

A Predictive and Scalable Health-Monitoring System for Handpumps



Heloïse Greeff

Balliol College

University of Oxford

Supervised by

Professor David A. Clifton

Trinity Term 2020

This thesis is submitted to the Department of Engineering Science, University of Oxford,
in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Declaration

The work contained in this thesis is entirely my own and, except where otherwise indicated, describes my own research.

Heloïse Greeff

Foremost, praise, and thanks be to Almighty God, the most gracious and merciful.

—

तमसो मा ज्योतिर्गमय ।

Acknowledgements

I am grateful for so many people who helped me during my DPhil journey. Firstly, to the person without who I would never have returned to Oxford to pursue another degree, Prof. David Clifton. You delivered the most delicate balance of attentive supervision, enthusiasm and freedom to allow me to become an independent researcher. Thank you to Prof. Robert Hope and Patrick Thomson for their relentless pursuit to bring the Smart Water Group together and for the many cold beers after long (frustrating) days in Shimba Hills. Thank you to Dr Achut Manandhar for sharing your knowledge and your snacks, which you always produced next to the most remote handpump.

Much of the work described here would not be possible without the support and interesting discussions with the staff of Rural Focus. In particular, thanks go to Dola, Nanga, George, Idd and Okoti for patiently listening and translating my ideas into Swahili, but more importantly for safely getting me to unmapped handpumps. Suleiman and Rueben proved robust for my emails from Oxford. Most of all, thank you to the water users of Kwale County who always welcomed me at their handpumps.

To the members of CHI Lab, old and new, and in particular neighbours on the Time Series Terrace, many thanks for always welcoming me back after long periods away for field work.

This work was funded by UNICEF, and I would like to thank Peter Harvey and Jessica Tripe in particular for their part in arranging this unique and pioneering funding arrangement. To Upgro and REACH, for not only funding a handpump installation in Oxford but welcoming me into your community. To the Rhodes Trust and the members of the selection committee, in particular, Dr Francois Bonnici, thank you for believing in me long before I did.

My deepest love and appreciation goes to my family and friends. To my friends and family far away, who have always encouraged me, thank you for your texts and calls to keep me going. Thank you to my brother, Louis, for always reminding me to laugh. To Vicky, Mienieke, and Pepper, your companionship made many long nights bearable. Of course, to you, Mamma and Pappa, your love, sacrifices, and unwavering support over the years is the foundation of every small success. Thank you for all the cheers and gentle nudges, at times.

To Ravi, thank you for supporting the patient monitoring aspect of my work but also for being a supportive and loving partner. I am so grateful for your patient and kind soul.

Baie dankie.

A Predictive and Scalable Health-Monitoring System for Handpumps

Heloïse Greeff

Balliol College

Thesis submitted for degree of Doctor of Philosophy

Trinity Term 2020

Abstract

The handpump remains the most reliable and low-cost method to access groundwater, making it a critical component of rural water supply for around 200 million people in sub-Saharan Africa to meet their daily water needs. Despite this, an estimated one in four handpumps does not work at any given time. This is primarily due to a lack of low-cost and reliable remote monitoring methods for predicting the condition of the rural infrastructure. Existing handpump maintenance models rely on manual reporting from the water users or use flow meter data without further analysis or insights into unique usage patterns. The investigations in this thesis describe the development and evaluation of a dynamic distribute inference (DDI) system, a novel method whereby a novelty score is produced in-situ and “intelligent” data summaries are transmitted to a cloud-based station for further assessment to improve overall fidelity. This work involved developing, manufacturing, and installing the bespoke sensors for collecting the high-rate vibration data from handpumps to create three unique open-access data sets that were used to conduct this work. An in-depth literature review combined with extensive first-hand field experience and expert interviews informed the proposed standardised functionality categories. Five lightweight classifiers are considered to operate within the limited-resources available onboard the embedded sensor: logistic regression (LR), naïve Bayes classifier (NBC), decision trees (DT), k-nearest neighbours (kNN) and neural networks (NN). All of these were deemed good predictors of handpump failure; however, the probabilistic framework of the LR allows dynamic adjustment of the novelty threshold. As such, LR was used to produce in-situ novelty scores and implemented in the field to collect novelty filter scores produced on the embedded system along with raw waveforms from 20 different handpumps during multiple field visits as well as ten handpumps over 18 months. Based on the novelty score, the cloud may request additional data from the embedded sensor (such as features or raw waveforms over novelty scores) to perform a further assessment in order increase the fidelity of the health-monitoring system. The data summaries from the sensors were retrospectively assessed on the cloud using two supervised machine learning methods: support vector classifier (SVC) and random forest (RF) and compared to conventional machine learning methods: multilayer perceptron (MLP) and 1-D convolutional neural network (CNN). The novelty filter threshold is adjusted based on the desired false positive rate (FPR). Transferability of the DDI platform was shown by evaluating the predictive performance on both vibration data from handpumps as well as more complex physiological patient data. In both cases, the implementation of the DDI improved classifier accuracy by more than 10%, while significantly reducing the cost of data transmission. These results represent a significant advancement in the possibility of low-cost sensors combined machine learning to deliver scalable and reliable quantitative insights that can inform priority setting for improved rural services toward achieving the sustainable development goals (SDGs).

Publications

Journals and Conferences

- **Greeff, H.**, Quoc, K.P.N., Bich, H.H., Tu, V.H.M., Thwaites, L., Clifton, D.A., The Use of Artificial Intelligence and Wearable Sensors to Manage Infectious Disease During Global Health Emergencies. *IET Appropriate Healthcare Technologies (AHT)*, London, 2020, pp. 1-2.
- Manandhar, A., **Greeff, H.**, Thomson, P., Hope, R., and Clifton, D.A., Shallow Aquifer Monitoring Using Handpump Vibration Data. *Journal of Hydrology X* 8, 2020, pp. 100057.
- **Greeff, H.**, Manandhar, A., Thomson, P., Hope, R., and Clifton, D.A., The Use of Artificial Intelligence to Support Rural Water Development. *Alan Turing Institute: Data Science for Development*, London, UK, 2020. [Postponed due to COVID-19]
- **Greeff, H.**, Manandhar, A., Thomson, P., Hope, R., and Clifton, D.A., AI for Rural Development: Remote Monitoring of Handpumps using Machine Learning. *Westminster STEM for BRITAIN*, London, UK, 2020.
- Manandhar, A., **Greeff, H.**, Thomson, P., Hope, R., and Clifton, D.A., Groundwater Monitoring Using Handpump Vibration Data for Rural Africa. *2019 Conference on Neural Information Processing Systems, AI for Social Good Workshop*, Vancouver, Canada, 2019, pp. 1-4.
- **Greeff, H.**, Manandhar, A., Thomson, P., Hope, R., and Clifton, D.A., Rural Infrastructure Health Monitoring System: Using AI to Increase Rural Water Supply Reliability. *2018 Conference on Neural Information Processing Systems, AI for Social Good Workshop*, Montreal, Canada, 2018, pp. 1-4.
- **Greeff, H.**, Manandhar, A., Thomson, P., Hope, R., and Clifton, D.A., *Distributed Inference Condition Monitoring System for Rural Infrastructure in the Developing World*, *IEEE Sensors Journal* 19(5):1820-1828, 2018.
- Thomson, P., Manandhar, A., **Greeff, H.**, Machine Learning for Aquifer Monitoring using Handpump Vibration Data, *45th International Association of Hydrologist Congress*, Seoul, South Korea, 2018.
- Colchester, F.E., **Greeff (Marais), H.**, Thomson, P., Hope, R., and Clifton, D.A.,

Accidental Infrastructure for Groundwater Monitoring in Africa, *Environmental Modelling and Software* 91: 241-250, 2017.

Patents

- **Greeff, H.**, Manandhar, A., Thomson, P., Hope, R., and Clifton, D.A.: Method and System for Monitoring a Remote System, GB 1819717.8, 2018, (Distributed Inference).

Published Data Set

- **H. Greeff**, A. Manandhar, and P. Thomson. Daily handpump accelerometer data and borehole water level data, Kwale County, Kenya (Dataset). British Geological Survey. 2020.

Related Awards

- Smart Handpump, **Oxford Vice-Chancellor's Innovation Award**, Joint Overall Winner and Building Capacity Category, 2018.

Abbreviations

Hydrogeology	
AFRIDEV	Africa Development (handpump type)
CBM	Community based management
GL	Ground level
PVC	polyvinyl chloride
VLOM	Village level operations and management
WDT	Water data transmitter
WRMA	Weather Risk Management Association
Clinical	
BP	blood pressure
BR	breating rate
DBP	diastolic blood pressure
HR	heart rate
ICU	intensice care unit
MET	Medical Emergency Team
NHS	National Health Services
RPM	Remote patient monitoring
SBP	systolic blood pressure
SDU	Step-down Unit
SpO ₂	blood oxygen saturation
UPMC	University of Pittsburgh Medical Centre
General	
GPS	Global Positioning System
GSM	Global System for Mobile communications
IC	Integrated circuit
ICT	Information and Communication Technology
IoT	Internet of Things
JMP	Joint Monitoring Programme
LMICs	low-and-middle income countries
MB	megabytes
M2M	Machine to machine
SDG	Sustainable development goals
SMS	short message service
UN	United Nations
UNICEF	United Nations International Children’s Emergency Fund
Novel abbreviations	
DI	Distributed Inference
DDI	Dynamic Distributed Inference

Condition monitoring	
FAR	false alarm rate
MAR	missed alarm rate
MEMS	micro-electromechanical systems
POD	Probability of detection
RUL	remaining useful life
SHM	Structural Health Monitoring
Methods	
AUC	Area under the (ROC) curve
CA	Classifier accuracy
CART	Classification and regression trees
CI	Confidence interval
CV	Cross-validation
CNN	Convolutional neural network
DIN	Distributed inference network
DT	Decision Tree
EWS	Early warning score
FFT	Fast Fourier transform
FIR	Finite impulse response
FN	false negative
FNR	False negative rate
FP	false positive
FPR	False positive rate
GP	Gaussian Processes
HPF	High-pass filter
IQR	interquartile range
KDE	kernel density estimates
kNN	k-nearest neighbours
LR	Logistic regression
LSTM	Long Short Term Memory
MAP	Maximum a posteriori
MLE	Maximum likelihood estimation
MLP	Multilayer Perceptron
NBC	naïve Bayes Classifier
NF	Novelty Filter
NLL	negative log-likelihood
NN	Neural Network
PC	Principle component
PCA	principal component analysis
PDF	Probability distribution functions
RBF	radial basis function
RF	Random forest
ROC	receiver-operating characteristic
SD	Standard deviation
SE	Standard error
SNR	signal-to-noise ratio
SSA	Singular spectral analysis
SVM	Support vector machine
TP	True positive
TF	True negative

Ubuntu: [ùbùntú]

Umntu Ngumuntu Ngabantu —“I am, because you are”

Contents

Abbreviations	vii
List of Figures	xiv
List of Tables	xix
List of Symbols	xxii
1 Introduction	1
1.1 Importance of Rural Water Point Monitoring	2
1.2 The Rural Water Challenge	4
1.3 The Challenge of Handpump Monitoring	6
1.4 The Future of Remote Health-Monitoring	7
1.5 Outline of Thesis	7
2 Rural Water Supply	9
2.1 Water Supply in Rural Kenya	9
2.2 Handpump Introduction	12
2.2.1 Types of Handpumps	12
2.2.2 Piston Handpump Operation	15
2.3 Conventional Handpump Maintenance Systems	17
2.4 Handpump Functionality and Continuity	19
2.4.1 Defining Functionality and Labelling Data	21
2.5 Preventable and Avoidable Failure Events	21
2.5.1 Failure Types	22
2.6 Remote Health-Monitoring Systems	29
2.6.1 Challenges of Remote Health-Monitoring	31
2.7 Conclusion	31

3	Literature Review	33
3.1	Our Previous and Related Work	34
3.2	Other Related Work	35
3.3	Context and Frameworks of Health-Monitoring Models	37
3.4	Vibration-Based Structural Health Monitoring	40
3.4.1	Level of Identification	41
3.4.2	Monitoring Using Accelerometry	41
3.4.3	Prognosis	43
3.4.4	Failure Alarm Threshold	44
3.5	Feature Extraction	46
3.5.1	Hand-crafted Features	47
3.6	Predictive Inference	48
3.6.1	Outcome Classification Framework	48
3.6.2	Clustering	49
3.7	Distributed Measurement Systems	51
3.8	Conclusion	52
4	Data Sets	53
4.1	Data Collection	54
4.1.1	Measuring Handpump Vibrations	54
4.1.2	Experimental Setting	56
4.1.3	Fault Experiments	57
4.2	Data Labelling	58
4.3	Kenya Original Database	59
4.4	Kwale Failures Database	60
4.5	Kwale Long-Term Database	62
4.6	Conclusion	63
5	Sensor-based Novelty Filtering	65
5.1	Lightweight Classifiers	66
5.1.1	Eager Learners	66
5.1.2	Lazy Learners	71
5.1.3	Classifier Comparison	72

5.2	Heavyweight Classifier	73
5.2.1	Support Vector Machines	73
5.3	Method	75
5.3.1	Onboard Pre-processing	75
5.3.2	Embedded System Considerations	79
5.3.3	Train-Test	81
5.3.4	Model Evaluation	81
5.4	Results	85
5.4.1	Visualisation of Data	85
5.4.2	Analysis of Spectral Data	88
5.4.3	Accelerometer Axis Selection	96
5.4.4	Feature Selection	97
5.4.5	Classifier Performance	104
5.5	Discussion	108
5.6	Conclusion	110
6	Cloud-based Dynamic Distributed Inference System	111
6.1	From Sensor to Cloud	112
6.1.1	Distributed Inference Model	114
6.1.2	System Requirements for Rural Monitoring	115
6.1.3	Intra- vs Inter-node Monitoring	118
6.2	Heavyweight Cloud-based Classifiers	118
6.2.1	Random Forests	119
6.2.2	Multilayer Perceptron	121
6.2.3	Convolutional Neural Networks	122
6.3	Assessment of Cloud-based Classifier	123
6.3.1	Method	123
6.3.2	Results	125
6.4	Dynamic Distributed Inference Platform Implementation	141
6.4.1	Method	141
6.4.2	Results	147
6.5	Discussion	158
6.6	Conclusion	162

7	From Handpump to Human	164
7.1	Remote Patient Monitoring	164
7.2	Early Detection of Deterioration	166
7.3	Data Description	168
7.3.1	Labelling Critical Events	169
7.3.2	Vital Sign Measurements	170
7.3.3	Description of Critical Event	173
7.4	Techniques	176
7.4.1	Kernel Density Estimator	176
7.4.2	Gaussian Processes Classification	178
7.5	Method	180
7.5.1	Pre-processing	180
7.5.2	Parameter Normalisation	181
7.5.3	Labelling	181
7.5.4	Feature Extraction	182
7.5.5	Train-Test	182
7.5.6	Novelty Filter Score Estimation	183
7.5.7	Central- or Remote-Monitoring Training Framework	184
7.5.8	Evaluation	184
7.6	Results	185
7.6.1	Labelling and Feature Selection	185
7.6.2	Performance of the Patient-based Novelty Filter	187
7.6.3	DDI Platform Performance	188
7.7	Conclusion	192
8	Conclusion and Future Work	194
8.1	Summary of the Thesis	196
8.1.1	Remote Health Monitoring	196
8.1.2	Distributed Inference	198
8.2	Future Work	199
8.2.1	Deployment for Impact	200
8.2.2	Prognosis for Planning	202

8.2.3 Clinical Decision Support with Wearables for Limited-Resource Set- tings	203
8.3 Conclusion	204
Bibliography	206

Figures

2.1	Map of study site in Kwale, Kenya, and locations of the original large-scale Afridev handpump deployment between 1983 and 1995. The matched water points (blue) indicate 471 of the original 580 found during a 2013 census. Unmatched water points (yellow) indicate additional handpumps installed later, likely by private donors or NGOs. (Source: [42])	11
2.2	Overview of handpump types as classified by water-lift method (displacement or direct) including VLOM capability. Standardised VLOM handpumps include the shallow well suction Tara pump (operating up to 10m) and the deep well lift pumps, India MK III and Afridev (operating over 10m). The Afridev is commonly used in Kenya.	13
2.3	Illustration of the standard assembly of the VLOM Afridev handpump. Pump rods connect the plunger to the handle, which moves in opposite directions. Positive displacement in the cylinder moves water up the rising main and out of the spout above the ground level (Source:[45]).	16
2.4	Schematic of the conceptual model of the factors affecting handpump functionality which illustrates the interaction between factors related to the failure (red) and the repair (green) (Source:[6]).	22
3.1	Updated categories of innovation areas in information and communication technology based on the flow of data/information (Adapted from [103, 104]). Data analysis is a crucial step currently omitted in the conventional 5-step data information flow for ICT.	36
3.2	Visualisation of the maintenance flow over the lifespan of rural infrastructure. The infrastructure, like a handpump, is either installed as routine development or as a response to an emergency. This is an iterative process aimed to maximise the operational time of the infrastructure.	39
3.3	A segment of a typical vibration level, here representing one failure event. The vertical and horizontal axes correspond to vibration level and time, respectively, in arbitrary units.	42
3.4	Two-stage failure delay time concept [129] attempts to separate normal conditions from failures by introducing the time delay, h , between the point, u , when the system becomes defective and when the failure occurs.	44
4.1	(a) A diagram showing the experimental set up of a vibration sensor attached to a surface-mounted handpump and the relevant orthogonal axes of accelerometry with a five-second interval of unprocessed accelerometer data of the same pump in a (b) normal and (c) abnormal condition in the X, Y, and Z dimensions (upper to lower plots, respectively) for the same user.	55

4.2	Example installation of a data logger with a specific sensor-based novelty filter version. The final placement is secured with bolts inside the handle. . .	55
4.3	Behind the scenes snapshots of the various handpump locations we visited and the remote villages in Kwale County where we lived for many months during the course of this thesis in order to ensure we collect the required data. Handpump locations are mapped here: https://goo.gl/Vo0eFx . . .	56
4.4	Overview of the representation of different handpumps in the data set grouped by: (a) condition, where C0 is abnormal and C1 is normal, and (b) depth of the well, where shallow is $< 25\text{m}$ and deep is $> 25\text{m}$. Each example is 128 samples, equivalent to 1.3 seconds of time-series data. . . .	62
5.1	A six-second interval of the final output of the pre-processing steps for the Y-axis accelerometry data.	78
5.2	A flow diagram of the classifier's validation process shows that both train and test sets are balanced before the test data evaluate the classifier's performance to identify the handpump as either "normal" or "abnormal". . . .	82
5.3	Sammon's mapping projection of 128-D feature vectors for handpump vibrations on the Y-axis. Clusters show different handpump conditions: (i) normal (dark blue), (ii) dry borehole (green), (iii) leak in the rising main (orange), and (iv) broken U-seal (light blue).	87
5.4	Projection of 128-D feature vectors for handpump vibrations on the Y-axis grouped by individual pump across all normal recordings in this data set. In each plot, the feature of a different individual pump is highlighted in red.	88
5.5	Projection of 128-D feature vectors for handpump vibrations on the Y-axis grouped by the depth of the well according to the number of rods for deep (orange) and shallow (blue) operating handpumps, respectively.	90
5.6	Spectra (in dB) for normal pumping conditions for a (a) shallow well and (b) deep well, and abnormal pumping conditions for a (c) shallow well and (d) deep well, respectively. The median amplitude of the spectral data are shown in (e)-(f) for shallow and deep, respectively. Distinct high-frequency components appear to take high energy for normal handpumps operating in deep and shallow wells.	92
5.7	Histograms of pumping speed for (a) normal (blue) and abnormal (red) handpump conditions and (b) sub-grouped according to varying water flows. The histograms are approximately pdfs as they integrate to unity. Each set of histograms is fitted with the maximum likelihood of gamma distributions.	94
5.8	A speed-based representation of spectral data for normal pumping conditions for (a) a shallow well and (b) a deep well, and abnormal pumping conditions for (c) a shallow well and (d) a deep well, respectively. The speed-frequency space alone does not provide sufficient evidence to support whether speed impacts the pumping characteristics.	95

5.9	Probability of pumping speed as a percentage of the maximum for varying the depth at which we classify the well, defined by the number of rods in the system, for shallow (blue) and deep (yellow), respectively.	96
5.10	The ROC plot comparing the proposed LR classifier's performance for different axes of the accelerometer shows that combining all three accelerometer axes does not provide additional information. Instead, a single axis can provide acceptable performance while preserving processing power.	98
5.11	Projection of 20-D feature vectors for handpump vibrations on the Y-axis: (a) FV_1 ; (b) FV_2 ; (c) FV_3 ; (d) FV_4 ; (e) FV_5 ; and (f) FV_6 . Labels are shown in Table 5.3.	102
5.12	Projection of 20-D feature vectors for handpump vibrations on the Y-axis: (a) FV_1 ; (b) FV_2 ; (c) FV_3 ; (d) FV_4 ; (e) FV_5 ; and (f) FV_6 . Labels are shown in Table 5.4.	103
5.13	ROC plot comparing the performance of the LR classifier for the proposed feature vectors, $FV_1 - FV_6$	105
5.14	ROC plot comparing the performance of the LR classifier for FV_4 grouped by depth for deep and shallow wells combined (blue), only deep wells (orange) and only shallow wells (yellow).	107
5.15	Predicted (black :) and actual (blue) outputs of the LR classifier for FV_4 for (a) mixed depth and (b) grouped by depth, respectively. Note that there is no ordering of data. The horizontal axis shows the index of feature vectors, which are independent and which are not consecutive in time.	108
6.1	Schematic diagram of the proposed DDI health monitoring system for a rural water supply network consisting of: (A) Handpump sensor node with related embedded sensors, (B) Onboard lightweight novelty filter producing data summaries, and (C) Gateway node and heavyweight cloud-based computing and analytics.	117
6.2	Lower dimensional PCA representation of Y-features in 2D for (a) a single deep operating handpump $\mathcal{D}_{d,2}$, (b) deep and shallow operating handpumps $\mathcal{D}_m$, (c) deep operating handpumps $\mathcal{D}_{d,1}$, and (d) the same group of deep operating handpumps $\mathcal{D}_{d,1}$ in 3D. The blue and red examples correspond to normal and abnormal handpump operations, respectively.	126
6.3	Comparison of the ROCs for novelty scores generated by the sensor-based novelty filter (blue) and novelty scores generated in the lab using the relevant raw data (red).	127
6.4	A comparison of the classifier performance for varying three different classifiers V1.1-1.3 in: (a) inter-handpump and (b) intra-handpump settings (left to right plots, respectively).	129
6.5	A comparison of the on-pump classifier performance for failures in inter-handpump systems: (a) a depth-specific classifier, $\mathcal{D}_{d,1}$, and (b) a mixed-depth classifier, $\mathcal{D}_m$. Novelty scores generated on the handpump (raw scores) are shown in addition to median filtered and moving averages. . . .	130

6.6	Probability density of on-pump novelty scores for normal (blue) and abnormal (red) handpump conditions in: (a) deep operating inter-handpump system, $\mathcal{D}_{d,1}$, (b) deep and shallow operating inter-handpump system, $\mathcal{D}_m$, and (c) deep operating intra-handpump system, $\mathcal{D}_{d,2}$	132
6.7	A comparison of the system-based novelty filter performance for different failures represented in the intra-handpump system: (a) broken pump rod and (b) leak in the rising main.	133
6.8	A comparison of different classifier performance given varying proportions of data for: (a) non-depth-specific inter-handpump case ($\mathcal{D}_m$), (b) deep-specific inter-handpump case ($\mathcal{D}_{d,1}$) and (c) deep-specific intra-handpump case ($\mathcal{D}_{d,2}$).	137
6.9	A comparison of classifier training and prediction times over a varying proportion of data for: (a & d) deep inter-handpump system, $\mathcal{D}_{d,1}$, (b & e) mixed inter-handpump system, $\mathcal{D}_m$, and (c & f) deep intra-handpump system, $\mathcal{D}_{d,2}$, (top and bottom, respectively).	138
6.10	A comparison of the classifier performance for varying number of features on two deep-operating models: (a) inter-handpump, $\mathcal{D}_{d,1}$, and (b) intra-handpump, $\mathcal{D}_{d,2}$, (left to right plots, respectively).	139
6.11	An overview of the proposed dynamic distributed inference system's parameter outputs for a single simulated iteration with 2000 samples. The simulated novelty filter scores from the LR classifier (shown in a) feed into the DDI system. For each sample, each subplot shows the corresponding output: (a) LR classifier output, (b) count of historical handpump state for a given failure event (indicated by the vertical dashed line), (c) predicted and true condition of the handpump, and (d) the type of data requested for the next sample based on the predicted state.	149
6.12	Pdf of the loss function for three different novelty filter boundaries such that $T_{B,l}$ (dashed black), $T_{B,m}$ (dotted red) and $T_{B,z}$ (solid blue), respectively.	150
6.13	Pdf of time to detect a failure event when adjusting the novelty filter threshold for a given boundary: (a) $T_{B,l}$, $T_{B,m}$ and $T_{B,z}$, respectively. A missed event is independently counted and indicated in red on the right of each pdf figure.	151
6.14	Performance of the early warning system given simulated data to investigate the change in threshold on the FPR. Mean, confidence interval (CI) and interquartile range (IQR) are calculated over 10,000 random iterations. The results show: (a) the FPR for a given threshold (as a percentage) and the subsequent proportion of missed events; (b) FPR for a given warning window including missed events, and (c) excluding missed events. In (b) and (c) the 95% CI (left) and IQR (right) are shown, respectively.	152
6.15	Comparison of the performance and cost of operating the proposed DDI system and traditional non-dynamic system transmitting only the novelty filter scores (dashed) or raw waveforms (dotted) from the handpump the central cloud.	154

6.16	Performance of the early warning alert by the DDI system along with the associated cost (measure in amount of data being transmitted) using real failure events from $\mathcal{D}_l$, given two different failure types from the same hand-pump.	157
6.17	Case using real data of true failure event: (a) scores generated by the system-based novelty filter and (b) performance of the early warning alert by the DDI system along with the associated cost (measure in amount of data being transmitted).	158
7.1	Vital sign measurements for a patient with C"-events and a patient without, top and bottom respectively.	173
7.2	Distribution of all the vital sign measurements for (a) all patients, (b) non-C"-patients, and (c) C"-patients.	174
7.3	Distribution of the total monitoring duration for (a) C"-patients, and (b) Non-C"-patients as a factor of the length of their stay in the SDU.	175
7.4	Distribution of the duration of C"-events for (a) total duration of the event, (b) amount of monitoring before the first event, and (c) the amount of monitoring after the first event.	175
7.5	Timeline for labelling the data as normal and abnormal for a given window size, W_1 and W_2 , given the patient's first C"-event.	182
7.6	Low-cost wearable sensors used for collecting mobile physiological data from patients in limited-resource settings. PPG sensor with ear clip and storage-battery unit (left) and PPG sensor with a finger probe and display unit (connected to tablet or smartphone via Bluetooth) (right).	183
7.7	Results for misclassification of the test set when varying the size of the post-C" window, W_2	185
7.8	Comparison of the misclassification rate for different feature vectors, FV_1 to FV_7	186
7.9	Projection of 25-D feature vector, FV_5 , for vital sign measurements for all C"- (red triangle) and non-C"-patients (black circle), respectively.	187
7.10	Example of the raw vital sign measurements (top) and associated features (bottom) for a C"-patient.	188
7.11	Comparison of the performance and cost of operating the DDI platform (dotted) to detect C"-events in patients over non-dynamic systems transmitting only novelty filter scores (solid) and features (dashed), respectively.	190
7.12	Two cases using individual patient data with C"-event to perform the early warning alert by the DDI system along with the associated cost (measured in amount of data being transmitted).	191
8.1	PC projection of different pump types: (1) Afridev (blue $\circ$), (2) Samrat (red $\times$), and (3) India MKII (black $\diamond$).	200

Tables

2.1	Technical data sheet comparing the three VLOM pumps considered in this study, for current and future work. Shallow operating handpumps offer greater discharge and typically serve fewer people as these tend to be more densely installed, such as in Bangladesh [43, 44, 48].	14
2.2	A qualitative framework for defining handpump functionality proposed by the Hidden Crisis Project addresses the challenges in identifying and monitoring long-term handpump reliability [56]. The tiered performance levels measure functionality as a working water point that yields at least 10L/min and experienced fewer than 30 day downtime in the last year.	20
2.3	Common handpump failures in deep well lift pumps (Afridev) [23, 2, 63, 64, 65].	26
3.1	Related ICT projects in operation and maintenance of rural water supply networks.	36
3.2	Relevant related work for vibration-based health monitoring systems applying different inference models.	50
4.1	Description of the mechanical condition and short-term water quantity classification labels assigned to the recordings of each pump, where condition one represents a functional or normal handpump and condition zero refers to a non-functional or abnormal handpump. Water flow refers to a qualitative interpretation of the handpump condition.	59
4.2	Overview of the three proposed handpump vibration data sets used in this thesis.	61
5.1	Comparison of the AUC and standard deviation for LR classifier output for the three accelerometer axes (Y, X, and Z), individually and combined. . .	98
5.2	Feature vectors $FV = \{FV_j\}$ extracted from the frequency domain of the data set D_1 . The notation $[i : j : k]$ describes a regularly shaped vector such that $[j, j + i, j + 2i, \dots, j + m \times i]$, where $m = 20 = \frac{k-j}{i}$	99
5.3	PCA labels within visualisation of normal and abnormal handpump conditions in Figure 5.11.	102
5.4	PCA labels within visualisation of normal and abnormal handpump conditions for different operating depth: (a) deep (46 m) and (b) shallow (15 m).	103
5.5	Comparison of LR and SVC classifier performance metrics [median (SD)] on a general data set, D_1 , for different feature vectors, $FV_0 - FV_6$, respectively. 106	

5.6	Comparison of classifier accuracy scores for subsets of deep, D_{deep} , and shallow, $D_{shallow}$, operating handpump groups. Classifier accuracy (CA) and AUC are represented by mean and standard deviation.	106
5.7	Performance metrics for the five different classification models proposed in this chapter. Results are presented with mean and standard deviation of 1000 iterations.	108
6.1	Estimated power consumption of essential components in proposed embedded system operating at 25°C.	116
6.2	Overview of the different sensor-based novelty filter versions that were deployed in the field during the course of this research.	128
6.3	Results for field-based, $Q_{n,i}$, and lab-simulated, $Q_{n,lab}$, sensor-based novelty filter classification performance, given the mean AUC of 20 iterations (one standard deviation).	131
6.4	Cloud-based condition classification using 100% of sensor-based processed data, given the mean AUC (one standard deviation) of 100 iterations. . . .	135
6.5	Results for lab simulated cloud-based condition classification of sensor-based processed data, given the mean AUC of 100 iterations (one standard deviation). Results show the maximum AUC achievable for a given proportion of data.	136
6.6	Performance evaluation of the proposed distributed inference method compared to other standard classifiers methods. Results show that the DDI method offers overall improved accuracy and precision.	140
6.7	Fault development states of active sampling in examined handpump conditions. The types of data to be transmitted from the handpump sensor to the cloud are: (1) the novelty filter (NF) scores generated by the embedded LR classifier,(2) the features used by the LR classifier, or (3) the raw accelerometry waveforms. $P(C_i)$ is the probability that the state is in a given condition (C_i).	146
6.8	Proposed sampling frequency and type of data to transmit for the various monitoring levels of a health-monitoring system given the observed condition of the handpump's health status.	147
6.9	Performance (AUC $\pm$ SD) of a RF classifier for different data types transmitted from the rural sensor to the cloud classifier. Median of AUC results of 1000 iterations are reported with one standard deviation.	153
7.1	Density of nurses and rural population in selected countries. (Source: [221, 222])	166
7.2	Clinical criteria for MET activation at UPMC Presbyterian Hospital [238]. .	170
7.3	Threshold for MET alarm and artefact rejection thresholds applied to patient data.	171

7.4	Count of patient files with less than one hour of data broken down by vital sign.	172
7.5	Count of the primary cause of the first C"-event.	173
7.6	Mean, μ_t , and standard deviation, σ_t , for each patient vital sign in the training data.	181
7.7	Proposed feature vectors, FV_1 to FV_7 , determined by the window, W_f , and step, S_f , combinations.	186
7.8	Performance metrics for the seven different classification models proposed to be embedded onboard the wearable device. Results are presented with mean and standard deviation for 20 iterations.	189
7.9	Performance evaluation of the proposed distributed inference method compared to other standard methods. Standard deviation calculated for 100 iterations.	189

Symbols

The following list contains several notations and symbols commonly used in the literature and in this thesis.

