A Data Driven Approach for Leak Detection with Smart Sensors

Bin Liang[1], Sunny Verma[1], Jie Xu[1], Shuming Liang[1], Zhidong Li[1], Yang Wang[1] and Fang Chen[1]

Abstract— **Preventing water pipe leaks and breaks has high priority for water utilities. It is a critical task for the utility to reduce water loss through leaks and breaks detection in water mains. The failure prediction and data analytics research have been conducted for an Australian water utility over the last few years to enhance the prediction of leaks and breaks detection in water mains. Intelligent sensing at sensitive locations with current research aids in prioritising investigation and prevention of potential breaks and leaks in water mains. The purpose of this work is to integrate the predictive analytics and intelligent sensing applications to identify high risk mains prior to failures. Predictive analytics and minimum night flow (MNF) analysis have been utilised to prioritise risky zones over the whole water network, and then risky pipes are identified to optimise sensors deployment. The sensing data is being collected for analysis and validation, and a machine learning model is being built based on the analysis results. This work is currently under progress and the planned outcomes will help the utility reduce water loss, improve leak detection, and enhance customer satisfaction by automating the process of leak detection using a data driven approach with smart sensors.**

I. INTRODUCTION

To ensure a high-quality supply of potable water to utility customers, it takes any disruptions to customers supplies (e.g. through pipe disruptions or impaired water quality) as high priority. For urban water utilities, the cost of maintaining aging water mains has become the major concern [3].

It would be highly beneficial for utilities to develop more reliable and targeted smart sensing tools (including mobile and fixed real-time sensors) to monitor risky areas to further validate and more accurately predict leaks and breaks, using reliable, easy to acquire data. Based on the breakthrough success of acoustic and pressure transient sensors detecting leaks in other industry areas [10], [5], these advances could be adapted and enable the development of a universally applicable "leak-before-break" model, which could significantly benefit water utilities to preventatively manage their buried critical water assets.

The emergence of Internet of Things (IoT) technology has allowed more modern methods to monitor water networks [8], [6]. The use of IoT enables real-time monitoring of an area by sensors for days or even several months. By analysing the monitoring data, people can eventually determine whether there are failures happening on the underground pipes, and then dig out the pipes for further inspection. Breakthrough acoustic techniques have been used by other industry for the detection of leaks and have the potential applications for the water industry. The sensors have been deployed in critical sections of the underground network as a real-time sensor input. For example, water utilities in Australia recently installed acoustic sensors for leak detection to help reduce and detect water main leaks and failures [1], [2]. Real time sensing of leaks over time will provide learning data with multiple modalities to pick up leaks and breaks prior to occurrence. Using machine learning and multi-modal data analytics for each of the utility study areas optimum site locations can be decided to get the best value of the real-time sensing data. The validated leak predictive model outputs, using real-time and mobile sensing data, facilitate the categorisation and prioritisation of pipes, which can be integrated with other developing condition assessment techniques to select the cohort of pipes to be replaced at their actual "end of life".

This paper summarises the ongoing research work with an Australian water utility for prioritising zone areas and pipes to reduce pipe breaks and leaks using the advanced data driven approach. The main contribution of this work are threefold: 1) An ensemble model is utilised to learn multi-source data for water pipe failure prediction; 2) Risky zones and pipes have been identified for optimised sensors deployment. In particular, methods for generating risk maps, and MNF adjustment are proposed and utilised to generate zone-level and pipe-level prioritisation lists; 3) Validations are being conducted on deployed sensors and prioritised pipes.The validation results demonstrate the effectiveness of the proposed data driven approach for leak detection with smart sensors.

The paper is organised as follows: Section II presents the ensemble machine learning model for water pipe failure prediction. Section III details zones prioritisation which includes MNF analysis and rank aggregation. Section IV discusses the leak detection and validation results, and Section V concludes the paper.

II. WATER PIPE FAILURE PREDICTION

For the urban water utilities, the cost of maintaining water mains has become the major concern. A collaborative work has been carried out with an Australian water utility to apply domain expertise and advanced machine learning techniques to achieve a cost-effective solution for water pipe failure prediction in the water network.

