SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 11, No. 2, (2025), pp. 609-623
Open Access. Online & Print

DOI: 10.5281/zenodo. 18231755

COST-EFFECTIVE RETROFIT PLANNING FOR AGEING FLUID PIPELINES: A SIMULATION-DRIVEN STOCHASTIC PROGRAMMING MODEL

Fengyang Yuan[1] and V. Ramani Bai[2*]

[1]Department of Civil Engineering, Faculty of Engineering and Built Environment, UCSI University, Kuala Lumpur, Malaysia, fengyangyuan2@gmail.com
[2]UCSI-Cheras Low Carbon Innovation Hub Research Consortium, Kuala Lumpur, Malaysia, ramaniucsi@gmail.com

Received: 24/07/2025
Accepted: 28/10/2025

Corresponding Author: V. Ramani Bai (ramaniucsi@gmail.com)

ABSTRACT

Ageing Heating, Ventilation, and Air Conditioning (HVAC) fluid pipeline systems pose significant operational and economic challenges due to degradation, energy inefficiency, and high retrofit costs. This study aims to develop a simulation-driven, machine learning-based framework to accurately predict retrofit costs and support cost-effective decision-making for ageing pipeline networks. A quantitative research methodology was adopted, where synthetic data were generated via stochastic simulation using Monte Carlo techniques in Python (NumPy and Pandas). Three supervised machine learning models, Random Forest, eXtreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN), were implemented and evaluated using RMSE, R^2, and MAPE as performance metrics. The results showed that XGBoost achieved the best performance with an RMSE of 221.19, R^2 of 0.929, and MAPE of 5.90%, followed closely by Random Forest, while ANN underperformed with an RMSE of 335.03. XGBoost and Random Forest are closely aligned with actual retrofit costs, indicating strong predictive accuracy. The study concludes that ensemble models trained on simulation-derived data offer a robust solution for proactive retrofit planning. The findings have significant implications for enhancing infrastructure resilience and optimising maintenance investment. However, the study is limited by the use of synthetic data and recommends future work with real-world datasets and expanded modelling techniques for multi-objective predictions.

KEYWORDS: Fluid, HVAC Pipeline, Network, Simulation.

1. INTRODUCTION

Managing and improving infrastructure for energy efficiency is challenging because of the issues with ageing HVAC pipelines. With ageing pipelines, their efficiency drops, resulting in increased costs, decreased energy use and higher risks for failure. With fluid pipeline networks receiving greater use in city infrastructure, it is crucial to retrofit them efficiently. The study proposes using a simulation model with machine learning to help make the best decisions on retrofitting, supported by predictive analytics.

HVAC (Heating, Ventilation and Air Conditioning) pipelines are the major points to consider providing the indoor quality of the environment and sustaining levels of energy efficiency through conducted chilled or heated water. The systems are based on a closed-loop piping and pumping circulation and heat exchangers to circulate thermal loads. The efficiency of HVAC pipelines deteriorates over time as the pipelines become clogged with internal corrosion, fouling, fatigue at joints, and sedimentation, which could slow down flow and adversely affect the thermal transfer performance of the pipelines (Liu et al., 2024). This causes more energy consumption, an imbalance in temperature regulations, and an increase in working expenses. To prevent such burdens and risks, the replacement or inefficiency of a piping system as old as HVAC is a significant problem both in the modern urban infrastructure, where energy efficiency is both an environmental and economic necessity.

Recent findings have suggested digital twins and the use of sensors in diagnosis, which in many cases involve large-scale data in real-time, which is not always available or arrives at a manageable cost (Shaheen et al., 2024). Hybrid methods using a combination of simulation and predictive modelling, as well as machine learning, have also displayed potential in more recent times. As an example, Taheri et al. (2021) used the deep learning model to predict the pressure drops in the HVAC systems, whereas Khosravian (2025) studied the ML and CFD-based models to estimate thermal efficiency in pipelines. Only a few studies, however, are available concerning merging simulation-based synthetic data and supervised ML algorithms specifically designed to work with cost-oriented retrofit planning. This paper fills that gap by proposing a framework which makes use of simulated performance degradation scenarios and machine learning in order to promote sound decision-making.

Significant infrastructure systems such as buildings, factories and district heating networks depend mainly on fluid pipelines in HVAC systems. Over the years, pipelines may wear and perform worse due to corrosion, scaling, thermal fatigue and similar causes (Behrooz, 2016). Improving pipeline networks through retrofitting is seen as a good way to restore their performance. Nevertheless, since retrofitting a building is both costly and complex, planning for it must use reliable data and be financially justified. Advancements in simulation and ML provide useful tools for planning infrastructure using data. With network simulations, planners can observe changes in fluid dynamics and practice retrofits with virtual tests (Park et al., 2023). Pairing simulated data with ML technology allows for building models that estimate the rate of decline, related costs and efficiency after various retrofit options are considered (Alrabghi & Tiwari, 2015).

Improper HVAC pipeline retrofitting may also increase its operation costs by 25% a year, as retrofitting poorly designed pipelines may struggle to handle the flow and cause pressure losses, resulting in a wastage of energy, along with capital expenses since retrofit investment is mismanaged (Behrooz & Boozarjomehry, 2017). This necessitates the absolute presence of well-based planning tools that could be guided by data.

1.1. Problem Statement and Literature Gap

The majority of models used for retrofitting fluid pipelines in civil and mechanical systems are still based on traditional and manual techniques (Lee et al., 2020). They often neglect the uncertain variations in pipeline performance and leave out uncertainties in economic aspects. Furthermore, some researchers argue that while simulation tools are popular in design and fault forecasting, incorporating their data with advanced supervised ML models for retrofit planning is still rare (Sani et al., 2025). Evaluating various machine learning models by using standard metrics is not commonly included in current literature, although this is necessary for judging which models are suitable and appropriate to use (Chen & Guestrin, 2016). Few studies have tackled using simulation-based datasets together with interpretable and scalable AI models for the maintenance of ageing HVAC systems (David, 2024; Kliangkhlao et al., 2024; Nashruddin et al., 2025). Therefore, the study introduces a reliable ML approach supported by simulations to assist decision-makers in identifying the most cost-saving retrofit strategies when there is uncertainty.

1.2. Aims and Objectives

This research aims to create a model that uses

simulations and machine learning to optimise cost-effective planning for the modification of older fluid pipeline networks. The specific research objectives are

To simulate the performance degradation and retrofit scenarios of ageing fluid pipeline networks under varying operational and environmental conditions, and generate a comprehensive dataset for predictive modelling.

To implement and compare advanced supervised machine learning models for predicting retrofit costs and performance outcomes, and evaluate their effectiveness.