η	Signal noise
λ	LR regularisation parameter
$\mathcal{D}$	Data set
μ	Mean
σ	Standard deviation
θ	Hyperparameters
C	SVM soft margin parameter
N	Number of observations
T	Novelty filter decision threshold
$\hat{y}$	Predicted output variables
$\mathbf{g}$	Gradient
$\mathbf{w}$	Weights
$\mathbf{X}$	Input variables
$\mathbf{y}$	Output variables
$\{\dots\}$	Set contents
$p(\mathbf{x})$	Probability density
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and (co-)variance Σ

Chapter 1

Introduction

Nipe maji, nipe maji ya uzima!

— Give me water, give me the
water of life!

Our Lady Of Fatima Kongowea

This thesis addresses how to use machine learning techniques in a frugal context to monitor the condition of equipment remotely. We shall demonstrate this for the case of water handpumps commonly used in rural water supply networks to extract groundwater. These equipment are suitable for this use case since it operates in limited-resource settings, deteriorates over time, and necessitate timely alerts since its failure can severely impact the communities who rely on the handpumps. We hypothesise that it is possible to distinguish normal handpumps from abnormal handpumps, with more than 80% precision, using machine learning techniques combined with suitable hardware. This problem can be broken into four key components to deliver a predictive and scalable health-monitoring system for handpumps:

1. Use low-cost, embedded sensors to acquire accelerometry data from routine daily use of handpumps and perform in-situ signal condition monitoring and pre-processing of the acquired data;
2. Use machine learning to analyse the resulting data and produce an informative

- assessment of the equipment health condition with at least 70% sensitivity;
3. Communicate with a cloud-based monitoring centre via the domestic telecommunications network to perform more advanced machine learning techniques on the cloud-based system that may increase the accuracy and certainty of the classified health condition of the equipment; and
 4. Demonstrate that the proposed system can identify the imminent failure in advance and is cost-effective such that it may scale across entire regions.

1.1 Importance of Rural Water Point Monitoring

Structural health-monitoring of rural infrastructures, such as manual handpumps, plays a vital role to support functioning and developing communities by meeting their needs to sustain productive social and economic activities. Reliable infrastructure is not only essential to improving the quality of life for rural communities but also a crucial aspect to drive economic growth, particularly in developing countries where rural areas remain mostly underdeveloped [1].

Despite what some may consider being outdated technology, the humble handpump [2] continues to be a very critical component of the rural water supply network and is used by nearly 273 million people in sub-Saharan Africa to access potable water daily [3, 4]. In reality, studies indicate that about 25% of the one million community-managed handpumps across rural sub-Saharan Africa are non-functioning at any given time [5, 6]. Although this is up from 36% a decade ago [7], it means that many counted as “served” are still left to rely on faulty or abandoned systems. A broken handpump in a remote village can force women or girls, who tend to be responsible for water collection in 8 out of 10 households [8], to walk up to 20 kilometres to find alternative sources, which may be

contaminated or expensive. Of those handpumps that break, 70% are repaired within one week [6], however, many remain broken for more extended periods, even up to 30 days; leaving entire communities without access to safe water for prolonged periods and forcing them to rely on other contaminated surface water sources.

Community management of rural handpumps attempted to address the lack of proper maintenance models and improve infrastructure sustainability [9, 10, 11, 12]. However, for economically poor communities, who rely on handpumps for their livelihood, even a small amount of money required at an unexpected time can be devastating given other budgetary responsibilities. The alternative is unsafe water sources, such as contaminated surface water or open wells. Water-borne infectious diseases, resulting from using unsafe water sources, annually result in 1.4 million child deaths from diarrhoea, 500 000 deaths from malaria and a near 5 million people seriously incapacitated by lymphatic filariasis [13]. In addition to the impact on local communities, these failures represent a sizable financial impact with an estimated investment of over USD 1.2 billion in the last two decades [14].

The perpetual water crisis in Africa is a result of both economic factors and physical water scarcity [15]. Although many African governments recognise that access to clean water is a fundamental human right, responding to the issue is not always their primary concern given the multitude of other development-related challenges they face. Improved water monitoring and management should improve both physical and economic water scarcity [16]. Access to handpump data leads to increased access to water for rural communities [17]. The use of predictive maintenance (whereby faults can be fixed in advance of handpump failure) has the potential to limit the substantial interruptions of weeks or more which are common across the rural water supply network [18]. In addition to enabling

faster repairs, harnessing mobile technology to monitor handpump usage also facilitate novel management models with transparent financial systems and contractual oversight which benefit the water user [19]. Reliability and sustainability of water supplies are essential to ensure healthy communities, societies, and economies in all regions of the world. Rural health-monitoring systems facilitate continuous feedback mechanisms in previously untapped rural water development programs that may stimulate stakeholder engagement and improve responsiveness. This increased accountability may further drive broader co-operation to ensure success in attaining poverty reduction and rural development goals.

1.2 The Rural Water Challenge

The new Sustainable Development Goals (SDGs) set the framework for a post-2015 global development agenda dedicated to water and proposed the ambitious global goal for “*securing sustainable water for all by 2030*” [20]. For the first time in 2016, the SDGs specified water security as one of the most pressing universal goals of our time, acknowledging that water crosses many other social sectors [21]. In particular, water scarcity in Africa relates to multiple aspects of human development that affect health, economics, agriculture, education, peace, and stability on the continent.

Although 89% of the overall global population has access to a basic drinking water service according to the 2015 SDG baseline, when broken down according to region, the access is severely skewed between developing and developed countries. Despite significant progress to increase access since the WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene (JMP) started in 1990, sub-Saharan Africa remains the region with the lowest access given that only 58% of the continent’s population has at least a basic water service [8]. The 2012 JMP reported that the coverage of proportion of the rural population in Africa with access to safely managed water sources improved by

just 10% since 1990, while the absolute number of increased from 37 million to 228 million people without access.

In rural areas where safely managed piped water schemes are limited or often non-existent, groundwater offers a reliable and safe alternative to contaminated surface water. An estimated 2 billion people worldwide still depend on groundwater to meet their daily water needs [22]. Since the early 1980s [23], the handpump has been the most widely-used technology to access groundwater, making it a fundamental component of the rural water supply network in Africa. However, given that one in four handpumps are non-functional at any given time, the return on the USD 1.2-1.5 billion infrastructure investment over the last two decades is uncertain [7].

When properly installed and maintained, a handpump can remain in a serviceable condition for up to 25 years [24]. Despite the widely known importance of infrastructure maintenance, it is often left to unskilled communities to perform routine and emergency repairs. The maintenance they perform is untimely or incorrectly executed, which can compromise the long-term structural integrity of the handpump and drastically reduce its remaining life. This is mainly due to three key factors:

1. **Lack of Finances:** Community management of handpumps shift the responsibility of maintenance to local communities without considering whether they have the organisational capacity to finance the necessary long-term support [12]. In reality, the majority of rural communities are unable to source or even sustain revenue collection through water committees [25].
2. **Lack of Equipment and Supplies:** The success of the supply chains for handpump spare parts compromises a lack of management and non-standardised procurement

policies [26]. Demand-driven supply chains, instead of those pushed by the private or public sector, result in spare parts often being hundreds of kilometres away from the people who need it most.

3. **Lack of Knowledge:** Trained pump mechanics who can perform major repairs are challenging to locate. Even though communities pay the mechanics, they often do not perform preventive maintenance but are instead called once a failure occurs [12]. Furthermore, they often have another profession to supplement their household income and may be delayed in responding to a failure report [27]. As such, maintenance occurs sporadically and is often not well documented, if at all.

1.3 The Challenge of Handpump Monitoring

The functionality of handpumps is a poorly-defined parameter that remains open to the judgment of the observer, is measured at a single point in time, is susceptible to changes in seasonality, and lacks a universally agreed definition [28]. Therefore, functionality alone is not an informative parameter for measuring the performance and sustainability of rural water services [29, 30]. NGOs and donor agencies have left a legacy of installing new infrastructure in rural regions instead of implementing long-term maintenance programs. Interventions which aim to ensure the long-term functionality and benefit of water services to their users should prioritise the reduction of downtime (both the time between failures and the duration of the breakdown), in addition to reducing the proportion of abandoned handpumps [28]. Regular inspection and monitoring of individual water points are vital to ensure continuity of water supply; however, this is unlikely to occur frequently in high-income countries, let alone in rural villages in low- and middle-income settings [31]. Currently, there is no way to collect reliable quantitative data of this nature to inform priority-setting for improved service delivery models [30].

1.4 The Future of Remote Health-Monitoring

Advances in the Internet of Things (IoT) and associated data analysis have accelerated the possibilities for remote health-monitoring systems. In particular, low-cost mobile technologies offer new and inclusive opportunities to approach the monitoring and maintenance of water services to support rural development toward achieving the Sustainable Development Goals [3]. The economic prosperity of rural communities reaches beyond the conventional trajectory of infrastructure development. IoT is allowing economies and industries worldwide to reimagine traditional business models and reinvent services. The widespread collection and storage of data have levelled the playing field and allowing developing countries to leapfrog the phases of development. By using machine-to-machine (M2M) communications to connect machines, devices, and objects to the internet, the IoT can turn them into “intelligent” assets. The prevalence of services based on mobile communications networks across Africa, and the application of low-cost embedded M2M services, has allowed the continent to advance rapidly in the connection of its people, places, and “things”. By retrofitting a simple and inexpensive device to a standard hand-pump handle, we can pro-actively monitor the condition of handpumps and ensure that millions of people can have access to a reliable water source.

1.5 Outline of Thesis

This thesis presents a method for dynamically inferring the health status of remotely monitored sensors by distributing the monitoring between the in-situ novelty filter and a cloud-based (fusion) centre. Chapter 2 explores the operation and maintenance systems of rural handpumps in the context of the study site in Kenya. It gives an overview of the historical and current monitoring methods of rural water supply networks. Chapter 3 presents an overview of previous works related to remote water point monitoring, along

with relevant methodological approaches used to deploy machine learning in remote health monitoring of systems. Chapter 4 describes the sensors that were designed and embedded in handpump handles to collect data for this thesis and also introduces the curated data sets. Chapter 5 investigates the use of lightweight machine learning methods to perform the initial novelty filtering onboard the embedded sensor by minimising the bandwidth and power requirements for classification. In Chapter 6, the best-performing novelty filter is used to develop a novel distributed inference system, which uses more advanced machine learning methods to analyse the data packages sent from the classifier on the node. The distributed inference health monitoring system is dynamically adjusted to reduce the subsequent proportion of raw data sent to the cloud base-station without compromising the fidelity of the health monitoring system. Chapter 7 uses the proposed dynamic distributed inference health-monitoring system to monitor patient health using key physiological parameters to demonstrate the flexibility of the system to adapt to different applications easily. Finally, Chapter 8 summarises the work proposed in this thesis, along with discussing its limitations and proposing ideas for potential future work.

Chapter 2

Rural Water Supply

Rural households around the world depend on the local water supply and ‘self-supply’ from surface waters, vendors, and groundwater-based sources [32]. Where water infrastructure does exist, the systems are decentralised (like community water collection points and pit latrines, in some cases) and often fall into disrepair due to mismanagement and neglect. A sustainable water source is one that functions appropriately, delivers an acceptable level of benefit, and remains reliable throughout the life cycle of the system [33].

This chapter will introduce an important tool used in the rural water supply to extract shallow groundwater, the surface-mounted handpump, and explore the complexities of its maintenance and limitations of the technology. The final section will explore the existing methods of maintenance and remote monitoring.

2.1 Water Supply in Rural Kenya

As recently as 2015, at least 20% of the total Kenyan population of nearly 48 million people still relied on a handpump for their daily water needs [34]. There are an estimated 32,000 handpumps installed across Kenya [35]. A survey found that 60% of rural households collect water from a source away from home and use on average four different water sources, depending on the reliability of the source and the intended use of the water [36].

During the rainy season, households with private pipe connections or wells collect more

than 55% of their water from rainwater, while households relying on public wells or shared piped connections collect 69%. Households using public boreholes or surface water as their primary sources have the most significant seasonal behaviour change and rely on rainwater for nearly 75% of their water needs during the rainy season. This is an important trend to note as those users relying on public handpumps tend to show the most seasonal behaviour change, which influences the demand and subsequent wear of handpumps during different seasons [37]. As such, handpumps tend to experience more failures during dry seasons as demand from water users increase when other sources dry up.

Kenya has an estimated groundwater potential of 619 million m^3 [38]. In 2012, the estimated abstraction rate of 7.21 million m^3/year was still within the safe abstraction rate of 193 million m^3/year [38]; however, this is a gross average estimation and in certain regions aquifers are already over-used. Over-abstraction occurs when the abstraction rate of a particular aquifer exceeds the recharge rate and results in the decline of water levels and a deterioration in water quality over time [39]. The Weather Risk Management Association (WRMA) use digital loggers to monitor water levels at eleven dedicated monitoring boreholes across the entire country of Kenya. Data related to water level and quality trends are collected manually every quarter, and more frequently for critical aquifers identified as being more intensively used [40]. Despite the magnitude of this effort, data points remain sparse across the region, making it difficult to accurately model changes in the aquifers. Over-exploitation in concentrated areas and saline intrusion into the water supply in coastal areas significantly limit groundwater use in Kenya. Also, Rift Valley aquifers in Eastern Africa are notorious for their high levels of fluoride, which affects the water quality [41]. Nonetheless, coastal zones, such as Kwale County (the study site for this project), rely exclusively on groundwater sources for domestic, commercial,

and industrial water needs.

In 1983, the first large scale deployment of the Afridev handpump took place in Kwale County, a rural region on the south coast of Kenya, resulting in the installation of 580 handpump-equipped boreholes recorded by the Kwale County Government between 1983 and 1995. The average borehole depth for these handpumps is $36.2 \pm 18.7\text{m}$, ranging from a minimum depth of 7m to 102m, which is well over the specified Afridev operating depth. A 2013 census of handpumps in the county mapped 571 sites in Figure 2.1 [42]. While most of the original installation sites could be traced (marked in blue as matched), 109 of the original water points were either converted or removed. Installations not on the county government record were likely funded by NGOs or private donors and marked as unmatched (in yellow).

During the census visit, 19 (3.3%) handpumps had been replaced with motorised pumps while only 314 (55%) were functional. The remaining 238 (41.7%) non-functional pumps is

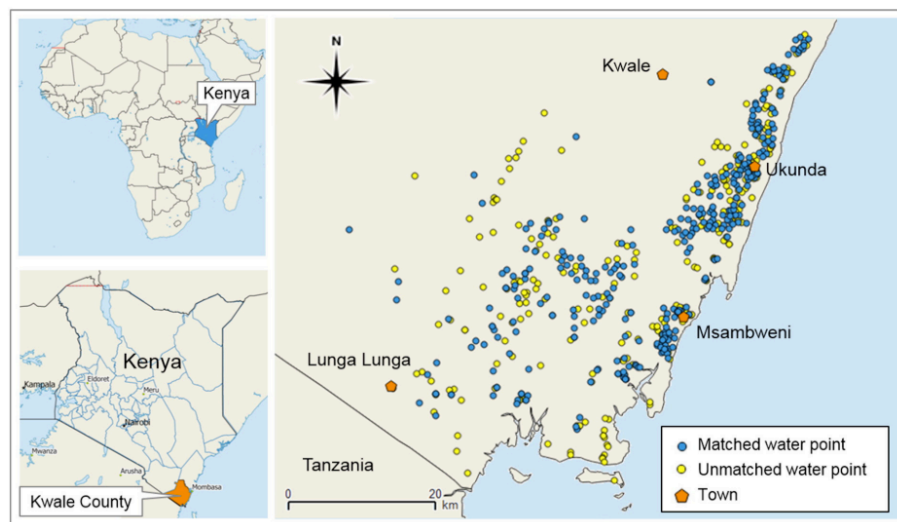


Figure 2.1: Map of study site in Kwale, Kenya, and locations of the original large-scale Afridev handpump deployment between 1983 and 1995. The matched water points (blue) indicate 471 of the original 580 found during a 2013 census. Unmatched water points (yellow) indicate additional handpumps installed later, likely by private donors or NGOs. (Source: [42])

much greater than the expected 25-30% reported in the literature. Unique hydrogeological and geographic factors of the area, such as high electrical conductivity of the groundwater and unconsolidated sands, are noted as the key contributing factors for the high failure rates of handpumps [42]. The deeper groundwater levels further drive this as many of these handpumps have been operating well beyond the rated depth for nearly three decades. In addition, the isolated location being further away from spare parts contributes to longer periods of non-function.

2.2 Handpump Introduction

Community handpumps, such as the Afridev [43] and India MK II [44], are designed to lift water from deeper sources than traditional ‘rope and bucket’ systems [45], which can only operate in shallow wells. The well is sealed during the installation, which dramatically reduces the risk of water contamination when the well is shared with livestock and other animals, as is often the case. A typical handpump supports 360 villagers and is used for up to 14 hours daily to extract water at approximately 0.7 - 1 m³/hour [46]. The handpump has evolved over the past 30-plus years to overcome the many challenges around contamination generally associated with traditional open wells, and yet it remains the most robust and affordable technology to access groundwater. This section will review the technical specifications of a handpump and discuss its operation.

2.2.1 Types of Handpumps

Over the years, many different types of handpumps have been developed to meet various demands. The design of handpumps has become standardised across various pumps, with the most crucial difference being the connection between the handle and the pumping rods [44, 47]. The introduction of open-top cylinders greatly simplified the design to make it possible for villagers to perform maintenance on the pumps. These are known as village

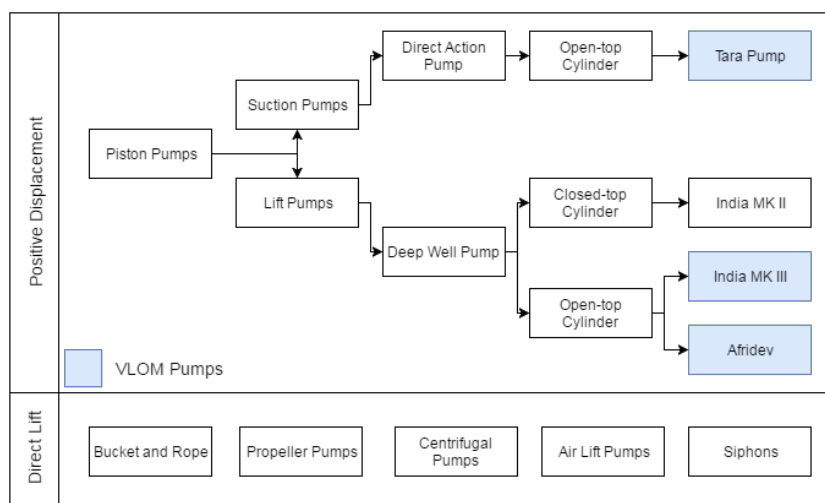


Figure 2.2: Overview of handpump types as classified by water-lift method (displacement or direct) including VLOM capability. Standardised VLOM handpumps include the shallow well suction Tara pump (operating up to 10m) and the deep well lift pumps, India MK III and Afridev (operating over 10m). The Afridev is commonly used in Kenya.

level operation and management (VLOM) handpumps. An overview of the different types of handpumps in Figure 2.2 highlights the evolution of the three most widely-used VLOM handpumps. The technical data sheet in Table 2.1 compares the functional specification of these three VLOM pumps that have become the standard for handpumps in most regions. Lift pumps operate at greater depths than suction pumps, which makes it more suitable for use in Kenya where groundwater is often located at depths greater than 10 metres. The physical lifting power of the water user determines the limitation on the operating depth of piston pumps.

Operating condition is the critical criterion for standardising the design of handpump types. In general, the most important factors to consider are [49]: (1) lift (shallow vs deep well), (2) user-group size, (3) groundwater conditions (e.g., factors leading to possible corrosion), and (4) community development. The water user would have different importance criteria from the installer and these could include: ease of repair, reliability of the handpump, user preference, availability of the spare parts, and cost. Design parameters of the

Table 2.1: Technical data sheet comparing the three VLOM pumps considered in this study, for current and future work. Shallow operating handpumps offer greater discharge and typically serve fewer people as these tend to be more densely installed, such as in Bangladesh [43, 44, 48].

Technical data	No.6 Pump	Afridev	India MKII
Depth to be used	0 - 6 m	10 - 45 m	45 m
Cylinder diameter (mm)	89	50	63.5
Maximum stroke	215 mm	225 mm	125 mm
Approx. discharge at 75 watt input (m^3/hour)			
At 5 m head*:	3.5		
At 10 m head:		1.4	1.8
At 15 m head:		1.1	1.3
At 20 m head:		0.9	1.0
At 30 m head:		0.7	0.8
Pumping lift	0 - 7m	10 - 45m	10 - 50m
No. of people served	50 - 100	300	300
No. of households served	5 - 10	30 - 50	30
Water consumption (litres/capita)	20 - 25	15 - 20	15 - 20
Type of well	collapsible tube or dug	borehole or dug	borehole

*pumping head is the maximum height (pressure) water can be raised

pump and the borehole drilling have direct cost implications when installing a small water supply system. Higher yields require deeper boreholes, bigger diameter boreholes, detailed site investigations, and intensive well development [50]. However, bigger rigs required to drill boreholes for these handpumps increase both the cost and the risk of failure of the project. The optimum installation criterion is a trade-off between these factors and should ultimately result in the long-term improvement in the livelihood of the community using the handpump. A simplified borehole with well yields of about $1 \text{ m}^3/\text{hour}$ is sufficient for a handpump that can serve up to 500 people [49].

Afridev Handpump

At our study site, the Africa Development (Afridev) handpumps are the most commonly used. Its development was a collaboration between UNICEF, UNDP, and World Bank in the early 1980s with the first drawing to be made available in public domain. The first prototypes were installed for field testing in Kwale County in late 1983 [2] - many of which

are still operating. Updated drawings and testing guidelines [43] remain freely available in the public domain to encourage standardisation and local production.

The Afridev is a locally produced conventional lever action handpump that uses hot-dip galvanised stainless steel for the handpump parts and u-PVC for the rising main (the “straw” used to suck the water from underneath the ground). Figure 2.3 shows the assembly of the critical handpump parts below and above the ground level (GL). The pump rods connect the plunger to the pump head and are available in 3-metre sections that interlock. The rods should ideally also be stainless steel to prevent corrosion [43]; however, the high upfront cost results in the cheaper alternative, mild steel, to be more commonly found. The depth of the well determines the total number of rods and therefore varies across handpumps. The 2013 census [42] and dataset overview in Chapter 4, both show that handpumps often operate at depths much greater than the original technical specifications of 10 - 45 m, where some go up to 100 m. As one of the VLOM designs, this handpump is easy to repair and requires only simple tools (although these are not always widely owned).

2.2.2 Piston Handpump Operation

The pump cycle of a handpump contains two distinct phases: the up-stroke and the down-stroke. The piston motion distinguishes the phase cycles. Due to the handle and rod coupling, the handle moves in the opposite direction of the piston. Most handpumps, including the *Afridev*, use reciprocating pistons and plungers to enable positive displacement in the pump rod. A non-return piston valve slides up and down the vertical cylindrical pipe fitted with a foot valve.

Up-Stroke

During the up-stroke, the piston assembly moves upward. In direct or chain coupled piston handpumps, this occurs when the handle moves toward the ground. The foot valve is open while the piston valve remains closed. The negative pressure in the cylinder cause water to push upward in the lower cylinder.

Down-Stroke

When the handle moves up, the piston moves down, closing the foot valve. Pressure in the lower cylinder forces the piston valve to open during this phase, filling the upper cylinder. As these phases repeat continuously, water is pushed up the rising main and flows out the spout.

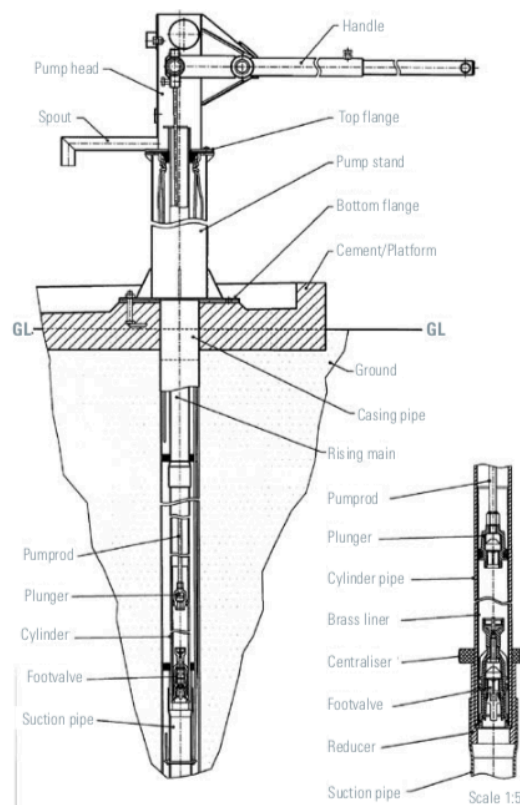


Figure 2.3: Illustration of the standard assembly of the VLOM Afridev handpump. Pump rods connect the plunger to the handle, which moves in opposite directions. Positive displacement in the cylinder moves water up the rising main and out of the spout above the ground level (Source:[45]).

2.3 Conventional Handpump Maintenance Systems

Community Based Models

Over the years, the lack of adequate maintenance has led to the decentralisation of handpump operation and maintenance; resulting in the widespread implementation of Community Based Management (CBM) models. In 1997, the World Bank OECD audit [46] identified the critical role of local community organisations and water committees to sustain handpump projects post-implementation. Community managed handpumps with increased ownership from users result in faster response times to problems.

However, the inability of CBM models to sustain revenue collection to finance handpump maintenance remains one of the greatest challenges for sustainable rural water supply [25] and alternatives to the dominant CBM models, such as cross-subsidisation, are emerging as viable alternatives [51]. The lack of local skills, tools, and access to spare parts further challenge the sustainability of VLOM handpumps [12].

Public-Private Models

In the public-private model, the maintenance and repair of handpumps are managed by private stakeholder and regulated by the government. The community is still responsible for financing the operation and maintenance costs. Typically, a community volunteer will be responsible for small maintenance tasks like cleaning the parts and will report any failures to the water committee. The committee will collect and maintain water service fees from individual community members and call on a pump mechanic or circuit rider once failure occurs.

The main advantage of this model is that it can address the skills and equipment shortcoming of the CBM model without taking ownership away from the local communities

[12]. Local government and NGOs play an essential role in supporting capacity building to ensure trained mechanics are available to water users. A shortage of local mechanics could lead to increased downtime of handpumps.

Private Models

In densely populated areas of South Asia, such as Bangladesh, where handpumps operate in close proximity, individual handpumps are often located on private land and used by a single household. Sparse distribution of handpumps across vast land areas in Kenya means that most pumps are community-operated but potentially privately owned by an institution or organisation. The existence of high-quality services in rural areas suggest that private operator models for handpumps can be successful [52]. The loss of income experienced by the private owner during downtime increase the urgency for faster repairs. As such, privately owned handpumps have higher levels of sustainability [53]. However, this model leads to disempowerment created by dependency and inequality when effective regulation is not in place [12].

Telemetry Solutions

Advances in communication technology are opening up a host of possibilities for the continuous monitoring of water services in remote villages that often lack infrastructures, making it difficult or unsafe to reach [54]. Novel smart technologies use sensors on rural infrastructures, such as handpumps and motorised boreholes, to monitor use. This type of remote telemetry solutions offers insights into real-time functionality to facilitate maintenance, and the usage patterns of water sources during changing seasons, such as rainy vs dry, or extreme weather conditions, such as droughts. Continuous monitoring also provides data about the time between repairs. Chapter 3 discusses examples of related work in detail.

2.4 Handpump Functionality and Continuity

Handpump functionality is a poorly defined metric in the literature, but an essential factor to determine the labelling of the data. Individual handpump *continuity*, or the average proportion of time it is functional, is estimated to be 80% just after installation and decreases linearly to 50% at year 10. As with any mechanical system, one cannot expect 100% functionality throughout the lifetime of the system. Instead, 85% operational performance at any given time over the entire lifetime of the handpump is proposed as an “ambitious but potentially realisable target” [28]. By this definition, the functionality of a group or region of handpumps can be modelled as [28]:

$$\text{Functionality, } f = (1 - a) - (1 - a) \frac{nd}{365} \quad (2.1)$$

where a is the fraction of overall water points permanently out of service or abandoned; n is the average number of breakdowns per water point per year (dimensionless), and d is the average breakdown duration (days).

Typically, the number of annual breakdowns n increases steadily over the lifespan of the pumps, from $n = 0.05$ in year 1 to 3.5 in year 10. As the severity of failure increases, so does the duration of breakdown d , which also increases steadily from $d = 5$ in year 1 to $d = 20$ in year 10. From equation 2.1, it becomes apparent that both the downtime and time between breakdowns are essential parameters to improve functionality. However, neither *functionality* nor *continuity* is a useful metric in a water risk management tool as neither offer great insight into the cause or the duration of the breakdowns [6]. For instance, the impact of 10 one-day breakdowns on communities will be significantly different to a single twenty-day breakdown, even though both cases report the same functionality score.

Table 2.2: A qualitative framework for defining handpump functionality proposed by the Hidden Crisis Project addresses the challenges in identifying and monitoring long-term handpump reliability [56]. The tiered performance levels measure functionality as a working water point that yields at least 10L/min and experienced fewer than 30 day downtime in the last year.

Performance Level	Assessment	Measurement
Basic	Water point working?	Yes/No
Snapshot	Basic AND sufficient yield on day?	10 L/min
Functionality	Snapshot AND is it reliable?	30 days downtime in last year
Water Quality	Functionality AND quality of the water?	WHO inorganic parameters, TCC Standards

Furthermore, this equation does not reflect the direct impact of non-functionality on human behavioural psychology. The critical downtime per breakdown that results in the ultimate abandonment of the water point remains unclear [55]. In 2017, the Hidden Crisis Project [56] attempted to address this shortcoming by introducing a qualitative framework to guide the broader definition of functionality, shown in Table 2.2. The guideline quantifies the assessment criteria for each performance level whereby a functional handpump is reliable when it has experienced less than 30 days downtime in the preceding year and delivers sufficient yield (10 L/min). However, the practical challenge is to reliably and continuously monitor each of these parameters for an individual handpump.