A. Feature Engineering with Domain Knowledge

Pipe failure behaviours are highly related to its characteristics such as materials, sizes, pressures, topography, soil types, etc and various kinds of information are managed from multiple sources. A high volume of historical data,

[1]All authors are with Faculty of Engineering and IT, University of Technology Sydney, 15 Broadway, Ultimo NSW 2007, Australia. {fistname.lastname}@uts.edu.au

pipe attributes, and operational data have been recorded by the water utility. Additionally, topographic information and soil information have also been collected from public data sources.

Feature engineering is fundamental to the application of machine learning, and is both difficult and expensive. To have a better understanding of multi-source data, interviewing with domain experts from the water utility has been actively performed. Domain experts explained their considerations of factors that affect pipe failures. For example, the difference of elevation between water mains could cause additional water pressure. Considering factors such as the direction of water flow and the distance to the water pump, the practical pressure for each pipe or joint between pipes is much complex, which is considered as one of the main factors causing pipe failures. Based on the pipe data and elevation data that we have collected, two topographic features have been extracted: *difference of ground level (DGL)* and *shape type of connection (STC)*.

As shown in Fig. 1, *DGL* feature is the mean elevation difference between the target pipe and pipes within the range of 100 metre from the centre of this target pipe, where the pipe which the feature is designed for is called target pipe. The *DGL* of the target pipe for each pipe line can be obtained by

$$DGL_{i=0} = \sum_{i=0}^{n} \frac{d_{(i,i+1)}}{100} DGL_{(i,i+1)} \tag{1}$$

where $\sum_{i=0}^{n} d_{(i,i+1)} = 100$, and $i = 0$ indicates the target pipe. $d_{(i,i+1)}$ is the horizontal distance between the central positions of pipe i and pipe $i+1$. $DGL_{(i,i+1)}$ is the difference of elevation between the central positions of pipe i and pipe $i+1$, where the elevation on the central position of each pipe is obtained by averaging the ground level of all nodes of this pipe. Then the *DGL* for the target pipe is the mean value of all $DGL_{i=0}$. This feature shows a positive correlation with the failure rate.

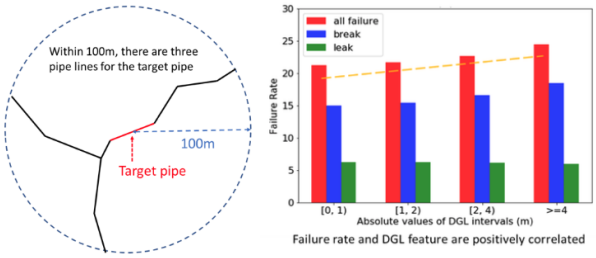

Fig. 1: An example of extracting the topographic feature *DGL* and its correlation with failure rate.

Another topographic feature, *shape type of connection (STC)*, is also designed in addition to *DGL* which only considers the difference of elevation. In the elevation direction, commonly there are three shapes of pipe connection as shown in Fig. 2. With the pipe layout information and ground-level data, this feature can be easily extracted. This feature shows a strong correlation with the pipe failure rate.

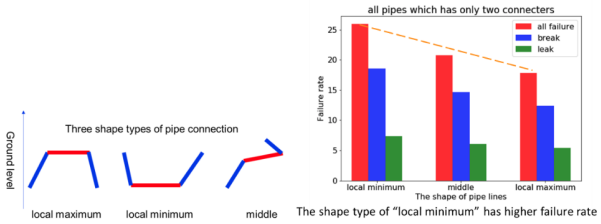

Fig. 2: Illustration of the feature *STC* and its correlation with pipe failure rate.

B. Ensemble Learning & Feature Importance

Ensemble learning is an advanced machine learning paradigm where multiple learners are trained to solve the same problem. In contrast to ordinary machine learning approaches which try to learn one hypothesis from training data, ensemble methods try to construct a set of hypotheses from multi-dimensional features and construct a strong model. This model enables to handle high-dimensional data for prediction. The underlying statistical principle employed here is gradient boosting technique [4]. The boosting technique consists in fitting sequentially multiple tree-based base learners in a very adaptive way: each model in the sequence is fitted giving more importance to observations in the high-dimensional features that were badly handled by the previous models in the sequence. Finally, the ensemble model is built based on a weighted sum of base learners.

