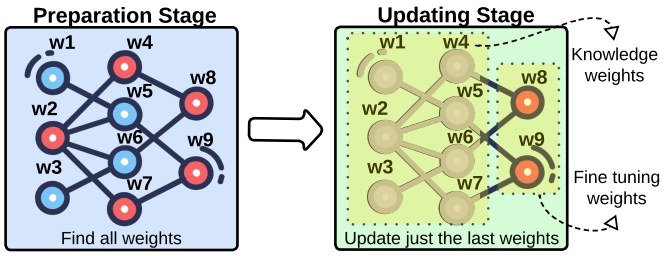

Figure 3. Example of how to do transfer learning.

Each of these two approaches have its own advantages and disvantages which should be analysed according to the process to model. Some of the main problems in the model update are:

- **Drifting**: Occur when models reduce its performance due to process changes through the time, losing precision in the predictions. This is a fundamental issue for digital twins;
- **Catastrophic forgetting**: Occur when the model after several updates overfit the new data learned and forgets the previous knowledge. This way, it can forget how to generalize the problem, becoming worst than the before the update.
- **Updating time**: Depending on the process it is vital the model to be available for desicion making. So, the time to update the model should be short enough to not hinder decision making.

Given the widespread use of neural networks for digital twin applications, the notoriety achieved by the CatBoost algorithm in machine learning problems in general and the incremental learning approach that KNN can offer, these three techniques were chosen for study in this work.

2. METHODOLOGY

Multi-Layer perceptron networks are a type of Artificial Neural Network, which are computational models of interconnected neurons that through mathematical operations seems to imitate human brain behavior such as pattern recognition and learning Bao et al. (2024). CatBoost, in turn, is a form of the gradient-boosting decision tree with enhanced capabilities, able to deal with nonlinear data Chelgani et al. (2024). K-nearest neighbors is an algorithm that classifies one element based on the categories of it's N nearest neighbors Zhu et al. (2024).

In order to compare different machine learning techniques applied to develop a digital twin and discuss the advantages and disvantages of each one, we choose three techniques to focus on: a Multi-Layer perceptron network in which only the last hidden layer will be updated, CatBoost with retraining and KNN with online learning as implemented by Montiel et al. (2021b).

In order to deal with this issue, each technique - ANN, CatBoost and KNN - has its own approach:

- **ANN**: it does the retraining using batches of data, having no support for incremental learning;
- **CatBoost**: create new trees to complete the previous learning;
- **KNN**: the KNN learning process is to store new data through memorization via queue.

2.1 Case study: setting the environment

To validate the modeling techniques proposed for developing a digital twin, it was considered an hydraulic network of the Laboratório de Eficiência Energética e Hidráulica em Saneamento (LENHS), UFPB, Brazil. This network has frequency converter-controlled feed pumps, pressure control valves and flow sensors and pressure sensors distributed throughout the pipeline. These hydraulic network is illustrated on Figure 4, in which PT means Pressure Transducer and FT means Flow Transducer.

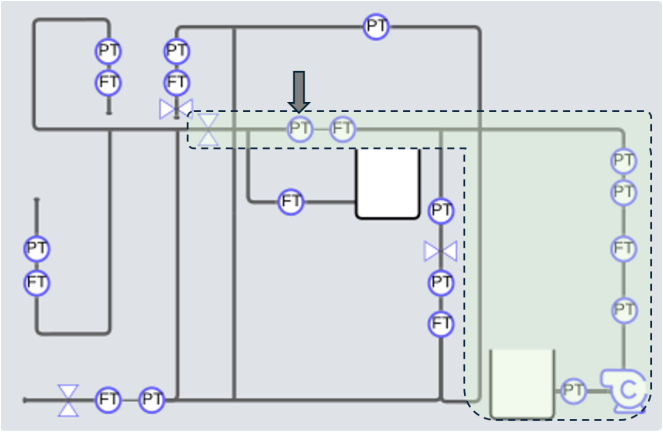

Figure 4. Illustration of the case study hydraulic network.