1.3. Significance of the Study

The study introduces a data-focused technique for retrofitting infrastructure that can benefit the industry. It brings together simulation modelling and machine learning to support decision-making by providing approaches for decision-makers on where to invest their resources. This study makes use of Google Colab, a remote cloud service, to ensure others can access and reproduce the methodology and use it in other types of pipeline networks aside from HVAC. Moreover, stochastic programming tools, after their use, enable the planners to consider risky elements, which create more solid decisions concerning infrastructure development.

2. LITERATURE REVIEW

The recent trend toward modernisation of ageing fluid pipeline networks, especially HVAC networks, has led to much investigation into innovative predictive and optimisation approaches. The infrastructure of fluid pipeline systems and HVAC networks is becoming outdated; scientists are paying more attention to the use of more sophisticated tools that would predict and enhance their effectiveness. One can now use modelling and ML to determine wear in the pipelines and chart appropriate retrofits that are not too expensive. In the current literature review, prior work that entails the simulation in data generation together with the application of ML in predicting the performance in the fluid infrastructure has been reviewed and evaluated in terms of their usefulness in retrofit projects. The review below thus discusses the research objectives for the simulation of aged pipelines, the application of machine learning techniques and the guidelines pertaining to the same.

2.1. Simulation of Ageing Fluid Pipeline Networks for Data Generation

The aged or updated networks that are speculated in fluid pipelines need to be simulated to get knowledge about the dynamics of the system and to generate data-driven models. When using the model examples, engineers may trust that the flow may be modelled with EPANET or OpenFOAM and the pressures distributed and losses of water energy measured in water systems (Arandia & Eck, 2018; Rettenmaier et al., 2019). They can also assist in duplicating the destruction that would be caused on the pipelines by scaling, corrosion or high temperatures. In their analysis, Kazi et al. (2024) stochastically modelled gas pipelines because the demand in the gas market is uncertain. The paper demonstrated that the probabilistic simulations can be useful, and they can indicate the alteration and uncertainty in the pipeline actions as experienced in the upgrade of the system. Moreover, Vilarinho et al. (2017) relied on simulation and used the results as a training set for ML models to optimise pipeline maintenance with dependability standards.

Park et al. (2023) examined the use of digital twin technology for the HVAC industry by processing data from sensors and combining it with simulations to keep a watch on system performance. The use of this approach suggests that the study can join degradation simulation data with other, real or synthetic, data to accurately represent training data. Even though advanced tools are available, a lot of the current research uses models that do not represent all the changes and uncertainty found in actual pipelines. There are still a few methods that rely on simulations to provide organized and labelled data used for developing further machine learning techniques. This study used current simulation programs to produce data that displays various ageing and retrofitting scenarios and helps train and validate predictive models under good conditions.

2.2. Application of Machine Learning Models for Predictive Retrofit Planning

Civil infrastructure uses widely supervised machine learning models for tasks like predicting faults, planning maintenance and estimating existing risks. Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Artificial Neural Networks (ANNs) are well-known as they perform well, handle nonlinear relationships and can work with large datasets. In the context of predictive maintenance, Random Forest can be used because it deals well with a lot of data and complicated relationships between the variables (Tang et al., 2018). RF has also been used by engineers in pipeline systems to predict leaks in pipes and to gauge the remaining life span of the installed facilities, and to predict flow losses. They introduce models that can be interpreted using RF

that are not sensitive to much variance (Sani et al., 2025). It was observed that XGBoost was more efficient than other algorithms in managing both regression and classification problems because of being able to efficiently manage missing values and utilising only a few traits (Biau et al., 2019).

According to Heymann and Schmitt (2023), the XGBoost was efficient in ranking the important degradation factors with respect to their feature importance to monitor the health of pipelines. A collection of Artificial Neural Networks has received high interest due to their ability to exhibit nonlinear relationships. Malek Mohammadi et al. (2019) noted that ANNs work well in both spatial and longitudinal predictions of infrastructure degradation if they are exposed to abundant training data. Simultaneously, they noted that this method is not always interpretable, which makes it difficult when the engineer would desire to explain decisions unambiguously.

Random Forest, XGBoost and ANN were chosen on the basis of the fact that they have achieved good results on structured tabular data. Ensembles, such as RF and XGBoost, are especially suitable in cases of nonlinearity, multicollinearity, and variable interactions, as well as low preprocessing versus models, such as Support Vector Machines (SVM). SVM and linear regressions were taken into account, but could not generalise well in non-linear form or high-dimensional artificial data that negatively affected their usability. ANN was also added to see how it performs on the deeper, non-linear interactions; however, interpretability and overfitting were a concern.

Most of the time, testing these models involves using both error-based and correlation-based metrics. To measure prediction errors, RMSE and MAPE are preferred, and the coefficient of determination (R^2) rates how much a model explains the variance in the output. The authors indicate that it is especially important to use these metrics for benchmarking models before using them to decide on infrastructure projects.

While lots of research work on ML for pipeline diagnostics exists, not many have compared various ML models using identical datasets computed with simulators. It is also true that most methods do not take uncertainty into account, even though this is essential for planning future building upgrades. By training and comparing RF, XGBoost and ANN models on data created by a simulation, this research uses RMSE, R^2 and MAPE to find the most reliable predictors for how much retrofit would cost and how it would perform.

3. THEORETICAL FRAMEWORK

The methodology is based on Stochastic Decision Theory and promotes decisions to be made under the probability and optimisation (Bertsekas, 2019). Such a theory can be used to formally assess potential retrofits by checking the most probable scenarios and the effects of the changes in usage, weather and damage to materials. The study executes the stochastic theory of the decision framework using simulation data on various situations of the pipeline network and learning the same. Such an approach not only assists planners to choose optimal retrofit strategies most of the time it also provides knowledge of the risk and the scope of options that can be expected. The focus on expected value and minimising errors in decision optimisation is supported by using model evaluation metrics.

The use of the stochastic decision theory is shown in the fact that uncertainty was added through Monte Carlo simulations in various pipeline degradation scenarios. It was then learned that the retrofit costs have an expected value through the ML models that are used in the supervised learning and are approximations of the value function in the stochastic decision theory. Although not encountered as a reinforcement learning problem, the predictive model presents a substitute that can be used to make decisions about retrofit strategies in the face of uncertainty.

Although the concept of stochastic programming is applied in this paper, it did not employ a formal structured description as it contains decision variables, objective functions, and constraints. The structure, however, can be augmented as a two-stage stochastic program in which the decision on strategies on retrofits is to be made in the first stage, and the second stage models retrofit cost as a random variable depending on the scenarios of pipeline condition. Subsequent efforts can refine this structure into one in which optimization is embedded as part of the planning.