Structural health-monitoring typically considers the performance of the mechanical system. However, handpump failure is not limited to the binary classification of the mechanical system’s performance alone. The primary attributes used to categorise handpump functionality are [28]: mechanical performance, water quantity or yield (short term), water quality, and seasonality (long term). Water quantity is a property of user acceptability. In most cases, users will tolerate a certain deterioration of yield until eventually abandoning the pump for an alternative water source that requires less effort for the same yield. Understanding the range of this tolerance is an essential factor in ensuring the health-monitoring system can flag pump failure at a point before mechanical failure but

not within user acceptance.

Similarly, water quality, particularly in areas of high salinity, could contribute to the mechanical wear of pump parts. Finally, seasonality is considered a long-term failure and refers to the health of the aquifer rather than the handpump system. A well-designed health-monitoring system for handpumps should be able to distinguish between a broken mechanical system or a drying well when flagging restricted flow at a water point. Doing so will identify handpumps relying on depleting aquifers, which requires more intense interventions such as re-drilling or replacing the well.

2.4.1 Defining Functionality and Labelling Data

Following the in-depth review of handpump functionality [57], we introduce two main attributes to classify handpump conditions for this work:

1. **Short-Term Water Quantity:** a pump is classed either as normal (C1) or abnormal (C0). A pump is normal when water flows from the spout while pumping and abnormal when no water flows from the spout while pumping.
2. **Mechanical Performance:** ten sub-categories are used to identify the mechanical attributes that describe the functionality and physical condition of the handpump along with relative yield associated with each.

The data overview in Chapter 4 (Figure 4.4) shows that there are fewer examples of the abnormal class compared to the relative frequency of data for the normal class. Details on the sub-categories are also detailed in Chapter 4.

2.5 Preventable and Avoidable Failure Events

There is a large body of evidence to show that lack of maintenance and high water demands at the handpump are not the only factors contributing to non-functioning parts or ultimate

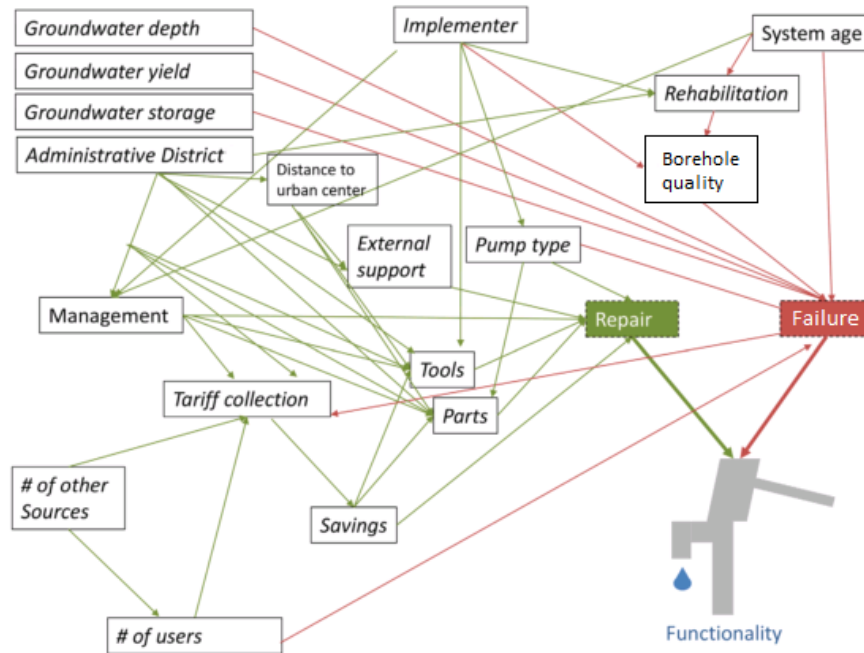


Figure 2.4: Schematic of the conceptual model of the factors affecting handpump functionality which illustrates the interaction between factors related to the failure (red) and the repair (green) (Source:[6]).

abandonment of the water point. Figure 2.4 shows a schematic summary of the interaction between multiple factors affecting handpump functionality, where green arrows indicate factors related to the rate of repair and red arrows to the rate of failure [6]. Regulatory and socio-economic factors mainly contribute to the complexities related to the repair time while geographic and hydrogeology greatly contribute to the failure rate of the handpumps.

2.5.1 Failure Types

In addition to maintenance and high demand, other condition variables contributing to failures in the system can be categorised into three vital aspects: (1) hydrogeology; (2) technical; and (3) socio-economic.

Hydrogeology

Water Conductivity

“The red water problem” is a well-known symptom of corrosion in a handpump that is

often encountered by the water user first thing in the morning or after being unused for some time [58]. Moderate to aggressive groundwater with a pH balance of less than seven can rapidly accelerate the wear of mild steel rods and cylinder parts in the pump due to increased corrosion. The estimated scale of impact of corrosion is much greater than anticipated; corrosion causes at least 66% of handpump failures in Ghana [58].

Furthermore, corrosion of the parts leads to the release of iron in the water source, which in turn alters the taste. In some areas of the Niger, nearly 30% of working handpumps are abandoned due to the poor taste of the water caused by excessive iron levels [55, 58].

Regions with aggressive water must factor in the additional wear associated with corrosion and anticipate that certain parts, such as pumping rods, require more frequent replacement. Hot-dip galvanisation of the mild steel components is not always sufficiently resistant to these conditions. As such, there is a strong motivation to use suitable materials, such as fibre reinforced plastic or stainless steel, to ensure that pumps are fully corrosion-resistant. Unfortunately, a trade-off exists in sacrificing durability (in the case of plastic) or increased up-front cost of installation (in the case of stainless steel), hindering the acceptance rate of these solutions.

Water Temperature

Handpump parts are designed to operate in water temperatures of 16 - 40°C [2]. In fluid dynamics, the temperature of the fluid is inversely proportional to the kinematic viscosity for laminar flow in a cylindrical tube. Thus, operating outside or at the higher end of this range would lead to faster degradation of the parts.

Water Level Fluctuations

Fluctuations in the general water table are predominantly seasonal as a result of climate,

rainfall intensity, and frequency. Leather cup seals, commonly used in suction pumps [49], and to a lesser extent rubber U-seals, commonly used in lift pumps, are particularly vulnerable under these conditions as they will deteriorate more rapidly if allowed to dry out thoroughly.

The draw-down in the borehole itself, as well as the presence of nearby boreholes, cause short term fluctuations, known as borehole interference [50]. In the case of a rural water supply network, where co-located boreholes are at intervals greater than 50 – 100 m, the impact of borehole interference is minimal to non-existent.

Screen

The natural recharge rate of the aquifer is a function of the permeability of the overlaying local soils and water table depth. Basement aquifers, with a transmissivity of less than $1 \text{ m}^2/\text{day}$, are more likely to experience limitations related to draw-down [59]. Precipitation and tidal-level variations are the main driving forces of recharge in coastal aquifers. Current forecasts using machine learning methods rely exclusively on piezometric height measurements [60, 61, 62].

Technical

Worn parts and a lack of regular (or any) maintenance are two of the primary causes of technical failure in handpumps. However, overuse is not the sole cause of these failures. Failures can also be due to [50]: (1) Design flaws; (2) Incorrect installation; (3) Manufacturing defects; and (4) Depth, location, and age of the borehole.

Borehole

The depth of the borehole determines the number of rods used in the system. Thus, we expect that the extra weight lifted in deeper boreholes may result in more frequent failures.

In sub-Saharan Africa drilling depths often range from 48 to 62 metres, resulting in the pump operated at settings outside the design specifications.

Often, but not always, boreholes are placed under high trees. This limits sun exposure and assists with the mechanical removal of the rising main. Therefore, boreholes drilled in direct sunlight leave the handpump vulnerable to heat exposure and may limit routine maintenance. Although not impossible, removing very deep rising mains without a nearby tree can be extremely difficult, which means that local mechanics try to avoid such a process.

Finally, pipes with PVC casings line the borehole such that it remains clear. Borehole maintenance is required every ten years to clean and re-develop the borehole [49]; however, this is rarely performed due to the high cost. Depending on the location of the borehole, drilling a single borehole can cost thousands of dollars.

Problems with Parts

The overall performance of the Afridev handpump lies within the technical specifications of typical part failure related to frequent use. However, manufacture non-compliance with the SKAT ¹ specifications [43], such as incorrect dimensions of U-seals and poor welding craftsmanship, substantially reduce the performance of handpump parts. Table 2.3 is a synthesis of the most common handpump breakdowns that occur as a result of a failure in a single part. This table draws on the literature and extensive field notes from Kenya taken over the years as part of this research project.

¹SKAT is an independent Swiss organisation working in the fields of development and humanitarian aid at the Swiss Centre for Development Cooperation in Technology and Management.

Table 2.3: Common handpump failures in deep well lift pumps (Afridev) [23, 2, 63, 64, 65].

Part	Common failure	Possible Cause	Fault Class
Foot valve	Foot valve becomes stuck on the cylinder lip.	Foot valve oscillates when dropped down the rising main during fitting.	Installation
Cup Seal	(1) Damaged cup seal. (2) Valve forced to remain open. Most often, the pump can continue to work but results in reduced water flow.	Foot valve bobbins that break on impact with the water or, if the well is dry, the foot valve receiver can cause the bobbins to break off and become lodged in the valve.	Design Flaw
Rod Centralisers	Worn centralisers cause additional friction during pump operation.	(1) Abrasion caused by siltation; (2) Misalignment of the rod and rising main eyelet.	Installation
Bent Pump Rods	Misalignment of rods or rod joints cause additional rubbing on rising main pipes.	Damage to the rods can occur during (1) transportation, (2) installation, or (3) maintenance and re-installation.	Installation or Maintenance
Corrosion of Pump Rods	Worn rods have weak spots and require more frequent replacement. It affects mechanical performance when rods break but also cause damage to the cylinder.	Moderate to aggressive groundwater ($\text{pH} < 7$). Galvanisation of mild steel is not resistant to these conditions.	Design Flaw or Maintenance
Rising Cylinder	Damage at the socket joint.	Misalignment of the socket joint during installation.	Installation
Plunger	Plunger legs break off and become lodged in the foot valve.	Cause is unclear but could be during impact with the water when excessive force is applied. Possible incorrect dimension, angle, or material of the legs.	Design flaw
U-Seals	Shorter than normal operating life (6 months) occurs when the seal rolls out of the location groove. Dislodged U-seals can block the cylinder and make it difficult to remove the foot valve with the “fishing” tool.	(1) Abrasion caused by siltation; (2) Incorrect dimensions; (3) Soft material; (4) Friction caused by withdrawal against the inside of the rising main.	Manufacture defect
Rising main	Crack in the casing cause loss of pressure in the system restricting or preventing water flow.	Aggravated by bent rods, removed rod centralisers, or poor joints.	Design Flaw or Maintenance
Handle	Handle is loose or “shaky” during use.	(1) Worn bush bearings. (2) Loose fulcrum or hanger pins caused by corrosion.	Design Flaw or Maintenance
Pump head and stand	(1) Head or stand is unstable during use. (2) Water is leaking from the pump body.	Loose flanges or a cracked platform cause instability during use.	Maintenance

Socio-Economic

Pump Usage Levels

Studies have found that a handpump with continuous high usage (more than 5m³ or 5000 litres per day) can experience a failure every 4-6 months for those parts experiencing the most stress, such as the U-seal [2]. The stress, S , experienced by the mechanical system, can be described as a function, f , such that [49]:

$$\text{Stress}, S \approx f(u^2, x^2) \quad (2.2)$$

where u is the number of users and x is the total pumping lift.

Thus, the number of users and pump lift are essential factors affecting the pump operating conditions, long term wear, and maintenance schedules. In the case of the Afridev handpump, pump lift increases as deeper well depths require more pumping rods in addition to the increased distance required to lift the water from below the ground level. Pumps designed for small user groups are less robust than those designed to serve large groups (>100 users).

Maintenance

The VLOM pumps were designed to allow users to perform minor maintenance tasks, such as cleaning the pump rods, in order to prolong the time between scheduled maintenance and major breakdown events. However, a lack of transfer of formal “ownership” of handpumps, insufficient flow of information, and training programmes often prohibit community participation in practice, which in turn increases the occurrence of unexpected breakdowns [23].

User Characteristics

There is no evidence in the literature to suggest that the characteristics (height, gender, age) of the water user directly contribute to the failure of the pump. However, extensive observations during field visits suggest that these vary so significantly, both at the same pump as well as across pumps, that we should not ignore this during the initial investigation. For example, we observed that user's height determines the range of motions in the handle during operation. Shorter pumping strokes are known to cause excessive wear in focal regions of the system. A decreased pumping range can either be a result of incorrect plunger timing, a broken foot valve, or children operating the pump [65].

Furthermore, we noted that gender, often but not always, determines the “pumping style”. Women often use both hands on one side of the handle T-bar while men hold one hand on either side of the T-bar. Repeated use in a specific handle range or one-sided pressure could lead to force imbalance that may result in faster wear of specific parts.

Most Common Failures

Following an in-depth review of the studies related to handpump failures, we conclude that the four most common breakdowns in Afridev handpumps occur in U-seals, plunger valves, foot valves, and the uPVC pipe (i.e., the cylinder of the borehole). An overview of the literature strongly suggests that the U-seal causes a disproportionate number of pump failures, followed by damage to the casing of the rising main. In addition to regular wear of the parts, soft materials and incorrect dimensions contribute to rapid wear of parts. Finally, Table 2.3 shows that failure is often due to the malfunction of more than one part. Early detection of wear in a single part may therefore prevent many other problems from occurring in other parts of the handpump.

2.6 Remote Health-Monitoring Systems

Health-monitoring is a data-led method used to measure, record, and analyse internal or external parameters of the system. The observations are collected using many internet-enabled sensors to provide a near real-time assessment of the condition of the system. The data collected also offer the opportunity to learn from failed projects and rehabilitate broken systems. In Africa, where mobile phone penetration is nearly 83% [66], GSM-enabled sensors allow the systems to monitor usage in even the most isolated locations remotely. This means that rural wells and handpumps can be monitored from a central location without the need for costly and time-consuming site visits. Fortunately, there has been a positive trend in recent years towards increased monitoring of both new and previously installed infrastructure projects. The ability to remotely determine the condition and usage status of a rural water source is a giant leap in the development of rural infrastructure delivery. The burden of maintenance can be shifted away from the VLOM structure to a utility maintenance service provider, similar to those that exist in more developed countries, and release the water user from the additional burden.

Moreover, monitoring plays a vital role in resource management and in the ability to track progress, if any, towards reaching targets. In 2010, the UN created the Joint Monitoring Programme (JMP) to track the progress towards the 2015 Millennium Development Goals, including target (7.C) of cutting in half “the proportion of the population without sustainable access to safe drinking water and basic sanitation”. The organisation, “Water for People”, created a monitoring protocol for tracking the post-construction sustainability, or continued operation, of water points. A case study across 655 water systems in five countries found an impressive 95% sustainability rate [67]. Unfortunately, the high cost of monitoring and intense human resource requirement made it impossible to scale this type

of intervention across vast regions as is desperately needed to track development progress.

The 2018 SDG Dashboard continues as part of the JMP to provide a data-driven tool to track and monitor progress towards the SDGs. The relative availability of low-cost solutions allows remote monitoring solutions to measure performance in addition to rudimentary feedback on whether or not a water point is working or not. In addition to lower operating costs, robust remote monitoring systems offer objectiveness that can overcome the bias often introduced by other survey methods, such as randomised control trials, frequently used in international development. An example of such bias is a 2013 study in rural Rwanda which showed that the actual measured use (73.2%) of cookstoves was significantly lower than the self-reported use (90.5%) [68].

The digital revolution is further advanced by the wide adoption of mobile money which allows the financing of assets and services in rural areas not previously possible. In 2018, the International Monetary Fund estimated the value of mobile money transactions in Kenya to be nearly Sh9.09 trillion (£64.6 billion) or 44% of the country's gross domestic product. The network-enabled sensors coupled with the ability for un-banked rural customers to instantly send money anywhere in the country at an insignificant cost provides the economic infrastructure for providing low-cost water services that enhance the rural water provision sector in developing countries [69]. The recent focus on sustainability in international development has pushed the existing paradigm beyond the conventional practice of conceiving and implementing an idea in the field to include methods of long-term monitoring and evaluation. As such, remote health-monitoring systems become an instrumental tool for the post-implementation monitoring of both governmental and non-governmental development projects.

2.6.1 Challenges of Remote Health-Monitoring

One of the challenges of remote monitoring in limited-resource settings remains access to reliable and cost-effective bandwidth to transmit the rural data stream to a central server where it can undergo further evaluation and review. Various approaches to rural and remote connectivity have been proposed over the last 20 years, many relying on dynamic routing strategies commonly found in wireless networks [70, 71, 72, 73]. The large-scale uptake of the proposed solutions remained costly compared to the rapid expansion of telecommunications networks. Fortunately, developments in telecommunications, along with the rapid mobile phone penetration in rural parts of sub-Saharan Africa and South-East Asia, are diminishing this challenge.

The second challenge with remote monitoring in limited-resource settings is that the infrastructure and data streams often do not exist at all. Providers who are entering this space must establish novel data streams, collection protocols and select suitable technology. These complex bespoke systems often result in higher development costs and longer lead times. As more examples of successful applications emerge, and off-the-shelf hardware evolves, the need for bespoke systems and the specialists required to maintain these are becoming obsolete.

Finally, the installation and maintenance of sensors are critical to ensure a continuous data stream. Unreliable sensors can result in poor-quality data which compromises the fidelity of the entire health-monitoring system.

2.7 Conclusion

This chapter describes the importance of groundwater in the rural water supply network by introducing the various types and operation of the handpumps used to extract groundwater

from the shallow aquifers. Furthermore, this chapter considers the existing community maintenance schemes intended to support the longevity of these systems. Finally, based on the mechanical operation described, the many challenges - both mechanical and social - that contribute to the lack in the reliability of the handpump infrastructure are described. The next chapter presents a review of the water point monitoring and inference methods previously used in rural, resource-constrained settings as well as the growing body of health-monitoring.

Chapter 3

Literature Review

Predictive health-monitoring is widely used in engineering applications to detect damage to infrastructure as early as possible. Forecasting failure rather than merely detecting failure once it occurs helps to reduce the downtime of systems, and, ideally, performing predictive maintenance can avoid downtime altogether. With this approach already widely used in many fields from commercial and military jet engines [74], through to patient monitoring in health systems [75], it can directly extend to monitoring the health status of handpumps in rural villages.

There is a growing trend to perform this monitoring remotely as the latest developments in cloud-based processing support plug-and-play systems while the cost of transmission and storage of large data sets continues to decline. Remote health-monitoring refers to the use of technology to collect and transmit electronic observations between the ‘node’ and the ‘observer’ to assess the condition of the system monitored. Over the years, IoT deployments for rural development have focused extensively on the use of wireless sensors networks for smart agriculture and food production, precision farming and irrigation [76, 77]. However, there exists a gap in the literature on the practical implementation and feasibility for the long-term return on investment associated with developing these application-specific networks [78].

This chapter provides an overview of previous work on monitoring systems used in

rural water management projects before describing the context and frameworks of health-monitoring models. Specific approaches to more general applications of vibration-based health-monitoring are introduced along with previous work on methods related to feature extraction and predictive inference.

3.1 Our Previous and Related Work

In 2011, a proof-of-concept study demonstrated the remote monitoring of handpumps using a simple microprocessor, accelerometer, and mobile communications (GSM) components [79, 80]. These “Smart Handpumps” provide basic hourly data related to usage [81]. The use of technology to flag failure retrospectively has shown that it is possible to reduce the average down-time in rural handpumps from 27 days to fewer than three days [82]. In 2014, our preliminary analysis used SVM to classify the high-frequency accelerometer data from handpump handles to show that these patterns contain useful information related to the aquifer underneath the handpump [83]. High-rate waveforms from the accelerometry data can be processed retrospectively using machine learning methods for shallow aquifer monitoring to estimate the groundwater level at the handpump [83, 84]. The original handpump vibration data [80] also showed the inverse correlation between handpump use and rainfall in our study region in Kenya [37].

Further work used probabilistic Bayesian methods, such as Gaussian Processes (GP), to capture important information about the heteroscedastic noise in the signal waveform [85]. The initial results showed that the data could be useful to predict aquifer levels below the handpump with relative accuracy given a well-curated data set. However, a small sample set collected over only a few weeks limits the study, and subsequent models are unable to account for seasonal variations in the aquifer not related to the handpump use.

A large-scale longitudinal data set containing both handpump vibration data from the handpump and water column data from a diver sensor in the borehole [86] showed that Multi-layer Perceptron (MLP) and Long Short Term Memory (LSTM) networks are suitable models for estimating changes in the aquifer level underneath the handpumps [87, 88]. Longitudinal data proved challenging to collect in a real-world resource-constrained setting and investigations of the data found that calibration of the system during repairs, sensor failure, and lack of accurate labels could impact the results. A regression model was used to calibrate data due to changes in the handpump dynamics resulting from repairs and GPs were shown to be useful to impute missing data that arose from broken sensors or broken handpumps [89]. The same handpump abstraction data was used along with hydrogeology time-series and socio-economic survey data to construct a model for predicting welfare using deep learning [90].

3.2 Other Related Work

The increase in rural mobile phone penetration in sub-Saharan Africa has a positive impact on the development of innovative water management systems, specifically the collection of water payments via mobile money to cover operation and management costs of water points [69, 19]. In addition to the Oxford “Smart Handpump” [80], several other organisations are working on innovations in handpump remote monitoring and maintenance systems. Table 3.1 provides a summary of past and present projects that are using these developments in technology to monitor rural water points. Earlier projects relied exclusively on users to manually report data related to the water points [91, 92, 93, 94, 95]. All the studies that incorporated automated, continuous data collection used flow sensors [96, 97, 98, 99, 100]. In most cases, the incoming data stream is used without any further analysis or mainly utilised for mapping of water points. This approach is quite common as the fidelity and

Table 3.1: Related ICT projects in operation and maintenance of rural water supply networks.

Project Name	Ref.	Year	Region	Infrastr. type	Data types	Input	Scale	Status
Akvo Flow	[91]	2006	70 countries	Water and Agriculture	Mobile App	User	258 organisations	Ongoing
Human Sensor Web	[92]	2009	Tanzania	Water Points	SMS	User	50 water points	Ended
mWater	[93]	2011	158 countries	Water and Health	Mobile App	User	40,000 users	Ongoing
Maji Matone	[94]	2013	Tanzania	Water Points	SMS	User	20 water points	Ended
Mobile 4 Water	[95]	2013	Uganda	Water points	SMS	User	400 water points	Ended
MoMo (Welldone)	[96]	2013	Tanzania	Pipe Water	Flow Sensor	Auto	Unknown	Pilot
Susteq Water Forever	[97]	2014	Kenya	Water Dispensers	Flow Sensor	Auto	12 sites	Pilot
Sweet Sense	[98]	2014	Rwanda	Handpumps	Flow Sensor	Auto	181 water pumps	Pilot
eWater Pay	[99]	2015	Tanzania, Ghana, Gambia	Water dispensers	Flow Sensor	Auto	335 sites	Ongoing
Water for People	[101]	2015	Uganda	Handpumps	Jet Meter	Manual	30 sites	Ended
MSABI Pump for Life	[102]	2015	Tanzania	Rope-pumps	Mobile App	Technician	175 water points	Ongoing
Charity:Water Sensors	[100]	2015	Ethiopia	Handpumps	Flow Sensors	Auto	2881 sites	Ongoing

frequency of these data were conventionally sparse.

However, recent increases in novel high-rate data that can be acquired and stored at lower costs necessitate the addition of a sixth step in the conventional 5-step data information flow in ICT. Figure 3.1 shows the proposed updated innovation categories with the addition of “analysis” before reporting takes place.

Of these projects, only two focused extensively on the analysis of the incoming data. Firstly, Sweet Sense was initially developed to monitor rural cookstoves [105]. However, the findings supported the idea that rural monitoring could lead to greater accountability of development projects [106], and the monitor was adapted to measure flow for interventions of the maintenance service models for handpumps in rural Rwanda [98]. The results show that remote monitoring using flow sensors can improve handpump reliability. The extreme environmental conditions led to a high failure rate of the sensors, resulting in a higher overall cost of the tested service delivery models. Following improvements to the sensor enclosure and battery lifetime, this work demonstrated that ensemble machine

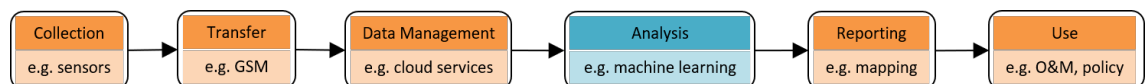


Figure 3.1: Updated categories of innovation areas in information and communication technology based on the flow of data/information (Adapted from [103, 104]). Data analysis is a crucial step currently omitted in the conventional 5-step data information flow for ICT.

learning, combined with efficient preventive maintenance service models, could increase handpump up-time to 99 per cent for the 42 handpumps monitored over 14 months [107]. The model relies on specificity (99.3%) rather than sensitivity (51.0%) when identifying independent failure events, which may be costly trade-off during large-scale implementation as false alarms have a high direct cost in remote health monitoring. Secondly, the ‘Charity:Water’ flow sensors applied Bayesian inference to develop a pump-scoring function to detect anomalies given hourly water usage data [108]. The data used to train the model lacked ground truth data for “normal” conditions and relied on the assumption that experts could select relevant periods of “normality” from the data with relative certainty, making it difficult to evaluate the performance of such an approach.

Despite the countless challenges and complexities to remotely monitor the assets that give valuable access to water, the usage and status data have proven incredibly valuable both to inform policy and to empower the water users responsible for managing scarce resources [109]. Real-time monitoring of handpumps, coupled with service maintenance models, lead to much greater increases in functionality than other rural water sources, like motorised boreholes [110]. As such, it is not only important to focus on handpump monitoring because many people still rely on these, but also because it will have the most significant impact on the rural water users who rely on them.

3.3 Context and Frameworks of Health-Monitoring Models

Condition or health-monitoring attempts to identify any deterioration that a system might experience as the state deviates from a “normal” state to an “abnormal” state. Sohn et al. (2003) [111] provide a general definition of damage:

“...changes introduced into a system that adversely affects its current or future performance. Implicit in this definition is the concept that damage is not meaningful without

a comparison between two different states of the system, one of which is assumed to represent the initial, and often undamaged, state.”

The monitoring of systems varies primarily due to the diversity of the type of nodes monitored. The two nodes considered as part of this project are condition monitoring of systems, such as rotating machines represented by handpumps, and health monitoring of human patients. Understanding the operational context of conducting this remotely is key for developing machine learning models that can accurately capture natural processes.

Figure 3.2 demonstrates that rural infrastructure may be commissioned as a routine or emergency installation. During the operational life-cycle of the system, various factors can contribute to the overall decline of the infrastructure quality. While some handpumps may become abandoned entirely due to repeated failure or high repair cost, maintenance can repair most handpumps. Different data types can be collected from the systems to enable timely monitoring of the condition at any given time. For example, water-related projects in Table 3.1 that apply remote monitoring use flow sensors to measure the volumetric flow of water point output while monitoring of structural and mechanical systems commonly use vibration characteristics [112].

Early warning scores (EWS) are widely used to monitor the deterioration of a system over time and their use on hospitalised patients are incredibly popular [113]. EWS can play a vital role in reducing downtime of systems; however, it is crucial to understand the limitations of EWS given the complexity and dependencies, both in the rate of failure and repair, that ultimately influence functionality. An important trade-off is the timeliness of the warning and the complexity of the monitoring hardware and data analysis procedure [114]. Ideally, condition-based monitoring should provide sufficient warning time to allow an appropriate response to correct any damage before the failure of the system can occur.

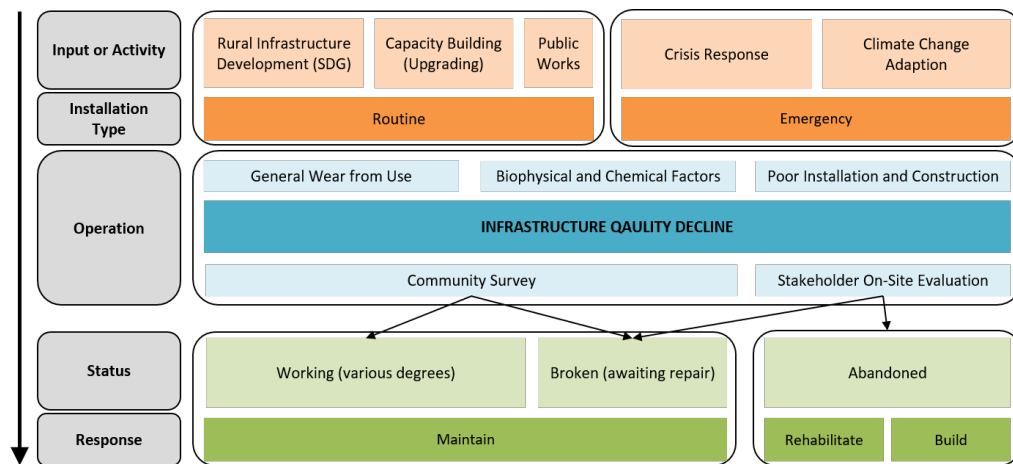


Figure 3.2: Visualisation of the maintenance flow over the lifespan of rural infrastructure. The infrastructure, like a handpump, is either installed as routine development or as a response to an emergency. This is an iterative process aimed to maximise the operational time of the infrastructure.

The use of remote monitoring of rural infrastructures in Africa lags in the uptake of different applications in other countries and contexts. Specifically, the lack of a supportive ecosystem and the high up-front cost are critical barriers to the widespread implementation and acute awareness of the potential applications and benefits [115]. However, the rapid pace of mobile phone penetration across all socio-economic groups in sub-Saharan Africa [66] along with the reduction in equipment cost, have paved the way for health-monitoring applications in this context.

Given the varied contexts of health-monitoring applications, the literature also varies on the outcome of prediction models based on the monitoring objectives. Many studies, particularly those related to aviation, value the probability of detection (POD). Some studies attempt to estimate the structural parameters while others apply statistical analysis to the data in order to detect damage [116]. Despite the different outcomes, the emphasis on the reliability of the models remains a crucial area for development across health-monitoring methods as part of successful operations, maintenance, and management of

remotely monitored systems.

3.4 Vibration-Based Structural Health Monitoring

Vibration-based structural health monitoring (SHM) is the analysis of specific structure characteristics (modal parameters) using pattern recognition to detect changes that may indicate extensive wear, degradation, or damage in the structural dynamic behaviour [114]. Specifically, condition monitoring focuses on the detection of damage in rotating machines [117]. Vibration amplitude or frequency is commonly used to measure the response of the degrading system — changes in the structural characteristics over time influence the vibration response characteristics of the structure [118].

Condition monitoring for fault diagnosis related to the “health” of structures and systems remain popular and widely applied in mechanical engineering infrastructure, civil engineering structures, oil industry platforms, aerospace, and composites [119]. The method is often so well-trusted that it has been applied to the monitoring of “high integrity” critical systems, such as jet engines and power-generation facilities [120]. The advancement of technology and widespread availability of nanosensors have resulted in an increase in SHM and condition monitoring for many new applications, including mobile health monitoring to measure the amount of physical activity in patients using either wearable devices or smartphones [121]. The traditional approach to maintenance is reactive and responds to breakage. This ineffective strategy results in decreased productivity due to unscheduled downtime and could have fatal consequences for specific systems. Novelty detection, described in section 3.6.2, enables predictive maintenance in systems, reduce downtime, and facilitate efficient use of the infrastructure.