A benefit of using gradient boosting is that after the boosted trees are constructed, it is relatively straightforward to retrieve importance scores for each attribute. Generally, importance provides a score that indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance. This importance is calculated explicitly for each attribute in the dataset, allowing attributes to be ranked and compared to each other. Information Gain (IG) determines which feature provides the maximum information about the prediction. It is based on the concept of entropy which is the degree of uncertainty, impurity or disorder. We investigate the multi-dimensional features by employing IG from the developed ensemble model. Based on the feature importance from the developed model, key features can be identified. Fig. 3 shows the top 5 features for the years of 2018 and 2019. In addition to pipe attributes, the variance of ground levels is also a key feature that impacts the pipe failures.

C. Model Evaluation

In order to evaluate the model, we have compared the prediction results with actual ground truth data by calculating the percentage of detected failures with respect to the percentage of prioritised pipes (when pipes are ranked in

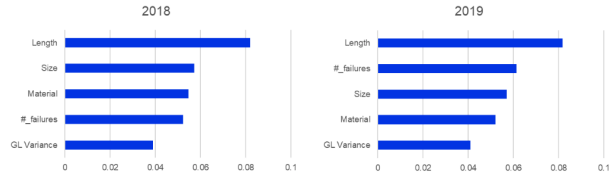

Fig. 3: Top 5 features for the years of 2018 and 2019. (GL Variance: Variance of ground levels).

descending order of the failure likelihoods). This is depicted in Fig. 4, where more than 60% of failures can be detected if the top 10% of the pipe are inspected. Results demonstrate that our model is capable to provide valuable assistance to forecast and plan water main renewals with more confidence via predictive analytics.

The developed pipe failure prediction model is able to produce failure likelihoods for the levels of pipes as well as zones. In order to achieve more reliable results, minimum night flow (MNF) analysis can be leveraged for zones prioritisation which will be discussed in Section III.

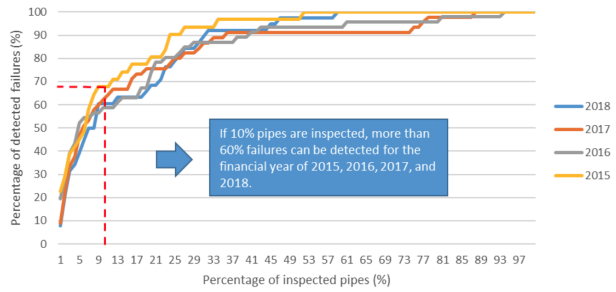

Fig. 4: Model evaluation based on historical data of different years.

III. MNF ANALYSIS & ZONES PRIORITISATION

In this section we present our work on prioritising zones using minimum night flow (MNF) analysis and rank aggregation.

A. Minimun Night Flow

Minimum night flow (MNF) for a pressure zone is defined as the water flow measured at night time. The motivation for choosing night time is that the amount of water consumption is minimal over the whole day. However, customers such as industrial and commercial units might still consume the water either at the standard consumption rate or at a slightly higher consumption rate during the nigh time. Hence a constant flow (or flow with slight variation) must be recorded for each day. If the recorded MNF for consecutive days deviates substantially from its historical records, it might be either due to occurrence of leak or valve is accidentally left open which is allowing water flow between neighbouring pressure zones. Besides, the MNF also helps in narrowing down pressure zones for Active Leak Detection (ALD). Therefore, its monitoring is a quick way to estimate the unreported leaks in pressure zones. An example of MNF monitoring is shown in Fig. 5

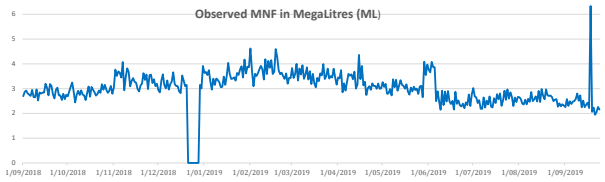

Fig. 5: An example of observed MNF

Although the monitoring of MNF is very beneficial in estimating unmetred water loss or identifying pressure zones where water-leak might be occurring. It is highly susceptible to system errors such as malfunctioning of water-consumption meter. Therefore, one needs to sieve zones where MNF is reliable(i.e., not showing large fluctuations throughout the data period). Once zones with reliable MNF recordings are identified, it still requires further processing to eradicate the influence of non-residential consumption.