3.1. Literature Gap

Although simulation and machine learning have long been used in the industry, most studies focus on them independently. Simulations help with testing systems under stress, whereas ML models are created using past data or data collected from sensors. This study combines generating data through simulation and advanced learning models to help guide the retrofitting of old HVAC pipelines. There is not enough literature covering the importance of models that are both accurate and can be easily explained in infrastructure planning. Even

though such models as RF, XGBoost and ANN have always been tested in different infrastructure domains individually, only a few studies compare these models using consistent data and a single way of measuring the results. This study helps by measuring the models and seeing which ones are more appropriate for designing retrofits. Another critical issue is related to measuring how uncertain the results are. This study applies stochastic models to both data collection and analysis to reflect the real-life challenges faced by infrastructure. Embedding these chance elements in simulation-ML frameworks allows this study to present a novel and efficient way to plan updates for old fluid networks.

4. RESEARCH METHODOLOGY

This section includes the process and methodology behind the creation of a simulation and machine learning framework for low-cost updating of aged fluid pipelines, particularly in HVAC systems. The approach is made to handle uncertainty that often affects pipes while making predictive models for assessing retrofitting. The quantitative approach leads the research to use simulation to create the data, machine learning for making predictions and performance measures to assess each model. Pretty much everything, including data processing, building models and visualising the results, is done using Python with Google Colab.

4.1. Research Method and Design

It uses various computational approaches and predictive methods in its research. Through simulation and the use of supervised machine learning, the design can predict both the cost and the performance of retrofits on ageing fluid pipelines. Simulation helps to prepare data that models can learn from, which is based on actual physics occurring in different situations (Alrabghi & Tiwari, 2015; Park et al., 2023). The use of numbers in the field makes it possible to measure objectively, analyse statistics and apply findings to a wide range of examples. The approach relies on stochastic modelling, as it deals with the fact that pipe ageing and related problems are mostly random and unpredictable. According to decision theory, actions for retrofitting systems are planned by looking at expected results that might change as conditions change (Busoniu et al., 2017).

The approach is based on the main principles of operational research (OR), especially the application of stochastic modelling as a means of making decisions in uncertain situations and optimisation of the distribution of resources. Simulation applied in the generation of a scenario is in line with other general methods on discrete-event models prevalent in OR. In addition, the models applied as machine learning fit into a bigger decision-support system based on prediction, which is within the tradition of the OR to merge analytical models and computational intelligence to solve strategic problems related to infrastructure investment. The approach involving the combination of simulation, uncertainty modelling, and predictive analytics reflects the current OR models of asset management and optimising maintenance.

The study's structure follows a sequence in Figure 1:

Figure 1: Study Framework.

4.2. Data Collection Techniques

In this study, stochasticity in a simulation model was achieved through Monte Carlo simulation of various degradation trends of HVAC pipelines in changing circumstances. Randomness was included in input values of the simulations on simulated age progression of the pipes, material degradation coefficient (they were sampled through a uniform sampling distribution) and changing operational parameters such as flow and external temperature. This enabled it to be able to create different scenarios that depict the uncertainty in the real world about pipeline performance. The stochastic framework made the resulting dataset more heterogeneous, including the variability not only in the sense of static inputs but also in the patterns in the degradation of the performance over time. This is a method of probabilistic training that enhances training machine learning models to generalise within a variety of possible future retrofit scenarios.

The data for this study were obtained through simulation modelling of pipelines affected by ageing. The simulations reflect how pipelines deteriorate over time due to different flow rates, temperatures, pressure drops and changes in materials (Bayani &

Manshadi, 2022). Every simulation scenario reflects a distinct arrangement, upgrading state or likelihood of failure. Main output variables are the pressure drop, the flow decrease, energy losses and the expected expenses for upgrading. For supervised learning models, the variables in the dataset are assigned as dependent variables (Hyndman & Koehler, 2006). This simulation model considers the pipeline diameter, its total length, its age, the material it is made of and the rate of flow inside it and the impact of different environmental factors on its performance (Malek Mohammadi et al., 2019). Using Monte Carlo techniques in simulations allows for including different cases of degradation within the dataset (Li & Guan, 2016). The acquired data is managed and prepared through Pandas and NumPy in Python. It is necessary to scale, encode (for categorical data) and find outliers to get the model ready for processing. After preparation, the final data is divided into an 80:20 training and testing split.

The simulations were done using Python 3.10, and stochastic sampling was done through NumPy, and data structuring through Pandas. Although physical CFD models, such as OpenFOAM, were not used, the process of simulation implemented degradation logic in terms of parameterised equations that approximated real-world hydraulic losses. Every simulated scenario was considered a different Monte Carlo iteration, and between 5,000 simulations were performed to provide statistical convergence. Boundary conditions enclosed the length of the pipes between 50-500 meters, diameter between 0.05-0.3 meters and a temperature that varied between -10 °C to 45 °C. Convergence was confirmed by tracking the stability of the distribution of output (e.g., average cost retrofit) of further iterations of the batch. There were no mesh sizes needed since the study and synthetic data generation were using instead of meshed geometries. Nevertheless, realism of simulations was also tested by comparing the sample outputs with reference to empirical degradation profiles reported in the literature (Bayani & Manshadi, 2022).

4.3. Data Analysis Method

Data scientists in Python use the following three machine learning models: Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Artificial Neural Networks (ANN). Using the dataset made from the simulations, the model was trained to estimate the cost of retrofitting and calculate energy savings and lower pressure loss. The following figure 2 indicates the analysis flowchart used in this study.

Figure 2: Analysis Flowchart (AI/ML-Integrated Predictive Modelling Framework).

Three metrics are used to evaluate the models: RMSE, R^2 and MAPE.

4.3.1. Random Forest Regression

Random Forest uses multiple decision trees to train its model and predicts by averaging the output of the trees (Breiman, 2001). Its prediction for a given input x is (in equation 1):

$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x)$---- (1)

Where f_t (x)the prediction of the t-th tree and T is is the total number of trees.

4.3.2. XGBoost

XGBoost produces regression trees step by step and applies a regularisation to generalise the trees. The model minimises the following objective (equation 2):

$L = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k)$-------(2)

Here l is the loss function (e.g., squared error), $\Omega(f_k) = \gamma T + 1/2\lambda\|w\|2$ is the regularisation term for each tree fk, with T as the number of leaves and w the weights (Chen & Guestrin, 2016).

4.3.3. Artificial Neural Network (ANN)

An ANN includes input, hidden and output layers and uses nonlinear activities (such as ReLU). The output y^ for input x is calculated as (equation 3):

$\hat{y} = \sigma(W_2.\sigma(W_1.x + b_1) + b_2$----- (3)

Where W1, W2 are weight matrices, b1, b2 are bias terms, and σ is the activation function.

4.3.4. Model Evaluation Metrics

To test how the Random Forest, XGBoost and Artificial Neural Networks models worked in this study, they were evaluated using Root Mean Square

Error (RMSE), R-squared (R^2) and Mean Absolute Percentage Error (MAPE).

Every metric sees the model's performance and predictions a bit differently, filling in the full picture. RMSE calculates the average size of errors made by the model, since squaring the differences between the predicted and real values gives more importance to large errors.

