# Rural Infrastructure Health Monitoring System: Using AI to Increase Rural Water Supply Reliability

**Heloïse Greeff[1], Achut Manandhar[1], Patrick Thomson[2], Rob Hope[2], David Clifton[1]**
[1]Department of Engineering Science, [2]School of Geography and the Environment
University of Oxford
`{heloise.greeff, achut.manandhar, david.clifton}@eng.ox.ac.uk`
`{patrick.thomson, robert.hope}@ouce.ox.ac.uk`

## Abstract

Remote health monitoring systems for rural infrastructure lack the advanced analysis offered by digital infrastructure due to operational constraints, such as limited data-transmission bandwidth and power supply, in these extreme locations. To overcome this, we present a novel approach of predicting failure (condition monitoring) in rural handpumps by distributing work between inexpensive systems embedded within the handpumps and a computationally-powerful cloud server that communicates with the handpumps [1]. Proof-of-concept work in rural Kenya correctly predict more than two thirds of failure events in handpumps where over 70,000 people in schools, communities, clinics and hospitals are benefiting from the implementation of this research. Performing anomaly detection at the rural node by applying a lightweight machine learning approach in the embedded system followed by more powerful machine learning algorithms in the cloud offer robust information without the need for expensive sensors embedded in situ – making the possibility of a large-scale rural monitoring system feasible.

## 1 Introduction: SDG 6.1 "safe drinking water for all"

According to the 2015 Sustainable Development Goal (SDG) baseline, 884 million people lack access to basic drinking water, rising to 2.1 billion without "safely managed" drinking water [2]. The ambition of the SDG target 6.1 of "safely-managed drinking water" by 2030 is unprecedented, particularly in rural areas where 4 out of 5 people live without access to drinking water. More than 200 million people in sub-Saharan Africa rely on the nearly one million rural handpumps to meet their daily water needs [3]. However, more than a third of these handpumps are broken at any time and often remain broken for up to 30 days [4].

Sustainable provision of reliable infrastructure hinges on the high standard of both installation and maintenance, the latter of which is often under invested in or completely neglected [5]. Downtime caused by system failure in rural settings can often last longer than in urban settings due to the practical challenges in the supply of spare parts combined with a lack of local skills. Novel remote monitoring systems have made it possible to track the operational and financial targets in rural water supply networks by sending timely data related to usage and payments [6], [7].

Predictive health monitoring is widely used in engineering applications to detect damage to infrastructure as early as possible to help reduce the downtime of systems, and, ideally, performing predictive maintenance can avoid downtime completely [8], [9]. A system which can virtually eliminate failure events for handpumps would generate sustained and significant impacts for the poor and be replicable across Asia and Africa, where this grand development challenge manifests in achieving "safe water for all".

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

## 2 System Requirements for AI-based Remote Health Monitoring

Transmitting raw observations from the sensing nodes in a large-scale sensor network can be costly and impractical as it requires large bandwidth, power, or both and in the case of rural monitoring can even be impossible. To overcome this limitation in the monitoring of rural infrastructure, we propose a "dynamic" approach whereby lightweight machine learning techniques, in the form of a logistic regressor (LR) novelty filter, be applied at the handpumps to perform initial processing of the raw observation prior to transmission [10]. The on-pump novelty threshold will define the proportion of original data to be transmitted based on the on-board assessment of the condition of the handpump.

The dynamic health-monitoring system should aim to: (i) use low-cost, embedded sensors to acquire accelerometry data from the routine daily use of handpumps, (ii) perform on-pump pre-processing of the acquired data, (iii) analyse the resulting data to produce informative assessments of the equipment health condition, (iv) communicate with a cloud server via the domestic telecommunications network, and (v) perform more advanced machine learning techniques on the cloud-based system to increase the fidelity of the classified health condition of the equipment [11]. The processing sequence for the proposed distributed inference system is shown in Figure 1 with three main sections:

**A. Sensor node** contains the sensor (accelerometer in the case of handpumps), battery, and data transmitter. A network could contain hundreds of nodes in a small geographical region.
**B. On-board novelty filter** performs real-time feature extraction of the data acquired during pumping and analyses the data using a LR novelty filter to produce data summaries that flag potential failure.
**C. Cloud server** performs more complex processing of the data summaries using advanced machine learning methods to increase prediction fidelity.