In 1996, Doebling proposed the first list of categories for damage detection supported by a comprehensive review of technical literature. The measured damage determines the

categories and includes methods based on natural frequency, mode shape, mode shape curvature or strain mode shape, dynamically measured flexibility, matrix update, non-linear methods, neural network methods, and other methods [112].

3.4.1 Level of Identification

Damage in the structure can range from changes in the mechanical, geometric, and material properties to boundary conditions or system connectivity that affect the current or future performance of the infrastructure. With a wide range of techniques, methods, and algorithms available to monitor the condition of systems, it is important to distinguish the level of classification required for a specific application. A classification method based on the level of identification required offers five levels [122] :

Level 1 - *Existence*: Determine that damage is present in the structure

Level 2 - *Location*: Determine where in the system is the damage

Level 3 - *Type*: Establish the kind of damage that is present

Level 4 - *Extent*: Quantify the severity of the damage

Level 5 - *Prognosis*: Predict the remaining service life of the structure

In order to determine the type and extent of damage, prior knowledge related to behaviour parameters of the undamaged state is required. The long-term observations involved with monitoring the system follows this operational evaluation.

3.4.2 Monitoring Using Accelerometry

Machines running under “normal” conditions have a signature frequency spectrum, equivalent to the horizontal section in Figure 3.3. Increased vibrations typically occur as the condition deteriorates, and failure becomes imminent. However, vibration levels increase when the machine is new or recently repaired before settling into the signature vibration in the running-in period. Establishing reasonable confidence intervals for these trip points

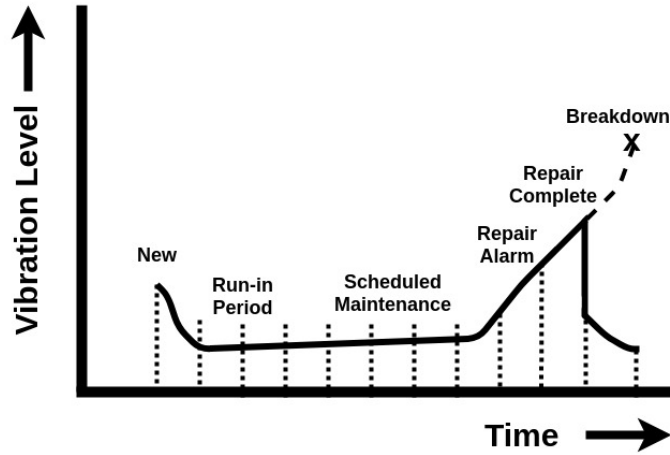


Figure 3.3: A segment of a typical vibration level, here representing one failure event. The vertical and horizontal axes correspond to vibration level and time, respectively, in arbitrary units.

will help to distinguish between these different states.

Accelerometers are the most basic and widely used instruments to measure the system vibration for rotating machines. The vibrations or harmonic motion can be described as the oscillation about a reference point and modelled as $x(t) = X \sin \omega t$. Monitoring vibrations require accelerometers that offer low noise and wide bandwidth. Tri-axial accelerometers use the piezoelectric effect, whereby gravitational acceleration forces stress microscopic crystal structures, to sense dynamic acceleration. Piezoelectric accelerometers are commonly used for condition monitoring applications because the sensors offer good linearity, signal-to-noise ratio (SNR), wide bandwidths and the ability to tolerate high temperatures.

However, recent advances in micro-electromechanical systems (MEMS) capacitive accelerometers offer a selection of new ultra-low powered devices with reduced size and weight as well as the potential for integrated intelligence. The digital nature of MEMS offer an advantage over piezoelectric sensors since the production of devices can be scaled to reduce the cost and offers greater versatility for interface and power supply. The rapid

advancement of industrial IoT, with an emphasis on utilising wireless technologies in order to reduce cabling, has favoured the development of ultra-low powered MEMS. These accelerometers are not only smaller and lighter than their piezoelectric counterparts but also are more shock tolerant (up to 10,000 g) and can offer more advanced features, such as spectral alarms. The main difference between these two accelerometers is bandwidth. MEMS typically offer bandwidths up to 3.3kHz while the more sensitive piezoelectric can monitor over 10kHz. As such, condition monitoring applications where equipment rotates at very low speeds, such as wind turbines, typically use MEMS while piezoelectric accelerometers are more suited to use for fast rotating equipment, such as electric motors.

Vibration analysis using accelerometers is one of the most effective ways to determine the mechanical condition of a device in the early stages of deterioration to allow predictive condition monitoring and timely corrective action.

3.4.3 Prognosis

The outcome of a health-monitoring system is always an alarm to notify that an unexpected event has occurred or, in the case of preventive monitoring, is imminent. The concept of delay time refers to the time between the system becoming defective and the point of failure [123]. Figure 3.4 shows how this two-stage deterioration process separates the “normal” operation from potential failure. Many condition monitoring approaches attempt to identify this initial point, u , when the system becomes defective such that the length of the delay time, h , is optimised [124, 125, 126, 127, 128].

Alarms are triggered when a variable goes outside the safe operating condition or a certain confidence range (for example, $\mu \pm 3\sigma$). Crossing the specific alarm threshold could either call to initiate action from the user or to activate the prognostic unit to estimate the remaining useful life (RUL) of the system. There are two main components

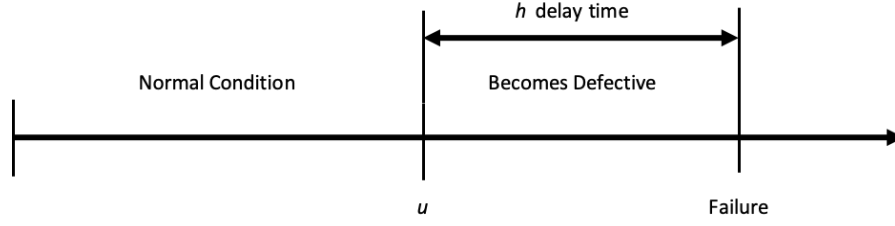


Figure 3.4: Two-stage failure delay time concept [129] attempts to separate normal conditions from failures by introducing the time delay, h , between the point, u , when the system becomes defective and when the failure occurs.

associated with the RUL prognosis: (1) the diagnosis from the health-monitoring system and (2) the uncertainty of the prediction. Often, as the likelihood of failure becomes more imminent, the accuracy of the diagnosed condition increases. If the diagnosis is the log mean of the true remaining life of the unit and log standard deviation of the uncertainty, then the RUL can be predicted by [130]:

$$RUL = \text{lognormal}(\text{Diagnosis}, \text{uncertainty}) \quad (3.1)$$

where $\text{uncertainty} = \ln(RUL_{prior}) * \sqrt{\text{uncertainty factor}}$ and $\text{uncertainty factor}$ is a design variable.

3.4.4 Failure Alarm Threshold

The widespread use of hardware and software in modern systems means that additional alarms add minimal or no additional cost. This resulted in poorly designed alarms with very low efficacy. The Engineering Equipment and Materials Users Association benchmark recommends less than six alarms per hour on a plant. Modern power, petrochemical and oil plants far exceed this, with an average of 48, 54, and 36 alarms per hour, respectively [131]. Early or frequent alarms to trivial problems or due to faulty sensors can lead to operator fatigue or disbelieved alarm information. Establishing the alarm threshold is a crucial step to reduce the number of false alarms. Failure to detect the damage has two

main groups: (1) *false-positive* occurs when the system suspects damage when not present, and (2) *false-negative* occurs when the system does not indicate damage present when in fact present.

Widely used techniques to reduce the false alarm rate include multivariate process monitoring, model-based process monitoring, threshold design, and data processing. Multivariate methods allow more information to be extracted from the individual variables with low signal-to-noise ratio while model-based methods attempt to model the predicted behaviour of the process. Regardless of which approach is used, the final alarm trigger is always binary. The alarm threshold determines the trigger of the prognosis module. There is a direct trade-off between the false alarm rate (FAR) and the missed alarm rate (MAR). The chosen alarm threshold determines these two critical factors. A small threshold will result in an increased FAR while a large threshold will reduce the FAR while simultaneously increasing the MAR. The vibration level threshold responsible for raising the repair alarm would vary between different systems and even between different units of the same type. Given that the normality may even vary within the same system type, some recommendations suggest that the threshold should be set at two to three times (6 to 10 dB above) the signature vibration level [132].

Threshold violations record key parameters related to the event value of the parameter that breached the threshold, the time of the event and the estimator over the limit. In each case, it is common to distinguish between the severity of the event in two categories: (1) warning and (2) alarm. The decision-making process for the attributes related to each of these two states should be determined based on the specific needs of the related health-monitoring application. For example, the importance of the threshold violation, such as those that might be potentially life-threatening, would require action in the warning state.

Careful design for each application is critical to ensure that operators respond accordingly [133]. Alarm thresholds are most often manually selected based on the acceptable FAR and MAR for the desired application. Processing the alarm variable to reduce noise can further reduce the FAR. The exponentially weighted moving average filter is commonly used in industrial data processing for an n -order filter to determine the filtered data, $y(i)$, such that [131]:

$$y(i) = \alpha x(i) + (1 - \alpha)y(i - n) \quad (3.2)$$

where $x(i)$ is the raw data and α is the filter parameter.

An automated alarm setting process describes the steps to select the threshold, T , based on the largest time constant or lowest probability such that [134, 135]:

$$T = \text{Alarm coefficient} \times \text{Baseline}_{\text{mean}(\text{deviation})} \quad (3.3)$$

where the alarm coefficient is indexed and the acceptable baseline for the low probability (6σ) [136] of having false alarms during normal operation is $k = \frac{\sigma}{\mu} < \frac{1}{6}$ for standard deviation, σ , and mean μ , assuming Gaussian distribution. The baseline criterion for the largest time constant is the slope of the first order step function response such that, $\frac{\partial c(t)}{\partial t} = \frac{1}{\tau} e^{-t/\tau} = m$ and $|m| < 0.017$.

3.5 Feature Extraction

The high-rate waveforms, such as vibration data, used to perform health monitoring, produce large quantities of data. Feature selection and extraction is an important step to compress the data and reduce the dimensionality to improve classification performance and speed as well as enhance the understanding of the generated predictions [137].

3.5.1 Hand-crafted Features

Domain expertise can be a powerful tool to guide the data pre-processing step to identify preliminary features from the input space. Additionally, this approach also allows the selection process to incorporate prior knowledge of the structure to enhance accurate estimates of the statistical distribution of the features.

Feature extraction methods for condition monitoring have five categories: (1) statistical time-domain features [138, 139], (2) piecewise approximation [140, 141], (3) frequency-domain transformations (such as Fourier transformation [142]), (4) time-domain transformations (such as principal component analysis [143]), and (5) time-frequency transformations (such as wavelet transformations [144]). Each of these methods offers the ability to extract information not available from the others.

Spectral analysis is widely used as an efficient method to study vibration responses of rotating machines [145, 146]. Fourier analysis transforms time-domain signals to the frequency domain, which allows the analysis of information related to frequency characteristics that are not otherwise observed. The vibration signals, requiring spectral estimation to predict the behavioural pattern of a random signal, reflect the stochastic nature of the grinding process that leads to damage in rotating machines [147]. The non-stationary waveforms of vibration signals are most often analysed using time-frequency methods (such as low-level features of local energy using empirical mode decomposition [148]) or wavelet (such as continuous wavelet transform [149]). Signal noise can result in features that are highly inconsistent due to either the variations in operating conditions or variations in the underlying assets. Related work addresses this feature sensitivity by reducing the noise, sparsity and irregularities in the sensor data.

3.6 Predictive Inference

Unsupervised learning can typically help to detect the existence and location of any damage that is present in the system. In order to further identify the level of structural damage that is present, supervised learning can be applied to determine the type and extent of the damage as well as the anticipated remaining useful life. Given that the objective of this project is to develop an early warning system, any models we use must be able to anticipate the useful life successfully. Table 3.2 provides an overview of the different classification models and inference used in remote health-monitoring applications of various infrastructure types presented in the literature. The following sections will discuss the application of each.

3.6.1 Outcome Classification Framework

One of the simplest linear classifiers is logistic regression, which is often used as a benchmark for more advanced prediction models in fault classification [150, 151]. Other conventional classifiers are also widely used in previous work to process the features. Principal component analysis can detect small shifts in frequency, often using fewer input features than the original set [152], and has proven versatile on a wide range of health-monitoring systems including those based on impedance [153], acoustics [154], as well as vibrations [155]. Support vector machines is a popular kernel method that has shown strong performance in fault detection [156, 157, 158], including real-time health-monitoring [159]. Artificial neural networks have proven robust in determining different kinds of failures as well as differentiating between single and multiple faults for failure events, especially for vibration-based structural damage [160, 161, 162, 163, 151, 164]. Deep learning methods are becoming increasingly popular in machine health-monitoring as computational power and the size of available data continue to increase in order to advance their adoption

beyond the early applications in areas of computer vision, speech/audio recognition and natural language processing. Despite the strong performance of deep learning techniques in detecting failures, the classification process can prove challenging to interpret.

3.6.2 Clustering

Since the “abnormal” class in the existing data set is not balanced or representative of all handpump failure types, it is unlikely that a single machine learning model can assess all possible object classes. Thus, non-model-based statistical pattern recognition approaches are required to complement existing model-based techniques [165]. Novelty detection models the distribution of data under “stable” conditions and estimates the probability of test data as belonging to that distribution. It is particularly useful for examples with large numbers of “normal” examples, and few examples of “abnormal” conditions [166], such as typically occur with high-integrity systems [167]. Examples of this one-class classification use Gaussian mixture models to monitor gearboxes [168] and extreme value theory for vibration spectra energy in jet engines [74].

K-means is a clustering method that calculates the distance between each centroid and the nearest cluster for several iterations to converge the clusters. It can detect anomalies or defects in structures and bearings [169, 151, 170, 171] without labelled data [172] and other applications, including solar systems [173]. Fuzzy c-means proved less successful on a bearing testbed [151] but useful for the discrete wavelet features of a high-speed mill [174]. A multisensory based fuzzy-logic clustering using both vibration and current signals proved effective to diagnose fault severity in bearings [175].

Table 3.2: Relevant related work for vibration-based health monitoring systems applying different inference models.

Authors	Ref.	Year	Infrast. type	Model	Inference
Vibration-based Health Monitoring Studies					
Wang et al.	[161]	2000	Cantilever beams	MLP, FLNN	Stochastic grad. desc.
Yan et al.	[150]	2004	Elevator Door	LR	max. likelihood
Malhi et al.	[152]	2004	Bearing test bed	NN, RBF	Softmax, Eclidean norm
Golinval et al.	[155]	2004	Electrical capacitors	PCA	Angle between subspace
Bulut et al.	[159]	2006	Bridge Model	SVM	min. generalization error
Park et al.	[153]	2007	Bolted Structure	K-means	RMSE
Saravanan et al.	[156]	2008	Spur-level Gearbox	SVM, PSVM	min. generalization error
Jayaswal et al.	[162]	2010	Rolling Bearings	Fuzzy logic, ANN	Stochastic grad. desc.
Perttunen et al.	[176]	2011	Road Surface	SVM with RBF kernel	g-means opt.
Konar et al.	[157]	2011	Induction Motor	SVM	min. generalization error
Yiakopoulos et al.	[169]	2011	Rolling Bearings	K-means	RMSE
Osornio-Rios et al.	[163]	2012	Trus Structure	MUSIC, ANN	Stochastic grad. desc.
Heyns et al.	[168]	2012	Gearbox	GMM	NLL
Soualhi et al.	[158]	2015	Rolling Bearings	SVM, SVR	min. generalization error
Torabi et al.	[174]	2016	High-speed mill	Fuzzy c-means	RMSE
Strączkiewicz et al.	[151]	2016	Rolling bearings	LR, KNN, ANN, k-means	RMSE
Pazouki et al.	[175]	2016	Rolling Bearings	Fuzzy c-means	max. belongness degree
Neves et al.	[164]	2017	Bridge structure	ANN, GP	RMSE
Other Relevant Rural/Remote Monitoring Studies					
Greitzer et al.	[160]	1999	Gas Turbine Engine	ANN	Rule-based
Tong et al.	[154]	2006	Tile-wall Bonding	NN	Stochastic grad. desc.
Varshney et al.	[177]	2015	Solar		
Garcia et al.	[178]	2015	Rainfall/Flood	RF	Presume: OOB
Mehra et al.	[173]	2016	Solar Systems	K-means; GMM	EM
Wilson et al.	[107]	2017	Handpump	GLM, Ensemble	Maximize AUROC
Borchers et. al	[179]	2017	Calving	LMM + NN	Stochastic grad. desc.
Vibration-based Patient Monitoring Studies					
Logan et al.	[180]	2007	Hypertension		MSD

Abbreviations: AUROC: Area under receiver operating curve, EM: expectation maximization, GMM: Gaussian Mixture Model, GLM: Generalized Linear Model, MSD: Mean standard deviation, ANN: Artificial Neural Network, OOB: out-of-bag, RBF: radial basis function RF: Random Forest, SVM: support vector machine, MUSIC: multiple signal classification, NLL: negative log likelihood

3.7 Distributed Measurement Systems

A condition monitoring system can be divided into two main parts: (1) the hardware, and (2) the software. In the past, research predominantly addressed the prior, investigating the angle of the sensors, attempting to optimise the trade-off between cost and accuracy by achieving lower-cost sensors with lower SNR and adjusting the configuration of the sensors to optimise performance in the case of single sensor failure. Review of the literature shows that the focus has shifted to addressing the practical implication of the software regarding the acquisition, collection and processing of the data.

The main developments include:

1. Synchronisation of dynamic data: simultaneous sampling is imperative for models extracted from multi-sensor data. However, aligning timestamps may not always be possible when sensor timers lag behind real-time, particularly in solutions involving GPS based time tagging.
2. Data merging of environmental data and sensor data: sensor data can no longer be used in isolation to estimate the condition or operation of a system. Environmental factors are essential to distinguish factors leading to failure, such as temperature, wind or humidity outside the operating threshold.
3. Centralised base-stations for remote analysis: wireless technology enables the local transfer of acquired data to a remote or cloud-based centralised station in isolated areas or reduces the need for the transmission of unprocessed raw data.

As such, the approaches used in this project will adapt the hardware learning from related water monitoring projects along with the software advances offered by novel distributed measurement systems to overcome over-dependence on single sensor units to

deliver reliable data streams.

3.8 Conclusion

The current chapter offers a review of the previous work related to rural water point monitoring and provides a context (Figure 3.2) of maintenance related to rural infrastructure. The most common implementation of ICT innovations for rural water monitoring is the use of flow sensors. To date, there is no other implementation of vibration monitoring of handpump handles. However, vibration analysis is one of the most common methods for structural health-monitoring and condition monitoring applications. The availability of high-rate waveforms allows the application of more advanced machine learning methods to monitor deterioration.

Furthermore, while remote monitoring has the potential to revolutionise the current management structure of rural infrastructure by shifting the maintenance burden and offering the ability to track progress towards development goals, the number of implementation remains sparse despite the rapid penetration of mobile data across rural sub-Saharan Africa. Thus, the primary goal of this thesis is to demonstrate the successful implementation and potential of bridging the chasm between rural monitoring and the addition of intelligent features offered by advances in machine learning. The next chapter reviews the sensors developed for data collection before presenting the resulting data sets used to develop and evaluate the novel handpump health-monitoring system developed in this thesis.

Chapter 4

Data Sets

This thesis investigation used subsets from three databases designed and collected as part of this study: Kenya Original ($\mathcal{D}_o$), Kwale Failures ($\mathcal{D}_f$) and Kwale Long-term ($\mathcal{D}_l$). The first database, $\mathcal{D}_o$, was used to train and validate the sensor-based novelty filter described in Chapter 5 and $\mathcal{D}_f$ was used to train and validate the distributed inference method proposed in Chapter 6. The third database, $\mathcal{D}_l$, has been used to demonstrate early detection of deterioration prior to known failure events. $\mathcal{D}_o$ used an off-the-shelf logger to capture the accelerometer waveforms connected via Bluetooth to a nearby laptop. $\mathcal{D}_f$ and $\mathcal{D}_l$ used a custom-designed and built logger to store the accelerometer waveforms locally. Each database contains the timestamps and waveforms from the three orthogonal axes of accelerometry. Additionally, $\mathcal{D}_f$ and $\mathcal{D}_l$ contain the accompanying sensor-based novelty scores produced by the embedded hardware which will be used for analysis and validation in Chapter 6.

The data used in this investigation are the first and largest of its kind to contain real-world high-frequency accelerometry waveforms from handpumps with accurate manual labels. The databases and code have in part been made publicly available ¹.

¹<https://www.robots.ox.ac.uk/~davidc/> where $\mathcal{D}_f$ and $\mathcal{D}_l$ are in part available via [86].

4.1 Data Collection

Data collection for this thesis took place during multiple fieldwork visits in Kwale County, Kenya, over three years between 2016 and 2018. As this is a novel data stream, the process included many iterations that improved with every data collection campaign. The *Afridev* is the most commonly-used handpump in Kenya and originally field-tested in Kwale County [2]. The eldest operating handpumps included in these databases were installed in 1989, along with the original installation certificate. Since the handpumps in this region were installed under the same programme in the early 1980s (discussed in Chapter 2), implementation fundamentals, such as siting, drilling techniques, and technology, as well as community participation are relatively constant. Data recorded from a small number of *India Mark III* handpumps installed in the study region; however, are not included in the databases used for this thesis but instead included in some comparative analysis in the discussion about future work.

4.1.1 Measuring Handpump Vibrations

To measure the vibrations in an operating handpump, a sensor, containing a consumer-grade IC-based accelerometer with a sampling frequency of 96 Hz, is attached to the handpump handle positioned closest to the pump body, as shown in Figure 4.1a and 4.2, without restricting the handle range of motion [80]. Long-term data collection sensors were bolted to a fixed position inside the hollow handpump handle. The vibration sensor records the resulting accelerometer data in three orthogonal dimensions related to the force applied to the piezoelectric material of the accelerometer that is directly proportional to the gravitational force applied during the movement of the handle. A five-second interval of data from a pump in the working and non-working condition is shown in Figure 4.1b-4.1c. From this it appears that there are more vibrations present under normal handpump

conditions in Figure 4.1b than abnormal conditions in Figure 4.1c.

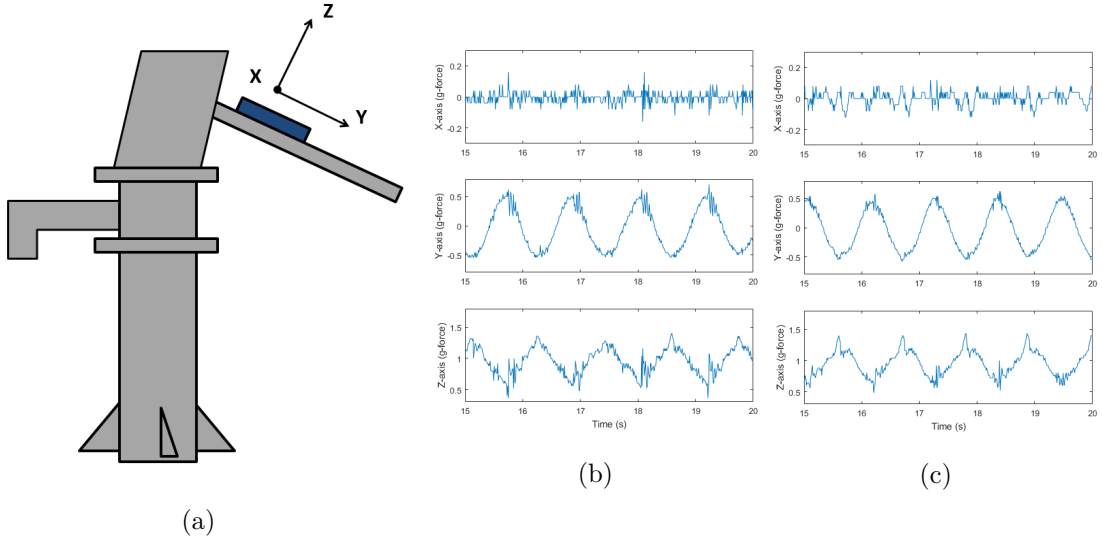


Figure 4.1: (a) A diagram showing the experimental set up of a vibration sensor attached to a surface-mounted handpump and the relevant orthogonal axes of accelerometry with a five-second interval of unprocessed accelerometer data of the same pump in a (b) normal and (c) abnormal condition in the X, Y, and Z dimensions (upper to lower plots, respectively) for the same user.

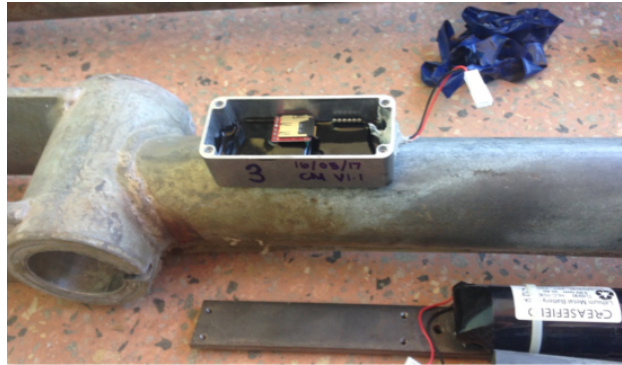


Figure 4.2: Example installation of a data logger with a specific sensor-based novelty filter version. The final placement is secured with bolts inside the handle.

Figure 4.3 provides a look at the setting of some of the handpumps we visited as part of this work. Our work often drew a small curious crowd and at those handpumps frequently visited, we made many friends and sadly said goodbye to some during this study. On every visit, we respectfully asked permission to use the handpump and sometimes had to wait a long time for village elders to be tracked down and consulted.



Figure 4.3: Behind the scenes snapshots of the various handpump locations we visited and the remote villages in Kwale County where we lived for many months during the course of this thesis in order to ensure we collect the required data. Handpump locations are mapped here: <https://goo.gl/Vo0eFx>.

4.1.2 Experimental Setting

The sampling strategy involved shortlisting suitable handpumps from the existing, albeit dated, Kwale County water point database. Clusters of potential handpumps were selected within a 1 to 2 km radius. Once shortlisted, a physical inspection during site visits could determine the suitability of the handpumps. Many shortlisted handpumps were unsuitable for data collection as the nearby community abandoned them after breaking down or drying up. Conversely, given the duration of this study, some of these pumps were repaired and included in later visits.

To ensure the data set represents sufficient variability in the pumping characteristics,

we include recordings from different pumps and different users. The data set contains recordings collected for both members of the project team as well as volunteers from the local community at 48 different Afridev pumps across the county (with the geographical locations mapped here: <https://goo.gl/Vo0eFx>).

The well depth, as determined by the number of rods in the pump, ranges from 7 to 46 m, which is the full operating range of a standard Afridev pump. In the database, 50.4% of the recordings are from deep wells, operating at a depth of more than 25 m below ground level, and the remaining from shallow wells, operating at a depth of less than 25 m below ground level.

4.1.3 Fault Experiments

In addition to the 48 pumps sampled during the field visit, long-term data loggers containing vibration sensors were installed at ten handpumps to track the deterioration of handpump mechanics leading up to the breakdown event over time. Retrospective labelling of these data could prove challenging as it would rely on manual and sometimes incomplete notes from the field mechanics. To mitigate the risk of missing failures or labels, $\mathcal{D}_f$ was collected to ensure the data set contains a comprehensive representation of many different failures.

By collaborating closely with a local Kenyan maintenance company, FundiFix, who are responsible for handpump repairs in the study area, we were notified of any reported failures. This collaboration gave us access to a valuable repository of data on known handpump failures. We worked closely with the local mechanic to respond to failure reports from the community, which provided examples of handpump abnormalities. Failures were usually reported first thing in the morning when people arrived to collect their water for the day. Over five months, we were often woken by a call from the mechanic at 5:30 AM to tend

to broken handpumps in order to record data pre- and post-repair. Although this meant we would not have recorded data for the deterioration of the handpump, we recognise that this limitation was preferred over the alternative of fault-seeding experiments (in which failures are artificially induced) as the data set is representative of real failures.

4.2 Data Labelling

The literature remains unclear about a unified definition for handpump functionality, as discussed in section 2.4, and currently there exists no standardised labelling system for handpump conditions. As such, we propose a category of labels based on combined insights from the literature summary in Table 2.3 and in-depth interviews with several handpump experts. The proposed labelling system considers the effect of mechanical performance on normality and uses the short-term water quantity as a “snapshot” in time to label the data, as shown in Table 4.1. As per the binary definition of handpump “functionality” in Chapter 2, a handpump condition with normal water flow is labelled one and those with no water flow are labelled zero. These are the condition labels used for the two-class classification.

The state of a handpump can be determined by binary classification as the pump being either functional ($Y = 1$) or non-functional ($Y = 0$). Thus, given input data, X , derived from the accelerometer mounted on the handpump handle, the probability that the inputs belong to the class ($Y = 1$) for the case when the handpump is functioning can be presented as:

$$P(X) \triangleq P(Y = 1|X) \tag{4.1}$$

The data were labelled using notes collected during in-person, contemporaneous observations. This level of labelling is limited in that it allows for only two classes. It is clear

Table 4.1: Description of the mechanical condition and short-term water quantity classification labels assigned to the recordings of each pump, where condition one represents a functional or normal handpump and condition zero refers to a non-functional or abnormal handpump. Water flow refers to a qualitative interpretation of the handpump condition.

Condition Description	Water Flow	Condition
Excellent working condition	High	1
Noisy but working	Average to Low	1
Dry borehole	None	0
Rising main leak	None	0
Broken U-seal	None	0
Worn U-seal	Average to low	1
Handpump body leaking	Average to low	1
Worn bush bearing	Average	1
Stiff handle	Average	1
Other	Average to low	1

that certain conditions, like those with average or low flow, are not entirely normal nor entirely abnormal. However, for this thesis, it is assumed that the proposed labels are adequately descriptive.

4.3 Kenya Original Database

In March 2016, the Kenya Original Database, $\mathcal{D}_o$, acquired short-term accelerometry data from 28 different handpumps over two weeks using a commercially available logger. Each recording is collected at 96 Hz for the duration required to fill a predefined water container, normally 20 litres, and on average has a duration from 30 to 180 seconds. From this database, which contains examples of all the labelled mechanical conditions across 28 pumps, a subset of the data were used to construct a training set for the sensor-based novelty filter described in Chapter 5. Given that the data were collected over two weeks, the data set contains very few examples of failure events (7.7 %), representing only three of the failures types identified in Table 4.1. These include normal pumps in excellent working condition and abnormal cases, including a dry borehole, a leak in the rising main and a broken U-seal. This resulted in a $M \times N$ matrix, where $M = 64$ is the number

of frequency bins per window and $N = 9221$ is the total number of windows across all recordings in $\mathcal{D}_o$.

Balancing the data set is an important step that helps prevent bias in the model by ensuring that one or more classes are not under-represented. Even though handpumps break frequently, users tend to discontinue use after failure, which results in limited data collection from broken handpumps while operated. To ensure the generalizability of the model to previously-unseen data, it is important that the final data set includes the same number of normal and abnormal points for each failure case as not to bias the results in favour of either condition. As a result, the final data set, $\mathcal{D}_{o,b}$, contains $N = 3036$ windows with equal-sized representation of normal and abnormal conditions, after sub-sampling the larger class. $\mathcal{D}_{o,b}$ was randomly subdivided into training and validation ($\mathcal{D}_{o,b1}$) and test ($\mathcal{D}_{o,b2}$). $\mathcal{D}_{o,b1}$ contains 2377 windows (78.3% of the balanced data) and $\mathcal{D}_{o,b2}$ contains 659 windows (21.7% of the balanced data).