Since great deviations can be a problem, this method helps predict infrastructure costs over time. When the RMSE is lower, it means the model's estimates are similar to the true values, which signifies the model is very accurate. According to RMSE, XGBoost was the most accurate model in this study.

This means R^2 tells us the part of the variation in the dependent variable that the independent variables can predict. It explains how successfully the model has extracted the main factors in the data. When the value is closer to 1, the model performs very well. R^2 values higher than 0.92 for both XGBoost and Random Forest mean that these tools are well suited for explaining how much the retrofit cost varies.

MAPE is found by computing the difference between predicted and actual value, which is divided by the actual value and the result is represented in terms of percentages. As a result, this is information that is highly understandable to the viewers.

5. DATA ANALYSIS

In this section, the number of simulation studies conducted to tune the process of updating worn-out HVAC fluid pipes is provided. The study uses Random Forest, XGBoost and Artificial Neural Network (ANN) machine learning algorithms in a synthetic data set to compare their performances. To compare the models, RMSE, R^2 and MAPE are considered. The goal is to select the algorithm that estimates retrofit costs more accurately to assist decision-makers in planning the building repair accordingly.

5.1. Data Simulation

Figure 3, which demonstrates the histograms of the individual variables, reveals the distributions of all variables. Most features seem to be uniformly distributed, including such features as Pipe_Age, Pipe_Length, Pipe_Diameter, Flow_Rate, and External_Temp, which displays the intended randomness of the simulation process. Tricky but a categorical variable such as Material_Factor presents three peaks of 1.0, 1.1, and 1.2, which verifies discrete sampling. The target variables, Supplanting non-uniformity: Retrofit_Cost is slightly right skewed with values like the middle going 2500-4000 units, and Energy_Loss is also right skewed, which means that most systems have moderate energy losses, but there are several extreme outliers.

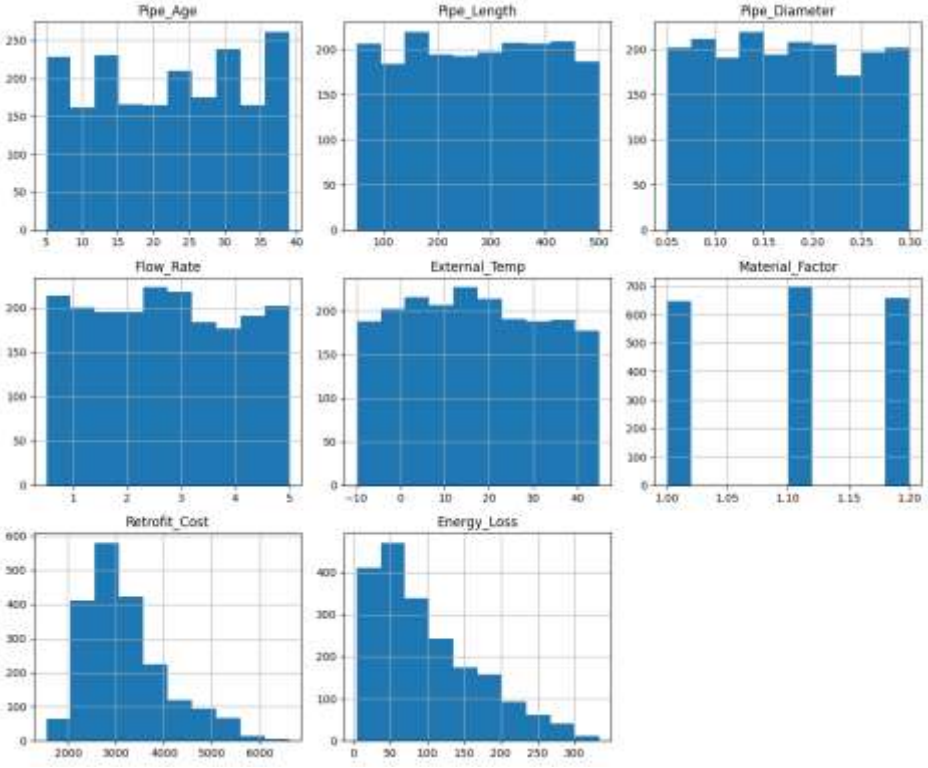

Figure 3: Visualisation of Dataset Variables' Distribution.

In this study, the dataset was artificially created to have similarities with the ageing HVAC fluid pipeline networks. It contains the dataset of 2,000 samples where each sample is a pipeline segment whose features include the age (in years), length (in meters), diameter (in meters), flow rate (in m 3 /s), the external temperature (in Celsius), and a factor that indicates how badly the material is worn. These variables influence two primary target outcomes: Retrofit Cost and Energy Loss. Table 1 provides a snapshot of the first five records from the dataset, showing the variability and realism built into the simulation process.

Table 1: Sample of Simulated Dataset.

Pipe_Age	Pipe_Length	Pipe_Diameter	Flow_Rate	External_Temp	Material_Factor	Retrofit_Cost	Energy_Loss
33	221.87	0.169	2.74	21.17	1.2	2995.28	164.80
19	142.56	0.138	4.29	9.49	1.0	2927.57	124.16
12	104.62	0.141	2.51	10.96	1.1	2801.98	47.74
25	326.76	0.251	3.69	12.65	1.0	2718.57	136.65
23	398.59	0.132	3.79	30.02	1.1	3954.24	142.58

To determine the retrofit cost, pipe age, inverse diameter, pipe length and stochastic noise were used in a combination that was not linear. As well, energy lost in the system was estimated because of pipe age, the flowing liquid amount and a factor specific to the pipe material.

Figure 4 is a pair plot matrix which we can use to get a graphical sense of interactions and possible nonlinearities. The factors are discernible as there is a significant negative trend between the two, as Pipe_Diameter and Retrofit_Cost tend to indicate a tendency of smaller diameters and larger cost of retrofitting. Likewise, Pipe_Age and Energy Loss are strongly linearly related, showing that there is a positive relationship between the increase of pipeline age and energy loss. Other feature relationships are either weak or random, and hence, it would be important to get machine learning models to take into consideration the complex interactions.