The collected MNF dataset consists of the observations from more than 300 pressure zones recorded from 2018 to 2019. Moreover, the MNF for each pressure zone is labelled either as reliable or not reliable by the water utility where the reliability is determined by observing fluctuations in the MNF. A zone with frequent fluctuations in the MNF observations is assumed to be unreliable as its MNF might be influenced by noises such as mis-calibrations in the apparatus or formula utilised to calculate the MNF observations. These observed MNF values for each pressure zone further need to be processed in order to estimate the reliable value for each day. In this regard, the observed MNF value for the whole period of time is added (the sign of the observed MNF is accounted while performing addition) and normalised by the number of connections in each zone. The obtained value is now converted into Megalitres per hour and then normalised over the whole year. This value now signifies litres/connections/hour (L/C/H), which can be utilised to conduct the MNF adjustment task.

B. MNF Adjustment

Customer consumption data is mainly used for MNF analysis and adjustment. Each customer connected to the network has a separate water metre that is read approximately every three months. The acquired dataset includes the historical water consumption for each pressure zone. This data has been interpolated to produce yearly estimates of the water used by each pressure zone. There are mainly two types of customers: residential and non-residential. Each type is further divided into several sub-types. For example, residential customers are further categorised into “single dwelling”, “flats and mixed developments”, etc. Non-residential customers include “commercial”, “industrial”, and “agricultural”, etc. Most residential customers normally use much less water at night, so non-residential customers consumption needs been further analysed. Based on our analysis,

non-residential consumption accounts for slightly more than one quarter of the total water consumption over the year of 2019.

The observed MNF values are influenced by both the residential consumption and non-residential consumption. While the residential consumption is assumed to be minimal at night, no such assumption can be drawn for non-residential consumption. Besides, the consumption cycle from non-residential consumers is dependent on their nature such as a poultry farm might consume more water in the daytime compared to RSL clubs which consume more water at night. Also, since the number of non-residential consumers and their types are unequal within the pressure zones, one cannot deduct a constant factor from each pressure zone to alleviate their influence. Therefore, estimation and eradication of non-residential water consumption from observed MNF becomes non-trivial. In this regard, we begin with summarising non-residential consumers according to their type and water consumption demand where the obtained frequencies of non-residential water consumption dictates that the paramount non-residential water consumption is due to commercial, industrial, and government institution and others consumers. This knowledge helps us in narrowing the type of non-residential consumers which might be affecting the MNF observations. One is still required to validate which of the consumption types are indeed influencing the observed MNF. To obtain such an insight we calculate the correlation between each consumption type and observed MNF. The motivation behind calculating correlation is that it will serve as an indicator to quantify the cumulative influence of non-residential consumption types on pressure zones. Positive correlation values indicate the strength of non-residential consumption type on MNF observations. Hence, we can select non-residential types with high correlation coefficients to model the non-residential usage in the observed MNF. We performed correlation analysis with the reliable pressure zones and five non-residential fields as show in Fig. 6.

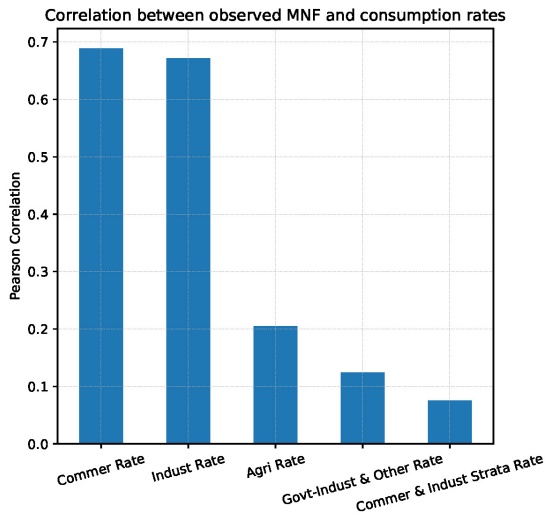

Fig. 6: Correlation between MNF and consumption rates.

It can be seen that commercial (Commer Rate) and industrial (Indust Rate) consumers highly influence the observed MNF values. Hence, these two types can be utilised to estimate the non-residential consumption or usage in the observed MNF. We now describe the leakage modelling utilised for MNF adjustment. The motivation is to alleviate the influence of non-residential consumption from MNF. In other words, we want to estimate the non-residential consumption in the MNF denoted as Q_{NRES} and deduct this value from the recorded value of MNF denoted as Q_{MNF}. This withdrawal gives us an estimate of the true value of MNF denoted as $L_{(zone@tMNF)}$, which is utilised to identify pressure zones with water leakage. Technically, the deduction of non-residential consumption from recorded MNF is as in Eq. 2.

$$L_{(zone@tMNF)} = Q_{MNF} - Q_{NRES} \tag{2}$$

where L is the water loss of zone at the MNF time, Q_{MNF} is the MNF, and Q_{NRES} denotes non-residential usage at the MNF time.