According to what is shown in the Figure 4, the LENHS's hydraulic network has a series of derivations in which the water can flow, and which could be studied. However, since our main goal in this work is to validate the modeling medodology to develop a digital twin, it was chosen to consider only the initial part of the system (filled in green color). This is a very important excerpt, because it includes the pump, which is the main process control element, so that the use of a digital twin in this system could be applied to monitor the changes caused by time, for example, when comparing different results of pressure and flow outcomes for the same pumping setpoint due to roughness increase on the pipeline or pump degradation. The PT sensor indicated by the arrow represents the output that the models used as a reference for learning.

2.2 Case study: setting the modeling

Given the excerpt chosed for study, it was chosen the input and output variables for the digital twin model considering the criteria: instrument measurement quality and its influence on output variables. For this study, it was chosen the flow measured right after the pump injection as target output, given the importance of knowing the volume of water injected into a distribution network.

On the other hand, for the inputs it was chosen the well calibrated instruments, in our case, the Pressure sensor highlighted in the Figure 4; For the system acting elements it was chosen the frequency in the frequency converter, based on tests done on different frequencies, for greater variability of process behavior. About the valves, it was chosen to keep them at a fixed value in a first moment and then vary them in the cycles of increase in the frequecy. To illustrate this, the operation variation is shown in the Figure 5.

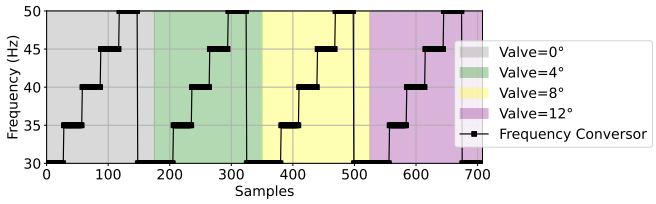

Figure 5. Frequency and valve's angle operation.

Regarding the issue of digital twin modeling, in the Figure 6 it is shown a diagram of inputs and outputs in which the valve angle was omitted. This was done because the goal of this work is to emulate network degradation in an water distribution network using a valve angle change to create a cargo loss. The impact of this change should be considered by the model through the relation with the other variables.

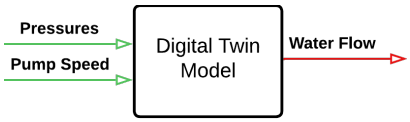

Figure 6. Input-output diagram of the modeling.

Regarding the data collection, it was done in a sample time of 1 second, in a 12-second window, increasing the rotation speed of the frequency conversor in 5 Hz each 3 minutes, generating 705 samples (during the acquisition process, some date was lost). The data was divided in 3 moments, according to Figure 7. The first moment represents a new network, in which the model was trained for the first time. In this first moment, 20 % of the data was used for validation; the second moment was applied to retrain the model and improve performance until the end. Finally, the test period is a period in which retraining based models (ANN and CatBoost) uses only for testing, while KNN continues to learn due to its incremental learning characteristic.

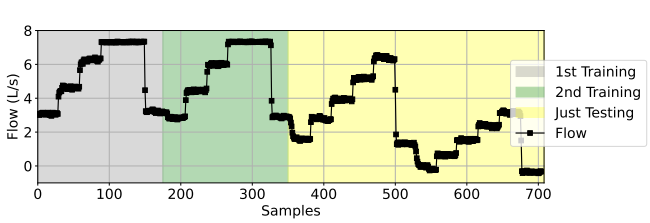

Figure 7. Division of training and testing data.

To ensure that the comparison between the algorithms would be fair considering their versions with optimized hyperparameters, the Bayesian hyperparameterizer was used, as implemented in Nogueira (2014–), which unlike approaches such as Random Search and Grid Search, the Bayesian hyperparameterizer is an optimization algorithm that performs a directed search based on iterations of previous parameter combinations.

2.3 Evaluating the models' performance

To compare the performance of machine learning models such as digital twins, 3 error metrics were considered: the Mean Absolute Error (MAE), Mean Squared Error (MSE) and the Mean Absolute Percentage Error (MAPE). These 3 metrics were chosen due to their characteristics, with MAE having a simpler interpretation, MSE more sensitive to outliers and MAPE for a more generic interpretation of the error. These 3 equations can be seen in Table 1, where $\hat{y}$ is the predicted value, y is the actual value and N is the number of samples.