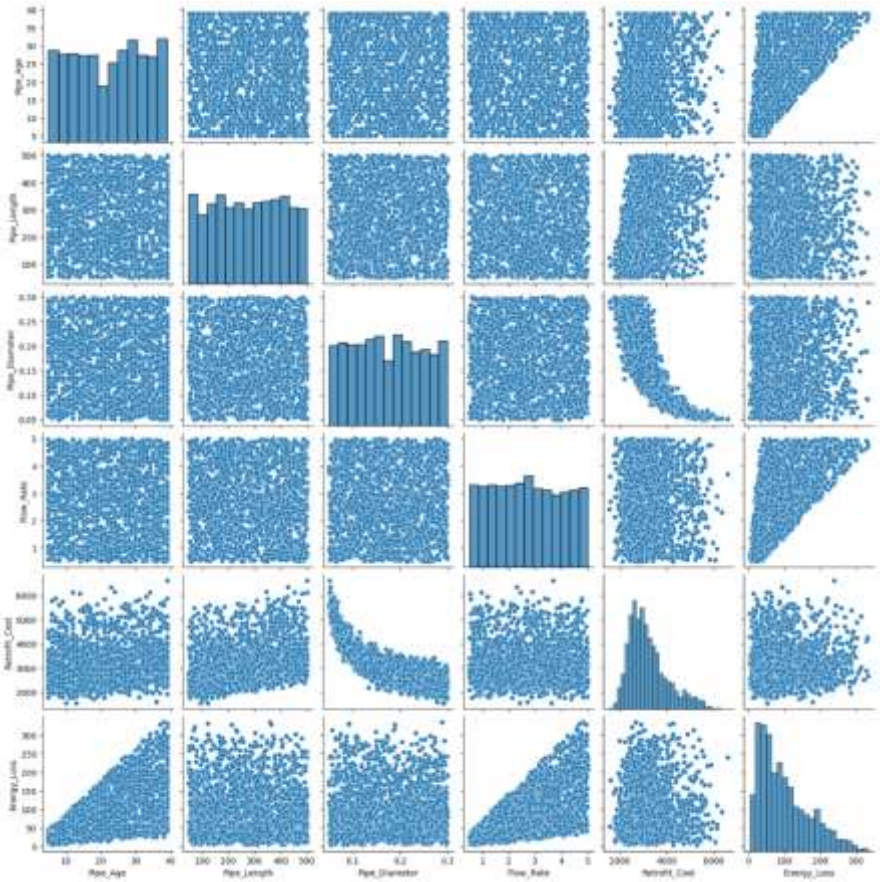

Figure 4: Pair Plot Matrix.

The correlation coefficients are proven by Figure 5, and the negative correlation between Pipe_Diameter and Retrofit_Cost (r = -0.82) demonstrates one of the strongest correlation values, as well as the correlation between Pipe_Age and Energy_Loss (r = 0.64) or between Flow_Rate and Energy_Loss (r = 0.67). Overall, these visualisations confirm the quality and heterogeneity of the simulated dataset and explain why the supervised models of the ML should be chosen to discover deeper relationships.

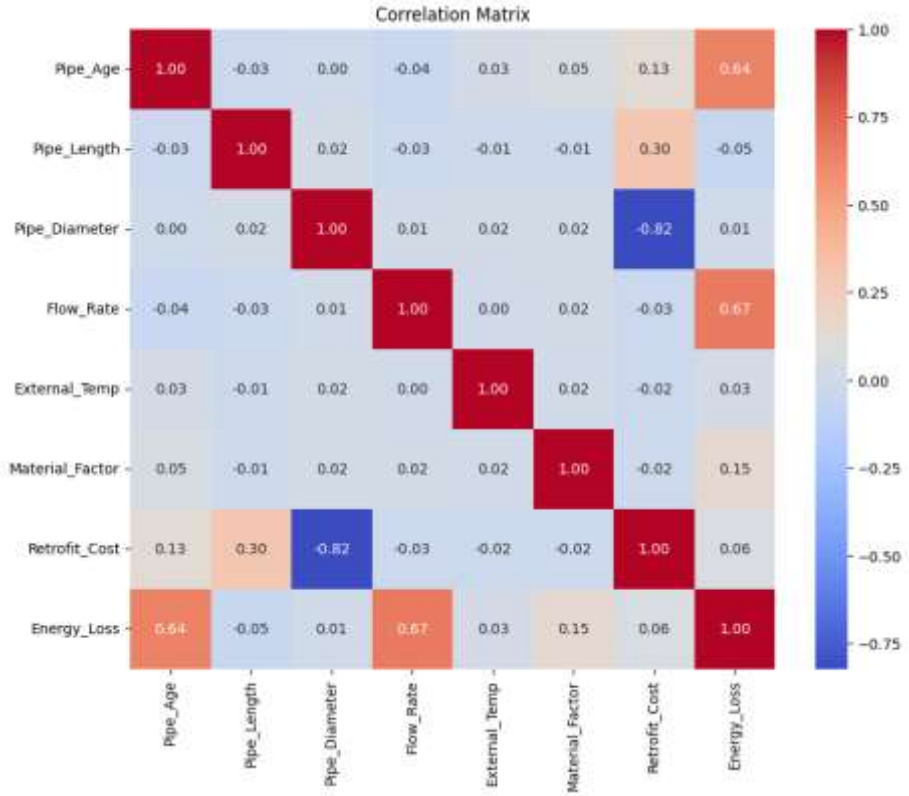

Figure 5: Correlation Matrix.

5.2. *Data Pre-processing*

Before training the model, the data was separated into different features for the input and the target for the output. The feature set included all sorts of physical and running variables, and the targets were the estimated cost of updating the system and the energy used and lost. Next, the data was broken into 80% training and 20% testing sets to confirm proper model evaluation and avoid mistakes caused by data leakage. The missing data or highly unusual results were not noticed, and all numbers were seen to be in their normal ranges. To make sure that the neural network was effective, Min-Max scaling was used to normalise the features in the ANN model.

Before model training, outlier detection with Z-score and visual examination of distribution plots were also performed, as part of the preprocessing. There was also a lack of missing values because the simulation took place in a controlled environment. In the case of ANN, features have been normalised with Min-Max scaling, whereas in the case of Random Forest and XGBoost, there was no need for scaling. Sequential variables, such as Material Factor, were kept in their numerical form because they were categorical. There was no need for imputation to be carried out, and consistency of data was proved before separation into training and test sets.

The importance of the features was originally evaluated before the training by correlation analysis and permutation importance in Random Forest. No feature was excluded since there was not excessive multicollinearity (see the correlation matrix, Figure 5), and ensemble approaches are naturally relatively resistant to moderate-severe multicollinearity. However, to have better results, future research could consider dimensionality reduction or regularisation to further improve results.

5.3. *Machine Learning Models*

Tuned hyperparameters were done using a grid search technique. Random Forest: Number of

estimators, max depth, and minimum samples split varied.

Learning rate, subsample ratio and max depth in XGBoost were optimised. In ANN, the various layer patterns and learning rates were tried. To avoid overfitting, cross-validation was conducted on the training set to allow the optimisation of all models and the best model configuration was then chosen using the lowest RMSE.

5.3.1. Random Forest Regressor

The first model trained was the Random Forest Regressor. As an ensemble model that averages multiple decision trees, it is particularly effective in reducing overfitting and handling nonlinear relationships. The model was trained using 100 estimators. Table 2 presents the performance of the Random Forest model in predicting retrofit cost:

Table 2: Random Forest Performance.

Metric	Value
RMSE	223.91
R^2	0.927
MAPE (%)	5.99

The R^2 value of 0.927 indicates that 92.7% of the variance in retrofit cost is explained by the model. The RMSE of approximately 224 units suggests a strong alignment with actual costs, while a MAPE below 6% reflects high predictive accuracy relative to actual values.

