

An explainable machine learning approach to the prediction of pipe failure using minimum night flow

Matthew Hayslep ^{a,*}, Edward Keedwell ^a, Raziye Farmani ^b and Josh Pocock^c

^a Department of Computer Science, University of Exeter, North Park Road, Exeter EX4 4QF, UK

^b Department of Engineering, University of Exeter, North Park Road, Exeter EX4 4QF, UK

^c South West Water, Peninsula House, Rydon Lane, Exeter EX2 7HR, UK

*Corresponding author. E-mail: mh989@exeter.ac.uk

 MH, 0000-0002-0767-0619; EK, 0000-0003-3650-6487; RF, 0000-0001-8148-0488

ABSTRACT

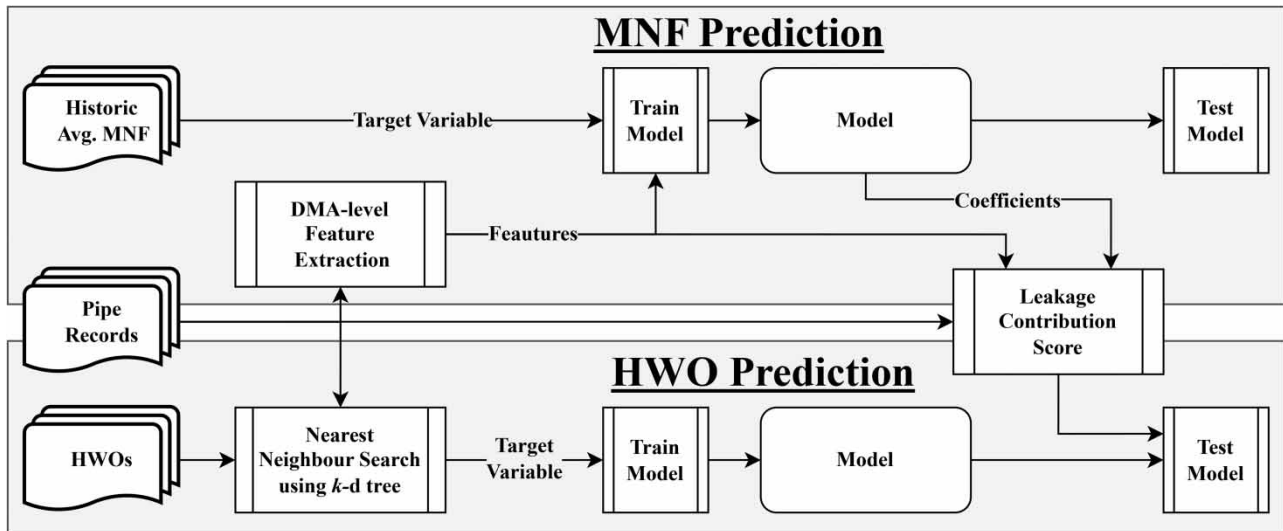
Both minimum night flow (MNF) and pipe failures are common ways of understanding leakage within water distribution networks (WDNs). This article takes a data-driven approach and applies linear models, random forests, and neural networks to MNF and pipe failure prediction. First, models are trained to estimate the historic average MNF for over 800 real-world DMAs from the UK. Features for this problem are constructed from pipe records which detail the length, diameter, volume, age, material, and number of customer connections of each pipe. The results show that 65% of the variation in historic average MNF can be explained using these factors alone. Second, a novel method is proposed to deconstruct the models' predictions into a leakage contribution score (LCS), estimating how each individual pipe in a DMA has contributed to the MNF. In order to validate this novel approach, the LCS values are used to classify pipes based on historic pipe failure and are compared against models directly trained for this. The results show that the LCS performs well at this task, achieving an AUC of 0.71. In addition, it is shown that both LCS and directly trained models agree in many cases on an example real-world DMA.

Key words: classification, machine learning, minimum night flow, pipe failure, regression, water distribution network

HIGHLIGHTS

- The use of explainable machine learning to estimate the ranked contribution of individual pipes to leakage based on minimum night flow measurements.
- The validation of the previous against machine learning models that were trained to classify pipe failures.
- Achieving the above using historic pipe asset data on real-world DMAs.
- Achieving the above in a way that such models are easily applicable to/for real-world DMAs.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Leakage is an important challenge within the water industry and has an economic, environmental, and sustainability impact (Puust *et al.* 2010). According to Ofwat, the UK regulator for the water industry, around 3 bn litres of water were lost to leaks every day across the UK in 2021 (Ofwat 2021). This translates to around 20% of clean water that is produced being lost to leakage (Farrow *et al.* 2017). These losses necessitate the production and transport of more clean water, increasing pumping and treatment costs. This carries an economic cost for companies as well as customers, in addition to increasing energy consumption and greenhouse gas emissions.

In the UK, to help understand and combat leakage, water distribution networks (WDNs) are subdivided into smaller district metered areas (DMAs) and flow meters are used to monitor the water input and output of DMAs (Romano 2021). The total difference between incoming water and outgoing water in a DMA is the total demand or consumption of water, which will include legitimate consumer demand as well as leakage. Minimum night flow (MNF) is the lowest flow rate during the night, usually between 12 am and 4 am, and is a commonly used proxy for estimating and analysing leakage in the industry (Farley & Trow 2005). The MNF approximates leakage under the assumption that legitimate water usage is lowest at night, and therefore, most of the water usage measured during this period is leakage. In the UK, MNF analysis is the most common form of leakage assessment (Farrow *et al.* 2017).

An alternative way of understanding leakage is to study known cases of historic leaks and bursts, i.e., pipe failures. This relies on the records of utility companies regarding engineering work carried out on the infrastructure. In this article, cases of leaks and bursts indicated by records of engineering work are called historic work orders (HWOs), to include any form of maintenance that involves a leakage component. The term HWO is preferred here over pipe failure to emphasise the importance and prevalence of smaller leaks and the proactive maintenance that water utility companies carry out in order to address existing leakage, rather than focusing on emergent bursts alone.

Explainable artificial intelligence (XAI) provides insight into the relationships between input features and outputs, furthering understanding of the problem itself and the factors that impact leakage the most. For water companies, this means additional factors can be considered, or better understood, when deciding on actions, company policies, and management. Therefore, XAI is important for water companies because it can support better decision making about leakage mitigation methods to implement and areas on which to focus resources. This does not mean that non-explainable methods should never be used, only that non-explainable methods should not be used without explanation. However, as this paper will demonstrate, explainability can also be leveraged to improve the capabilities of the model and to estimate a contribution score of individual pipes to MNF.

In real-world DMAs, there are numerous factors that affect the total amount of leakage. Pressure, pipe age, pipe material, weather conditions, soil conditions, time of year, and pipe condition are some commonly described factors among many

others (O'Day 1982; Boxall *et al.* 2007; Barton *et al.* 2019). Many of these factors are chiefly of concern when considering bursts and the extent to which they affect background and unreported leakage is uncertain. In this article, only the physical properties of pipes are considered. This constitutes the simplest version of the problem and is reliant on data that is widely available for most water companies and is thus generalisable to many types of systems around the UK and even globally. This article takes a data-driven approach to predicting both historic average MNF and HWOs and proposes a novel method for approaching both problems with a single model which does not require pressure or flow estimations typically provided from a calibrated hydraulic model that are not always available for many DMAs. A linear regression model, which is trained to predict the historic average MNF, is used to predict HWOs by creating a leakage contribution score (LCS) which indicates the effect an individual pipe has on the MNF. The LCS models are compared to other models which are trained to predict HWOs directly.

2. LITERATURE REVIEW

Pipe failure records and MNF are widely used in the literature to understand leakage. However, these are not often studied together. For many reasons, one is usually chosen over the other. More weight is often given to investigating bursts due to them being singular events which can severely impact customer service. Although not trivial, it is relatively easy for a human to spot a potential burst from pressure or flow data. While bursts are important, unreported and background leaks can also account for a significant proportion of leakage (Farley & Trow 2005). These types of leaks tend to go undetected for significant periods of time, with some potentially going undetected until the infrastructure is replaced or repaired, usually for some other reason. These types of leaks can also increase in severity over time and in some cases eventually lead to a burst.