4.4 Kwale Failures Database

From February to June 2017, we acquired the largest database of real-time failures $\mathcal{D}_f$ using a bespoke logger with the novelty filter on-board. Following the original investigation and hardware design in Chapter 5, a large-scale, long-term data collection campaign followed to establish a database with more representative data identified during the early investigation of this study. To understand the features and performance of various classifiers, the data within set $\mathcal{D}_f$ were subdivided into groups depending on the handpump operating depth, described in table 4.2. The three proposed data sets ($\mathcal{D}_m$, $\mathcal{D}_{d,1}$, and $\mathcal{D}_{d,2}$) represent typical scenarios for condition classification using a novelty filter designed for both inter- and intra-handpump operation and deep- and shallow-operating handpump groups.

The first data set, $\mathcal{D}_m$, represents a general inter-handpump system consisting of twelve

Table 4.2: Overview of the three proposed handpump vibration data sets used in this thesis.

Set	Description	Range
$\mathcal{D}_m$	Eleven handpumps (both deep and shallow)	6-53m
$\mathcal{D}_{d,1}$	Eight deep operating handpumps	33-54m
$\mathcal{D}_{d,2}$	One deep operating handpump (broken pump rod)	54m

different handpumps of varying operating depths, ranging between 6 m to 53 m, and was included to establish the baseline performance of a general classifier. The second data set, $\mathcal{D}_{d,1}$, represents a deep-operating inter-handpump system consisting of eight different handpumps operating at depths between 33 m to 54 m. The third data set, $\mathcal{D}_{d,2}$, represents a deep-operating intra-handpump system of one handpump operating at 54 m. Although the implementation of a region-wide intra-handpump system is unfeasible, this data set was selected to investigate the influence of different failure types, while controlling for the handpump.

The failures database $\mathcal{D}_f$ contains recordings from eight different common handpump failure types. As with $\mathcal{D}_o$, the data sets were balanced and randomly divided into a training-and-validation set (80%) and a test set (20%).

Investigations from the previous database, $\mathcal{D}_o$, showed enhanced performance for data containing shallow wells exclusively. It may be that the features of these wells are more distinct since the physical location will be closer to the sensor than those in deeper wells. Considering these findings and given that deep operating handpumps are more likely to break due to the additional force required to lift the additional rods, the focus for our extended data collection campaign was to prioritise data collection of failures from deeper wells. Figure 4.4b shows the disparity in the representation of deep and shallow wells in the additional data set, where the majority of handpumps are deep wells. The depth

range for shallow wells is 6-21m and 33-54m for deep wells. As with all the other data sets and novelty detection in general, the failure class is under-represented in the data set, shown in Figure 4.4a. The abnormal class (C0) contains data from only two of the eight handpumps, while the normal class (C1) has more than four times the number of samples. This imbalance is often the case in health-monitoring problems where failures are infrequent or result in the disuse. Finally, following the shortcoming of $\mathcal{D}_o$ to include data from deep operating pumps, which tend to break more often, this database $\mathcal{D}_f$ focused on collecting more data from deeper wells, as shown in Figure 4.4b.

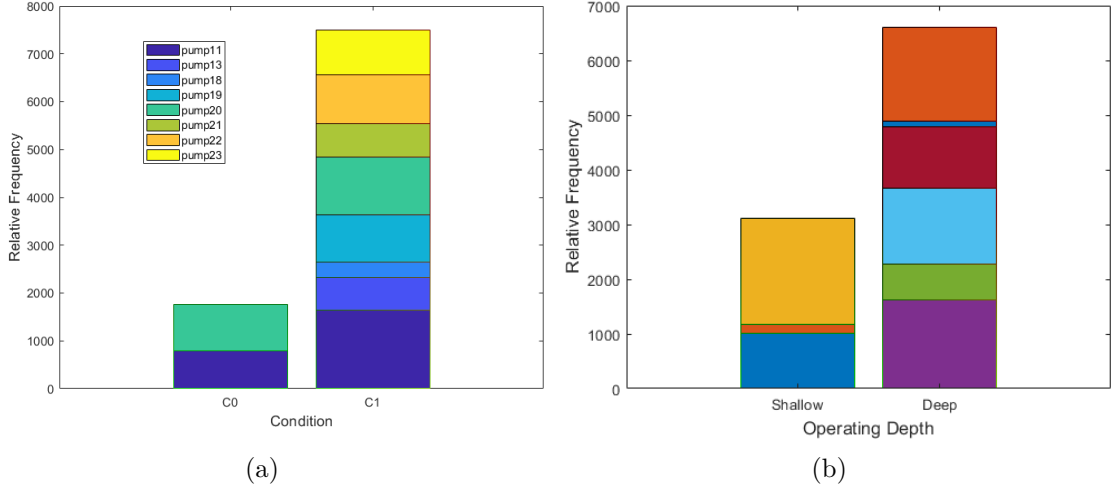


Figure 4.4: Overview of the representation of different handpumps in the data set grouped by: (a) condition, where C0 is abnormal and C1 is normal, and (b) depth of the well, where shallow is < 25 m and deep is > 25 m. Each example is 128 samples, equivalent to 1.3 seconds of time-series data.

4.5 Kwale Long-Term Database

After June 2017, we selected sampling clusters of co-located handpumps for long-term monitoring using the bespoke loggers, which continued until early 2019. Each cluster consisted of three handpumps within a 1 km radius to control for variations in the shallow aquifer feeding the wells. The $\mathcal{D}_l$ data set contains ten handpumps fitted with permanent loggers for 18 months. The performance and longevity of the bespoke loggers proved

rather challenging under the operating conditions. The extreme heat and moisture in this highly humid region meant that the sensor and other parts were sometimes operating outside the recommended conditions, which resulted in malfunctioning of the equipment in what seems random times. Of the ten handpump-mounted sensor systems, three needed replacing during the long-term data collection campaign. Additionally, these data had to be retrospectively labelled using the notes from the local maintenance company, which means that the time and date of failure events are only approximate rather than exact. Fortunately, given the insight gained from acquiring data set $\mathcal{D}_f$, it was possible to label the failure events with a sufficient level of certainty. The labelled events from this data set were used to validate the distributed inference monitoring method proposed in Chapter 6.

4.6 Conclusion

This chapter presents the three handpump databases that were created and used in this thesis: Kenya Original ($\mathcal{D}_o$), Kwale Failures ($\mathcal{D}_f$), and Kwale Long-term ($\mathcal{D}_l$). Since the data collection happened in parallel to the analysis of this thesis, it was possible to amend the sampling strategy as the research evolved. The databases include a broad subject range from a vast geographical region to include different geological compositions for the aquifers feeding the handpumps and also an accurate representation of the well-known major failure types.

The Kenya Original database was balanced to create $\mathcal{D}_{o,b}$, representing 28 different handpumps and three different failure types. It was used to train and validate the sensor-based novelty filter on the bespoke logger used to collect the subsequent two databases. The Kwale Failures database, $\mathcal{D}_f$, was designed to collect data with more representation from deeper wells and different failure types. It contained examples from eight different failures types and was used to create three data sets, $\mathcal{D}_m$, $\mathcal{D}_{d,1}$, and $\mathcal{D}_{d,2}$, based on the op-

erating depth of the well. These data sets were used to design and validate the distributed inference platform described in Chapter 6. Finally, the Kenya Long-term database contains continuous usage data from ten handpumps over 18 months. This data set was used to validate the early warning score from the distributed inference platform on real-world data. Each of these databases contains waveforms from the three axes of the accelerometer and accompanying timestamps. In addition, $\mathcal{D}_f$ and $\mathcal{D}_l$ contain the novelty filter scores for every 128 samples. The following chapters will describe how these data sets were used to develop and validate the sensor-based novelty filter and the subsequent cloud-based distributed inference system for the health monitoring of handpumps.

Chapter 5

Sensor-based Novelty Filtering

The previous chapters demonstrated the operational need for robust remote monitoring of rural systems, given the unique operating environment and limited resources that make monitoring these systems particularly challenging. An extensive review of the literature provided an overview of the use of vibration monitoring in conventional remote condition monitoring. It revealed that existing methods rely either on transmitting a large amount of data from the sensor to be processed by a remote client-server or performs direct classification by logic rules and thresholds. With the aim to develop affordable and scalable predictive maintenance technologies for rural systems, the following chapters of this thesis design, test, and validate a novel method for monitoring the condition of a rural node, which can be implemented without the need for expensive hardware.

This chapter describes the development of a sensor-based technique to classify the condition of the system by using a classifier embedded in the rural node to analyse the different vibration features. This technique should be low in computational cost, because (i) the CPUs onboard are usually small and (ii) preserve the battery. Such a system should be sufficiently reliable, yet low-cost and scalable, to be deployed across entire system networks in rural regions. The five sensor-based classifier techniques considered are based on: logistic regressor (LR), naïve Bayes classifier (NBC), decision trees (DT), k-nearest neighbours (kNN) and neural networks (NN). These binary classifiers mapped

the outcomes of the function as either “normal” or “abnormal”. The performance of the techniques is compared to support vector machines (SVM) that are used as the gold standard, assuming no hardware processing constraints. However, since these are all feed-forward classifiers, generating informative features is essential. Thus, the first section of this chapter will visualise the data, analyse the spectral data to investigate any structure that may be present, select a suitable accelerometer axis and perform principal component analysis (PCA) to select the best features. The final section of this chapter will compare the performance of the different classifiers.

5.1 Lightweight Classifiers

Classifiers map the feature space to the correct label or category. The algorithms are either discriminative classifiers, such as LR and SVM, that use the training data to learn a mapping function which predicts the corresponding labels or generative classifiers, such as naïve Bayes, that model the joint probability of the input features and the labels. Some classifiers construct the models before receiving new test data, known as *eager learners*, while others defer the decision until the test query by storing the training data, known as *instance-based* or *lazy learning*. Eager learners tend to require more time for learning and less time for prediction; making them faster during prediction and thus ideal for applications that require fast response times or offer limited memory during run-time. To understand the other advantages and limitations of each, this section will review the classifiers based on being eager or lazy.

5.1.1 Eager Learners

Eager learners construct a classification model using a given set of training data with the main advantage being that this speeds up the classification of new test data. Although this has the advantage of requiring less time for predicting, models have to commit to a single

hypothesis to cover the entire instance space, which could make them less accurate than lazy learners. Given that this meets the criteria for a sensor-based condition monitoring system, four of the proposed methods are eager learners, including LR, NBC, DT and NN.

Logistic Regression

Logistic regression (LR) is a linear classification method used in statistics to categorise the dichotomous dependent variable and predict the probability (0,1) of membership of one class (e.g., *true/false*) in a two-class setting. In machine learning, LR can be used to model the posterior probability of input variables, $\mathbf{X}$, associated with a class by fitting a linear model to the feature space. This approach is quick to train and very fast to classify new records, making it suitable for applications with limited processing power and bandwidth. A linear decision boundary restricts the natural probabilistic views of class predictions; however, this does help to prevent over-fitting of the model and improve the generalisability of the model.

LR can be formulated using the sigmoidal hypothesis function, $\mu_i = \sigma(\mathbf{w}^T \mathbf{x}_i)$, to determine the probability that a given example is of class 1 [181] where σ is a sigmoid:

$$p(y = 1|\mathbf{x}, \mathbf{w}) = \text{Ber}(y|\sigma(\mathbf{w}^T \mathbf{x})) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}_n}} \quad (5.1)$$

where σ is a sigmoid and $\mathbf{w}$ is a set of weights assigned to each input feature, $\mathbf{x}_n$. The decision threshold, T , is used to assign a given example to class 1 based on whether the hypothesis function is greater than or less than T . This will initially be set to $T = 0.5$; however, Chapter 6 will investigate dynamically adjusting the threshold.

The logistic function is a cumulative logistic distribution used to measure the relationship between the dependent (output) variable and the independent (input) variables. The

function assumes a logistic distribution of errors. Given that the dependent variable is categorical (binary), it follows that the conditional distribution $y|x$ is a Bernoulli distribution rather than a Gaussian distribution.

By using maximum likelihood estimation (MLE), we calculate the cross-entropy by minimizing the negative log-likelihood (NLL):

$$\text{NLL}(\mathbf{w}) = - \sum_{i=1}^N \left[y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i) \right] \quad (5.2)$$

where the gradient, $\mathbf{g} = \sum_i (\mu_i - y_i) x_i$ and Hessian, $\mathbf{H} = \sum_i \mu_i (1 - \mu_i) x_i x_i^T$. Given that $\mathbf{H}$ is positive definite, NLL is convex with a unique global minimum which can be found using gradient descent, $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta_k \mathbf{g}_k$, for some learning rate, η_k .

The coefficients should be optimised to minimise the error in the predicted probabilities; however, minimizing the NLL leads to over-fitting of the model as the predictions match the truth. To discourage the weights from over-fitting to the data, large values in $\mathbf{w}$ are penalised by adding the ℓ_2 regularisation term such that the new objective can be described as:

$$f'(\mathbf{w}) = \text{NLL}(\mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w} \quad (5.3)$$

where λ is a hyperparameter that governs the degree of regularisation. This results in the new gradient, $\mathbf{g}'(\mathbf{w}) = \mathbf{g}(\mathbf{w}) + 2\lambda \mathbf{w}$, and Hessian, $\mathbf{H}'(\mathbf{w}) = \mathbf{H}(\mathbf{w}) + 2\lambda \mathbf{I}$.

The ℓ_2 regularisation is equivalent to placing a non-zero Gaussian prior on the feature weightings from a Bayesian perspective.

Naïve Bayes Classifier

The probabilistic naïve Bayes classifier (NBC) algorithm is inspired by the Bayes Theorem,

$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$, and offers a simple, fast classifier based on the assumption that

predictors contribute equally and are independent. By deriving the maximum posterior probability for the maximal $P(C_i|X)$ and applying the naïve assumption that all variables are conditionally independent for class-conditional distribution, $p(\mathbf{x}|y = c)$, such that the NBC model can be described by [181]:

$$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^D p(x_j|y = c, \boldsymbol{\theta}_{jc}) \quad (5.4)$$

where the multivariate Bernoulli naïve Bayes model is created in the case of binary classification by using the Bernoulli distribution, $p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^D \text{Ber}(x_j|\mu_{jc})$, and μ_{jc} is the probability that feature j is in class c .

As before, we use MLE to fit the NBC model. For binary classification where $x_j|y = c \text{ Ber}(\theta_{jc})$, the MLE can be described by:

$$\hat{\theta}_{jc} = \hat{P}(x_j = 1|y = c) = \frac{N\hat{P}(x_j = 1, y = c)}{N\hat{P}(y = c)} = \frac{N_{jc}}{N_c} \quad (5.5)$$

where N_c is the number of class examples and N_{jc} is the number of class examples where feature j turns on.

Even though predictors in real-world data are rarely entirely independent, NBC has proven to work quite well in complicated real-world settings [182].

Classification and Regression Trees

Classification and regression trees (CART) is a non-parametric method that uses a sequence of decision rules inferred from prior data to categorise any new data. Features are repeatedly split into binary subsets or nodes until no longer possible, where the terminal node is assigned to one of the classes [183]. Each axis split creates a corresponding region, R , that can be associated with a class label in each leaf to create a piecewise constant

surface defined by [181]:

$$f(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] \sum_{m=1}^M w_m \mathbb{1}(\mathbf{x} \in R_m) = \sum_{m=1}^M w_m \phi(\mathbf{x}; \mathbf{v}_m) \quad (5.6)$$

where w_m is the class response for each of the m regions and $\mathbf{v}_m$ captures the threshold value and splitting variable from the root to the m 'th leaf.

Splits are based on achieving the maximum information gain in order to minimize the cost or node error, defined as the proportion of misclassified classes at that node [181]:

$$\frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathbb{1}(y_i \neq \hat{y}) = 1 - \hat{\pi}_{\hat{y}} \quad (5.7)$$

where $\mathcal{D}$ is the data in the leaf and the most probable class label is $\hat{y}_c = \operatorname{argmax}_c \hat{\pi}_c$. The depth of the tree, T , is limited in order to prevent overfitting [184]. The tree-like structure of the model makes it very intuitive to interpret.

Feedforward Neural Network

Neural networks (NN) are parameterised models that are optimised through continuous learning and based on the universal approximation theorem. A series of logistic regression models is known as feedforward neural network or multi-layer perceptron (MLP) [185]. For a neural network with a single hidden layer, one input layer, and one output layer, forward-propagation is used to calculate the output, $\hat{y}$, of each layer for a given input, X , such that:

$$\hat{y} = A^{[2]} = \sigma(Z^{[2]}) \quad (5.8)$$

where the intermediary output of the first hidden layer, $Z^{[1]} = W^{[1]}X + b^{[1]}$ such that the final output of the first hidden layer be $A^{[1]} = \sigma(Z^{[1]})$ and the sigmoid activation

function is $\sigma(x) = \frac{1}{1+e^{-x}}$. The intermediary output of the output layer can be defined as $Z^{[2]} = W^{[2]}A^{[1]} + b^{[2]}$. W is the randomly initialised weight associated with the neuron and b is the bias associated with the neuron, initialised as zero. Finally, gradient descent is used to maximise the logistic regression log likelihood [186]:

$$L(W, b) = \sum_{i=1}^m (y^{(i)} \log a^{[2](i)} + (1 - y^{(i)}) \log(1 - a^{[2](i)})) \quad (5.9)$$

Also known as training the neural network.

5.1.2 Lazy Learners

Lazy learners are algorithms that require little or no training time since they defer processing until after the query. As such, these are also known as memory-based or instance-based classifiers since they store the training data until the instance inference is required to generalise the data, rather than learn any model. Although these learners require more time to predict, the overall accuracy can often be higher than eager learners since this method tends to use an enriched hypotheses space rather than selecting a single hypothesis. Although the memory requirement of lazy learners makes them unsuitable for onboard processing, we include one lazy learner for reference.

K-Nearest Neighbours

K-Nearest Neighbours (k -NN) is a non-parametric classifier that determines the object of the class based on the majority vote of its k nearest neighbours, where instances are represented in the Euclidean space. The approach is rather straightforward as the input variables are the features used to classify each object [187]. The K points in the training set that are nearest to the input, $\mathbf{x}$, are counted as members of the class in this set such

that [181]:

$$p(y = c | \mathbf{x}, \mathcal{D}, K) = \frac{1}{K} \sum_{i \in N_K(\mathbf{x}, \mathcal{D})} \mathbb{1}(y_i = c) \quad (5.10)$$

where $N_K(\mathbf{x}, \mathcal{D})$ are the indices of the points, K and the indicator function is used to determine the probabilities for this distance metric:

$$\mathbb{1}(e) = \begin{cases} 1 & \text{if } e \text{ is true} \\ 0 & \text{if } e \text{ is false} \end{cases} \quad (5.11)$$

This approach is robust to noisy training data but has a high computation cost since voting takes place at run-time [188].

5.1.3 Classifier Comparison

The LR model can easily be updated with new data when using the online gradient descent method and is not impacted by the correlation between the features. The probabilistic framework allows the classification threshold to be dynamically adjusted or choose not to classify at all when the classifier is unsure based on the calculated confidence intervals.

NBC assumes conditional independence between features and therefore requires very little training data to converge, especially in cases when the naïve Bayes conditional independence is true. The assumption of independence allows each distribution to be estimated as a one-dimensional distribution, which can help overcome the curse of dimensionality in many cases. However, this over-simplified assumption of independence is also the main disadvantage as the classifier is unable to learn interaction between features, which often makes it a very poor estimator.

CARTs are very intuitive and easy to interpret. Given their non-parametric nature, the models tend to perform well even when presented with outliers. The main disadvantage

of the decision tree is that this approach would not allow online learning, but instead requires re-learning whenever presented with new data.

Although NNs tend to outperform most machine learning algorithms, the “black-box” nature of the process means it is not very well suited for applications where interpretability is a priority. NNs perform better when trained with more data points; however, training these is computationally intensive and thus not able to retrain on the embedded system. Wide NNs can perform sufficiently well to approximate any function; however, these are not widely generalisable.

All these algorithms are fast to train and feasible to implement on an embedded system with limited processing capabilities and power constraints. In the results section of this chapter, the performance of each method will be compared using the selected data set before proceeding to discuss the final selection, taking into account the advantages and limitations of each method highlighted here.

5.2 Heavyweight Classifier

We use the support vector machine as a comparison to the proposed *lightweight* methods.

5.2.1 Support Vector Machines

The support vector machine (SVM) is a classification algorithm used to recognise patterns for analysis by forming a decision boundary in some feature space [189, 190]. The hyperplane is optimised to maximise the margin between the data points near the boundary in high-dimensional feature space in which the decision boundary is linear [191]. The kernel trick can enable the classification with a maximum-margin hyperplane in some high-dimensional feature space, F [192, 193], with a set of labels $\ell_1 \dots \ell_N, \ell \in \{-1, 1\}$.

The classifying SVM estimates the value of the hyperplane, $y(\mathbf{x}_n)$, at each of the

training points, $\mathbf{x}_n$, such that the linear hyperplane is of the form [194, 195]:

$$y(\mathbf{x}_n) = \mathbf{x}_n \cdot \mathbf{w} + b \quad (5.12)$$

where $\mathbf{w}$ is the support vector weights and b is the bias. The distance between each data point and the hyperplane is described as $\frac{|y(\mathbf{x}_n)|}{\|\mathbf{w}\|}$. For the case where the classes of data are linearly separable in $\mathbf{X}$, this decision boundary is defined by $y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1$.

Given that “real-world” data are not always noise-free or linearly separable, a *soft-margin* SVM can allow mislabelled data while still maximising the margin. The “slack variable”, ξ_n , measures the degree of misclassification and is introduced such that $y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n$ for $\xi \geq 0$. If the data point, $\mathbf{x}_n$, lies on or within the correct boundary then $\xi_n = 0$, otherwise $\xi_n = |\ell_n - y(\mathbf{x}_n)|$. The cost hyperparameter, C , is introduced to penalise misclassification and acts as a trade-off parameter of the model between the margin size and the amount of error tolerated in training. The value of C as a penalty factor is analogous to a regularisation parameter and plays an important role in the fitting of the model. A large value C for non-separable points will over-fit while a low penalty will result in under-fitting of the model since decision boundaries are more slowly-varying in the feature space, with less curvature.

The hyperparameters of the soft margin SVM are optimised by minimising the primal form of the classifier:

$$f(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi_n \quad (5.13)$$

where minimising $\|\mathbf{w}\|^2$ is equivalent to maximising the distance between the hyperplane and the closest data point, written as $\|\mathbf{w}\|^{-1}$. K-fold cross-validation is used to find the optimal combination of the slack variable, ξ , and kernel parameter, C , which will result in

high accuracy without over- or under-fitting of the model.

In order to use the “kernel trick” to project data into a high-dimensional space where two classes may be linearly separable, the primal form is considered in the dual form:

$$f(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{j,k=1}^N \alpha_j \alpha_k \ell_j \ell_k \mathbf{x}_j^T \mathbf{x}_k \text{ for } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^N \alpha_i \ell_i = 0 \quad (5.14)$$

where α is another vector of weights, with $\mathbf{w} = \sum_{i=1}^N \alpha_i \ell_i \mathbf{x}_i$. A kernel function, describing the features in relation to each other, is used rather than the actual features vectors that define each $\mathbf{x}_n$. The most commonly used kernel, the Gaussian radial basis function (RBF) kernel, interprets the similarity between two points and can be described as:

$$k(\mathbf{x}, \mathbf{x}') = \exp \left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2} \right) \quad (5.15)$$

Implementation was performed in MATLAB using the public open-source LIBSVM environment [196].

5.3 Method

5.3.1 Onboard Pre-processing

Pre-processing includes several steps to prepare the data for feature extraction and ensure hardware compatibility. Time-domain recordings are transformed to generate frequency-domain features that model handpump conditions, where the mechanical failures are most likely to be obvious in the frequency domain [197, 198]. In the following section, we describe the steps taken to convert the time signal to the frequency domain. Since these steps are performed onboard the embedded system, wherever applicable, this section will also discuss the steps taken to overcome the limitations imposed by such a system.

(1) Vectorisation

Each recording is summarised by a *shape vector* x to describe the underlying sinusoidal waveform. Each vector consists of a set of points $(t, x) = (t_i, x_i)_{i=1}^N$, which represent the measured gravitational force x_i at time t_i such that $X = [x(i)_1, x(i)_2, x(i)_3]$ for each orthogonal axis of the accelerometer. The signal waveform contains vibration noise such that each $x = f(t) + \eta$, where $f(t)$ is the function describing the underlying periodic waveform of the pumping motion and η is the noise in the signal. Unlike other signal processing applications which assume the noise to be independent of the signal, the heteroscedastic nature of the noise is shown in Figure 4.1, where the signal appears to have more noise on the downward turn of the curve in the Y-axis and similarly on the upward turn of the curve in the Z-direction when compared to rest of the signal.

(2) Peak and Trough Detection

After creating the shape vector for the waveform of each recording, the first pre-processing step was to extract the peaks and troughs of the pumping waveform and use them to perform a subsequent assessment of data integrity. For waveforms on the Y- and Z-axes, we chose periods with a difference of $\Delta t > 0.25$ s between peaks and troughs and $\Delta t > 0.015$ s for the signal measured in the X-direction. This removes any motion detected not associated with the regular use of the handpump, such as removing the handle in the case of repairs or children playing on the handpump.

(3) Filtering

The second pre-processing step was to filter each unique vibration waveform in order to separate the underlying waveform from the vibration noise. Each waveform is filtered using a phase-corrected 4-point moving average finite impulse response (FIR) filter to represent the *shape* of the recording, which we remove from the original signal. Trail-

and-error optimised the window size by minimising the length to retain the detail in the signal while maximising the window to smooth the filtering. This process acts as a high-pass filter (HPF) to remove the low-frequency components associated with the pumping tempo that is not indicative of handpump failure, and which retains an estimate of the fast-moving trend components of the noisy vibrations. We hypothesise that changes in pump conditions are not affected by the component of the signal caused by the relative low-frequency motion of the user moving the pump handle.

A moving average filter is easy to implement, making it suitable to use on resource-constrained hardware applications. The filter calculates the average of several points from the input signal such that each point of the output signal, y , is calculated as [199]:

$$y[i] = \frac{1}{M} \sum_{j=0}^{M-1} x[i+j] \quad (5.16)$$

where x is the input signal and M is the number of points used in the moving average.

The moving average of the training data was used to filter the test data as well. This ensures that the testing process closely simulates the onboard processing performed on the embedded system. This simple FIR filter will allow continuous filtering in real-time. Figure 5.1 shows an example of the output of the filtering step by comparing the unprocessed pumping waveform (red), the low-passed moving average (black), and the final high-passed output of the vibrations (blue).

(4) Data Integrity Verification

In addition to pre-processing, we implemented vital steps to ensure that all unsuitable or low-quality windows were discarded. A waveform that varies significantly could, for example, indicate unusual use of the pumps, such as children playing with the handle or

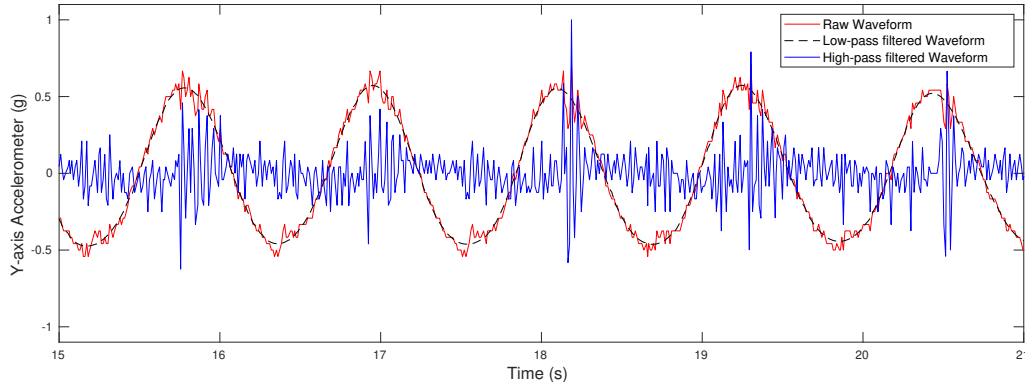


Figure 5.1: A six-second interval of the final output of the pre-processing steps for the Y-axis accelerometry data.

disassembling of the system during maintenance or repair.

The first data verification step was to verify that the data vectors contained non-zero values. The second step was to clip the recordings by removing the first and last periods of each recording to ensure only full waveforms are included.

Next was to identify recordings with irregular or abnormal pumping rates. A typical user settles at a natural pumping rate of 0.5 Hz to 1 Hz (1 to 2 seconds/stroke). The period, calculated as the time between subsequent troughs, was used to evaluate the pumping rate during regular use. We removed waveform periods that were 1.5 times larger or smaller than the median period.

The last data verification step was to identify a full pump stroke, which is a deviation in the pump handle of at least $+25^\circ$ above the horizontal and at least -15° below the horizontal (parallel to the ground). Periods of incomplete pump strokes were removed.

(5) Fast-Fourier Transforms

The final pre-processing step was to convert the time-series signal to the frequency domain. The fast Fourier transform (FFT) is commonly used in signal processing to decompose a time-based signal into a sum of sinusoidal basis functions. The output of the trans-

formation is used to describe the frequency content within a time-series waveform [200]. The most common implementation of the Discrete Fourier transform, X_k , for a discrete time-signal x_m will be used in this thesis and defined as:

$$X_k = \sum_{m=0}^{N-1} x_m \exp \left\{ -2\pi i \frac{mk}{N} \right\} \quad (5.17)$$

where k denotes the frequency domain ordinal and m represents the time-domain ordinal. N is the number of samples to be transformed.

To perform FFT on the waveforms used in this thesis, 1.3 s windows are created from each recording with a 50% overlap. This creates 128 samples per window, transformed to 64 frequency components with a resolution of 0.75 Hz per component for a sampling frequency of 96 Hz. Figure 5.1 shows that high-pass filtering removes the underlying periodic trend of the signal associated with the pumping tempo. As a result, the finite window of the non-periodic signal will result in a truncated waveform with discontinuous endpoints (“edge effects”), which appear as high-frequency components in the FFT not present in the original continuous-time signal [201, 202], known as spectral leakage. Windowing can reduce this effect by reducing the amplitudes of the discontinuities at the boundaries of each finite window. A 128-point Hamming window function was therefore applied to each of the 128-sample windows in the time-domain. The final result after FFT is a 128-point feature vector with 64 frequency components per window, up to the Nyquist rate, $f_N = 48$ Hz.

5.3.2 Embedded System Considerations

To ensure the model is suitable for implementation within the available hardware and to limit the use of resources (such as battery life and data-transmission), we took two

additional steps.

(1) Handling Floating Point Values

First, the original floating-point values were scaled by a factor of 10^3 to convert the values to fixed-point representation, which is easier and quicker to store and process when using an embedded microprocessor.

(2) High-pass Filter

The high-pass filtering step performed as part of the pre-processing in section 5.3.1, was done due to system constraints. The HPF of the signal was calculated by subtracting the low-pass filtered signal from the original signal, where the low-pass filter is approximated by the moving average, a simple step that could be implemented on the limited system.

(3) Sampling Resolution

Sub-sampling the recorded signal reduces the sampling rate, F_s , in order to reduce the data rate and size of the data. This was considered to optimise resources on the embedded system and to ensure compatibility with the hardware constraints of the embedded microprocessor. However, the initial analysis of the spectral data showed high intensities in the 20 to 30 Hz band for some abnormal cases. Thus, downsampling by a factor of two on the signal sampled initially at 96 Hz (to 48 Hz) would only allow analysis up to the Nyquist rate of 24 Hz, which could negatively affect our subsequent spectral analysis. The signal above 24 Hz would appear as an artefactual “alias” below 24 Hz, potentially confounding the analysis.