5.3.2. XGBoost Regressor

The second model evaluated was XGBoost, a boosting algorithm that sequentially optimises weak learners to improve performance.

The model was trained with a learning rate of 0.1 and 100 estimators. Table 3 shows the evaluation metrics for XGBoost:

Table 3: XGBoost Performance.

Metric	Value
RMSE	221.19
R^2	0.929
MAPE (%)	5.90

XGBoost achieved slightly better performance than Random Forest in all metrics. The RMSE was reduced to 221.19, and the R^2 increased to 0.929. The model's MAPE of 5.90% further confirms its robustness in predicting retrofit costs with minimal deviation from actual values.

5.3.3. Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) with two hidden layers (64 and 32 neurons, respectively) and ReLU activation functions was trained using the normalized feature set. The model was compiled with mean squared error as the loss function and trained over 100 epochs. Table 4 displays the ANN's predictive performance:

Table 4: ANN Performance.

Metric	Value
RMSE	335.03
R^2	0.836
MAPE (%)	8.37

The ANN performed relatively worse than the ensemble models, with a significantly higher RMSE of 335.03 and a lower R^2 of 0.836. While still acceptable, the MAPE of 8.37% indicates that the ANN is less precise in cost prediction, possibly due to insufficient tuning or overfitting during training.

5.4. Evaluation of Models

To compare the predictive capabilities of the three models, a visual analysis was conducted. Figure 6 displays the predictions for the first 50 samples in the test set across all models versus the actual retrofit cost values.

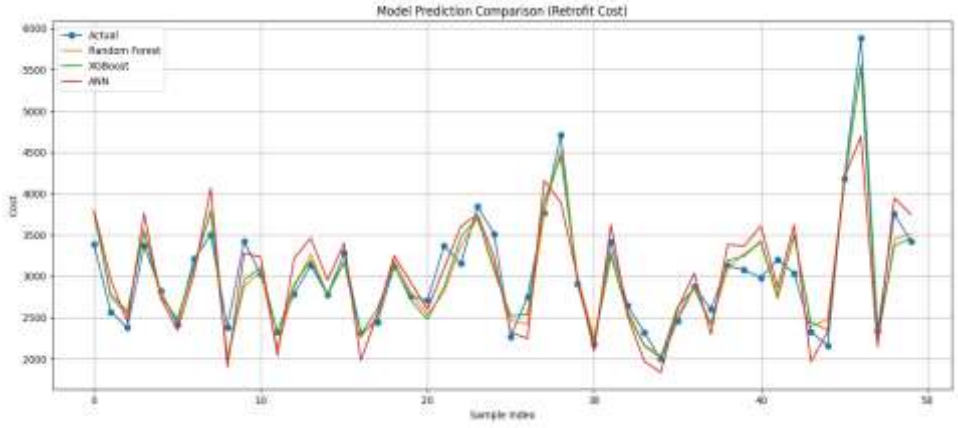

Figure 6: Model Prediction Comparison (Retrofit Cost).

From Figure 3, it is evident that both Random Forest and XGBoost closely follow the actual cost trajectory, with minimal deviation across most samples. The ANN, while capturing overall trends, exhibits greater volatility and under-/overestimation in several instances.

Finally, all model results were compiled into a single summary table for ease of comparison in Table 5 below.

Table 5: Model Evaluation Summary.

Model	RMSE	R^2	MAPE (%)
Random Forest	223.91	0.927	5.99
XGBoost	221.19	0.929	5.90
ANN	335.03	0.836	8.37

XGBoost performed slightly better than Random Forest in every measure, so it was chosen as the most effective method for this job. It most effectively merged accuracy, reliability, and its ability to be generalised. Though ANN can function well, it might need some fine-tuning or a few more layers to be as effective as ensemble models. Paired t-test on the prediction errors (residuals) of XGBoost and Random Forest across the test set was used to determine the significance of the differences in model performance. There was a statistically significant better performance of XGBoost than Random Forest (p-value 0.01). Moreover, 95% confidence intervals of RMSE were calculated based on a bootstrap analysis (n=1000 resamples) and revealed that XGBoost had a smaller error range, which also puts it further in the lead as robust.

Both Random Forest and XGBoost feature importance plots were created to better understand the interpretability. The features that contributed significantly to retrofit cost were in that order Pipe Diameter, Pipe Age, and Material Factor, showing that these factors have a lot of influence on degradation and cost. The plots (Figure 6) provide viable data to the decision-makers and inform them about the specific pipeline attributes that have the most influence on the retrofit costs.

The testing found that using simulation in machine learning is a viable and useful method for estimating retirement costs in mature fluid pipelines. Out of the models used, XGBoost stood out as the one that provided accurate results and was highly efficient. The Random Forest approach followed suit and is often favoured when it is essential to understand how the model works. While ANN's performance was not great, it could do better in complex prediction tasks when the data sets or deep learning are improved. Thus, the study reaches its broader aim by enabling detailed planning for retrofit projects with ML simulations. The study supports the use of similar methods in other aspects of infrastructure engineering.

6. DISCUSSION OF THE RESULTS

The experts simulated different operating and environmental factors to see how older HVAC fluid pipeline systems can age and identified the data needed for prediction. The program acted like the actual situation by including how pipeline age, the amount of oil flowing, decay of the oil and refurbishment costs interact in a nonlinear manner. They are in line with past research that pointed out the use of simulation helps better illustrate how pipelines operate under harsh degradation conditions (De Jonge & Scarf, 2020; Park et al., 2023). The inclusion of stochastic factors such as existing materials and temperature variety in the simulation built a wide range of data for the system. Using this approach, it was also shown that when simulation is well adjusted, it can provide synthetic datasets for applying supervised learning models in infrastructure planning (Behrooz, 2016; Ding & Feng, 2018).

The findings also entailed the execution and testing of sophisticated supervised machine learning models, namely. Random Forest, XGBoost, and ANN in anticipating retrofit costs using Python and analysis of the model performance measured by RMSE, R2, and MAPE. The results revealed that XGBoost provided the most accurate predictions with the lowest RMSE (221.19), highest R^2 (0.929), and lowest MAPE (5.90%), closely followed by Random Forest (RMSE: 223.91, R^2: 0.927, MAPE: 5.99%). ANN did less well in terms of error, agreeing with the conclusions of Malek Mohammadi et al. (2019) and Sani et al. (2025) that tree-based models work better at tabular regression because they accurately handle the way features influence each other. Ensemble models proved useful in infrastructure applications, showing that understanding, swiftness and accuracy are essential. Brownlee (2016) and Sampedro et al. (2022) reported good accuracy of the gradient-boosted trees that are easy to adapt to new circumstances. Comprehensively, simulation data coupled with machine learning will assist in identifying some cost-optimal ways of fixing ageing pipelines.