There are many studies that use flow data, sometimes only during the MNF period, in order to detect anomalies. Examples such as Mounce & Boxall (2010), Romano *et al.* (2014), and Farah & Shahrouz (2017), among many others, use a variety of methods (both statistical and machine learning) to analyse time series data. However, many studies using this approach are focused specifically on bursts (Wan *et al.* 2022).

Studies that focus more broadly on leakage often look at longer periods of time. In addition, the prediction of leakage levels or similar metrics often includes features about some of the important factors previously mentioned that are not considered in this article: pressure, weather, time of year, etc. The rate of leakage is estimated by Jang *et al.* (2018) and Jang & Choi (2017, 2018) in a case study of over 150 DMAs from a city in South Korea. These studies used neural networks to estimate the leakage ratio (between legitimate demand and leakage demand) of DMAs. The studies found that pipe length was the most important factor by far. However, the overall accuracy of the models was low, potentially due to the limited number of samples.

Leakage rates were predicted by Kizilöz (2021) using neural networks. In this study, average pressure was used to allow 67 DMAs to have different predictions for different months, i.e., treating each month DMA combination as a different sample. However, the data were split randomly into training and test sets, meaning that the training set is likely to have had at least one example from each DMA. As a result, the task may have been better interpreted as predicting the leakage rates of different months for different DMAs. Despite this, the study reported an R^2 of 0.86 for the largest neural network which had a single layer of 30 neurons.

Alkassseh *et al.* (2013) used linear regression to predict the average MNF of 30 DMAs using the number of customer connections, total pipe length, length-weighted mean age, and mean pressure. The results showed an R^2 of 0.713 and the features and coefficients were then analysed statistically.

Finally, in a previous article (Hayslep *et al.* 2023), the authors used genetic programming (GP) to create a small optimal set of features for the historic average MNF prediction problem. The aim of this work was to create a set of explainable DMA-level features to estimate the historic average MNF. A multi-objective multi-gene GP approach was taken to evolve a set of up to nine features while minimising total tree size and maximising R^2 . The results found that five features – corresponding to sum pipe length, sum plastic pipe age, sum non-metal pipe diameter, max age, and min diameter (inversely) – could account for 65% of the variation in historic average MNF between different DMAs. These features are used in Sections 4.2 and 4.3 for comparison and are referred to as evolved equations.

2.1. Pipe failure

Most research on burst/leak localisation and pinpointing seeks to understand these pipe failures using temporal data (Romano 2021). The use of flow or pressure time series is common (Wan *et al.* 2022), both with real-world and simulated

case studies. There is significant research on the topic of optimal sensor placement (Romano 2021), which improves the information contained within the flow and/or pressure signals, allowing more accurate detection and pinpointing. These approaches focus on understanding WDNs as hydraulic systems and therefore often employ simulations as case studies. The use of simulations brings with it many assumptions and simplifications, but in studies using real-world data there is also often an implicit assumption that DMAs are under ‘normal operation’ outside of detected anomalies (Wan *et al.* 2022).

Instead, many studies have sought to understand why failures happen, which indirectly helps to understand how to localise and pinpoint leaks/bursts. Boxall *et al.* (2007) used a Poisson generalized linear model (GLM) to estimate the burst rate of pipes based on diameter, length, and age. Two datasets were split based on which of two materials (cast iron or asbestos cement) the pipes were made of, for a total of four different models. The resulting models were analysed and discussed to give insight into the factors that affect bursts and a further analysis considering soil conditions (corrosivity and shrink/swell) was also undertaken.

Pipe deterioration is modelled by Berardi *et al.* (2008) using evolutionary polynomial regression (EPR). Pipes from 48 DMAs are separated into three groups based on age and diameter criteria. Additional aggregated features are calculated for each of these groups and EPR is used to describe the relationship between the total number of pipe bursts for each group and its features. An R^2 of 0.822 is achieved using age, total group length, and diameter. Another model found an R^2 of 0.550 with the number of property connections as the only feature.

Kakoudakis *et al.* (2017) also used EPR to predict pipe bursts in a real-world case study. Pipes were grouped based on diameter, age, and soil type. A separate EPR model for each group was developed to predict the total number of failures based on total length, diameter, and age. The best set of models achieved an R^2 of over 0.85. However, by grouping the pipes together, predictions are made for the entire group of pipes. This is based on the assumption that homogeneous pipes have the same failure rate.

Giraldo-González & Rodríguez (2020) compare various statistical and machine learning models that predict pipe failure in a case study of a densely populated area. The authors consider numerous factors including the number of previous failures, valves, weather, and soil as well as pipe age, length, and diameter. The statistical models perform a failure rate prediction with the pipes clustered into groups based on length, age, and diameter. The machine learning models perform a classification task with separate models trained for the two different materials studied and the authors showed good results when including environmental and operational factors.

Liu *et al.* (2022) used random forest and logistic regression for pipe failure prediction as a classification problem. The authors compare sampling methods to address the class imbalance of pipes with and without failures. The results suggest that random forest performs better than logistic regression, and that the most important factors of those considered are length and diameter.

These studies further our understanding of pipe failures and why they happen, and, using that knowledge, attempt to understand where pipe failures are likely to happen. These approaches have found success at predicting pipe failures even when they do not directly consider flow, pressure, or any other environmental or operational factors.

3. METHODS

The aim of this article is to demonstrate a novel method for decomposing DMA-level predictions of MNF into an LCS which can be used to rank pipes by their effect on the MNF. This novel method has two requirements: an explainable model for estimating MNF and a set of DMA-level features which are calculated from pipe-level features. This results in a pipe-level LCS which is validated in a classification task for pipe failures. This article builds on previous studies and leverages a large dataset to not only understand pipe failures (or HWOs in this article) but to also understand MNF with the same model, by utilising the LCS. This is accomplished by treating MNF as a measure of leakage, in the same way that HWOs are a measure of pipe failures, and by predicting MNF in a regression task. Ideally, this would be accomplished by knowing the MNF of each individual pipe, then the relationships between pipe characteristics and MNF could be learned. However, the resolution of available MNF data is too low for this to be possible because MNF data is only known at the DMA-level. Therefore, the first part of this article trains a linear regression model to predict the historic average MNF of DMAs using the aggregated features of pipes. In the second part, the linear regression model is decomposed, exploiting the knowledge of how the features were constructed and the coefficients of the linear model, to create the LCS for each pipe, which corresponds to how much that pipe contributes to the MNF. In the final part of this article, the LCS is validated by using it to predict HWOs (since there is no pipe-level MNF data).

The dataset covers 806 DMAs in the UK. Together, these DMAs represent over 12 million metres of pipe, ranging from dense urban networks to large sparse rural networks. The dataset consists of over 290,000 pipe records detailing the DMA, length, diameter, volume (calculated from length and diameter), age, material (grouped into ‘metal’, ‘plastic’, and ‘other’), and an estimation of number of customers connections (based on distance to the pipe). As has been consistently noted in the literature, pressure is also an important factor for leakage. Unfortunately, this data was unavailable for this dataset, and so it is not included in this work.

Historic daily MNF records cover the year 2022 for each DMA. On average, each DMA has 328 MNF records for this period. The prediction target for the MNF prediction problem in this article is the average MNF of each DMA for this entire period. Therefore, this is a regression problem rather than a time-series prediction problem. This simplification is required because the predictor features do not vary with time. The model is tested on DMAs that are not seen during training. For this purpose, the full dataset of 806 DMAs is randomly split into a training set, consisting of 70% of DMAs, and a test set, consisting of the remaining 30% of DMAs. The test set is used only for final evaluation of the models.