It was shown, however, that a frequency resolution of 0.75 Hz per component offered the same level of detail than that of a finer resolution of 0.375 Hz per component. Given the resource constraints of the embedded system, it was deemed more favourable to opt for a coarser resolution without the loss of information.

5.3.3 Train-Test

In machine learning, it is standard methodology to hold out some subset of the data to be used as unseen data to test the trained model and evaluate the ability of the model to generalise “out-of-sample”. Thus, 20% of the data set was held out to remain unseen for testing the final model. In each train-test set, the data set is balanced such that it contains the same number of normal and abnormal examples. Figure 5.2 shows a flow diagram of the train-test method for the classifier that was used in this thesis.

Each feature vector in the training data are normalised by subtracting the mean of the training set and then dividing by the standard deviation. The test data are normalised by using the same mean and standard deviation values from the training data.

For those classifiers, such as LR and DT, that have no hyperparameters to control model flexibility, there is no validation step (e.g., 10-fold cross-validation) required to optimise its parameters. Hyperparameters of the SVM (cost parameter, C , and slack variable, ξ), kNN (numbers of neighbours, k) and NN (learning rate) models were optimised using cross-validation (CV). K-fold CV was used to partition the available balanced data set, D_1 , into K subset “folds” [203], where $K = 5$ as commonly used in machine learning. Each data point in the set is randomly assigned to a subset such that the subsets are of equal size. The final accuracy score of the trained models is the average of K outcomes using unseen data. In the case of SVM, grid search was used on each fold to find the optimum combination of the two hyperparameters.

5.3.4 Model Evaluation

The performance of the classification models was evaluated based on the accuracy and the robustness measures defined in this section. Given the trade-off between accuracy and speed of the proposed classifier methods, the sensitivity and specificity are also considered.

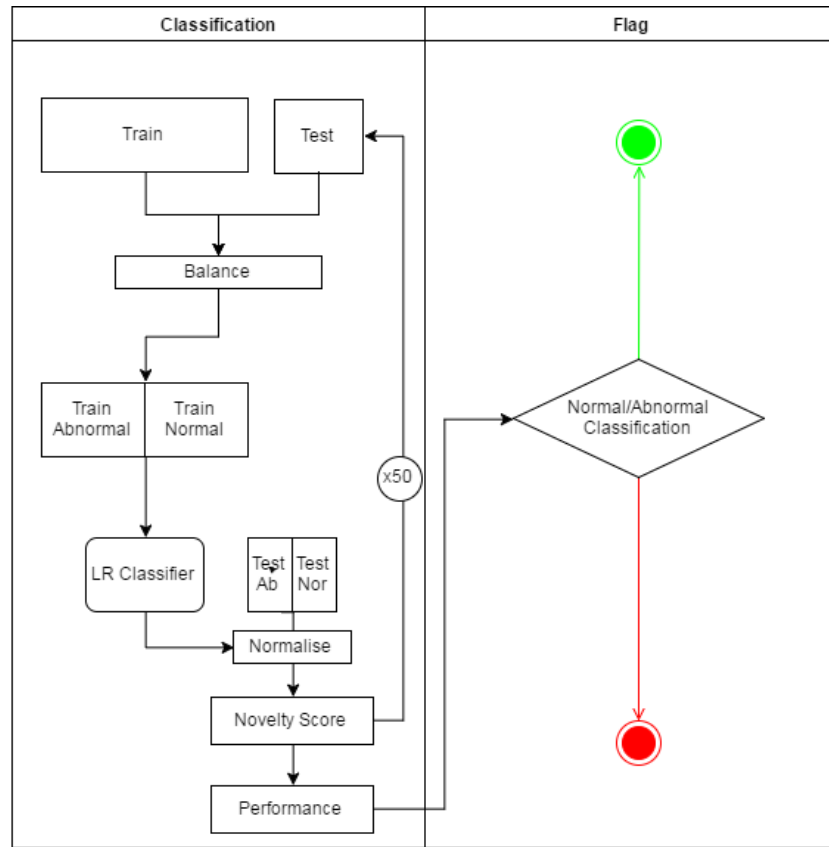


Figure 5.2: A flow diagram of the classifier’s validation process shows that both train and test sets are balanced before the test data evaluate the classifier’s performance to identify the handpump as either “normal” or “abnormal”.

In this thesis, the classifier accuracy is defined by how well the model can detect deterioration in the system’s health during operation by assessing the quality of the vibrations.

Accuracy Measures

A true positive (TP) is an instance when the model correctly labels the system as abnormal (broken) in the case of a failure, and a false positive (FP) is an instance when the model incorrectly labels the system as abnormal (broken handpump) in the case of functioning handpump. Inversely, any instances the model labels as normal (functioning handpump) when the system is working are called true negative (TN) or false negative (FN) when failure cases are labelled as functioning by the model. Using these classes, the classification accuracy (Acc) of the model can be defined as:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \quad (5.18)$$

Specificity and sensitivity values describe the proportion of cases that are correctly classified within all the negative and positive cases, respectively. Specificity (*Spec*) is defined as:

$$Spec = \frac{TN}{TN + FP} \quad (5.19)$$

Similarly, sensitivity (*sens*) or true positive rate (TPR) is also known as the recall and can be defined as:

$$Sens = \frac{TP}{TP + FN} \quad (5.20)$$

Precision, also known as the positive predictive value, describes the proportion of positive cases that were correctly identified and can be defined as:

$$P_r = \frac{TP}{TP + FP} \quad (5.21)$$

The receiver operating characteristic (ROC) is used to compare the performance of the proposed classifier for different feature vectors to verify condition monitoring reliability. This metric compares the actual and predicted output for each class. For the case of a two-class classifier, the outcomes are either labelled as positive (normal), *p*, or negative (abnormal), *n*.

The true positive rate (TPR), or *sensitivity*, of a classifier is defined in Eq. 5.20. The false positive rate (FPR), or *fall-out*, is defined to be the probability of a false alarm such

that $1 - Spec$ and is defined by:

$$FPR = \frac{FP}{FP + TN} \quad (5.22)$$

The ROC curve is defined as the plot of *Sensitivity vs (1 - Specificity)* for all possible combinations of some threshold or operating point, o . A diagonal line joining the left-bottom corner and the right-top corner of such a plot is equivalent to random guessing, such that area under the curve $AUC = 0.5$. Minimising the loss function, L , can optimise the operating point for this curve such that:

$$L(o) = v(1 - Sens(o)) + (1 - v)Spec(o) \quad (5.23)$$

where the cost of false negatives are represented by v such that $v = 0.5$ when the equal relative importance of FP and FN are assumed. Optimising the area under the ROC (AUC) will maximise failure detection while simultaneously minimising false alarms, which can be costly in real-life. Missed alarms can be costly as vulnerable communities may be without a reliable potable water source. Similarly, false alarms increase the cost of the maintenance service company when sending a team member to repair a functioning handpump. In the ideal case, the classifier would be very sensitive ($TPR = 1$) with no false alarms ($FPR = 0$).

Robustness Measures

To ensure the robustness of the model, the accuracy metrics described above were determined using unseen test data and taken as the median of 50 iterations using different randomly selected balanced training sets. The variability in performance between subsequent iterations was measured by observing the interquartile range, $IQR = Q_3 - Q_1$,

of the accuracy scores and found to be negligible (less than 0.02) in most models. After establishing that no major outliers influence IQR, the standard deviation for all iterations was reported along with the mean, such that:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x - \bar{x})^2}{N}} \quad (5.24)$$

where x is the accuracy metric reported and N is the number of iterations.

5.4 Results

When using “lightweight” classification methods, feature selection is a crucial step to ensure the performance of the classifier. Selecting a large number of features may improve prediction accuracy; however, this often leads to over-fitting and makes the model difficult to interpret. Conversely, using a low number of features will reduce the risk of over-fitting and make the model easier to interpret but could lead to reduced overall prediction accuracy. There are various methods available to rank the importance of features or reduce the dimensionality; however, interpretability of the model is a priority at this early stage of the project to ensure an in-depth understanding of the performance of any selected model used as the foundation for ongoing work. As such, the results will begin with a thorough analysis and visual inspection of the data to guide common sense and intuition for any future design choices. It will then use PCA to reduce the number of features, before finally investigating the performance of the proposed classifiers on different feature subsets.

5.4.1 Visualisation of Data

Visualisation is a powerful tool to analyse a data set by visually considering a summary of the main characteristics of different pump types, operating conditions, and failures. This is performed by reducing the data dimensionality into a “latent space” while aiming to

preserve the structure of the data in the original (higher-dimensional) data space.

Sammon’s Mapping

One such tool, Sammon’s mapping, is a non-linear transformation that tries to conserve the original structure by minimising the distance, in this case Euclidean, between each pair of points in the original multi-dimensional data space in the two-dimensional latent space [204]. Sammon’s mapping presents a measure of how well the result represents the structure in the original data set, rather than providing a function that represents the transformation itself. The Sammon stress metric provides a measure of the quality of the transformed data set as the difference between the original and transformed structure:

$$E_{sam} = \frac{1}{\sum_{i < j} [d_{ij}^*]} \sum_{i < j} \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*} \quad (5.25)$$

where d_{ij}^* is the distance between two points i and j in the original latent space and d_{ij} is the distance in the low-dimensional space.

Sammon’s mapping was used to project the feature vectors, corresponding to each FFT 128-component window extracted from the accelerometry data of the handle, in 2 dimensions for visual inspection. The most relevant Sammon’s mapping result is shown in Figure 5.3. Four groups are coloured in this figure: (i) normal working handpumps (dark blue) and broken handpumps with three common failures, including (ii) dry borehole (green), (iii) leak in the rising main (orange), and (iv) broken U-seal (light blue).

The resulting projection shows that “abnormal” classes for specific failure types intermix with the “normal” class. However, classes of individual abnormal conditions form separate clusters. The projected points of the abnormal classes form tight clusters, respectively, indicating that their shape vectors are very similar. It appears that the presence

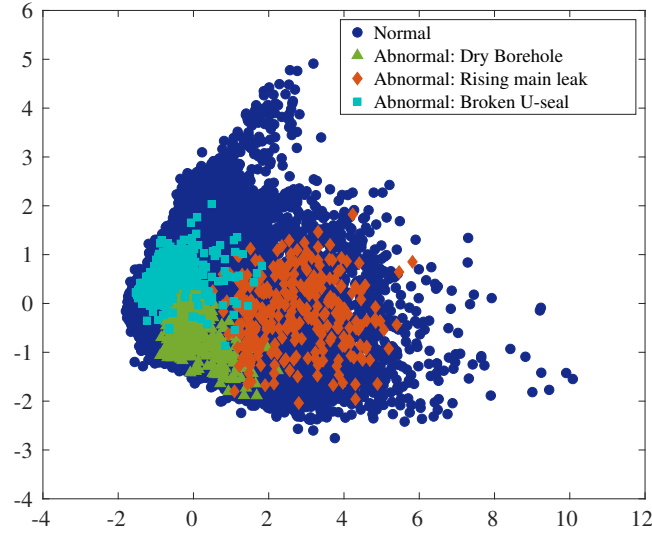


Figure 5.3: Sammon's mapping projection of 128-D feature vectors for handpump vibrations on the Y-axis. Clusters show different handpump conditions: (i) normal (dark blue), (ii) dry borehole (green), (iii) leak in the rising main (orange), and (iv) broken U-seal (light blue).

or absence of water in the system for normal and abnormal classes separate into distinct clusters when projecting the feature vectors. Although all three these failures result in the handpump not drawing water from the borehole, the main discriminating factor between these is the location at which each occurs, with the broken seal furthest from the sensor. Based on this analysis, we investigate the influence of the operating depth of the system, which will be discussed in section 5.4.2.

Figure 5.4 shows the patterns projected for individual pumps, where each of the 28 sub-plots highlights a single normal operating handpump in the data set. The spread of tight clusters for individual pumps show the apparent variability that occurs across individual pumps of the same type. Some level of deviation from the normal is expected in most systems and is especially common in vibrating systems that, as was illustrated in Figure 3.3, require some time before reaching the signature frequency spectrum during run-time. This state of increased vibration can occur either immediately before a failure or immediately after a repair.

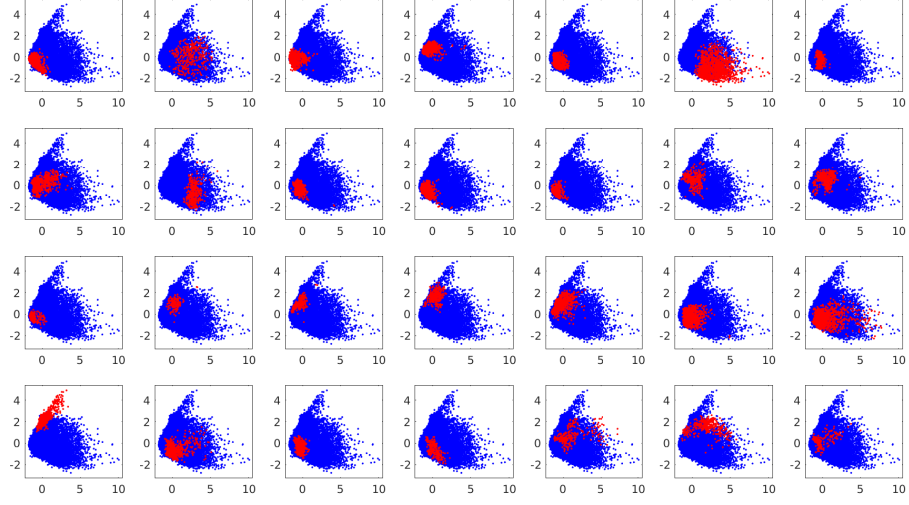


Figure 5.4: Projection of 128-D feature vectors for handpump vibrations on the Y-axis grouped by individual pump across all normal recordings in this data set. In each plot, the feature of a different individual pump is highlighted in red.

5.4.2 Analysis of Spectral Data

For each frequency component, the corresponding power in the vibration signal is known as the power spectral density. The power spectral density of the accelerometry waveform is calculated using the FFT by multiplying each frequency bin by its corresponding complex conjugate to determine the real spectrum amplitude in g^2 and normalising the amplitude to the frequency bin width to achieve the unit of g^2/Hz . The power spectral density of a signal can be defined by:

$$P_{xx}(\omega) = \sum_{\tau=-\infty}^{\infty} r_{xx}(\tau) e^{-j\omega\tau} \quad (5.26)$$

where $P_{xx}(\omega)$ is the energy present in the signal, τ is the sample displacements, r_{xx} is a auto-correlation sequence of several signals or in the signal itself from a random sequence to identify patterns in the signal.

The normalised spectral bands are used to compare vibration levels in different signals, irrespective of the length, and those bands with the largest magnitude are considered to

have the most interesting vibration content.

As part of the preliminary analysis of the original data set, a series of FFTs are used to represent approximations of the power spectral density of the handpump vibration for the accelerometer sensor output. A spectrogram illustrates how the spectrum changes in the frequency domain over time. The spectral data are then analysed to investigate the effect of the weight in the system and the speed of pumping; we hypothesised that these are the two main contributing factors to creating vibrations in the system.

(1) Effect of Weight in the System

Two factors determine the weight inside the handpump system being lifting with each pump stroke: (1) the number of rods in the pump determined by the operating depth and (2) the presence or absence of water determined by the condition of the system.

Depth: Number of Rods

The depth of the borehole determines the number of rods in the pump, connecting the handle and the plunger just above the foot valve. Each rod used in an Afridev handpump is 3 metres in length and made of either mild or stainless steel. Although stainless steel is recommended to ensure the handpump is more resistant to corrosion, mild steel is more often used during installation and later maintenance due to the lower upfront cost. In addition to the rods, the plunger is connected to another one-metre rod. Thus, the typical Afridev handpump has between 2 and 23 rods for an operating depth of 7 to 47 metres below ground level.

In the data set considered, there was only one handpump that contained only stainless steel rods, a small number that contained a mix of stainless and mild steel rods and the majority contained only mild steel rods. Although mild steel is heavier than stainless steel, it is treated as similar in this thesis (for simplification). Similarly, the weight of rods

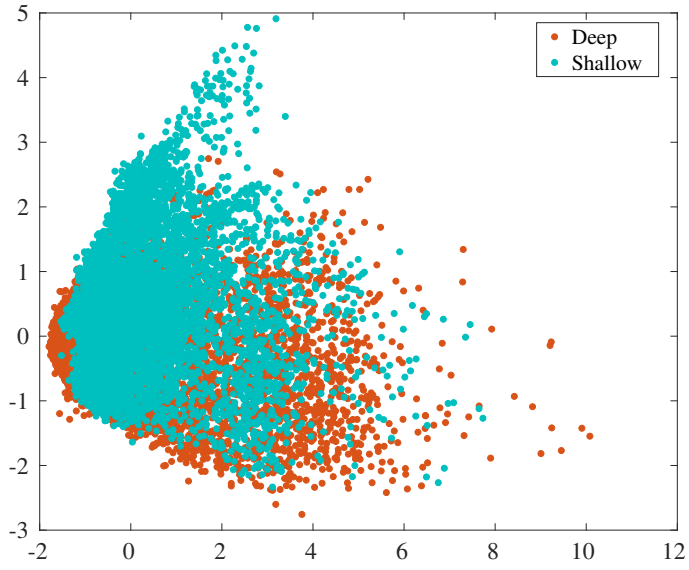


Figure 5.5: Projection of 128-D feature vectors for handpump vibrations on the Y-axis grouped by the depth of the well according to the number of rods for deep (orange) and shallow (blue) operating handpumps, respectively.

from the same material may vary slightly as some may be more worn out than others. Instead, we assume that there is only a significant increase in weight in the system with the addition of entire rods rather than the weight difference between different material types.

The Sammon’s mapping representation of the feature vectors grouped according to the well depth, as determined by the number of rods in the pump, is shown in Figure 5.5. Two groups are coloured in this figure: (i) deep wells (orange), operating at a depth greater than 25 m or more than 7 rods, and (ii) shallow wells (blue), operating at a depth of less than 25 m or fewer than 7 rods. From this projection, it appears there is some subset of feature vectors that is unique to deep wells, while feature vectors from shallow pumps vary more widely. The increased weight in deeper wells may contribute to the unique clustering of these features.

Normality: Water or No Water

For the spectra density of normal handpump conditions, distinct high-frequency components appear to take a high energy, shown in Figures 5.6a and 5.6b for shallow and deep wells, respectively. However, in the spectra density for abnormal handpump conditions at the same pumps, shown in Figure 5.6c and 5.6d, these bands of high energy are less pronounced than under normal operating conditions. This may suggest that an increase in weight lifted by the system could intensify the vibration signatures measured by the sensor in the handle.

Furthermore, when we compare the median amplitude of the magnitude of the spectra of a shallow and deep operating handpump in Figures 5.6e and 5.6f, it can be seen that these high energy bands vary depending on the depth at which the handpump is operating. For a shallow handpump (a), operating at 15 m, this can be seen between approximately 22 and 30 Hz and for a deep handpump (b), operating at 47 m, this spectral band is visible approximately between 30 and 37 Hz. This may be due to users applying less force on the handle at shallower wells in order to lift the water from the well. From this, we can conclude that the weight in the system, due to water and the number of rods, are essential features to consider. Based on this analysis, we will continue to consider the difference in deep and shallow operating wells.

(2) Effect of Pumping Speed

In essence, the handpump, while operated, is a rotating machine, where the handle is a lever rotating around the fulcrum and changing direction between up- and down-strokes. To investigate the effect of the speed at which a handpump operates, this section will investigate the relationship between pumping characteristics and speed.

If the arc is defined as the angle of the handle away from a horizontal position, the

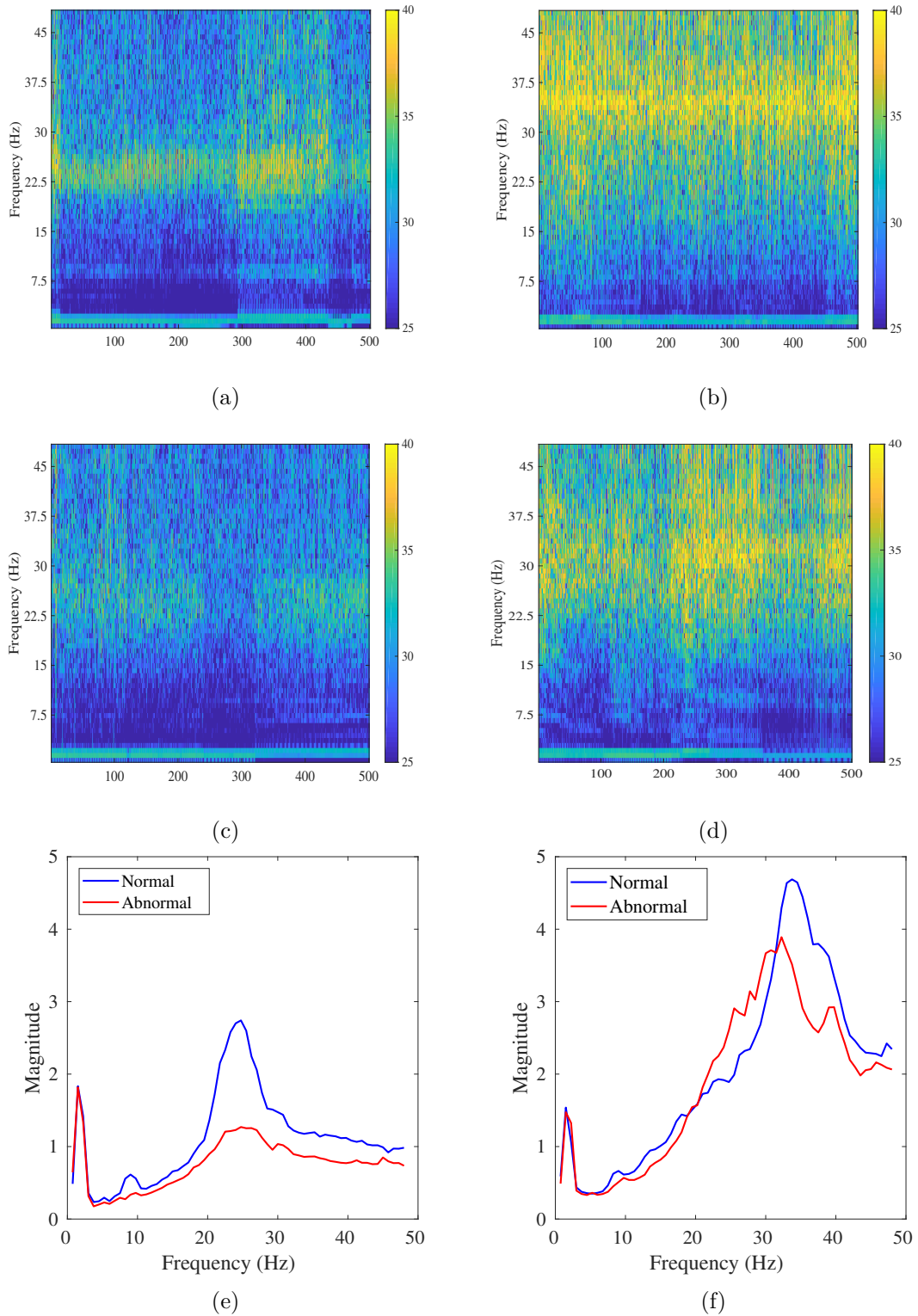


Figure 5.6: Spectra (in dB) for normal pumping conditions for a (a) shallow well and (b) deep well, and abnormal pumping conditions for a (c) shallow well and (d) deep well, respectively. The median amplitude of the spectral data are shown in (e)-(f) for shallow and deep, respectively. Distinct high-frequency components appear to take high energy for normal handpumps operating in deep and shallow wells.

maximum-recorded handle rotation above the horizontal is 61.3° and -34.5° below the horizontal. For a full pump stroke with a median period of 1.1 seconds, the rotational speed of the handle would then be 174.2 degrees/second. However, pump strokes vary by the sex and height of the user. The median arc on the down-stroke (handle moves above horizontal) is 25.3° , and the median arc on the up-stroke (handle moves below the horizontal) is -15.5° . The median pumping speed across all recordings is 86.6 degrees/second, and the maximum pumping speed is 264.5 degrees/sec, recorded as the median of the instantaneous speed for each 128-sample window.

For the entire data set, D_1 , the probability of recorded speeds, along with fitted gamma distributions (often used to model asymmetric distributions of speed for rotating machines [74]), for normal (blue) and abnormal (red) conditions are shown in Figure 5.7a. The figure shows that the distribution of speed is approximately gamma for both normal and abnormal conditions, but reduced in distribution $\text{Gam}(x|a, b)$, where a and b are location and scale, for abnormal handpump conditions. The change observed in the probability density function (pdf) of speed for normal (blue) and abnormal (red) handpump conditions indicates that there is a slight increase in pumping speed under normal handpump conditions. This may be due to the smoother operation of a fully functioning machine, making it easier to move the handle at greater speeds.

Figure 5.7b considers the influence of speed for the case where the data are grouped according to a qualitative assessment of the water flow, such that abnormal handpumps are those with (0) no water flow. In contrast, normal handpumps are grouped as (1) restricted, (2) average, (3) good, or (4) excellent water flow, where, for consistency, these conditions were labelled by the same observer during data collection. The probability density function is approximately gamma and greatly reduced in location and shape for the case of restricted

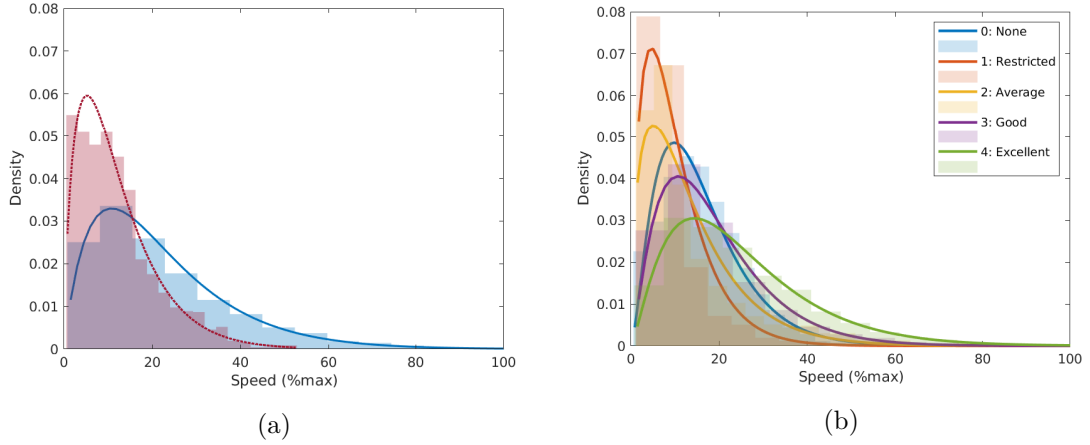


Figure 5.7: Histograms of pumping speed for (a) normal (blue) and abnormal (red) hand-pump conditions and (b) sub-grouped according to varying water flows. The histograms are approximately pdfs as they integrate to unity. Each set of histograms is fitted with the maximum likelihood of gamma distributions.

and average flow when compared to good and excellent flow. We could consider the pdf of abnormal handpumps with no water flow as a baseline between these two groups as there is no weight in the pump at this point. For those handpumps with restricted or average water flow, operating speed tends to be lower than the baseline while those handpumps with good or excellent water flow are operated at speeds faster than the baseline. This may suggest that the pumping speed may also be a function of the energy exerted or motivation of the user attempting to extract water.

Figure 5.8 considers a speed-based representation of the spectral data in Figure 5.6, based on work modelling jet engine vibration spectra [74]. The speed-frequency space is quantised into $N_s \times N_f = 10 \times 10$ bins, where N_s is the number of bin related to speed and N_f is the number of bins related to frequency. The horizontal axis shows speed bins ranging from 70 to 160 degrees/second as a percentage of the maximum. There appears to be some variation for speed, particularly for the case of the deep well, but this could also be a feature of a single person pumping at a different point in the speed range rather than being a generalised characteristic of the speed-based vibration signatures. Based on this

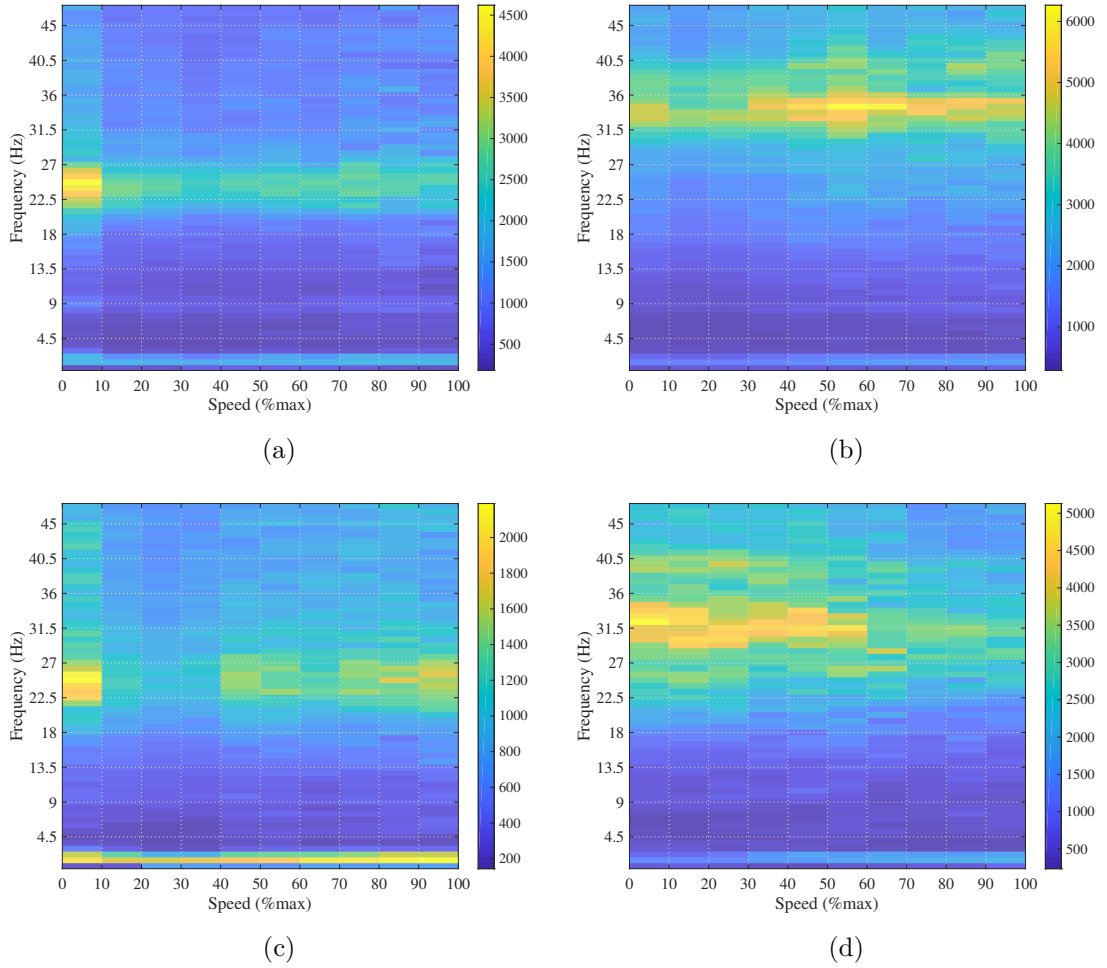


Figure 5.8: A speed-based representation of spectral data for normal pumping conditions for (a) a shallow well and (b) a deep well, and abnormal pumping conditions for (c) a shallow well and (d) a deep well, respectively. The speed-frequency space alone does not provide sufficient evidence to support whether speed impacts the pumping characteristics.

preliminary data representation, there is insufficient evidence to support whether speed affects the vibration of the handpump.