Table 1. Models' Performance Metrics.

Metrics	MAE	MSE	MAPE
Equations	$\frac{1}{N}\sum_{i=0}^{N}\lvert y-\hat{y}\rvert$	$\frac{1}{N}\sum_{i=0}^{N}(y-\hat{y})^2$	$\frac{1}{N}\sum_{i=0}^{N}\frac{y-\hat{y}}{y}$

3. RESULTS AND DISCUSSIONS

After searching for the optimal parameters using the Bayesian hyperparameterizer for each of the proposed algorithms, the ideal hyperparameters were obtained for each of the models, i.e. ANN, CatBoost and KNN. Table 2 shows the optimum values for each of them:

Table 2. Model Parameters

ANN	CatBoost	KNN
- nº hidden layers: 3 - layer 1: 17 - layer 2: 19 - layer 3: 11	- bagging temperature: 16.535 - depth: 9 - l2_leaf_reg: 3.754 - random strength: 7.260	- aggregation method: 1.50 - nº neighbours: 8 - p: 1.197

The first analysis to be done is about the model performance considering only a static phase, i.e, evaluating only the performance at the first phase previously separate, while the valve was 100 % open. In this context, the graphics of Figures 8, 9 and 10 show the results for each machine learning model proposed.

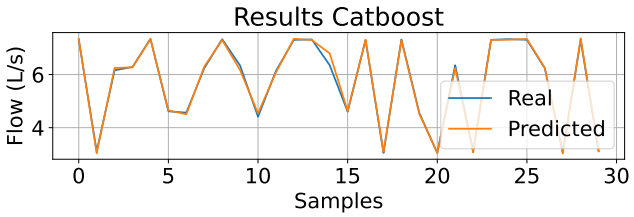

Figure 8. CatBoost's Performance from the 1st period of data.

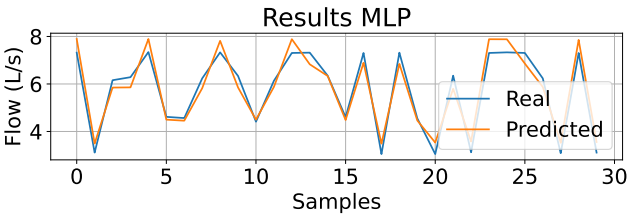

Figure 9. MLP's Performance from the 1st period of data.

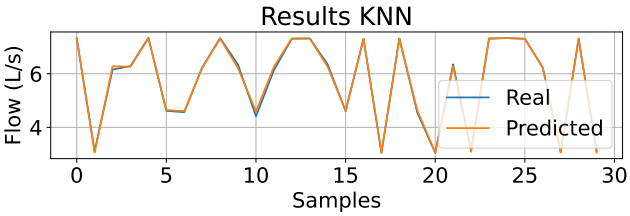

Figure 10. KNN's Performance from the 1st period of data.

In this first evaluation it is visually discernible that CatBoost and KNN had better performance, while ANN had in some points a perceptible distance from the real value. In the Table 1 it is compared the performance using the metrics MAE (Mean Absolute Error), MSE (Mean Squared Error) and MAPE (Mean Absolute Percentage Error).

Table 3. Comparing Models by Error Metrics.

Models/Metrics	MSE	MAE	MAPE
MLP	0.18	0.389	7.33%
CatBoost	0.01	0.058	1.09%
KNN	0.004	0.0429	0.8%

In this comparison, the KNN had the best performance, being more accurate than the other techniques, which is not so common due to the simplicity of its inferring method.

This good performance of KNN ocurred too in the second test phase, in which the fitness of the model for use as a digital twin was evaluated. In the Figures 11, 12 and 13, it can be noted that the algorithms based on retraining lose quality vary fast in spite of the retraining have increased its repective performance, specially the ANN, while the KNN due to its incremental training approached the system real value quickly. However, it is worth observing a negative point that is the presence of spike-type predicions, to which the model was sensitive. This can be caused by some variation in the input data or even the predominance of old patterns in the memory that influenced with great strength at the time of prediction.