### 2.1 Lightweight On-pump Novelty Filter

Data is pre-processed on-board the system embedded in the handpump and consist of: peak and trough detection, high-pass filtering, windowing in the time-series domain [12], and transforming to the frequency domain [13]. Finally, it performs feature selection by sampling across frequency bins, discarding low frequency components that represent the pumping motion of the user. As a first layer of in situ health monitoring, the on-board classifier should be highly sensitive but not specific. Given the limited processing power (8-bit microprocessor) of the embedded system, logistic regression is used to calculate a novelty score by predicting the probability of membership of one class (e.g., Normal/Abnormal).

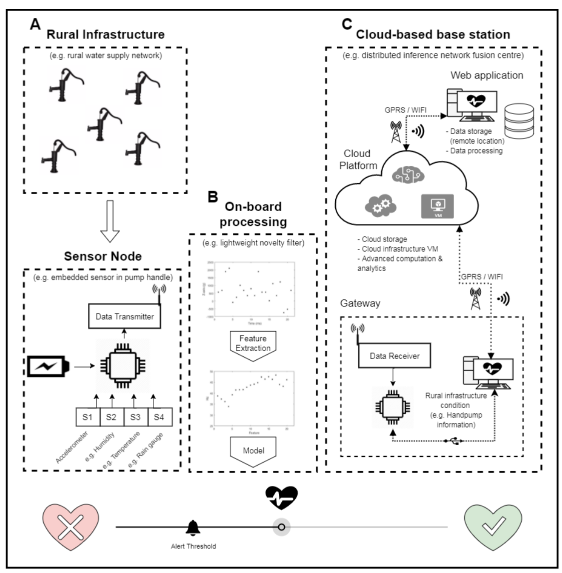



Figure 1: Rural infrastructure health monitoring system: (A) Sensor nodes with embedded system, (B) On-board novelty filter producing data summaries, and (C) Gateway node and cloud-based computing and analytics.

### 2.2 Heavyweight Cloud-based Classifier

The next stage of health monitoring involved performing more advanced machine learning processing on the sub sets of data flagged by the on-board novelty filter. The on-pump novelty filter is used to ensure that under normal operating conditions the vast majority of data is not transmitted and only as the on-pump model suspects the condition is degrading will data be transmitted to the cloud, which means that in most cases the cloud server will only be receiving higher-fidelity data related to abnormal conditions. Support Vector Machine (SVM) [14] and Random Forest (RF) [15] classifiers are applied to the data packages received by the centralised cloud server.

# 3 Results with Real World Data

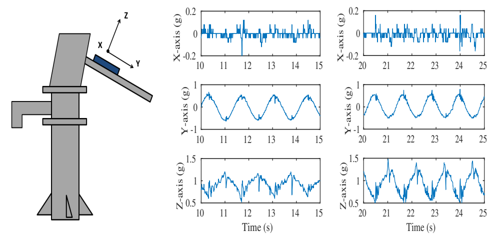
## 3.1 Data Collection and Data Sets

We consider two data sets from pumps in our study site in Kwale, Kenya. Recordings contain high-frequency (96 Hz) three axes accelerometery readings from a logger mounted inside the handle of handpumps, as shown in Fig. 2a. A 5 s interval of data from a handpump in a normal and abnormal condition is shown in Fig. 2b and 2c, respectively.

Typically, deep handpumps are the primary source of drinking water for nearby households, making timely repair even more crucial. Thus our main focus was deep wells, operating at depths greater than 25 m. Fig. 3 shows the difference in the spectra of deep and shallow handpumps.

The first data set, $\mathcal{D}_{d,1}$, represents a deep-operating inter-handpump system consisting of eight different handpumps operating at depths between 30 m to 55 m below ground level. The second data set, $\mathcal{D}_{d,2}$, represents a deep-operating intra-handpump system of one handpump operating at 54 m.

The data was labeled during in-person, contemporaneous observations and contained examples from eight different common handpump failures. All the data sets were randomly divided into a training-and-validation set (80%) and a test set (20%).

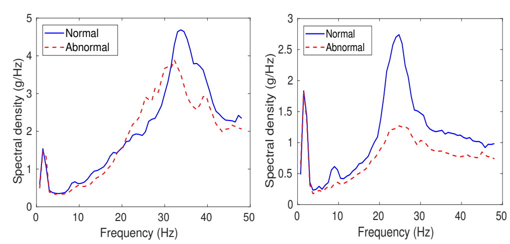



(a) (b) (c)

Figure 2: (a) Embedded system mounting and a 5 s interval of accelerometer data of the handpump in a (b) normal and (c) abnormal condition in the X, Y, and Z dimensions (upper to lower plots, respectively).