HWOs also cover 2022 with nearly 19,000 records. Of all field work that water companies undertake, HWOs are a subset defined as having a water loss component, measured in equivalent service pipe bursts (ESPBs) (Lambert & Morrison 1996). Each HWO has a GIS location. Using this location, each HWO is assigned to the nearest pipe in the dataset using a k -d tree (Bentley 1975). If there are no pipes within 50 m of the HWO, then it is discarded. This approach has some drawbacks. For example, in cases where work is done on a junction which might involve several leaking pipes, the HWO is only assigned to one. This could be overcome by associating each HWO with the pipes in a defined radius in addition to the nearest neighbour search. However, this article uses the nearest neighbour approach because it is both straightforward to implement and understand. In this article, this problem is approached as a classification problem, where the positive class consists of pipes that had at least one HWO during 2022 and the negative class consists of every pipe that did not have an HWO during this period. The resulting dataset contains nearly 11,000 pipes in the positive class (with HWOs) resulting in an HWO rate of 3.7% over all pipes.

This work is divided into three parts: (1) the prediction of historic average MNF, (2) the calculation of the novel LCS, and (3) using the LCS to predict HWOs. This article takes an explainable AI approach to the prediction of historic average MNF and HWOs. In all tasks and for all models, the features are scaled using the mean and standard deviation during pre-processing; the specific value for the mean and standard deviation is found on the training set of the models in order to prevent the leaking of information on the test set back to the model. Figure 1 shows the overall methodology and process followed in this article. Specifics of the MNF prediction, LCS calculation and HWO prediction follow in Sections 3.1–3.3, respectively.

3.1. MNF prediction

Pipe-level features are combined with different aggregation and filtering methods to create 240 DMA-level features. Four aggregation methods are used: sum, max, min, and mean. These are combined with the numerical pipe-level features (age,

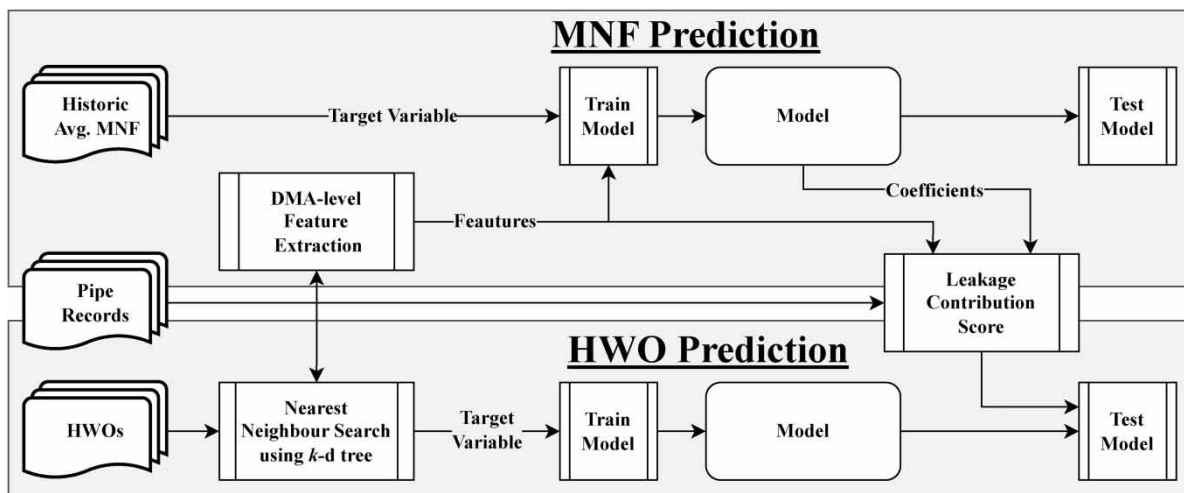


Figure 1 | A simplified scheme for the overall methodology.

length, diameter, volume, and number of customer connections) to create features such as ‘max age’ or ‘min diameter’. Additional features are created by filtering for pipes of a specific material (metal, plastic, and other), creating features such as ‘sum metal length’ or ‘max plastic age’. Finally, two filter functions, argmax and argmin, are used to create combinatorial features where the value of one pipe-level feature is used as a filter. This creates features such as ‘sum length of argmin diameter’ which represents the total length of all pipes with the minimum diameter in this DMA. The argmin and argmax are calculated with regard to the pipes within a DMA, rather than across all DMAs. Figure 2 shows pseudocode for the feature creation process. The aim of these features is to represent various relationships within the pipe-level data for each DMA and this method creates 240 DMA-level features, resulting in high-dimensional data.

This section tackles the same problem as Hayslep *et al.* (2023) but without using GP. This is done to emphasise that the novel aspect of this work (covered in Section 3.2) can be applied to any DMA-level feature calculated from pipe-level features, evolved by GP or otherwise. Therefore, this article will apply the novel LCS method to two different sets of features that describe the same dataset for comparison. Sections 3.1 and 4.1 detail the methods and results for the features described above.

The accuracy of the MNF prediction model reveals how much of the variation in leakage between different DMAs can be solely attributed to the physical properties of the pipes. Three different regression methods are shown: linear regression via elastic net, random forest, and neural networks. However, as will be seen in the next section, the linear regression model is the most important, as the LCS is calculated using the coefficients of the model. The other, non-linear, models are shown for comparison.

The elastic net (Zou & Hastie 2005) method is used to train a linear model. This method is especially useful in this case because there are many groups of correlated features. While other linear regression methods, such as Ridge or Lasso, might select only one of a group of correlated features to assign importance to, elastic net will instead spread that importance among the group (Zou & Hastie 2005). This improves the robustness of the model. The elastic net method achieves this by using both l_1 regularisation – a penalty to the magnitude of coefficients – and l_2 regularisation – a penalty to the number of non-zero coefficients. These penalties are applied to the objective function, which minimises the error of the linear model, so that a better set of coefficients can be found (Zou & Hastie 2005).

Random forests (Breiman 2001) are also used to train an ensemble of decision trees which linearly splits different features at a series of thresholds. The resulting prediction of each decision tree is then averaged. The random forest method is a simple,

```

Input: set of numerical features  $N$ ,
         set of materials in  $M$ ,
         set of aggregation functions  $F$ ,
         set of filter functions  $G$ 
Output: set of functions  $Y$  that when input with pipe records for a DMA
 $x$ , construct the feature for that DMA

for each numerical feature  $n$  in  $N$  do
  for each aggregation function  $f$  in  $F$  do
     $Y \leftarrow f(n, x)$ 
    for each material  $m$  in  $M$  do
      let  $x_m$  be the set of pipe records with the material  $m$ 
       $Y \leftarrow f(n, x_m)$ 
    for each numerical feature  $n'$  in  $N$  do
      for each filter function  $g$  in  $G$  do
        if  $n \neq n'$  then
           $Y \leftarrow f(n, g(n', x))$ 
return  $Y$ 

```

Figure 2 | Pseudocode for the feature creation process. In this case, the numerical features are length, diameter, volume, age, and number of customer connections, the materials are ‘metal’, ‘plastic’, and ‘other’, the aggregation functions are sum, mean, min, and max, and the filter functions are argmax and argmin.

efficient, and effective method and has been demonstrated to perform well in numerous fields and problems, including leakage.

Finally, neural networks are also examined (Bishop 1995). Neural networks are black-box methods, meaning that it is difficult to understand how predictions are being made. Despite this, neural networks have also seen wide use and success in numerous fields and problems, including leakage. In this article, feed-forward neural networks are used and trained via back-propagation. Various configurations, numbers of hidden layers and hidden neurons in each layer, are shown to demonstrate different levels of complexity. In this article, dropout (Srivastava *et al.* 2014) is used to prevent overfitting, or more specifically, to prevent overfitting from reducing test accuracy. This works by randomly disabling hidden neurons during training which results in increased robustness in the model.