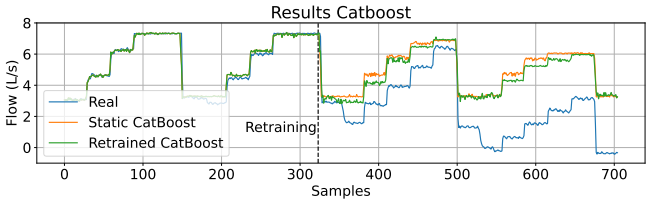

Figure 11. CatBoost's Performance from the all period of data.

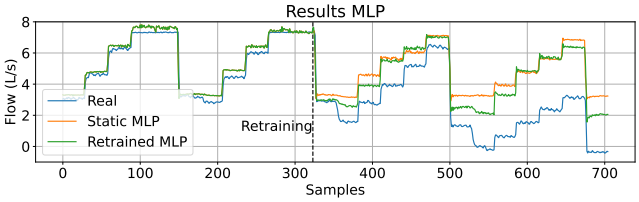

Figure 12. MLP's Performance from the all period of data.

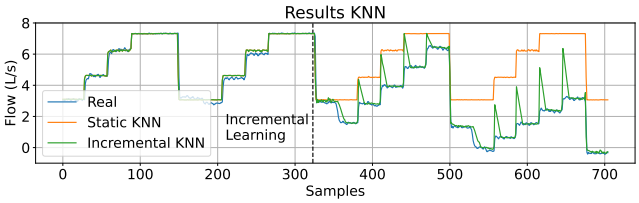

Figure 13. KNN's Performance from the all period of data.

Evaluating only the test period, in the table 2 it is the performance comparison of static and dynamic models. In this table it is clear the performance improvement with the addition of new data, and at the same time it can be noted the difference of prediction between techniques. The numbers of the MAPE metric can be highlighted, because they had high values. These values can be atributted to these metric nature, for it has a division and when the divider tends to zero, the metrics value tends to explode, as it is in the present case for the flow value approaches zero.

Table 4. Final comparison between models.

Models/Metrics	MSE	MAE	MAPE
MLP (Static)	4.35	1.434	366.1%
CatBoost (Static)	4.08	1.381	387%
KNN (Static)	5.03	1.503	364.2%
MLP (Retrained)	1.0	0.700	127%
CatBoost (Retrained)	3.46	1.254	379.2%
KNN (Incremental Learning)	0.27	0.231	28.1%

4. CONCLUSION

It this work, 3 different techniques were tested for use as digital twin, emulating the change in a water supply network causing loss of cargo through the closing of a valve. It was shown that the 3 models were capable of learn the system behaviour while the valve angle was keep fixed, for the absolute errors were lower than $0.38 L/S$. Moreover, KNN had the best performance since the first phase.

When the valve angle was changed, emulating the effect of the natural cargo loss suffered by a real network, the models that need retraining increased its performance but they still fell far short by no longer being retrained after subsequent changes in the system, while the KNN, which practice incremental learning, managed to obtain the best performance in addition to rapid monitoring of the new dynamic (that is, needing only a few samples), showing itself as an interesting technique for application with digital twins.

For future work, we suggest evaluating KNN against other more complex algorithms, such as more robust ANN topologies, as well as other techniques with an incremental learning approach that could outperform KNN.

In addition, the performance of this approach can be verified for different industry problems in order to validate its use for a variety of applications.

ACKNOWLEDGMENT

We would like to thank the VIVIX glass factory for encouraging research and the UFPB postgraduate program.

REFERENCES

Anele, A., Todini, E., Hamam, Y., and Abu-Mahfouz, A. (2018). Predictive uncertainty estimation in water demand forecasting using the model conditional processor. *Water*, 10, 475. doi:10.3390/w10040475.

Bao, W., Cao, Y., Yang, Y., Che, H., Huang, J., and Wen, S. (2024). Data-driven stock forecasting models based on neural networks: A review. *Information Fusion*, 102616.

Chelgani, S.C., Homafar, A., Nasiri, H., et al. (2024). Catboost-shap for modeling industrial operational flotation variables–a "conscious lab" approach. *Minerals Engineering*, 213, 108754.

Grieves, M. (2014). Digital twin: manufacturing excellence through virtual factory replication. *White paper*, 1(2014), 1–7.