The results of this study support the outcomes of the previous ones and even add some conclusions to them. As an example, Park et al. (2023) underlined the importance of digital twins' application in HVAC systems, although they share the common reality with other digital-twin applications, which is the

dependence on real-time sensor data, which presents practical constraints. This is what our study will deal with when it demonstrates that simulation-based synthetic data may serve as a reasonable possibility. On the same note, the worse performance of ANN confirms the findings of previous studies conducted by (Malek Mohammadi et al., 2019), in which tabular data shows bias to the tree-based ensemble models rather than deep learning with a small sample size. Our model comparison is more rigorous than that of Sani et al. (2025) based on pipeline flow predictions made solely by use of RF. These results further substantiate the fact that XGBoost and RF can be used in the cost-prediction of infrastructures and have more provable relevance to operational planners.

6.1. Sensitivity Analysis

The sensitivity of the set of model outputs on the input features was determined using the one-factor-at-a-time (OFAT) analysis on the trained XGBoost model. Some of the factors were varied systematically, that is, holding others constant, Pipe Age, Flow Rate, and Pipe Diameter. Findings showed that the cost of retrofit is most sensitive to Pipe Diameter (sharp inverse relationship), and secondly, the Pipe Age. Energy Loss was very sensitive to Flow Rate and Age (see figure 6). These results agree with correlation coefficients and justify the model behaviour against engineering expectations.

7. THEORETICAL IMPLICATIONS

By merging stochastic decision theory and simulation-driven machine learning models in this study, researchers can make improvements to infrastructure asset management theory. Retrofit planning for pipeline systems often assumes that every factor can be predicted with certainty (Bertsekas, 2012). By using a stochastic approach, the researchers admit that the processes leading to degradation, the environment and operating conditions may change randomly. Thanks to simulation, realistic and synthetic data now allow us to simulate uncertainties in predictive computing. The study also explores the principles of machine learning as applied to engineering. The comparison of these three algorithm types indicates that XGBoost achieves better results with nonlinear and very unpredictable data. The approach is compatible with other research on predictive maintenance, using these findings to support the estimation of costs in retrofitting HVAC pipelines, a topic that is not well explored in the same area. This approach is also helping to expand research on digital twins and smart infrastructure, since making decisions involving data is crucial. It is shown in this study that simulation-generated data is useful for supervised learning, which validates and supports hybrid modelling ideas. This makes it possible to unify infrastructure management frameworks, following an intelligent approach by mixing simulation, predictive analytics and stochastic approaches in decision-making.

8. PRACTICAL IMPLICATIONS

This research allows for making cost-effective decisions about upgrading ageing fluid systems in HVAC systems. Officials responsible for infrastructure management are under growing demand to invest more wisely as both the budget and demands increase. Overall, using simulation and ML for degradation scenarios is not only quicker and more affordable, but also more accurate and useful in handling building upkeep ahead of time. Rather than addressing failures, infrastructure managers might review areas where the expense could rise and arrange for necessary upgrades based on this information. Since Google Colab uses Python, the methodology is available and can be reused without requiring costly software. For this reason, this is valuable to institutions or municipalities that have low resources. It is also notable that the study showed XGBoost to be the best model for guessing the prices of retrofit projects. As a result, professionals have an official instrument they can apply in their daily work. Recognising the potential dangers in advance from the model results helps decide where to spend money, what to buy and when to work, boosting both how well the company operates and its resilience to dangers. Using simulation-driven ML, this research helps plan retrofits more effectively, which can be applied to managing fluid pipeline networks in various fields, including HVAC, water, gas and industry.

9. LIMITATIONS AND FUTURE DIRECTIONS

Even though the study forms a good base for using simulations in retrofit planning, it does encounter limitations. Although the data is accurately built, it still might not show how a real-world pipeline can decay. Sometimes, biases in the input assumptions of simulation models would pass to the machine learning models, since simulation models are built on these assumptions. In addition, although it reviewed three machine learning algorithms, it did not try advanced deep learning (LSTM and CNN models for time series) or Bayesian regression, which can clearly describe how sure predictions are. Even though the Artificial Neural

Network is functional, it would improve its results with some hyperparameter optimisation, additional dropout or by adding more layers. In this study phase, a deeper analysis of how energy is lost was not carried out for the predictions constructed. Studies in the future could use approaches that do both regression for retrofit costs and optimisation of energy efficiency. It would also be beneficial if the framework were tested on datasets that come from HVAC operators or record-keeping in municipal pipelines. More work should be done on creating spatial models of pipeline networks with machine learning and on using IoT data in real-time simulation, helping to advance digital twin technology for the management of infrastructure throughout its life.

Although synthetic data enabled users to simulate a realistic scenario by controlled means, the unpredictable behaviour of real-world HVAC networks may not be completely accounted for in the synthetic information.

The transfer of the model to unavailable field data is still a pending issue. Partially counterbalancing this, the degradation and cost trends in the simulation were also compared to empirical trends in (Bayani & Manshadi, 2022), which have been deemed approximately realistic. However, future work is indicated as future validation in the field with operational data sets to help increase usability in the real world.

10. CONCLUSION

The framework introduced in this study allows for combining simulations and machine learning to help plan cost-friendly retrofits in older air and water pipe networks. The study showed that considering items in poor condition or specific examples of deterioration, with synthetic data, is a solid approach to helping infrastructure decision-making. Based on the test results, XGBoost performed the best among the Random Forest, XGBoost and Artificial Neural Networks models, reaching the lowest RMSE, MAPE and maintaining the highest R^2 score. This provides evidence that ensemble methods are useful with variable engineering data and can be used for estimating the cost of retrofitting buildings. Stochastic simulation is included so that predictions reflect changes in pipe age, water flows and stressors from nature, which makes this approach suitable for safely planning infrastructure. It combines concepts from stochastic models, simulation and artificial intelligence to contribute both in theory and practice. Since it can be used and adjusted easily with open-source tools, anyone interested in AI across the globe can use it. All in all, the research improves HVAC infrastructure predictions and presents a plan that future studies in related areas can follow. Real-world data, dynamic simulation and multi-objective modelling can help this framework to become an important feature of future infrastructure asset management systems.