For all models, the hyperparameters were chosen after a period of trial and error using grid search and k -fold cross-validation. As previously discussed, the elastic net linear model is the most important because it exposes coefficients, which are needed to calculate the LCS.

3.2. Decomposition to pipe-level LCS

The LCS is a numerical value that, in relation to the LCS value of other pipes, represents a particular pipe's contribution to the leakage of a DMA, as measured by the MNF in this case. In order to create LCSs for the pipes in this dataset, this article uses three pieces of data:

1. Coefficients of a linear model (trained to predict MNF).
2. Feature values for a DMA (as the trained model would see them, i.e., after pre-processing).
3. The contribution of a pipe to a particular feature. This can be calculated using:
 - a. Pipe dataset (length, age, material, etc.) and
 - b. Knowledge of how all the features (that have non-zero coefficients) are calculated.

Together, parts 1 and 2 describe the relevance of different features (relative to different DMAs) to the MNF. Part 3 is the proportion to which a pipe accounts for a feature value. For example, consider a feature which is the sum of all pipe lengths in a DMA. Each pipe is responsible in proportion to its length. In a DMA with two pipes, pipe a with the length of 20 m, and pipe b with the length of 80 m, pipes a and b contribute 0.2 (20%) and 0.8 (80%) to the feature, respectively. All such features, which are summations, are calculated this way. For all non-summation features (e.g., the maximum pipe age in a DMA), all relevant pipes are given equal contribution, proportional to how many relevant pipes there are. Each pipe whose age is equal to the maximum age is responsible in proportion to the number of pipes with the maximum age. In a DMA with three pipes: a with age 8 years, b with age 10 years, and c with age 10 years, pipe a contributes 0.0 (0%), pipe b 0.5 (50%), and pipe c 0.5 (50%).

The LCS of a pipe i , which is part of a DMA j , is calculated as follows:

$$\text{LCS}_{ij} = \sum_{f \in F} \beta_f x_{jf} r_{if} \quad (1)$$

where F denotes the set of all features, β_f is the coefficient of the linear model for a feature f , x_{jf} is the value of the feature f for the DMA j , and r_{if} is the contribution of the pipe i to the feature f . This equation is similar to the equation for the linear regression model where the prediction of the target y is:

$$y_j = \beta_0 + \sum_{f \in F} \beta_f x_{jf} \quad (2)$$

where β_0 is the intercept of the linear model. Therefore, as r_{if} is a proportion whose sum over j (the DMA) is equal to 1, it follows that:

$$y_j - \beta_0 = \sum_{i \in j} \text{LCS}_{ij} = \sum_{f \in F} \beta_f x_{jf} \quad (3)$$

Therefore, the LCS is an estimation of how each pipe contributes to the MNF (leakage). However, because of the resolution of MNF data, the accuracy of this estimation cannot be directly assessed. In any case, because some of the coefficients of the linear models are negative it is possible for the LCS of a pipe to be negative. Therefore, the LCS is an indication of how a pipe contributes to MNF, relative to the LCS of other pipes, rather than a concrete estimation of the severity of leakage from a pipe.

An LCS could be calculated from any model that uses coefficients and where the contribution of a pipe to a feature can be assessed. An LCS may also be calculable for many equation-based methods such as EPR. In this article, two different models are used: the elastic net-based model described here, as well as the equations found from a previous work using GP (Hayslep *et al.* 2023) to find a minimal set of features. In order to validate the LCS and determine whether it is an effective measure, the LCS is assessed as a way of predicting HWOs and is compared to models that were trained directly on that task. It should be expected that the models trained directly on HWO prediction should perform better. However, the performance of the LCS on this task, considering that it is a score of contribution to MNF, will demonstrate that the LCS is a useful measure. In addition, the LCS is also useful in places without HWO data, as an alternative to predicting where leaks might be located.

3.3. HWO prediction

As previously discussed, the HWOs are associated with the nearest pipe within 50 m. The dataset covers the same period as the MNF prediction (2022). In this article, HWO prediction is a classification task to determine whether a pipe was subject to an HWO during this period or not. The models will be trained on a subset of pipes and tested on a different, unseen, subset of pipes.

The test–train split is done by using the same DMAs used in the MNF prediction. A pipe is in the HWO prediction training set if it belongs to a DMA which was in the training set for the MNF prediction task. Therefore, the models for both MNF and HWO prediction are trained on the same set of pipes. As a result, the different models and LCSs are directly comparable because they have learned from the same set of pipes. The resulting training set consists of roughly 70% of all pipes in the dataset. As previously discussed, the features for the pipe records are length, diameter, volume, age, material, and number of customer connections. The material feature is converted into one-hot encodings (Albon 2018) of each material, i.e., separate ‘metal’, ‘plastic’, and ‘other’ features, the value of each being either 1 or 0 depending on the material of the pipe. The one-hot encoded features are used for every model.

Around 3% of the pipes had an HWO in this period, creating an imbalanced classification task. Sample weighting is used to deal with this problem. This approach weights the minority and majority classes according to their frequency in the dataset. The weight is then used while calculating the error of different models to make up for the class imbalance.

Three different types of models are trained for this prediction task: logistic regression, random forest, and neural networks. In order to directly compare the LCS based on the elastic net linear model described above, an elastic net is used as the linear model for the logistic regression. As a result, when comparing the LCS and logistic regression model the main difference is the approach: predicting MNF and calculating LCS or predicting HWOs. Both the underlying training data and model are essentially the same.

Logistic regression using elastic net operates similarly to linear regression, except that the output of the linear model is used to determine whether a sample is part of a particular class or not, i.e., linear regression is for regression while logistic regression is for classification (Albon 2018). Broadly, the random forest and neural network models are described previously in Section 3.1. The only difference is these models are used for classification, which mainly changes the error that is being minimised in both cases, and the activation function of the output layer for the neural networks.

4. RESULTS AND DISCUSSION

This section details the results of each task in turn, starting with Section 4.1 covering historic average MNF prediction, then Section 4.2 shows a comparison of the models used for the LCS, and finally, Section 4.3 covers the HWO prediction for different models and comparison with the LCS.

4.1. MNF prediction

In order to compare the different models fairly, 5-fold cross-validation is used on the entire dataset (all 806 DMAs). The purpose of using 5-fold cross-validation is to fairly demonstrate and compare the different models. The hyperparameters of the different models were found after a period of trial and error using grid search. The average results from each fold of the 5-fold cross-validation are shown in Table 1. This table shows that the elastic net method performs better than more complex models. The elastic net methods also show less overfitting and have the best generalisation compared to the other models.

Table 1 | Metrics for the 5-fold cross-validation

| Model | R^2 | | Mean Absolute Percentage Error (MAPE) | | Mean Absolute Error (MAE) | | Root Mean Square Error (RMSE) | |
|----------------------------|-------|--------------|---------------------------------------|--------------|---------------------------|--------------|-------------------------------|--------------|
| | Train | Test | Train | Test | Train | Test | Train | Test |
| Elastic net | 0.692 | 0.659 | 0.343 | 0.364 | 2.877 | 3.031 | 4.101 | 4.297 |
| Random forest | 0.939 | 0.627 | 0.158 | 0.379 | 1.304 | 3.113 | 1.830 | 4.488 |
| $3 \times 1,024$, Sigmoid | 0.942 | 0.514 | 0.188 | 0.395 | 1.295 | 3.517 | 1.784 | 5.139 |
| 3×512 , ReLU | 0.963 | 0.512 | 0.107 | 0.379 | 1.042 | 3.64 | 1.382 | 5.137 |
| 128, 64, 32, 16, ReLU | 0.882 | 0.492 | 0.201 | 0.389 | 1.926 | 3.697 | 2.509 | 5.246 |
| 32, 16, LeakyReLU | 0.808 | 0.577 | 0.254 | 0.374 | 2.323 | 3.383 | 3.237 | 4.780 |

This table shows the average of all folds. The best results in each test case are shown in bold. The last four rows show different neural network models and their architecture, with the numbers indicating the number of hidden neurons (3×512 indicates three layers of 512 neurons) followed by the activation function used.