## 3.2 Evaluating Performance

The ability of the on-pump novelty filter and subsequent cloud-based classifier to verify reliability was assessed using the receiver operating characteristic (ROC) to compare the actual and predicted outputs for each class. The true positive rate (TPR), or *sensitivity*, is defined to be the probability of detection such that, $TPR = \frac{\sum TruePositive}{\sum ConditionPositive}$ and the false positive rate (FPR), or *fall-out*, is defined to be the probability of a false alarm such that, $FPR = \frac{\sum FalsePositive}{\sum ConditionNegative}$. Optimising the area under the ROC (AUC) will maximise handpump failure detection while minimising false alarms, which can be costly in real-life.




(a) (b)

Figure 3: Median amplitude of the spectral data for (a) deep and (b) shallow operating handpumps.

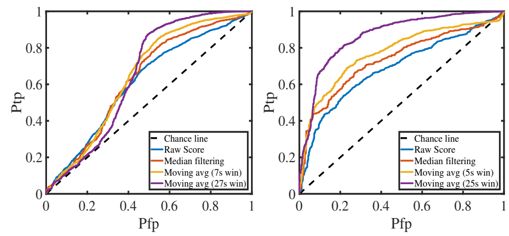
## 3.3 Performance of On-pump Novelty Filter

Once a robust model was trained, the algorithm is deployed in the embedded handpump systems, such that new vibration data resulting from new pumping generates an on-pump novelty score in real-time. The novelty score is a scalar number that increases with probability of the pump being "abnormal" with respect to the machine learning model.

Figure 4 compares the receiver operator curve (ROC) for a deep-operating (a) inter-handpump classifier, and (b) intra-handpump classifier with a broken rod (common failure). The intra-handpump classifier achieves an AUROC of 86.21% while the inter-handpump classifier correctly predict two thirds of failure events (65.7% AUROC). In both cases, the lab simulated results were a slight improvement of the real time performance, 89.01% and 69.80% respectively.

### 3.4 Performance of Cloud-based Classifier during Active Sampling

The next stage of condition monitoring involved performing more computationally advanced machine learning methods on the data flagged by the on-pump novelty filter. Figure 5 compares the AUROC for the cloud-based classifiers using the novelty scores identified by the on-pump novelty filter. In both cases, the variance of the predictive accuracy of the LR and SVM classifier is reduced as the test set size increased. Conversely, however, in all cases the RF classifier achieves the highest AUROC with a relatively small proportion data and does not benefit from additional data both in improving prediction accuracy or decreasing prediction variance.



(a) (b)

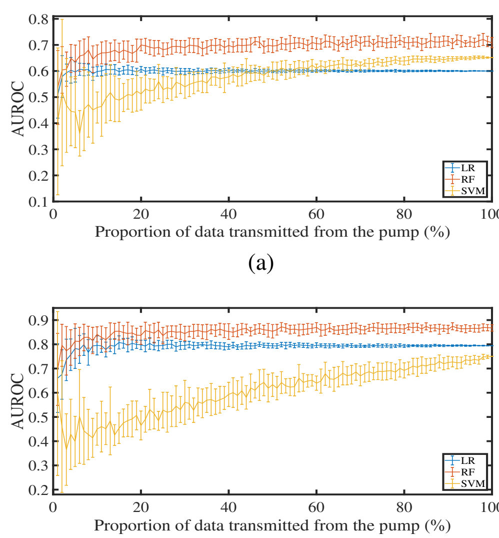
Figure 4: ROC of on-pump novelty filter classifier performance for a deep operating: (a) inter-handpump classifier trained using $\mathcal{D}_{d,1}$, and (b) intra-handpump classifier trained using $\mathcal{D}_{d,2}$.

Overall, additional machine learning methods applied to the data summaries from the on-pump novelty filter offer a 10 per cent improvement from the raw on-pump generated novelty scores. The two cases show that there is a trade-off between accuracy and specificity. Whilst the RF classifier may offer a higher overall prediction accuracy, both LR and SVM can dramatically reduce the variability in predictions as the proportion of data supplied is increased.