We cannot conclude from this alone if pumping characteristics depend on speed.

To further verify the effect of speed on the pumping characteristics of the diagnostic labels, we consider the PDFs, shown in Figure 5.9, for two groups labelled either as “shallow” or “deep” for different depth thresholds to distinguish each group. The depth threshold increases by increments of 3 m as each time the group is split according to the number

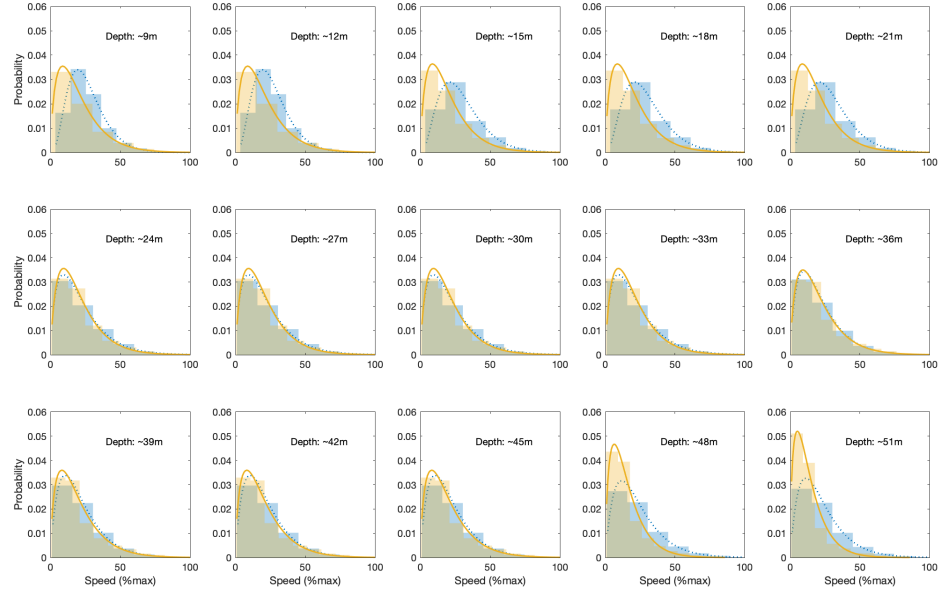


Figure 5.9: Probability of pumping speed as a percentage of the maximum for varying the depth at which we classify the well, defined by the number of rods in the system, for shallow (blue) and deep (yellow), respectively.

of pumping rods present. For depths less than 21 m, it appears that distributions are reduced in shape for those pumps labelled as deep. However, at a depth of approximately 25 m, there is no difference in the distribution of pumping speeds for the two groups. As such, any handpump operating at a depth of less than 25 m is grouped as “shallow” and handpumps operating at a depth greater than 25 m are grouped as “deep”. From this analysis, we conclude that the proposed grouping should be sufficient to control for any variation due to pumping speeds between handpumps operating at different depths.

5.4.3 Accelerometer Axis Selection

The accelerometer provides three data channels; one for each of the three orthogonal axes of the gyroscope and aligned with the handle of the handpump previously shown in Figure 4.1a. The data from each axis could either be used separately or in combination with one or both of the other two axes. In signal processing, the simplest method to combine a data

stream is the Euclidean norm that computes the sum of squares such that a combination of all three axes would be, $X_{123} = \sqrt{x_1^2 + x_2^2 + x_3^2}$.

To determine which input would be the most beneficial, Figure 5.10 compares the AUC of a simple LR classifier for each axis input, using the seven different combinations shown in Table 5.1. It appears that axes Y and Z combined provides the best classification performance. However, given the orientation of these two axes being perpendicular and parallel to the handle, respectively, it might be that the pre-processing favoured these two axes during HPF or it could be that the X-axis does not contain valuable vibration signatures.

From this analysis, it is unclear if one combination is superior to the others given the data set, D_1 . A larger data set could help to determine the correct combination. Since all the combinations performed sufficiently well, no single combination would be detrimental to the investigation; in fact, correctly identifying the optimal axes at a later stage using a more extensive data set could only improve future results. As such, for simplicity Y-axis will be used in the ongoing analysis.

5.4.4 Feature Selection

Feature selection is used to reduce the dimensionality of the data set by selecting a parsimonious representation, such that classification accuracy is improved while storing a reduced summary of the data. The choice of feature vectors hinges on prior knowledge of the system obtained during the investigation earlier in this chapter.

Feature vectors of 20 points (or frequency components) were extracted from the original data set, D_1 . Feature vector FV_0 is the original feature vector with 64 samples and is included as the gold standard to evaluate the performance when selecting a subset of these features. For feature vectors $FV_1 - FV_3$, frequency components with indices in

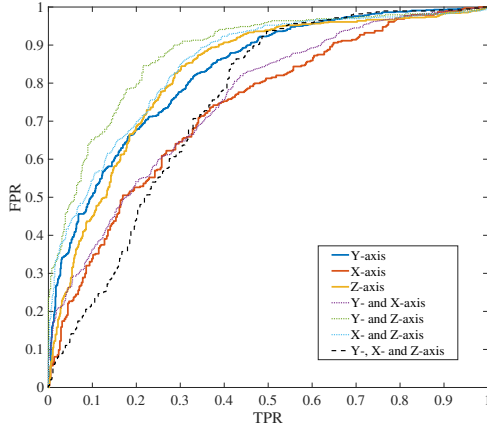


Figure 5.10: The ROC plot comparing the proposed LR classifier's performance for different axes of the accelerometer shows that combining all three accelerometer axes does not provide additional information. Instead, a single axis can provide acceptable performance while preserving processing power.

Input	AUC (SD)
Y-axis	0.827(0.01)
X-axis	0.759(0.02)
Z-axis	0.834(0.01)
Y- and X-axis	0.766(0.01)
X- and Z-axis	0.846(0.01)
Y- and Z-axis	0.892(0.01)
Y-, X- and Z-axis	0.762(0.01)

Table 5.1: Comparison of the AUC and standard deviation for LR classifier output for the three accelerometer axes (Y, X, and Z), individually and combined.

the range from 3 to 64 (corresponding to frequencies of 2.25 to 48 Hz) were sampled step-wise in 3 equal-sized intervals for the lower, middle, and upper range, respectively. Feature vector FV_4 is constructed by uniformly sampling across the entire frequency range. FV_5 selects the 20 frequency components with the greatest absolute difference between the median power spectra of normal and abnormal handpump conditions. Finally, FV_6 randomly selects 20 frequency components across the entire frequency range in bins 3 to 64. Definitions of the feature vectors are described in Table 5.2, along with the respective bin index and frequency range.

The first three frequency components, equivalent to 0 - 2.25 Hz, were not considered for the construction of features FV_1 to FV_6 in order to avoid dominant vibration features within the median pumping frequency (associated with the manual operation of the pump) of 1.1 Hz. Figure 5.6 showed that the moving average filter is not successful in removing all the frequency components in the lower range, and as such, it was decided not to consider

Table 5.2: Feature vectors $FV = \{FV_j\}$ extracted from the frequency domain of the data set D_1 . The notation $[i : j : k]$ describes a regularly shaped vector such that $[j, j + i, j + 2i, \dots, j + m \times i]$, where $m = 20 = \frac{k-j}{i}$.

	Description	Freq. Bin Index	Freq. (Hz)
FV_0	Full range	[1:1:64]	[0.75:0.75:48]
FV_1	Lower range	[4:1:23]	[3:0.75:17.25]
FV_2	Mid range	[24:1:43]	[18:0.75:32.25]
FV_3	Upper range	[44:1:63]	[33:0.75:47.25]
FV_4	Uniformly sampled	[3:3:60]	[2.25:2.25:45]
FV_5	Hand-picked	[11,26:1:36, 45:1:51,53]	[8.25,19.5:0.75:27, 33.75:0.75:38.25,39.75]
FV_6	Randomly selected		

these frequency bins.

Principal Component Analysis

Principal component analysis (PCA) is another statistical analysis tool used to reduce the dimensionality of a data set by transforming $\{x_{n,1} \dots x_{n,D}\}$ to $\{x'_{n,1} \dots x'_{n,M}\}$ where $M \leq D$. The first principal component (PC) encompasses the maximum variance to describe all the observations in the data [205]. By simplifying complex data in this way, it is possible to observe the underlying hidden structure in the parsimonious form when considering the dominant principal components. Each succeeding PC accounts for as much of the remaining variability and is orthogonal to all the previous PCs.

The respective PCs can be mathematically solved either via the maximal variance formulation or the minimum-error formulation [205]. In this work, the maximal variance formulation was used where the first PC is selected to such that it accounts for as much variability in the data as possible. Each subsequent PC is then selected to optimise the maximal variance in the data while it is: (1) orthogonal to all other PCs, and (2) the number of PCs does not exceed the dimensionality of the original data set. If the variance between a set of observations, $\mathbf{x}_n$, is maximised when projected onto a D-dimensional unit vector, $\mathbf{u}_1$, where $\mathbf{u}_1 \mathbf{u}_1^T = 1$, then $\mathbf{u}_1$ is the first PC. In order to maximise the variance of

the projected data, $\mathbf{x}_n$, the covariance matrix, S , is defined by [205]:

$$S = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T \quad (5.27)$$

where $\bar{\mathbf{x}}$ is the sample set defined by:

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \quad (5.28)$$

The first PC is found by maximising the variance $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$ with respect to $\mathbf{u}_1$, given that $|\mathbf{u}_1| < \infty$. The Lagrange multiplier, λ_1 , is used to enforce this constraint by maximizing:

$$\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 - \lambda_1 (1 - \mathbf{u}_1 \mathbf{u}_1^T) \quad (5.29)$$

The derivative of this equation with respect to $\mathbf{u}_1$ at the zero-crossing will result in:

$$\mathbf{S} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1 \quad (5.30)$$

which can also be written as:

$$\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 = \lambda_1 \quad (5.31)$$

Equation 5.31 shows that the variance is optimised when $\mathbf{u}_1$ is equivalent to the eigenvector with the largest eigenvalue, λ_1 .

PCA was used to analyse the selected feature vectors visually. The resulting projections of the first two principal components from each 20-D feature vector are shown in Figure 5.11(a)-(f). In all cases, the selected features perform well in explaining variability in the data. Inter-mixing of “normal” and “abnormal” conditions is seen across all the proposed

feature vectors. Some clustering of specific conditions can be seen in Figure 5.11d and 5.11f.

A similar projection from each feature grouped by the operating depth of the handpump is shown in Figure 5.12. For the case of shallow operating pumps, separate clusters of blue and green appear in Figure 5.12d, 5.12e and 5.12f. For the case of deep operating pumps, similar clusters of black and red points are seen in Figure 5.12e, 5.12f, and 5.12d.

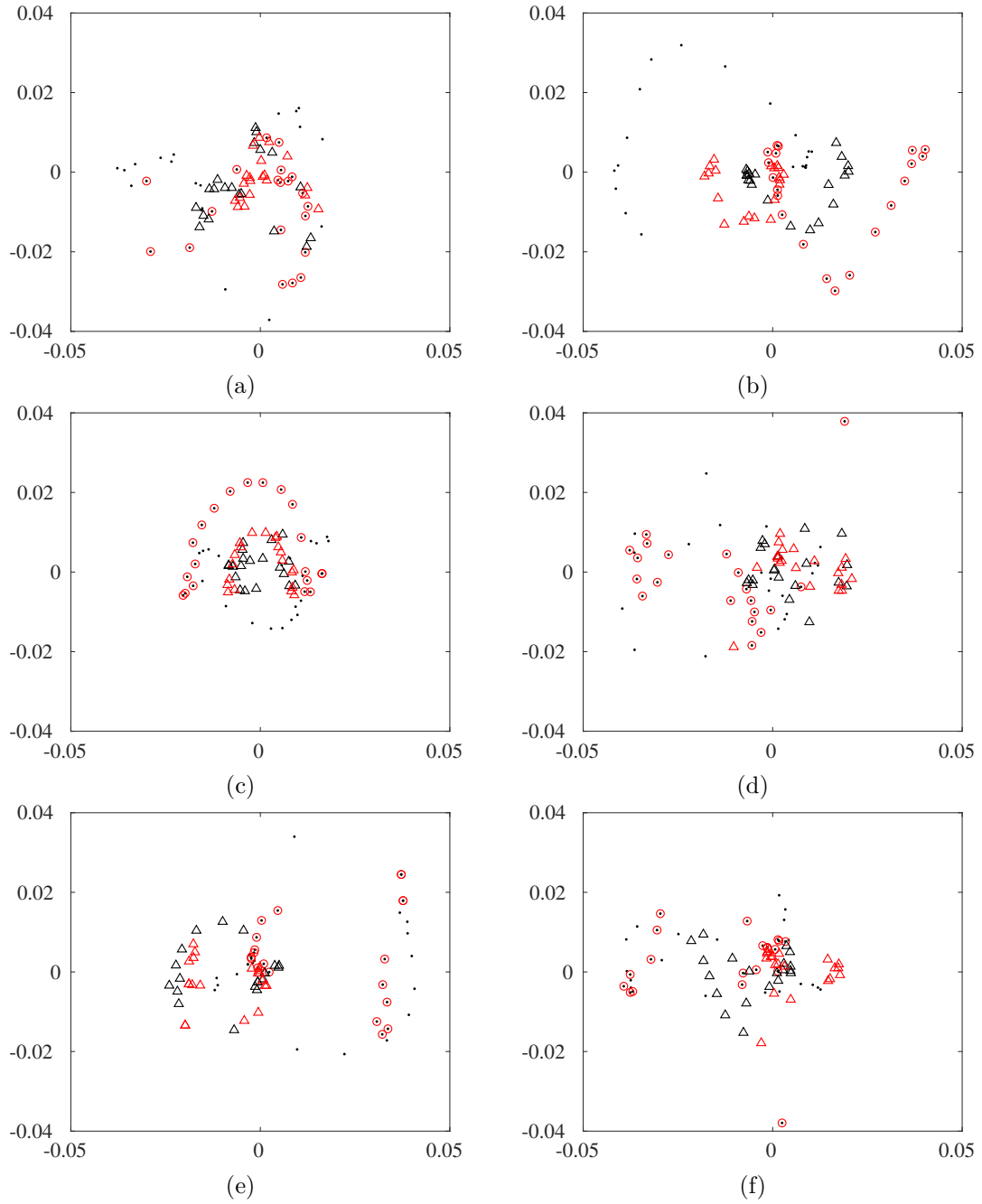


Figure 5.11: Projection of 20-D feature vectors for handpump vibrations on the Y-axis: (a) FV_1 ; (b) FV_2 ; (c) FV_3 ; (d) FV_4 ; (e) FV_5 ; and (f) FV_6 . Labels are shown in Table 5.3.

Symbol	Description
Black ●	Normal Train
Red ⊙	Abnormal Train
Black △	Normal Test
Red △	Abnormal Test

Table 5.3: PCA labels within visualisation of normal and abnormal handpump conditions in Figure 5.11.

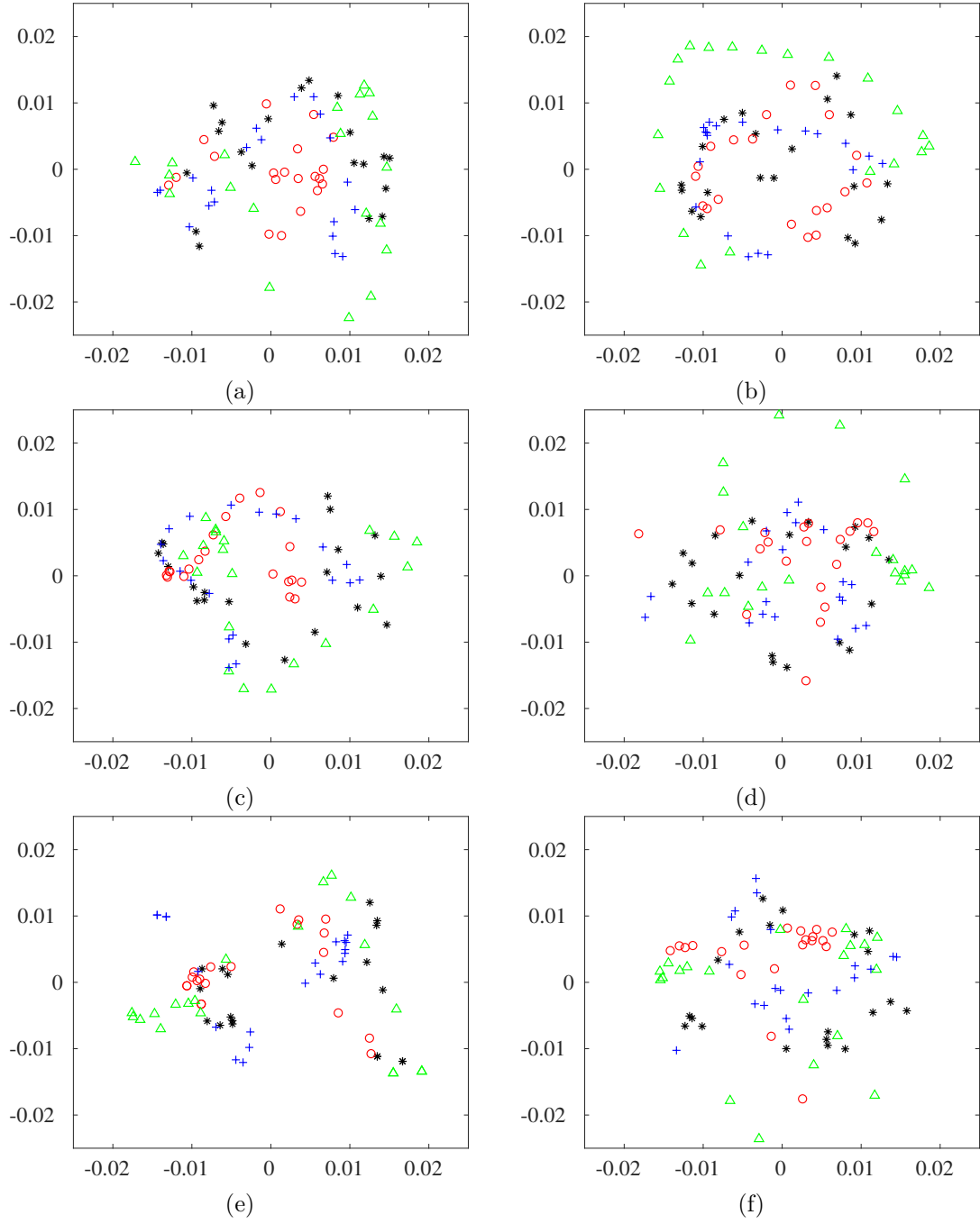


Figure 5.12: Projection of 20-D feature vectors for handpump vibrations on the Y-axis: (a) FV_1 ; (b) FV_2 ; (c) FV_3 ; (d) FV_4 ; (e) FV_5 ; and (f) FV_6 . Labels are shown in Table 5.4.

Symbol	Description
Black *	Deep Normal
Red O	Deep Abnormal
Green Δ	Shallow Normal
Blue +	Shallow Abnormal

Table 5.4: PCA labels within visualisation of normal and abnormal handpump conditions for different operating depth: (a) deep (46 m) and (b) shallow (15 m).

In both cases, it appears that FFT bins across the entire frequency range are essential to capture variability in the data. Further investigation will verify this work with larger data sets.

5.4.5 Classifier Performance

This section analyses the performance of classifying the handpump condition using the different vibration features that were proposed in the previous section. The section begins by comparing the performance of binary classifiers to map the outcomes of the function as either one (abnormal) or zero (normal) when considering the various sets of feature vectors FV_1 to FV_6 using the entire data set, D_1 . The SVM is used as the gold standard to assess the performance of a logistic regressor (LR) classifier, assuming no hardware processing constraints. Based on this analysis, the optimal feature vectors are selected and used to investigate the influence on the classifier performance when considering subsets of data, D_{deep} and $D_{shallow}$, based on the operating depths of the pumps within each group. Finally, this section concludes with a comparison of the five proposed simple classifiers given the selected axes, feature vectors, and mixed data set containing both deep and shallow handpump data.

Figure 5.13 shows the ROC plot for the LR classifier trained using the different feature vectors, $FV_1 - FV_6$, with the relevant results detailed in Table 5.5. The reduction in the number of feature vectors from the original 64 bins in FV_0 to the proposed 20 reduces the accuracy with more than 15%. For the proposed feature vector sets, $FV_1 - FV_6$, there is only marginal variation in the accuracy. Uniform sampling across the entire frequency range, FV_4 , appears to be the optimal set of features given the data set, D_1 , with a random selection of 20 in FV_6 being a close second. This may suggest that a further reduction in the number of features could be considered and will be discussed in Chapter 6 where we

consider measures to optimise the transmission of data.

Overall, the LR classifier does not sacrifice performance when compared to the more complex SVM. Most notable for the proposed application of a sensor-based novelty filter is the advantage of minimal computational resources required by the LR classifier, which is far more suitable for an embedded application than the SVM. The median time required to train the LR classifier is 0.42 seconds compared to 9.04 seconds required to train the SVM for the same feature set. This is an important advantage for our embedded application with highly limited memory and power resources should we choose to do online learning for the dynamic system proposed in Chapter 6.

Table 5.6 compares the classifier performance for subsets of data, D_{deep} and $D_{shallow}$, grouped according to the depth for feature vectors $FV_0 - FV_6$. The boxed values show that the best performing feature vector in the case of shallow and deep wells, respectively, improve on the highest accuracy for the case of mixed wells in Table 5.5 by 5-7%. The slight increase in the magnitude of the standard deviation may be due to the smaller data

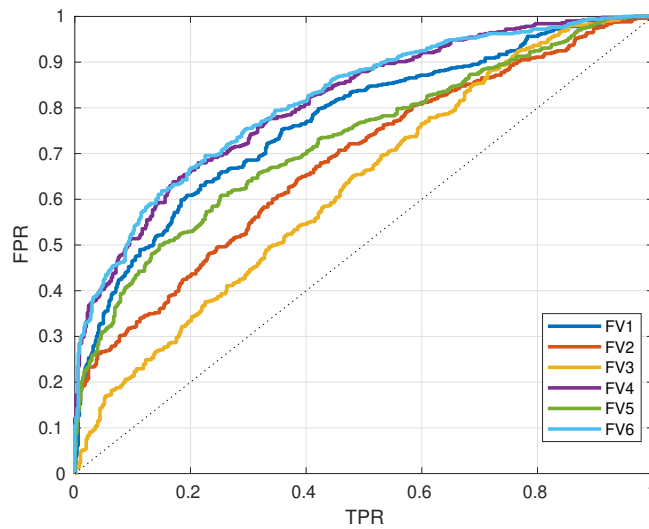


Figure 5.13: ROC plot comparing the performance of the LR classifier for the proposed feature vectors, $FV_1 - FV_6$.

Table 5.5: Comparison of LR and SVC classifier performance metrics [median (SD)] on a general data set, D_1 , for different feature vectors, $FV_0 - FV_6$, respectively.

	Logistic Regression		SVM	
	CA	AUC	CA	AUC
FV_0	0.919(0.01)	0.756(0.01)	0.921(0.02)	0.874(0.02)
FV_1	0.713(0.01)	0.775(0.01)	0.774(0.03)	0.800(0.03)
FV_2	0.637(0.02)	0.697(0.02)	0.738(0.04)	0.691(0.03)
FV_3	0.587(0.01)	0.634(0.01)	0.738(0.06)	0.623(0.06)
FV_4	0.753(0.01)	0.809(0.01)	0.802(0.03)	0.831(0.03)
FV_5	0.691(0.01)	0.748(0.01)	0.752(0.05)	0.754(0.05)
FV_6	0.727(0.01)	0.806(0.01)	0.762(0.03)	0.801(0.03)

set rather than an indication of reduced robustness. FV_4 , uniformly sampled across the entire frequency range, remains the optimal set of features for shallow wells while deep operating handpumps favour FV_5 . Furthermore, Figure 5.14 shows that the sensitivity of the classifier improves when the LR model is trained using data sets grouped according to operating depth compared to a more general model that is trained using the mixed data set.

Finally, we use the general data set, D_1 , with the selected Y-axis and FV_4 , uniformly sampling 20 features across the entire frequency range to train the five simple classifier models proposed to embed in the handpump sensor. Table 5.7 shows a comparison of the performance metrics for the trained classifiers using the unseen test data. KNN is the

Table 5.6: Comparison of classifier accuracy scores for subsets of deep, D_{deep} , and shallow, $D_{shallow}$, operating handpump groups. Classifier accuracy (CA) and AUC are represented by mean and standard deviation.

	Shallow		Deep	
	CA	AUC	CA	AUC
FV_0	0.870(0.01)	0.936(0.01)	0.871(0.01)	0.936(0.01)
FV_1	0.764(0.02)	0.842(0.02)	0.751(0.02)	0.837(0.02)
FV_2	0.773(0.02)	0.856(0.02)	0.699(0.02)	0.766(0.02)
FV_3	0.658(0.02)	0.711(0.02)	0.784(0.02)	0.893(0.02)
FV_4	0.806(0.02)	0.881(0.02)	0.807(0.02)	0.891(0.02)
FV_5	0.783(0.02)	0.862(0.02)	0.821(0.02)	0.903(0.02)
FV_6	0.793(0.02)	0.884(0.02)	0.809(0.02)	0.895(0.02)

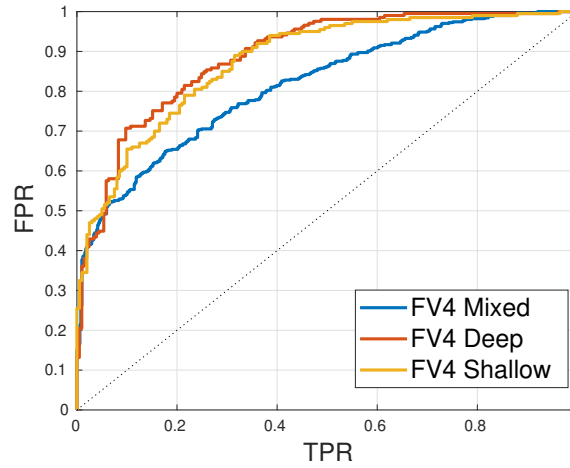


Figure 5.14: ROC plot comparing the performance of the LR classifier for FV_4 grouped by depth for deep and shallow wells combined (blue), only deep wells (orange) and only shallow wells (yellow).

best performing model both in accuracy and in robustness, which was expected since lazy learners tend to use an enriched hypothesis space rather than a single hypothesis. However, the memory requirement of an instance-based method is unsuitable for an embedded classifier. NBC is the worst-performing model which may be a result of its oversimplified assumption that predictors contribute equally and are independent, which is in direct contrast with the findings described earlier in this chapter. The LR and CART models have similar accuracy and robustness. Single-layer NN lacks the accuracy and robustness that deeper networks would offer. This poor performance may be due to the limited size of the data set. Overall, one of these models would offer reasonable accuracy and robustness to meet the criteria required for the first stage of in-situ novelty detection embedded in the 8-bit microprocessor.

Figure 5.15 shows the final output of the trained LR classifier for the unseen test data compared to the correct labels. Figure 5.15b shows the output for a model considering only data from handpumps operating at similar depths (>25 m) and Figure 5.15a shows the output when features of deep and shallow pumps are mixed. In both cases, there is an

Table 5.7: Performance metrics for the five different classification models proposed in this chapter. Results are presented with mean and standard deviation of 1000 iterations.

Model	CA	Sens.	Spes.	AUC
LR	0.725(0.01)	0.731(0.02)	0.720(0.02)	0.724(0.01)
KNN	0.814(0.01)	0.834(0.02)	0.793(0.02)	0.814(0.01)
NBC	0.651(0.02)	0.462(0.04)	0.840(0.02)	0.651(0.02)
CART	0.726(0.01)	0.734(0.02)	0.718(0.02)	0.726(0.01)
NN	0.707(0.05)	0.708(0.17)	0.706(0.08)	0.707(0.05)

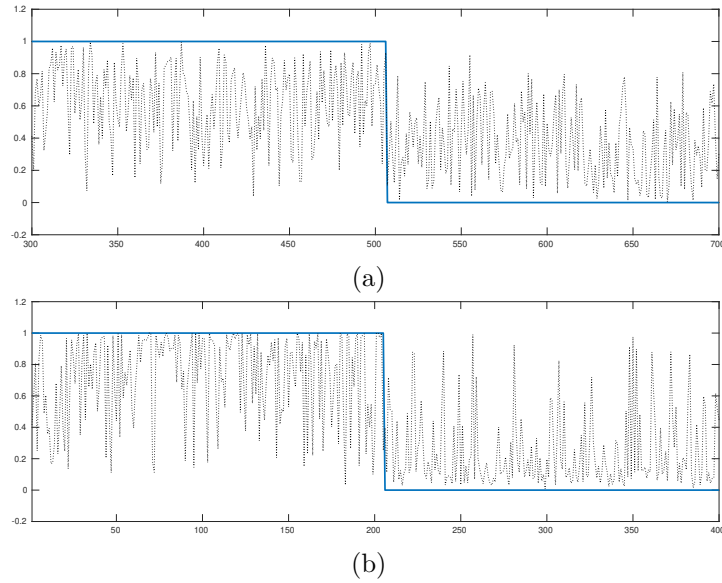


Figure 5.15: Predicted (black :) and actual (blue) outputs of the LR classifier for FV_4 for (a) mixed depth and (b) grouped by depth, respectively. Note that there is no ordering of data. The horizontal axis shows the index of feature vectors, which are independent and which are not consecutive in time.

evident change in the trend of predictions for normal and abnormal conditions, showing that the LR classifier can be used to classify the condition of a handpump given the selected feature set. However, depth-specific models, training using data from only deep handpump are more sensitive.

5.5 Discussion

The results described in this chapter show that it is possible to determine the condition of a rural handpump based on the vibration signatures collected from a sensor retrofitted inside the handle. However, sensitivity and accuracy can be improved by including knowledge

of the specific system, such as the operating depth.

Fourier analysis was used to extract frequency-based features that describe the characteristics of the data, while additional steps, such as floating-point handling and reducing the sampling resolution, enable implementing the models on the resource-constrained hardware. It was shown that separate clustering of normal and abnormal data occurs while variability exists across individual pumps. Further separation of clusters occurs among abnormal data for different failure types (Figure 5.3). The effect of the weight in the system and pumping speed were considered for the selection of the most informative vibration features. The weight in the system and subsequently, the operating depth of the handpump was shown to be factors that potentially affect the distribution of feature vectors. However, at 25 m the distribution of deep and shallow labelled groups appear to be unaffected by the pumping speed (Figure 5.9).

The use of simple classifier models requires well-selected features. As such, an in-depth analysis of the proposed features was presented before testing each of the models. Six sets of feature vectors, FV_1 - FV_6 , were proposed and investigated. Of the set of feature vectors, FV_4 (uniformly sampled across the entire frequency range) was the most accurate to describe the variability in the given general data set. However, all the proposed sets, FV_1 to FV_6 are sufficiently accurate and robust to meet the criteria for a sensor-based novelty filter. Reducing the number of feature vectors from 64 to 20 points, dramatically reduced the classifier accuracy. However, given that performance has already been sacrificed to gain a reduction in data transmission, a reduction in the number of features could be considered to optimise the final system and will be discussed in the next chapter.