Due to the low number of samples (for a regression task), particularly during k -fold cross-validation, it is impractical to use a validation set for early stopping of the training of the neural network. Instead, as previously discussed, dropout in combination with a low number of epochs is used to prevent overfitting (Srivastava *et al.* 2014). Specifically, this can prevent the test accuracy from dropping while training. The large size of the neural networks considered allows for complex relationships between the large number of features for this problem. However, the low testing accuracy of these models suggests such relationships do not appear to be present, or are at least not beneficial to the problem, in this case.

The results of the random forest and neural networks are presented to demonstrate that more complex or non-linear models do not have better testing accuracy than the simple linear model in this particular problem, although their training accuracies are higher. The relatively high test accuracy of the linear model compared to the more complex and non-linear models may be due in part to the combinatorial features created, which may extract many non-linear relationships from the data in a way that the linear model can make the best use of. It is also clear that the more complex models experience significant overfitting.

Figure 3(a) shows the prediction of the elastic net linear model. Samples with a prediction ≤ 10 account for 40.7% of all samples and samples with a prediction of ≤ 20 account for 91.4% of all samples. Figure 3(a) shows a handful of samples with high observed values but lower predicted values and Figure 3(b) shows these points have a slightly above-average absolute percentage error compared to the rest of the sample set. Further analysis has shown that these points whose observed

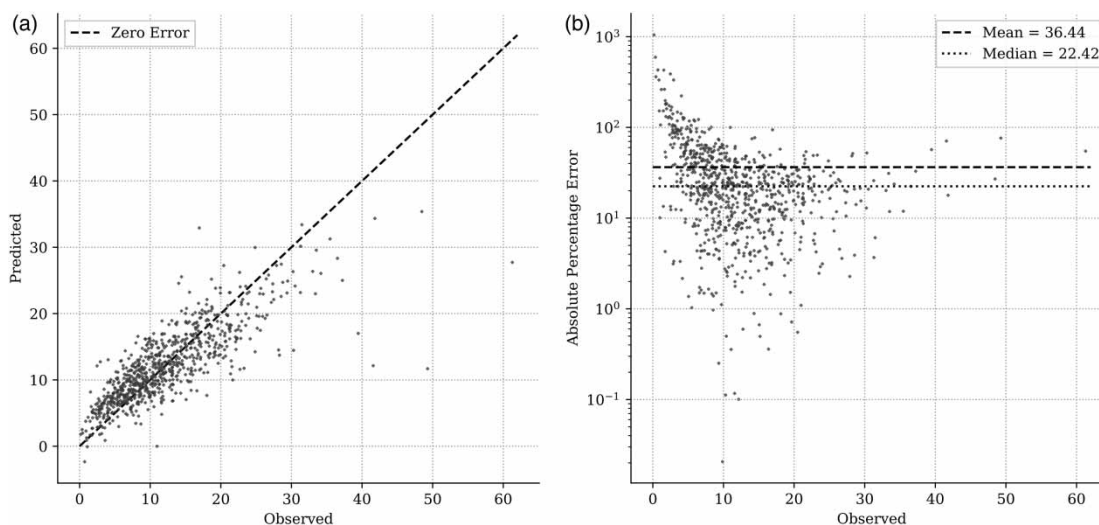


Figure 3 | The elastic net linear model during 5-fold cross-validation, (a) predictions and observed values, with the dashed line showing the line of no error, and (b) absolute percentage error of predictions, where the dashed line shows the mean absolute percentage error, i.e., the MAPE, and the dotted line shows the median absolute percentage error. Predictions were recorded when each sample was in the test set.

value is greater than 38 are all in the top 25% of DMAs with the most missing historic daily MNF records. The average number of missing MNF records for this subset is 115. This suggests that these points are outliers due to the number of missing MNF records. However, when considering all of the DMAs, there is no overall trend between the number of missing observations and the observed value or the prediction error.

As discussed in Sections 1 and 2, there are numerous factors that contribute to MNF, and MNF is made of both legitimate demand and leakage. This article uses simple, static data to make predictions of historic MNF. However, the results suggest that this is enough for reasonable accuracy. For a linear model, the R^2 can be interpreted as the percentage of variation that is explained by the model, and therefore, the data. This means that over 65% of the variation between average MNFs of the different DMAs in 2022 can be explained by a linear combination of the pipe records alone. In other words, 65% of the variation in historic MNF can be attributed to the physical pipe characteristics, without considering pressure. Considering that MNF is not a perfect metric and will include some legitimate demand, this is a large amount of variation that can be explained using readily available, static data. It is possible that, in aggregate, over the course of a full year, legitimate demand during the MNF period becomes a predictable constant based solely on the number of customer connections in a DMA. Arguably, this could suggest that a significant proportion of this variation is due to total customer demand rather than leakage. However, using similar data and methods, but without customer connections as a feature, Hayslep *et al.* (2023) also found that 65% of the variation between DMAs could be explained using just the length, age, diameter, and material of the pipes. Therefore, we can conclude that either: (1) the number of customer connections is not relevant to the variation of average MNF between DMAs *as a measure of legitimate demand* or (2) length, age, diameter, and material also encapsulates the same information about legitimate demand that the number of customer connections does. Similarly, the same conclusions can be reached when considering the number of customer connections as a measure of service pipe leakage; either it is not relevant, or the other features encapsulate the same information.

Other studies, such as Berardi *et al.* (2008), Kakoudakis *et al.* (2017), Giraldo-González & Rodríguez (2020), and Giraldo-González & Rodríguez (2020), which consider pipe failure, rather than MNF, have consistently found that length, age, and diameter are important factors for predicting pipe failure. Therefore, it is logical that a related measure, such as MNF, would share many of these factors. However, there is no 'legitimate demand' for pipe failure and pipe failure does not account for background or unreported leakage as opposed to MNF which will include all these effects. This increases the uncertainty of the MNF prediction problem and explains why the results shown in Table 1 are generally lower than those of the articles referenced. However, it is difficult to directly compare the results of these works with the results shown here due to the difference in problem definition.

4.2. Decomposition to pipe-level LCS

The elastic net linear model used for the LCS is trained using a random subset of 70% of the DMAs in the dataset, as described in Section 3. The remaining 30% is used as the test set. The train–test split is the same for all models. The results are similar to those shown in Table 1 which used 5-fold cross-validation. The evolved equations refer to the GP features (and subsequent linear model) evolved in Hayslep *et al.* (2023). Therefore, this table shows the difference between two separate sets of features: (1) the 240 DMA-level features constructed in this article and (2) the 5 evolved features from Hayslep *et al.* (2023). Due to the slight differences in data used between this article and Hayslep *et al.* (2023) the intercept of the equations from the latter is refit using the training data, the coefficients remain as described in the previous article. Changing the intercept has no bearing on the LCS and is purely for the comparison shown in Table 2. Note the better performance of both models on the test set, indicating that the particular split used has placed many of the outliers shown in Figure 3 into the training set. It is uncertain whether this is beneficial or detrimental to the performance of the LCS on the HWO prediction task. Both models are then used to

Table 2 | Metrics of the linear models used to calculate the LCS

| Model | R^2 | | MAPE | | MAE | | RMSE | |
|-------------------|-------|--------------|-------|--------------|-------|--------------|-------|--------------|
| | Train | Test | Train | Test | Train | Test | Train | Test |
| Elastic net | 0.672 | 0.721 | 0.342 | 0.380 | 2.924 | 2.882 | 4.218 | 3.944 |
| Evolved equations | 0.596 | 0.647 | 0.431 | 0.484 | 3.266 | 3.249 | 4.677 | 4.437 |

Metrics shown are for the MNF prediction task. The 'evolved equations' refer to those from Hayslep *et al.* (2023).

calculate two different LCSs for comparison in the next section. As a result of this process, each pipe has an LCS value which estimates its contribution to leakage. In the next section, the LCS values are validated by using them as an indication of how likely a pipe was to have an HWO in the year 2022. Figure 5 visually shows the LCS values of an example DMA as well as the predictions of the models presented in the next section. Note that the MAPE of these models, shown in Table 2, is influenced by the large percentage error, but small absolute error of DMAs with low historic average MNF.