Hoi, S.C.H., Sahoo, D., Lu, J., and Zhao, P. (2018). Online learning: A comprehensive survey. URL `https://arxiv.org/abs/1802.02871`.

Huang, Z., Soh, K., Islam, M., and Chua, K. (2022). Digital twin driven life-cycle operation optimization for combined cooling heating and power-cold energy recovery (cchp-cer) system. *Applied Energy*, 324, 119774. doi:https://doi.org/10.1016/j.apenergy.2022.119774. URL `https://www.sciencedirect.com/science/article/pii/S0306261922010546`.

Jara-Arriagada, C. and Stoianov, I. (2021). Pipe breaks and estimating the impact of pressure control in water supply networks. *Reliability Engineering & System Safety*, 210, 107525.

Jara-Arriagada, C. and Stoianov, I. (2024). Pressure-induced fatigue failures in cast iron water supply pipes. *Engineering Failure Analysis*, 155, 107731.

Jones, D., Snider, C., Nassehi, A., Yon, J., and Hicks, B. (2020). Characterising the digital twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29, 36–52. doi:https://doi.org/10.1016/j.cirpj.2020.02.002. URL `https://www.sciencedirect.com/science/article/pii/S1755581720300110`.

Kumar, J.S., Anuar, S., and Hassan, N.H. (2022). Transfer learning based performance comparison of the pretrained deep neural networks. *International Journal of Advanced Computer Science and Applications*, 13(1).

Montiel, J., Halford, M., Mastelini, S.M., Bolmier, G., Sourty, R., Vaysse, R., Zouitine, A., Gomes, H.M., Read, J., Abdessalem, T., et al. (2021a). River: machine learning for streaming data in python.

Montiel, J., Halford, M., Mastelini, S.M., Bolmier, G., Sourty, R., Vaysse, R., Zouitine, A., Gomes, H.M., Read, J., Abdessalem, T., et al. (2021b). River: machine learning for streaming data in python.

Nogueira, F. (2014–). Bayesian Optimization: Open source constrained global optimization tool for Python. URL `https://github.com/bayesian-optimization/BayesianOptimization`.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Shafto, M., Conroy, M., Doyle, R., Glaessgen, E., Kemp, C., LeMoigne, J., and Wang, L. (2010). Modeling, simulation, information technology and processing roadmap.

SNIS (2022). tratabrasil.org.br. `https://tratabrasil.org.br/wp-content/uploads/2024/06/Estudo-da-GO-Associados-Perdas-de-Agua-de-2024-V2.pdf`. [Accessed 06-08-2024].

Sun, C. and Shi, V.G. (2022). Physinet: A combination of physics-based model and neural network model for digital twins. *International journal of intelligent systems*, 37(8), 5443–5456.

Weiss, K., Khoshgoftaar, T.M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1). doi:10.1186/s40537-016-0043-6. URL `http://dx.doi.org/10.1186/s40537-016-0043-6`.

Yang, S., Kim, H., Hong, Y., Yee, K., Maulik, R., and Kang, N. (2024). Data-driven physics-informed neural networks: A digital twin perspective. *Computer Methods in Applied Mechanics and Engineering*, 428, 117075. doi:https://doi.org/10.1016/j.cma.2024.117075. URL `https://www.sciencedirect.com/science/article/pii/S0045782524003311`.

Zeng, G., Wang, J., Zhang, L., Xie, X., Wang, X., and Chen, G. (2023). Multi-domain modeling and analysis of marine steam power system based on digital twin. *Journal of Marine Science and Engineering*, 11(2). doi:10.3390/jmse11020429. URL `https://www.mdpi.com/2077-1312/11/2/429`.

Zhou, D., Gao, H., Wang, W., Cao, J., Yang, W., Zeng, R., and He, Y. (2022). Application of three-flow fusion technology based on modelica in thermal power digital twin. *IEEE Journal of Radio Frequency Identification*, 6, 715–723. doi:10.1109/JRFID.2022.3205855.

Zhu, M., Lin, J., Cao, G., Zhang, J., Zhang, X., Zhou, J., and Gao, Y. (2024). Prediction of constitutive model for basalt fiber reinforced concrete based on pso-knn. *Heliyon*.