To estimate Q_{NRES} we utilised a simple linear regression model without the intercept term as in Eq. 3. The motivation for excluding the intercept term is to force the regression error to be non-zero as the value of this error is a measure of the strength of our regression estimate. In other words, we want mean of error term to be equal to the mean of the observed MNF with only residential consumption. This constraint will enforce the adjusted term to reflect the residential consumption as shown in Fig. 7.

$$Q_{NRES} = \beta_{COM} \times COM_{Ratio} + \beta_{IND} \times IND_{Ratio} \tag{3}$$

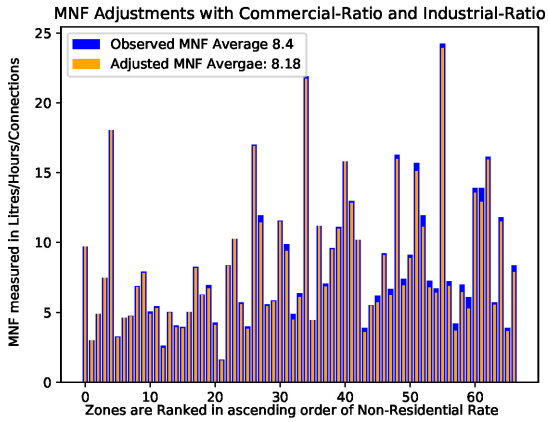

Fig. 7: MNF Adjustment Results.

The adjustment result in Fig. 7 demonstrates that the mean of adjusted MNF with our regression scheme is able to estimate the residential consumption of pressure zones with negligible non-residential consumption. The zones are ordered in ascending order of non-residential consumption and illustrate that the two values overlaps with negligible difference between the observed and adjusted means.

C. Zones Prioritisation

In order to prioritise zones/areas for sensors deployment, risky zones need to be prioritised. Since the MNF has been adjusted based on our model, the zones can be ranked. Furthermore, zones prioritisation can also be obtained based on the pipe failure failure prediction. Specifically, risk score of the pressure zone is aggregated by summing up all failure likelihoods of the pipes in that zone. Finally, the analysis at zone level can be aggregated by utilising failure prediction and adjusted MNF to achieve more reliable outcomes.

We rank the pressure zones based on their aggregated risk scores in descending order. The pressure zone with maximum risk probability is allocated rank 1 and the second maximum risk probability zone is allocated rank 2 and so on. Similarly, we rank all pressure zones based on the adjusted MNF values in descending order. These two rankings of pressure zones are then aggregated with various rank aggregation schemes such as Borda-Count (L2 norm, Geometric Mean, etc.) and Markov-Chain based methods [7]. The rank aggregation schema is illustrated in Fig. 8 and the rank aggregation results are shown in Fig. 9.

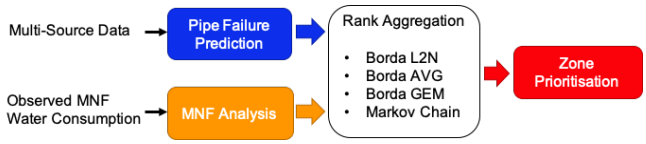

Fig. 8: Rank aggregation for zones prioritisation.

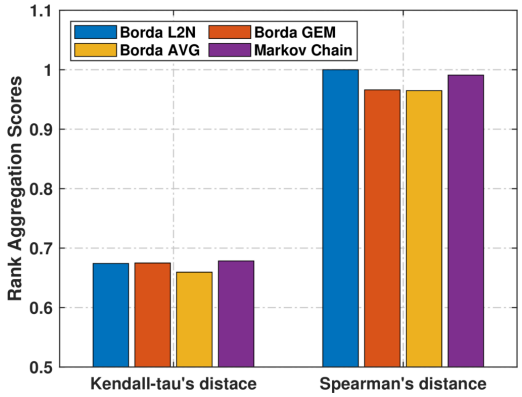

Fig. 9: Rank aggregation scores with different ranking schemes.