4.3. HWO prediction

Table 3 shows the results of each model for the HWO prediction task. This table shows that the neural network with three hidden layers of 512 neurons using ReLU activation performed the best according to AUC, with the logistic regression performing the worst according to AUC. However, according to the F1, the largest neural network performed the best with the logistic regression coming second. Overall, the performance of each model is relatively close with no model outperforming the other by a significant margin or in all metrics. It is possible that this demonstrates the skill ceiling of the particular data and features being used for this task. As previously discussed, there are many factors that contribute to bursts and leakage, many of which are not possible to consider here. Therefore, this may be close to the best possible results using these features, with this particular definition of the problem. Nevertheless, the models show good accuracy at this task. A big challenge in this kind of prediction, where the minority class is the positive class, is reducing the number of false positives, especially when the number of true positives is very small in comparison. These results are comparable to previous studies on this problem. Liu *et al.* (2022) achieved results in the range of 0.7211 and 0.8669 AUC for a random forest and logistic regression method with different sampling methods to deal with the imbalanced classes. Tang *et al.* (2019) achieved an AUC of 0.79 and 0.84 for two different learning methods of Bayesian networks. Finally, Debón *et al.* (2010) achieved an AUC of 0.8278 for a GLM. While it is difficult to undertake a direct comparison due to the differences in datasets, these results show that the directly trained models shown here perform similarly compared with other studies.

Figure 4 shows Receiver Operating Characteristic (ROC) curves, which visualises the performance of different models. Specifically, the ROC curve shows the true positive and false positive rates at different thresholds. For the models trained directly on the HWO prediction problem, the thresholds are the probabilities of predicting the positive class. For the LCS models, the thresholds are the LCS values, whose calculation was described in Section 3.2. This allows comparison not just at a single decision threshold, but across all decision thresholds. Figure 4(a) shows the models trained on the HWO data and the LCS. This figure shows that the Logistic, Random Forest, and 3×512 ReLU have equivalent performance. In addition, the other neural network models also have equivalent performance to the Logistic, Random Forest, and 3×512 ReLU, but are not shown for clarity. Figure 4(a) also shows that the LCS, considering that it is based on a model that was trained to predict MNF, has reasonable accuracy, with the elastic net-based LCS achieving an AUC of 0.71 and the much simpler evolved equation-based LCS achieving an AUC of 0.69. While clearly less effective than the directly trained models, the LCS method is able to create similarly accurate predictions, without ever having been trained on HWOs.

At a false positive rate of around 0.09, the ROC curves of the elastic net based and evolved equation-based LCSs cross over, as can be seen in Figure 4(b). For thresholds lower than this, the evolved equation-based LCS is superior, though only by a slim margin. However, considering the imbalance of the classes and an understandable preference to reduce false positives in

Table 3 | Metrics for the different models for HWO prediction

| Model | F1 | | Accuracy | | Log Loss | | AUC | |
|----------------------------|-------|--------------|----------|--------------|----------|---------------|--------|---------------|
| | Train | Test | Train | Test | Train | Test | Train | Test |
| Logistic (Elastic net) | 0.204 | 0.210 | 0.803 | 0.804 | 0.5436 | 0.5439 | 0.8179 | 0.8174 |
| Random forest | 0.212 | 0.198 | 0.772 | 0.771 | 0.4530 | 0.5148 | 0.8709 | 0.8295 |
| $3 \times 1,024$, Sigmoid | 0.212 | 0.218 | 0.814 | 0.817 | 0.5365 | 0.5395 | 0.8293 | 0.8296 |
| 3×512 , ReLU | 0.170 | 0.176 | 0.700 | 0.706 | 0.5134 | 0.5109 | 0.8279 | 0.8297 |
| 128, 64, 32, 16, ReLU | 0.183 | 0.190 | 0.741 | 0.749 | 0.5081 | 0.5113 | 0.8320 | 0.8295 |
| 32, 16 LeakyReLU | 0.177 | 0.184 | 0.722 | 0.727 | 0.5087 | 0.5110 | 0.8396 | 0.8288 |

The best results in each test case are shown in bold. The last four rows show different neural network models and their architecture.

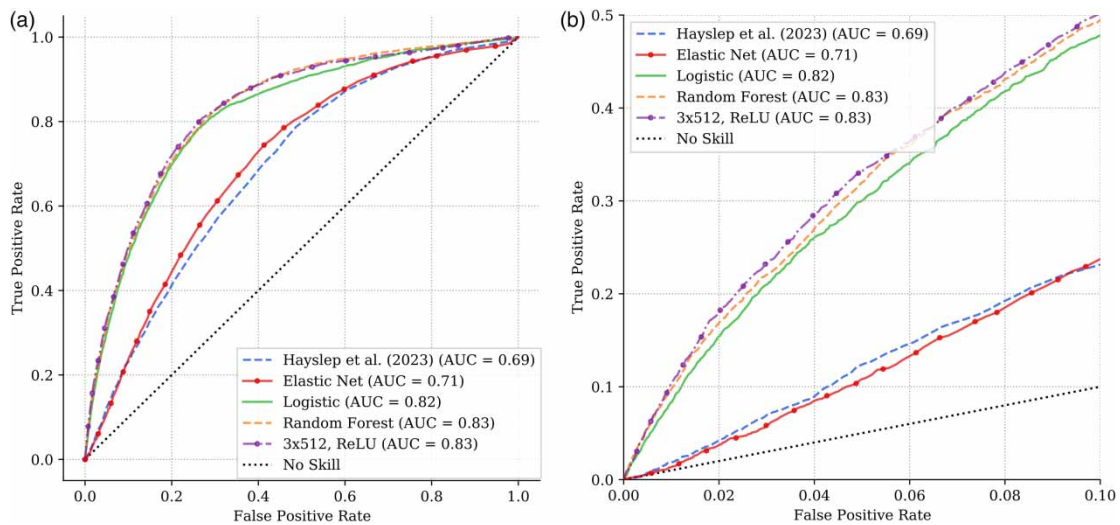


Figure 4 | ROC curves for work order prediction on the test set using trained models. (a) The entire ROC curve while (b) shows only the segment where the false positive rate is less than 0.1.

this kind of problem, this is the most important area of the ROC curve. With an HWO rate of 3.7%, a false positive rate of 0.1 would mean more than 2.5 times as many false positives as *all* true cases. This is one of the primary hurdles for methods such as these to be used in the real world.

The equation of the underlying linear model of the logistic regression for a pipe i is:

$$y_i = 0.5755l_i - 0.2123d_i - 0.0619v_i + 0.6975c_i + 0.2740a_i - 0.0236m_i + 0.0129p_i + 0.0175o_i - 0.5388 \quad (4)$$

where l_i is the length, d_i is the diameter, v_i is the volume, c_i is the number of customer connections, a_i is the age, and finally, m_i , p_i , and o_i are 1 if the pipe is *metal*, *plastic*, or *other*, respectively, for a pipe i , and 0 otherwise. This equation broadly aligns with other research in this area such as Boxall *et al.* (2007), Barton *et al.* (2019), and Berardi *et al.* (2008) demonstrating: (1) the inverse relationship between diameter and HWOs (Berardi *et al.* 2008), (2) the positive relationship between length and HWOs, which has been suggested to be non-linear (Boxall *et al.* 2007), (3) the positive relationship between customer connections and HWOs (Berardi *et al.* 2008), which could be a proxy for the total length of supply pipes, which tend to have a smaller diameter, and (4) the relationship between age and HWOs, which here is shown to be simply positive, but has been shown to be more complex and to depend on other factors such as material and environmental conditions (Boxall *et al.* 2007; Barton *et al.* 2019). However, the relatively low importance of material does contrast with the aforementioned articles. It has been suggested that the effects of material are both seasonal and weather dependant (Boxall *et al.* 2007; Barton *et al.* 2019). Therefore, it is possible that because the study period is an entire year, these effects have been factored out.