## 4 Discussion and Conclusion

As proof of concept, we used a LR classifier on the embedded system to produce a novelty score related to the condition of the pump. These scores are then sent to a cloud server where we showed that SVM and RF classifiers can be used to further increase the fidelity of the health monitoring system. Ongoing work has already started to replace these cloud-based classifiers with deep learning techniques.

This work demonstrates that a distributed inference health monitoring system for rural infrastructure offers a number of advantages over existing remote condition monitoring systems that are both energy and bandwidth heavy. Incorporating more advanced machine learning methods on a cloud-based platform have been shown to increase the system's overall positive predictive value by more than 10 per cent when "intelligent" subsets of flagged data from the rural node is transmitted.



(b)

Figure 5: AUROC comparison the cloud-based classifier using novelty filtered data subsets for: (a) a deep inter-handpump classifier trained using $\mathcal{D}_{d,1}$, and (b) a deep intra-handpump classifier trained using $\mathcal{D}_{d,2}$.

Importantly, this AI-based system uses a "dynamic" approach that distributes work between inexpensive systems embedded within the handpumps and a computationally-powerful cloud server that communicates with the handpumps. Proof-of-concept work in Kenya correctly predicts more than two thirds of failure events in handpumps. Such novel data are able to inform operational innovations that has reduced "downtime" from a month or more, to a guaranteed service of fewer than three days of outage [16]. Following the success of this pilot in Kenya, future work aims to apply transfer learning to translate this approach to different handpump types in Bangladesh.

## References


[1] H. Greeff, A. Manandhar, P. Thomson, R. Hope, and D. A. Clifton, "Distributed Inference Condition Monitoring System for Rural Infrastructure in the Developing World," IEEE Sensors Journal, [In Press].

[2] WHO/UNICEF, "Progress on Drinking Water, Sanitation and Hygiene: 2017 Update and SDG Baselines," Geneva, 2017.

[3] E. Baumann and S. G. Furey, "How Three Handpumps Revolutionised Rural Water Supplies: A Brief History of the India Mark II/III, Afridev and the Zimbabwe Bush Pump," Rural Water Supply Network, St. Gallen, Switzerland, 2013.

[4] E. Baumann, "May-day! May-day! Our handpumps are not working!," RWSN Perspective, vol. RWSN/A/200, no. 1, 2009.

[5] S. G. Furey, "RWSN Handpump Survey 2013: Summary of Findings," SKAT Foundation, St Gallen, Switzerland, 2013.

[6] P. Thomson, R. Hope, and T. Foster, "GSM-enabled remote monitoring of rural handpumps: a proof-of-concept study," J. Hydroinformatics, vol. 14, no. 4, p. 829, Oct. 2012.

[7] C. Nagel, J. Beach, C. Iribagiza, and E. A. Thomas, "Evaluating Cellular Instrumentation on Rural Handpumps to Improve Service Delivery-A Longitudinal Study in Rural Rwanda," Environ. Sci. Technol., vol. 49, no. 24, pp. 14292–14300, 2015.

[8] L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko, "Predictive Monitoring of Mobile Patients by Combining Observations With Data From Wearable Sensors," IEEE J. Biomed. Heal. Informatics, vol. 18, no. 3, pp. 722–730, 2014.

[9] D. A. Clifton and L. Tarassenko, "Novelty Detection in Jet Engine Vibration Spectra," Int. J. Cond. Monit., vol. 5, no. 2, pp. 2–7, 2015.

[10] V. V. Veeravalli and P. K. Varshney, "Distributed inference in wireless sensor," Phil. Trans. R. Soc., vol. A, no. 370, pp. 100–117, 2012.

[11] S. Funiak, C. Guestrin, and M. Paskin, "Distributed Inference in Dynamical Systems," in NIPS, 2006.

[12] A. Testa, D. Gallo, and R. Langella, "On the Processing of Harmonics and Interharmonics: Using HanningWindow in Standard Framework," IEEE Trans. Power Deliv., vol. 19, no. 1, pp. 28–34, 2004.

[13] R. G. Lyons, Understanding Digital Signal Processing, 3rd ed. Prentice Hall, Upper Saddle River, NJ, 2011.

[14] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," IEEE Intell. Syst., vol. 13, pp. 18–28, 1998.

[15] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[16] J. Koehler, S. Rayner, J. Katuva, P. Thomson, and R. Hope, "A cultural theory of drinking water risks, values and institutional change," Glob. Environ. Chang., vol. 50, pp. 268–277, 2018.