The rank aggregation results suggests that in general Kendall-tau distance as a metric is better than SpearMan's distance for our rank aggregation task. Moreover, the Borda-Count scheme with rank aggregation as average achieves the best aggregation among both the distance metrics. These aggregation results lead to prioritised zones, which guides the water utility to deploy sensors in top ranked zones.

IV. LEAK DETECTION & VALIDATION

With the prioritised pipes and zones, sensors have been deployed. This section shows the validation results of pipe prioritisation against the detected leaks.

A. Acoustic Monitoring for Leak Detection

There many sensing options for leak detection. Static acoustic sensors are one option as a permanent or semi-permanent solution for leak detection. These sensors are attached onto a pipeline surface for sensing real information about pipeline resources [9]. For our work, different sensor models have been deployed in different zones. Fig. 10 shows one type of the deployed sensors. Each static acoustic sensor model has its own data recording capabilities. However they have been configured to make ambient noise recordings between 2 am and 4 am, during which water usage is often low. Under such a setting, environmental noise from traffic and water usage are reduced to an extend that recorded sensor data across multiple days can be compared for changes.

Fig. 10: A Von Roll acoustic sensor

Leak alerts are raised from interpreted acoustic sensor data for further investigation by water utility crews. Specifically leak alerts are raised automatically by proprietary interpretation systems via a threshold based approach. Generally, simple noise level thresholds, proven to lead to large percentage of false positive leak alerts, are used by the interpretation systems to determine if a leak has occurred. By taking the advantage of acoustic monitoring, we can evaluate and then improve our machine-learning driven models for leak and break prediction.

B. Pipes Prioritisation & Validation Results

We trained our machine learning model using the pipe failure records till the end of 2019. The failure records include various failure types, including leaks, breaks, main-to-metre failures. We then applied the model to generate prioritised pipes in Zone A, B and C of the utility's water network. The pipes are ranked as per their probabilities of future failures. For the validation, the location of the prioritised pipes are compared with the locations of the detected leaks which will be described in Section IV-B.

In the first half of 2020, a total of 20 leaks have been detected by the acoustic sensors. Amongst those detected leaks, 19 have been confirmed as true positives. When examining the locations of those true positives, we found that 15 of them are overlapped with the prioritised pipes, resulting in a total matching rate of 75%. It demonstrates the effectiveness of data-drive solution for leak prediction. Details of the validation statistics are presented in Table I.