Figure 5 shows an example DMA from the test set and the prediction probability or LCS for each of the models shown in Figure 4. This figure shows some level of agreement between the different models, as demonstrated by the labelled points A and B. This is particularly interesting considering the LCS-based approach is based on models that are trained to predict average MNF. This may be evidence of the importance of background and unreported leakage (as measured by the MNF) to the occurrence of bursts and leaks (as measured by HWOs).

The purpose of the LCS is to understand how much individual pipes contribute to leakage. The results in this section suggest that this is a valid approach. There is a difference between predicting HWOs and historic leakage. However, as shown in this section, the LCS is still able to show good results, and on inspection of the example DMA in Figure 5, shows that the LCS models and the prediction models agree on many occasions. Overall, these results validate the LCS as a sound concept while also validating the particular implementation of it in this article.

The results in this article, for both tasks, are for the year 2022. The same features are likely to be able to explain a similar amount of variation in any other year, or combination of years. For example, Hayslep *et al.* (2023) used data covering 2018–

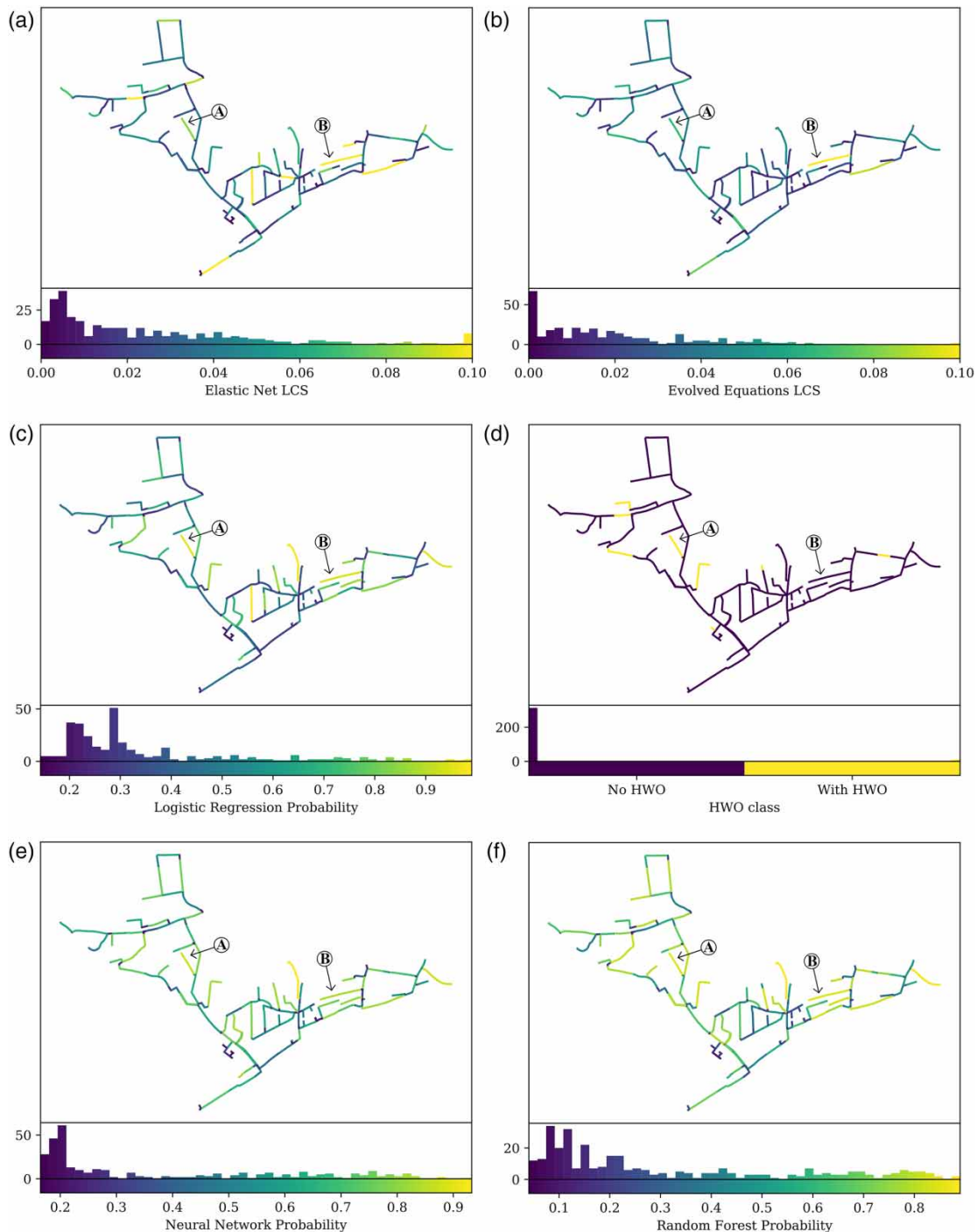


Figure 5 | Example of a real-world DMA from the test set with pipes coloured according to each model's prediction. (a, b) The different LCS-based methods. (c, e, and f) The prediction probability of each regression model. Finally, (d) the true class of each pipe. Labels A and B show two pipes which all models agree have a high likelihood of having an HWO, A shows correct agreement, while B shows an incorrect agreement. The histogram shows the number of pipes along the probability, LCS, or class axis. The LCS in (a) and (b) is clipped between 0 and 0.1 to show the main distribution in more detail. Points outside this range are in the first and last bins of the histogram, respectively.

2022 and explained a similar amount of variation. However, the importance of each feature might vary between different periods, and this may encapsulate some information about changes in environmental or operational conditions. Future work will examine the extent of this variation.

As previously discussed, each HWO is assigned to the nearest pipe within 50 m, meaning each HWO is associated with a maximum of one pipe, and as discussed in Section 3.3, this is a classification task for whether or not a pipe had an HWO in 2022. As a result, near-misses where, for example, a pipe that is within 1 m of an HWO, but is not the closest pipe, counts as a false positive if labelled positive by the model. In reality, such pipes may have been involved in the HWO. Such cases may overly penalise the LCS which has never seen the particular delineation of HWO onto pipes that was chosen in this article. Figure 5 shows such a case, where pipes with a high LCS are adjacent to pipes with a high number of work orders. Depending on how common an occurrence this is, the performance of the LCS (which is based on MNF prediction) at predicting HWO may be better than illustrated here.

Another limitation is that, due to the way in which the LCS is validated by using ROC curves, only the ordering of the pipes by their LCS values is assessed. Although this is also true for the prediction probability, it means that there is no evidence, using this particular analysis, that the actual values or magnitudes of the LCSs have any particular meaning outside of how they order a collection of pipes. Future work will attempt to address this limitation. Despite this issue, the LCS provides a good prediction of HWOs, especially considering that the models used to calculate it were trained to predict historic average MNF. This may raise additional possibilities about the potential uses of transfer learning, especially considering that MNF is much more commonly available, even in DMAs where there might be no HWOs to study otherwise. Future work could also replace the linear regression with a non-linear method such as EPR.

5. CONCLUSIONS

This article has trained models to predict two different measures of historic leakage: average MNF and HWOs. The regression models showed good R^2 for the MNF prediction task and show that a significant proportion of variation in historic average MNF can be accounted with the pipe-based features used. Then by decomposing the model used to predict MNF, a numerical value for each pipe's contribution to leakage is calculated and named the LCS. To validate this score, it is compared against models trained to predict the HWOs. The two different LCSs perform well at this task, as shown by their AUC metrics, and the results validate the LCS as a useful tool for further understanding leakage and MNF. This approach could be especially useful in cases where there is very little or no historic pipe failure data.