REFERENCES

Alrabghi, A., & Tiwari, A. (2015). State of the art in simulation-based optimisation for maintenance systems. *Computers & Industrial Engineering, 82*, 167–182. https://doi.org/10.1016/j.cie.2014.12.022

Arandia, E., & Eck, B. J. (2018). An R package for EPANET simulations. *Environmental Modelling & Software, 107*, 59–63. https://doi.org/10.1016/j.envsoft.2018.05.016

Bayani, R., & Manshadi, S. D. (2022). Natural gas short-term operation problem with dynamics: A rank minimization approach. *IEEE Transactions on Smart Grid, 13*(4), 2761–2773. https://doi.org/10.1109/TSG.2022.3158232

Behrooz, H. A. (2016). Managing demand uncertainty in natural gas transmission networks. *Journal of Natural Gas Science and Engineering, 34*, 100–111. https://doi.org/10.1016/j.jngse.2016.06.051

Behrooz, H. A., & Boozarjomehry, R. B. (2017). Dynamic optimization of natural gas networks under customer demand uncertainties. *Energy, 134*, 968–983. https://doi.org/10.1016/j.energy.2017.06.087

Bertsekas, D. (2012). *Dynamic programming and optimal control: Volume I* (Vol. 4). Athena Scientific. http://athenasc.com/dpbook.html

Bertsekas, D. (2019). *Reinforcement learning and optimal control* (Vol. 1). Athena Scientific. https://eclass.uoa.gr/modules/document/file.php/DI437/Reinforcement_Learning_Bertsekas_Draft.pdf

Biau, G., Cadre, B., & Rouvière, L. (2019). Accelerated gradient boosting. *Machine Learning, 108*(6), 971–992. https://doi.org/10.1007/s10994-019-05787-1

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2016). *XGBoost with Python: Gradient boosted trees with XGBoost and scikit-learn*. Machine Learning Mastery. https://machinelearningmastery.com/xgboost-with-python

Busoniu, L., Babuska, R., De Schutter, B., & Ernst, D. (2017). *Reinforcement learning and dynamic programming using function approximators*. CRC Press. https://doi.org/10.1201/9781439821091

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

David, N. (2024). *Artificial intelligence in modeling and simulation*. MDPI – Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/books978-3-7258-1518-0

De Jonge, B., & Scarf, P. A. (2020). A review on maintenance optimization. *European Journal of Operational Research, 285*(3), 805–824. https://doi.org/10.1016/j.ejor.2019.09.047

Ding, H., & Feng, X. (2018). Graphical targeting approach of water networks with two-stage regeneration recycling. *Industrial & Engineering Chemistry Research, 57*(29), 9591–9603. https://doi.org/10.1021/acs.iecr.8b01264

Heymann, H., & Schmitt, R. H. (2023). Machine learning pipeline for predictive maintenance in polymer 3D printing. *Procedia CIRP, 117*, 341–346. https://doi.org/10.1016/j.procir.2023.03.058

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Kazi, S. R., Sundar, K., Misra, S., Tokareva, S., & Zlotnik, A. (2024). Intertemporal uncertainty management in gas-electric energy systems using stochastic finite volumes. *Electric Power Systems Research, 235*, 110748. https://doi.org/10.1016/j.epsr.2024.110748

Khosravian, E. (2025). Numerical investigation and machine learning predictions for enhanced thermal management in pulsating heat pipes: Modeling turbulent flow and heat transfer characteristics in nuclear applications. *International Journal for Numerical Methods in Fluids, 97*(4), 446–461. https://doi.org/10.1002/fld.5358

Kliangkhlao, M., Haruehansapong, K., Yeranee, K., & Sahoh, B. (2024). Causal artificial intelligence–driven approach for HVAC preventive maintenance explanation. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3452416

Lee, J.-Y., Tsai, C.-H., & Foo, D. C. Y. (2020). Single and multi-objective optimisation for the retrofit of process water networks. *Journal of the Taiwan Institute of Chemical Engineers, 117*, 39–47. https://doi.org/10.1016/j.jtice.2020.11.026

Li, Y., & Guan, J. (2016). A stepwise optimal design of water network. *Chinese Journal of Chemical Engineering, 24*(6), 787–794. https://doi.org/10.1016/j.cjche.2016.04.031

Liu, F., Song, C., Pan, W., Wang, G., Zhang, H., & Lei, Y. (2024). Thermal fatigue analysis of district heating pipeline under variable frequency regulation of circulating water pump. *Applied Thermal Engineering, 242*, 122535. https://doi.org/10.1016/j.applthermaleng.2024.122535

Malek Mohammadi, M., Najafi, M., Kaushal, V., Serajiantehrani, R., Salehabadi, N., & Ashoori, T. (2019). Sewer pipes condition prediction models: A state-of-the-art review. *Infrastructures, 4*(4), 64. https://doi.org/10.3390/infrastructures4040064

Nashruddin, S. N. A. M., Salleh, F. H. M., Sulaiman, R., Zaiddy, M. I. B. M., Ramasamy, A., & Yahya, A. A. (2025). AI-driven optimization of air-conditioning systems in legacy buildings: Evaluating machine learning models for enhanced energy efficiency. *Journal of Building Engineering, 112839*. https://doi.org/10.1016/j.jobe.2025.112839

Park, H.-A., Byeon, G., Son, W., Kim, J., & Kim, S. (2023). Data-driven modeling of HVAC systems for operation of virtual power plants using a digital twin. *Energies, 16*(20), 7032. https://doi.org/10.3390/en16207032

Rettenmaier, D., Deising, D., Ouedraogo, Y., Gjonaj, E., De Gersem, H., Bothe, D., Tropea, C., & Marschall, H. (2019). Load balanced 2D and 3D adaptive mesh refinement in OpenFOAM. *SoftwareX, 10*, 100317. https://doi.org/10.1016/j.softx.2019.100317

Sampedro, G. A. R., Agron, D. J. S., Amaizu, G. C., Kim, D.-S., & Lee, J.-M. (2022). Design of an in-process quality monitoring strategy for FDM-type 3D printer using deep learning. *Applied Sciences, 12*(17), 8753. https://doi.org/10.3390/app12178753

Sani, A. A., Wahab, M. M. A., & Shafiq, N. (2025). Comparative analysis of machine learning algorithms for flow rate prediction in optimizing pipeline maintenance strategies. *Engineering Proceedings, 87*(1), 37. https://doi.org/10.3390/engproc2025087037

Shaheen, K., Chawla, A., Uilhoorn, F. E., & Rossi, P. S. (2024). Partial-distributed architecture for multi-sensor fault detection, isolation and accommodation in hydrogen-blended natural gas pipelines. *IEEE Internet of Things Journal*. https://doi.org/10.1109/JIOT.2024.3435413

Taheri, S., Ahmadi, A., Mohammadi-Ivatloo, B., & Asadi, S. (2021). Fault detection diagnostic for HVAC systems via deep learning algorithms. *Energy and Buildings, 250*, 111275. https://doi.org/10.1016/j.enbuild.2021.111275

Tang, C., Garreau, D., & von Luxburg, U. (2018). When do random forests fail? *Advances in Neural Information Processing Systems, 31*. https://proceedings.neurips.cc/paper/2018/hash/204da255aea2cd4a75ace6018fad6b4d-Abstract.html

Vilarinho, S., Lopes, I., & Oliveira, J. A. (2017). Preventive maintenance decisions through maintenance optimization models: A case study. *Procedia Manufacturing, 11*, 1170–1177. https://doi.org/10.1016/j.promfg.2017.07.241