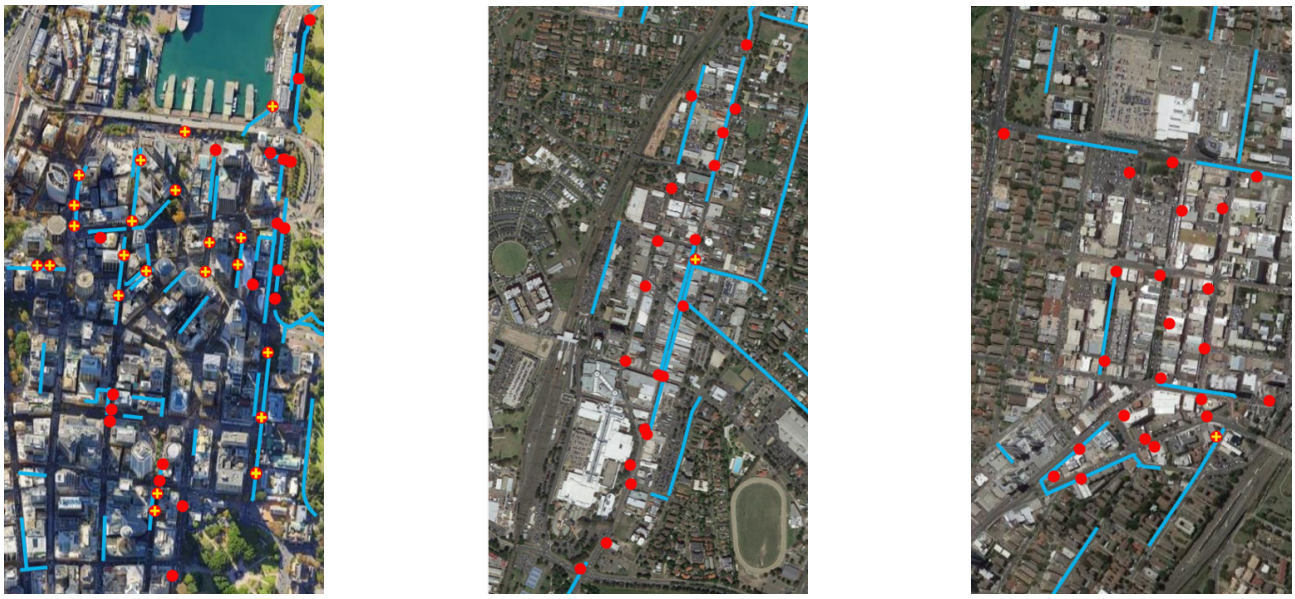

Fig. 11: Validation maps for Zone A, B and C (from left to right). Blue lines indicate prioritised pipes, red dots are deployed sensors, and red dots with yellow crosses are true positive alarms.

TABLE I: The validation statistics

Zone	#sensors	#detected leaks	#confirmed leaks	#overlaps
A	40	17	17	13
B	25	1	1	1
C	23	2	1	1

Note that Zone A is in metropolitan area, whilst Zone B and C are in rural areas. Thus it is expected that the pipe network in Zone A have a higher coverage density than those in B and C. Also there are more sensors deployed in Zone A than Zone B and C. As such sensors in Zone A are expected to have better coverage on the pipe network than those in B and C. That could be one reason for Zone A to contain the most detected leaks. Relatively low coverage in Zone B and C results in lower leak detection numbers.

The validation results for all zones are plotted in Fig. 11. As shown in the figure, prioritised pipes are depicted in light blue, while deployed sensors are illustrated in red dots, and red dots with yellow crosses mean true positive alarms. We count the overlaps by checking if the true positive alarms are overlapped with the prioritised pipes on each zone.

V. CONCLUSIONS

As leaks and breaks becomes susceptible in ageing water mains. Its maintenance has become a major concern for water utilities across globe, where with detecting leaks has achieved substantial breakthrough through utilisation of sensing and analytic technology. In this paper we have presented our data driven approach for leak detection with smart sensors. The approach encapsulate procedures for water pipe failure prediction, minimum night flow (MNF) analysis, and zones prioritisation by consolidating failure records and adjusted MNF. We have also validated our leak detection approach with acoustic monitoring data. The validation results show that our approach has achieved a matching rate of 75% by overlapping confirmed leaks with our prioritised pipes.

In the future we will implement a Web portal that sources sensor data from acoustic models for signal consolidation. More zones will be set up as test beds for pipe leak and break prediction. We will also investigate more sensor models in the new test beds.

REFERENCES

[1] "Acoustic sensing project to help sydney water prevent pipe breakage," https://pacetoday.com.au/acoustic-sensing-project-help-sydney-water-prevent-pipe-breakage/, accessed: 2020-07-09.

[2] "Continued sensor success for sa water's smart network," https://www.sawater.com.au/news/continued-sensor-success-for-sa-waters-smart-network, accessed: 2020-07-09.

[3] BITRE, "Yearbook 2014: Australian infrastructure statistics report," 2014.

[4] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[5] P. Karkulali, H. Mishra, A. Ukil, and J. Dauwels, "Leak detection in gas distribution pipelines using acoustic impact monitoring," in *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2016, pp. 412–416.

[6] D. Koo, K. Piratla, and C. J. Matthews, "Towards sustainable water supply: schematic development of big data collection using internet of things (iot)," *Procedia engineering*, vol. 118, pp. 489–497, 2015.

[7] S. Lin, "Rank aggregation methods," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 5, pp. 555–570, 2010.

[8] T. Perumal, M. N. Sulaiman, and C. Y. Leong, "Internet of things (iot) enabled water monitoring system," in *2015 IEEE 4th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2015, pp. 86–87.

[9] L. Wong, R. Deo, S. Rathnayaka, B. Shannon, C. Zhang, J. Kodikara, W. Chiu, and H. Widyastuti, "Leak detection and quantification of leak size along water pipe using optical fibre sensors package," *Electron. J. Struct. Eng*, vol. 18, pp. 47–53, 2018.

[10] R. Xiao, Q. Hu, and J. Li, "Leak detection of gas pipelines using acoustic signals based on wavelet transform and support vector machine," *Measurement*, vol. 146, pp. 479–489, 2019.