The work presented supports a broader approach to leakage by understanding both MNF and HWOs in the same context. Both pipe failure and MNF are strongly linked to leakage, and therefore, each other. The concept of LCSs links these different approaches to understanding leakage, through the use of explainable models. This article has also demonstrated the potential of practical and applicable techniques for the water industry that are usable on a wide scale with data that is easily available to water companies and without the installation of new sensors. By understanding leakage through MNF at the pipe-level, water utility companies will be able to make better and more informed decisions on how and where to tackle leakage.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support from South West Water in collaboration with the University of Exeter Centre for Resilience, Environment, Water and Waste (CREWW).

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict of interest.

REFERENCES

- Albon, C. 2018 *Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning*, 1st edn. O'Reilly Media, Sebastopol, CA.
- Alkassab, J. M. A., Adlan, M. N., Abustan, I., Aziz, H. A. & Hanif, A. B. M. 2013 *Applying minimum night flow to estimate water loss using statistical modeling: A case study in Kinta Valley, Malaysia*. *Water Resources Management* **27**, 1439–1455. <https://doi.org/10.1007/s11269-012-0247-2>.
- Barton, N. A., Farewell, T. S., Hallett, S. H. & Acland, T. F. 2019 *Improving pipe failure predictions: Factors affecting pipe failure in drinking water networks*. *Water Research* **164**, 114926. <https://doi.org/10.1016/j.watres.2019.114926>.

- Bentley, J. L. 1975 **Multidimensional binary search trees used for associative searching**. *Communications of the ACM* **18**, 509–517. <https://doi.org/10.1145/361002.361007>.
- Berardi, L., Kapelan, Z., Giustolisi, O. & Savic, D. A. 2008 **Development of pipe deterioration models for water distribution systems using EPR**. *Journal of Hydroinformatics* **10**, 113–126. <https://doi.org/10.2166/hydro.2008.012>.
- Bishop, C. M. 1995 *Neural Networks for Pattern Recognition*. Oxford University Press, Inc, USA.
- Boxall, J. B., O'Hagan, A., Pooladsaz, S., Saul, A. J. & Unwin, D. M. 2007 **Estimation of burst rates in water distribution mains**. *Proceedings of the Institution of Civil Engineers – Water Management* **160**, 73–82. <https://doi.org/10.1680/wama.2007.160.2.73>.
- Breiman, L. 2001 **Random forests**. *Machine Learning* **45**, 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Debón, A., Carrión, A., Cabrera, E. & Solano, H. 2010 **Comparing risk of failure models in water supply networks using ROC curves**. *Reliability Engineering & System Safety* **95**, 43–48. <https://doi.org/10.1016/j.res.2009.07.004>.
- Farah, E. & Shahrou, I. 2017 **Leakage detection using smart water system: Combination of water balance and automated minimum night flow**. *Water Resources Management* **31**, 4821–4833. <https://doi.org/10.1007/s11269-017-1780-9>.
- Farley, M. & Trow, S. 2005 *Losses in Water Distribution Networks: A Practitioners' Guide to Assessment, Monitoring and Control*. IWA Publishing, London, UK.
- Farrow, J., Jesson, D., Mulheron, M., Nensi, T. & Smith, P. 2017 *Achieving Zero Leakage by 2050: The Basic Mechanisms of Bursts and Leakage*. UK Water Industry Research Limited, London.
- Giraldo-González, M. M. & Rodríguez, J. P. 2020 **Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks**. *Water* **12**, 1153. <https://doi.org/10.3390/w12041153>.
- Hayslep, M., Keedwell, E. & Farmani, R. 2023 **Multi-objective multi-gene genetic programming for the prediction of leakage in water distribution networks**. In: *Genetic and Evolutionary Computation Conference (GECCO '23)*, ACM, July 15–19, Lisbon, Portugal, p. 8. <https://doi.org/10.1145/3583131.3590499>.
- Jang, D. & Choi, G. 2017 **Estimation of non-revenue water ratio for sustainable management using artificial neural network and Z-score in Incheon, Republic of Korea**. *Sustainability* **9**, 1933. <https://doi.org/10.3390/su9111933>.
- Jang, D. & Choi, G. 2018 **Estimation of non-revenue water ratio using MRA and ANN in water distribution networks**. *Water* **10**, 2. <https://doi.org/10.3390/w10010002>.
- Jang, D., Park, H. & Choi, G. 2018 **Estimation of leakage ratio using principal component analysis and artificial neural network in water distribution systems**. *Sustainability* **10**, 750. <https://doi.org/10.3390/su10030750>.
- Kakoudakis, K., Behzadian, K., Farmani, R. & Butler, D. 2017 **Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with K-means clustering**. *Urban Water Journal* **14**, 737–742. <https://doi.org/10.1080/1573062X.2016.1253755>.
- Kizilöz, B. 2021 **Prediction model for the leakage rate in a water distribution system**. *Water Supply* **21**, 4481–4492. <https://doi.org/10.2166/ws.2021.194>.
- Lambert, A. & Morrison, J. a. E. 1996 **Recent developments in application of 'bursts and background estimates' concepts for leakage management**. *Water and Environment Journal* **10**, 100–104. <https://doi.org/10.1111/j.1747-6593.1996.tb00017.x>.
- Liu, W., Wang, B. & Song, Z. 2022 **Failure prediction of municipal water pipes using machine learning algorithms**. *Water Resources Management* **36**, 1271–1285. <https://doi.org/10.1007/s11269-022-03080-w>.
- Mounce, S. R. & Boxall, J. B. 2010 **Implementation of an on-line artificial intelligence district meter area flow meter data analysis system for abnormality detection: A case study**. *Water Supply* **10**, 437–444. <https://doi.org/10.2166/ws.2010.697>.
- O'Day, D. K. 1982 **Organizing and analyzing leak and break data for making main replacement decisions**. *American Water Works Association* **74**, 588–594.
- Ofwat 2021 *Service and Delivery Report 2020–21 (Comparative Reports and Data), Service and Delivery Report*. Ofwat.
- Puust, R., Kapelan, Z., Savic, D. & Koppel, T. 2010 **A review of methods for leakage management in pipe networks**. *Urban Water Journal* **7**, 25–45. <https://doi.org/10.1080/15730621003610878>.
- Romano, M., 2021 **Review of techniques for optimal placement of pressure and flow sensors for leak/burst detection and localisation in water distribution systems**. In: *ICT for Smart Water Systems: Measurements and Data Science, the Handbook of Environmental Chemistry* (Scozzari, A., Mounce, S., Han, D., Soldovieri, F. & Solomatine, D. eds). Springer International Publishing, Cham, pp. 27–63.
- Romano, M., Kapelan, Z. & Savić, D. A. 2014 **Automated detection of pipe bursts and other events in water distribution systems**. *Journal of Water Resources Planning and Management* **140**, 457–467. [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000339](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000339).
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. 2014 **Dropout: A simple way to prevent neural networks from overfitting**. *The Journal of Machine Learning Research* **15**, 1929–1958.
- Tang, K., Parsons, D. J. & Jude, S. 2019 **Comparison of automatic and guided learning for Bayesian networks to analyse pipe failures in the water distribution system**. *Reliability Engineering & System Safety* **186**, 24–36. <https://doi.org/10.1016/j.res.2019.02.001>.
- Wan, X., Kuhanestani, P. K., Farmani, R. & Keedwell, E. 2022 **Literature review of data analytics for leak detection in water distribution networks: A focus on pressure and flow smart sensors**. *Journal of Water Resources Planning and Management* **148**, 03122002. [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0001597](https://doi.org/10.1061/(ASCE)WR.1943-5452.0001597).
- Zou, H. & Hastie, T. 2005 **Regularization and variable selection via the elastic net**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.