Given the trade-off between accuracy and computation cost, the best-performing classifier may not be the ideal choice for a sensor-based novelty filter. A sensor-based model

requires a high level of sensitivity with the minimum processing during run-time. While the KNN model is the best performing on accuracy and robustness alone, the delay in inference comes at a high computation cost which is not suitable for the intended application of a sensor-based novelty filter. As a result, other models should be considered for the final selection.

5.6 Conclusion

This chapter validated the accuracy and robustness of a simple classifier embedded onboard the system to distinguish between normal and abnormal handpump conditions for the proposed feature vectors, FV_1 to FV_6 . Uniformly selecting 20 points across the entire frequency range was shown to be the most robust feature set. An SVM classifier was used as the gold standard to compare the performance of five proposed classifier models. Given the data set and features considered, the LR classifier was able to perform condition monitoring with a reasonable level of sensitivity, $AUC = 0.725(0.01)$, using unseen data with mixed depth characteristics. Eliminating deterministic features, such as the operating depth of the handpump, that can be learned by the model improve the performance of the trained model. Finally, the LR classifier performed condition monitoring with a satisfactory level of sensitivity, $AUC = 0.821$ and 0.806 , for deep and shallow operating handpumps, respectively. Overall, the LR model performs sufficiently well and offers the ability to update the model with new data. The probabilistic framework allows for the dynamic adjustment of the classification threshold, making it ideal for a distributed inference system that is proposed in Chapter 6. In the next chapter, the description and analysis of dynamically adjusting the LR threshold to transmit a small package of data to a cloud-based classifier is discussed. This approach will determine the trade-off between accuracy and cost of transmission in a dynamic system.

Chapter 6

Cloud-based Dynamic Distributed Inference System

The previous chapter presented the development of a model to perform novelty filtering onboard the embedded system installed inside a handpump handle. Given the hardware constraints, the approach proposed five “lightweight” machine learning methods as potential suitable sensor-based classifiers: LR, KNN, NBC, CART, and NN. Visualisation of the data indicated that certain spectral bands contain more interesting vibration content than other bands. The investigation considered six different feature vectors, $FV_{[1...6]}$, containing the twenty most prominent frequencies. The comparison showed that FV_4 , uniformly sampled across the entire frequency range, was the set of features that yielded highest overall accuracy, both for a generalised model and depth-specific models. Analysis of the classifiers indicated that all the methods were sufficiently accurate and robust for a sensor-based application, while KNN was the best and NBC was the worst-performing. We concluded that the most appropriate method for the proposed application was LR, given its ability to use gradient descent to update the model online and implement an adjustable threshold.

A two-step approach is used in this chapter to investigate more advanced machine learning methods, which are not bound by the limited resources in-situ, and to build on

the results from the previous chapter. This chapter presents a framework for improved predictive accuracy, where (1) the sensor-based novelty filter identifies and flags subsets of data as potential failures, (2) the “intelligent” summaries of these data subsets are sent to a “cloud-implemented” base station, and (3) the base station performs more advanced machine learning approaches. A dynamic framework is developed for measuring, processing, and transmitting data between rural nodes and a central cloud-based server such that the processing and transmission can be updated based on the current health status of the system. This approach would not only increase the specificity of novelty detection but more importantly, optimise the allocation of limited computational resources by strategically scheduling new measurements and data transmission (known as “active sampling”).

This chapter presents the development of the cloud-based dynamic distributed inference (DDI) framework based on four relatively heavyweight (in terms of computational resource) machine learning classifier methods: support vector machines (SVM), random forests (RF), multilayer perceptrons (MLP) and convolutional neural networks (CNN). This chapter will first review the real-world performance of the selected sensor-based novelty filter used to create the “intelligent” summaries, then describe the performance of the cloud-based methods and finally evaluate the results of the proposed dynamic distributed inference (DDI) framework during active sampling.

6.1 From Sensor to Cloud

The analyses in Chapter 5 showed that resource-constrained devices could use lightweight, onboard machine learning approaches to perform anomaly detection in the embedded system to improve the use of its limited resources. Distributing inference between the embedded system at the pump and powerful cloud-based machine learning algorithms promises to offer robust information without the need for expensive hardware or sensors

embedded in situ – making the possibility of a large-scale (and perhaps even continent-wide) monitoring system feasible - as is the ultimate goal of our work.

Moreover, a large-scale predictive and scalable health monitoring system for hand-pumps will create a data stream that provides reliable data to connect the water user, the environment, and infrastructure investors in order to overcome unequal water access and uneven distribution of resources. Real-time water data will allow evaluation of locations to drive business and financial transactions far beyond what is currently achievable. The sustainability of a community handpump reaches beyond the community and the pump to include the aquifer, the well, the maintenance system, and associated finance [2].

Without a platform to host the data (and store, and post-process early warning flags generated by the sensor-based health monitoring loggers) it is challenging to reach the potential and scale required by a region-wide remote health monitoring system. Updating the system health status to adjust processing and transmission dynamically is an important step to improve the overall specificity while reducing the computational needs. This is important because the need for extremely low-cost sensing in these rural areas results in a low signal-to-noise ratio of the processing hardware which severely compromises the performance of machine learning models fitted to this type of data. Incorporating more heavyweight condition monitoring methods on the cloud-based platform will increase the system’s positive predictive value.

In this novel dynamic platform, the sensor, as the first point of contact with the system, will continue to perform the lightweight health monitoring introduced in the previous chapter. Given the likelihood that the estimate is accurate, the sensor could adjust the amount of data being transmitted to the central servers accordingly. For higher likelihoods, it would only need to transmit novelty scores, but as the likelihood of the estimated

accuracy decreases, sending either features or raw waveforms will allow a cloud-based server to perform more heavyweight estimation algorithms.

This will require research into the “optimal” means of representing the data, which should increase the fidelity as the risk of deterioration of the handpump is deemed to be increasing. The “optimal” data packages can be obtained by distributing inference between the embedded sensor (low complexity, high sensitivity, low specificity) and the cloud (high complexity, improved sensitivity, improved specificity). This interaction between, for example, the LR models at the pumps and more heavyweight machine learning processes in the cloud, is a novel research field which will be described in more detail in the following sections.

6.1.1 Distributed Inference Model

Distributed Inference Networks (DIN) rely on the information available from the distributed sensing nodes across large-scale wireless sensor networks to detect, estimate, and track changes by performing probabilistic dynamic inference about a given phenomenon of interest [206, 207]. Traditional frameworks of centralised inference networks are widely used in medical, environmental, and surveillance monitoring applications by transmitting raw observations from the individual nodes to a fusion centre where global inference regarding the network is made. However, transmitting raw observations can be costly and impractical as these require substantial bandwidth, energy, or both and in the case of rural monitoring can even be impossible. DINs overcome this limitation by transmitting compressed summaries obtained by performing local processing techniques of the raw observations before transmission to the fusion centre [208].

A parallel DIN configuration has neither any feedback from the central network back to the sensors, nor is there any communication between individual sensors within the

network. Local sensors and the base station (fusion centre) apply decision rules to optimise decision performance. The Neyman-Pearson formulation and the Bayesian criterion are most widely used to design decision rules [209]. For the proposed dynamic framework, feedback exists between the local sensors and base station to allow adjusting the LR threshold or requesting alternative data packages, like feature vectors rather than novelty scores.

6.1.2 System Requirements for Rural Monitoring

In the case of rural water supply networks, the hypothesis is that some lightweight machine learning techniques, in the form of an LR novelty filter (as shown in Chapter 5), can be applied at the node (i.e. handpumps) to perform initial processing of the raw observation before transmission. The sensor-based novelty filter threshold can be used to define the proportion of data to be transmitted from the sensor given the onboard assessment the handpump health-status.

Observing and monitoring individual handpumps in the network presents several challenges, although not unique to this type of network, and can often be made worse by the extreme operating environment of these sensor nodes. These failures result in corrupted data caused by three leading causes: (1) channel/sensor noise, (2) faulty sensors (both hardware and firmware failure), and (3) handpump failure (resulting in disuse of the handpump and thus no vibration data generated by the handpump).

The DDI health-monitoring system should aim to: (i) use existing low-cost, embedded sensors to acquire accelerometry data from the routine daily use of handpumps, (ii) perform on-pump signal condition monitoring and 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-based base station via the domestic telecom-

munications network, and (v) perform more advanced machine learning techniques on the cloud-based system to increase the accuracy and certainty of the classified health condition of the equipment. Figure 6.1 shows an overview of the proposed system consisting of three main sections:

A. Sensor node contains the embedded sensor, battery, and data transmitter. In this study, the embedded system attached to the handpump handle to record vibrations consists of three essential elements: (a) an IC-based, 96 Hz accelerometer; (b) an 8-bit microprocessor; and (c) a GSM modem. The estimated power consumption of this node in run and sleep mode is compared in Table 6.1. Selected components are sufficiently efficient to operate given the power restrictions of the proposed application. Alternatively, nodes could make provision for other sensors depending on the infrastructure type that is being monitored, like power meters in solar home units, heart rate monitors in patients or temperature and humidity sensors in agriculture. This network could contain hundreds of nodes in a small geographical region covered by the local telecommunications network and managed by the local maintenance delivery team.

B. Sensor-based (on-pump) novelty filter performs real-time pre-processing of the data acquired during pumping and analyses the data using an LR novelty filter to produce “intelligent” data summaries that aim to flag potential handpump failure.

Table 6.1: Estimated power consumption of essential components in proposed embedded system operating at 25°C.

Component	Run Mode	Sleep Mode
Microprocessor (8-bit at 8 MHz)	20 mA	5 mA
Accelerometer at 96 Hz	200 μ A	0.5 μ A
GSM Module (for SMS)	250 mA	40 μ A

*Approximation only. Actual consumption will depend on the components selected and operating conditions.

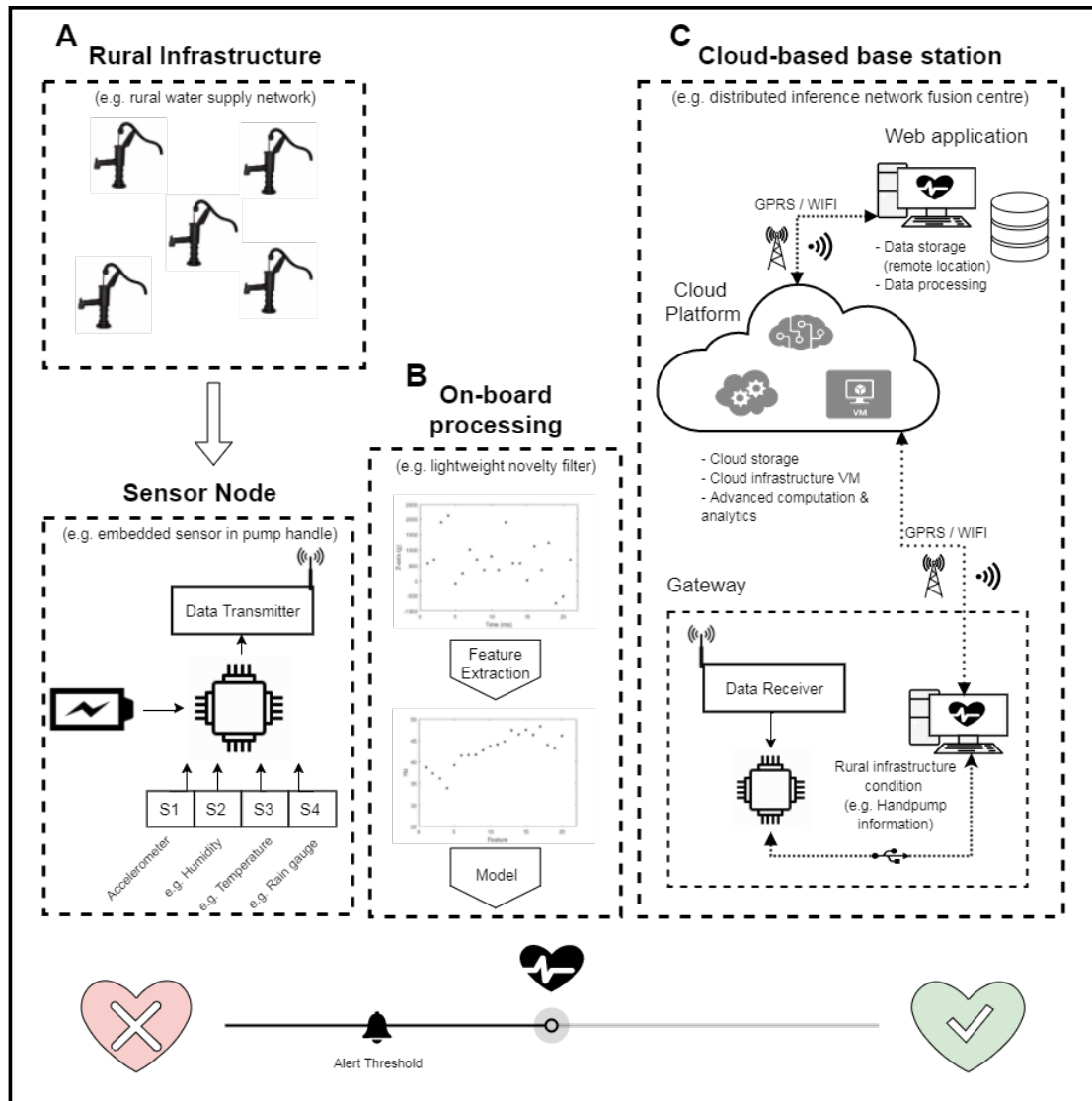


Figure 6.1: Schematic diagram of the proposed DDI health monitoring system for a rural water supply network consisting of: (A) Handpump sensor node with related embedded sensors, (B) Onboard lightweight novelty filter producing data summaries, and (C) Gateway node and heavyweight cloud-based computing and analytics.

C. Cloud-based distributed inference system performs more complex processing of the data summaries using advanced machine learning methods that are not constrained by the same limited processing capabilities of the sensor-based classifier.

Finally, the alert threshold is a user-defined boundary at which point the DDI suspects deterioration to the point that requires action to be taken, such as a visit by a mechanic.

The existing system already performs essential communication by using the domestic telecommunications network; however, this is limited to simple text summaries of hourly usage based on the number of handpump strokes performed [80]. The new sensors used in this thesis contain larger processors and onboard memory to collect and store high-frequency data for this proof-of-concept study.

6.1.3 Intra- vs Inter-node Monitoring

Node (or handpump) specific monitoring requires that a unique model is constructed for each node in the system using only data previously seen from that specific node, such an approach can be referred to as an intra-node monitoring system. This is in contrast to an inter-node monitoring system using population-generic data from many nodes to construct a model. This approach has the advantage that both the model, $M(\theta)$, and the novelty threshold can be determined before deployment for an entire population albeit it at the cost of performance accuracy. We will consider the results for both where $\mathcal{D}_m$ and $\mathcal{D}_{d,1}$ are used to train inter-handpump models and $\mathcal{D}_{d,2}$ is used to train an intra-handpump model.

6.2 Heavyweight Cloud-based Classifiers

The previous chapter presented the development of lightweight classifier methods that are suitable for use onboard the sensor node. In this chapter, the proposed DDI framework will transmit intelligent data packages to a cloud server where processing power and batter

is less restricted. The availability of resources in the cloud allows the system to use more advanced classifier methods to increase the fidelity. As outlined previously, the four proposed methods for the cloud-based classifier include two supervised machine learning methods: support vector machines (SVM) and random forests (RF) and two deep learning methods: multilayer perceptron (MLP) and convolutional neural network (CNN). The SVM was introduced in Chapter 5.

6.2.1 Random Forests

Random forests (RFs) [210] are an ensemble learning method which predicts the class based on the votes of a multitude of individual decision trees. The decision tree structures are a branching composite of binary decisions used to separate data into classes. Although individual trees are easy to interpret, the single tree often over-fits the training data, giving it high variance and resulting in inaccurate predictions. However, over-fitting to the training set can be minimised by constructing a multitude of individual decisions and reduce the variance in prediction. However, each tree, at each node, only has access to a subset of the data and a randomly selected percentage of features, resulting in many “weak learners”.

RFs are built on two concepts: boosting and bagging. **Boosting** [211] is a greedy algorithm approach to improve the prediction power of the decision tree which relies on the idea that many uncorrelated weak classifiers can construct a robust classifier. **Bagging** or **bootstrap aggregating** [212], used in this study, is the approach of repeatedly sampling from a set of training points (bootstrapping) without replacement and then determines the result by taking the majority of individual votes per class (aggregating) by calculating the ensemble [181]:

$$f(\mathbf{x}) = \sum_{m=1}^M \frac{1}{M} f_m(\mathbf{x}) \quad (6.1)$$

where f_m is the m^{th} tree.

Three hyperparameters of the RF algorithm that should be optimised include: (1) the number of trees, (2) the number of feature vectors in each decision tree, and (3) the proportion of training data to be bootstrapped.

In RFs, a random selection of a subset of features, Θ_k , and a random subset of the training data, $D(t)$, is used to grow each decision tree, T . At each node, t , of the tree, the split $s_t = s^*$ to separate the input vector, X , is chosen to minimise the impurity, $i(t)$, in the class labels [213]. Impurity is measured in multiple ways; in this thesis the *classification error* is used to minimise misclassification such that:

$$i_E(t) = 1 - \max\{p_c\} \quad (6.2)$$

where p_c is the probability of a class c . Minimising the impurity is equivalent to maximizing the information gain at each node.

The importance of the variable input feature X to predict the output depends on their weighted impact on decreasing the impurity of that node for all N_T trees in the forest:

$$Imp(X_m) = \frac{1}{N_T} \sum_T \sum_{t \in T: v(s_t)=X_m} \frac{N_t}{N} \Delta i(s_t, t) \quad (6.3)$$

where v_s is the variable used in split s_t .

Although each random vector generated is independent of past random vectors, the distribution of all random vectors $\Theta_1 \dots \Theta_k$ is the same [210].

6.2.2 Multilayer Perceptron

A multilayer perceptron (MLP) is a deep neural network consisting of multiple perceptrons and is an evolution of the feed-forward neural network described in the previous chapter. However, the approach now has multiple hidden layers between the input and output layers rather than only one. Similar to supervised learning, MLPs train on input-output pairs in order to learn the correlations between each. Backpropagation or gradient descent minimises the error by adjusting the weights and biases. In backpropagation the input vector, x_n , is applied to a general feed-forward network to compute the weighted sum of the inputs, $a_j = \sum_i w_{ji}z_i$. A non-linear activation function $h(\cdot)$ is used to transform this sum such that the activation of unit j becomes [205]:

$$z_j = h(a_j) \quad (6.4)$$

The outputs are evaluated using the error-function derivatives, $\delta_k = y_k - t_k$, using the chain rule for partial derivatives such that:

$$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i \quad (6.5)$$

where $\delta_j = \frac{\partial E_n}{\partial a_j}$ and $z_i = \frac{a_j}{\partial w_{ji}}$.

The δ_j of each hidden unit is obtained by using the backpropagation formula:

$$\delta_j = h'(a_j) \sum_k w_{kj} \delta_k \quad (6.6)$$

The required derivatives are updated using:

$$\frac{\delta E_n}{\delta w_{ji}} = \delta_j z_i \quad (6.7)$$

This process repeats until the error is minimized.

6.2.3 Convolutional Neural Networks

Convolutional Neural Networks (CNN) [214, 215] is a type of MLP that builds translational invariance into the structure such that it can classify consistent patterns. CNN offers the ability to capture temporal dependencies in the input data by applying relevant filters or convolutional layers before the conventional fully connected layers, such as in the standard MLP. CNNs are typically easier to train than fully connected networks, given that it has fewer parameters than a fully connected network with the same number of hidden layers. The convolutional layer has a feature map that takes only subregions from the original input, which are combined using the weights and the bias, and transformed by [205]:

$$y(\mathbf{x}, \mathbf{w}) = f \left(\sum_{j=1}^M w_j \phi_j(\mathbf{x}) \right) \quad (6.8)$$

where $f(\cdot)$ is a nonlinear sigmoid activation function. This mechanism is known as local receptive fields and derives from the idea that local features are dependant only on a small subregion. Local receptive fields in the hidden units tie the weights, reducing the overall number of features. Furthermore, if weight sharing greatly reduces the number of weight parameters required based on the assumption, then a local feature map can be shifted to different locations to detect the same pattern.

The CNN accepts a volume of size $W_1 \times H_1 \times D_1$ to produces a volume of size $W_2 \times H_2 \times D_2$, where $W_2 = H_2 = \frac{W_1 - F + 2P}{S + 1}$ by symmetry and $D_2 = K$.

While 2D-CNNs are popular for image analysis, the use of 1D-CNN is appropriate for the analysis of time sequence data, such as our accelerometer data acquired from handpumps. A 1D-CNN is useful for smaller segments of the entire data set and where the feature location is irrelevant. Accelerometry data has two dimensions, the time-steps and secondly the gravitational force of the three axes.

6.3 Assessment of Cloud-based Classifier

Results in Chapter 5 has shown that it is possible to create “intelligent” data summaries on-board the handpump using lightweight machine learning techniques. However, to increase the fidelity of the classifier, the DDI platform requires more advance machine learning techniques to process these summaries in the cloud. The present analysis aims to design and validate a suitable heavyweight machine learning method that can be used in the cloud of the DDI platform.

6.3.1 Method

The goal of the cloud-based classifier is to enhance the DDI platform fidelity without the need for extra resources. This section describes the method used to design and validate the final DDI system by: (1) evaluating the novelty scores generated onboard the handpump as input to the cloud-based classifier, and (2) evaluating the performance of the proposed cloud-based methods given the novelty filtered as input.

Novelty Filter Score Estimation

Following the analysis in Chapter 5, a novelty score is produced in-situ at the remote node, $Q_{n,i}$, onboard operating handpumps. Due to the lightweight processing requirement of the system-based novelty filter used during this feasibility study, the temporal dependence of the accelerometer observations was not considered at the node. Instead, it was done during post-processing by aggregating the classifier scores over consecutive examples to varying

degrees by applying a moving average window and increasing the size of the window from 7 s to 27 s. This produced three lightweight condition scores, $Q_{n,i}$, with $i = 1...3$ where $i = 1$ is equivalent to the raw novelty scores generated on the sensor hardware without temporal dependence, $i = 2$ is equivalent to post-processed novelty scores using 7 s moving average window, and $i = 3$ is equivalent to post-processed novelty scores using 27 s moving average window.

The in-situ condition score, $Q_{n,1}$, was then used to filter the transmitted data such that data summaries sent to the cloud-based classifier contain only abnormal examples, as labelled by the system-based novelty filter. Finally, the novelty filtered data was used to produce condition scores for each of the four cloud-based classifier methods (SVM, RF, MLP and CNN).

Cloud-based Classifier Training Framework

Five-fold cross-validation was used to optimise the hyperparameters required for each of the classifiers which include: the LR regularisation parameter, λ (Eq. 5.3), the SVM soft margin parameter, C (Eq. 5.13), the SVM RBF width, σ (Eq. 5.15), the number of RF trees, T , and proportion of features selected. Once these parameters were selected for each classifier, the final algorithm was trained.

To create the unseen test data set, 20% of the data was held out to validate the final trained model and confirm the ability of the model to generalise “out-of-sample”. In each case, the data set is balanced such that it contains the same number examples for each class (i.e. abnormal and normal handpump condition). This is repeated for each classifier algorithm. The mean and standard deviation of the area under the receiver operator curve (AUROC) is calculated and recorded over multiple iterations, as shown in Algorithm 1.

Evaluating Classifier Performance

The receiver operating characteristic (ROC) is used to compare the performance of the different classifiers to verify reliability. This metric compares the actual and predicted outputs for each class. The true positive rate (TPR), or *sensitivity*, of a classifier 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 pump failure detection while simultaneously minimising false alarms, which can be costly in real-life. In the ideal case, the classifier would be maximally sensitive ($TPR = 1$) with no false alarms ($FPR = 0$).

6.3.2 Results

Visualisation of Data

Principal component analysis (PCA) [216] was used to reduce the dimensionality of the data set to allow visual inspection of the underlying structures. This was done by projecting the 20 feature vectors, corresponding to each 128-component FFT window extracted from the accelerometry wavelet, in two- and three-dimensional space. The resulting PCA

Algorithm 1: Classifier Learning Framework

Input: Data set of features and class labels

for $i \leftarrow 1$ **to** 100 **do** // For each of the iterations

 Balance and split data into train and test;

for $f \leftarrow 1$ **to** 5 **do** // For each of the five folds

 Optimise the hyperparameters for every classifier;

 Train model using training set;

 Evaluate model performance using test set;

end

end

Take the average performance and standard deviation of for each iteration;

Result: AUC average and standard deviation for each classifier type

plots show tight clustering of specific conditions for a deep intra-handpump system, $\mathcal{D}_{d,2}$, in Figure 6.2a. In a 2D PCA, both deep ($\mathcal{D}_{d,1}$) and deep/shallow ($\mathcal{D}_m$) inter-handpump sets appear to have limited clustering of the conditions, as seen in Figure 6.2b-6.2c. When viewed in 3D, however; apparent separate clustering of conditions appear in Figure 6.2d.

Sensor-based Novelty Filtering with Real Field Data: Lab vs Real World

In addition to recording the raw accelerometry data at 96Hz, the sensor-based LR classifier also produces a novelty score (between zero and one) for every 128 samples, which is equivalent to one FFT bin. The scores produced on the handpump are compared to

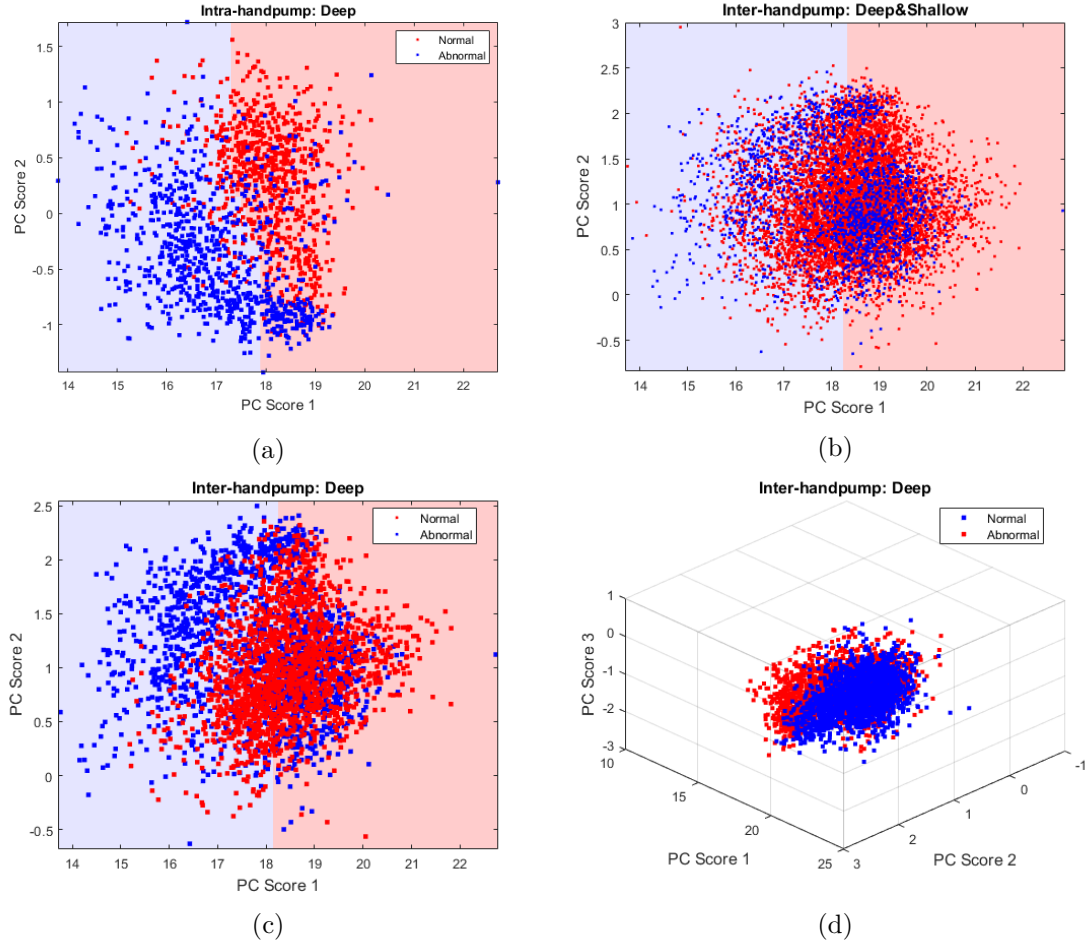


Figure 6.2: Lower dimensional PCA representation of Y-features in 2D for (a) a single deep operating handpump $\mathcal{D}_{d,2}$, (b) deep and shallow operating handpumps $\mathcal{D}_m$, (c) deep operating handpumps $\mathcal{D}_{d,1}$, and (d) the same group of deep operating handpumps $\mathcal{D}_{d,1}$ in 3D. The blue and red examples correspond to normal and abnormal handpump operations, respectively.

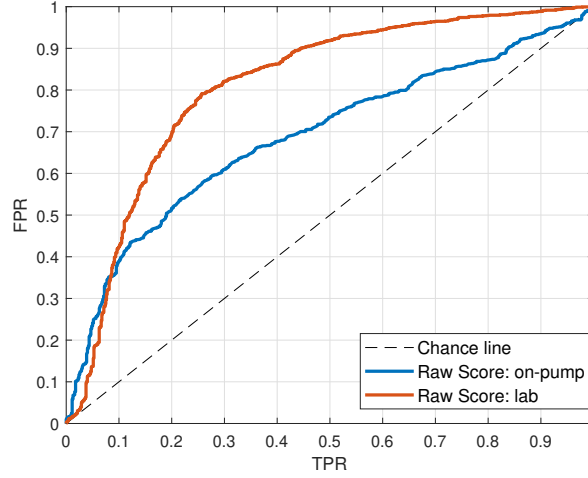


Figure 6.3: Comparison of the ROCs for novelty scores generated by the sensor-based novelty filter (blue) and novelty scores generated in the lab using the relevant raw data (red).

the equivalent scores produced in the lab using the relevant raw data, under ideal circumstances without the limitations of the sensor-based processor. Figure 6.3 confirms that the on-pump performance ($AUC_{on-pump} = 68.9$) is slightly compromised by the limited onboard processing power when compared to the performance in the lab environment ($AUC_{lab} = 81.0$). However, the on-pump novelty filter performs sufficiently well despite the limitations and given the requirement for low complexity to create data packages for the cloud-based processor.

Sensor-based Novelty Filtering with Real Field Data: Depth-specific Models

During this investigation, three different versions (V1.1 - 1.3) of the proposed on-pump LR classifiers were deployed in the field. The details of these are summarised in Table 6.2. During different field visits throughout six months, each version was manually re-trained and updated as data collection was ongoing.

Figure 6.4 compares the output of the field-generated classifier V1.2 from the embedded sensor on the handpumps with lab-generated scores from different classifier versions for

Table 6.2: Overview of the different sensor-based novelty filter versions that were deployed in the field during the course of this research.

Version	Training Set	Description
V1.1	$\mathcal{D}_m$	Shallow and deep operating handpumps
V1.2	$\mathcal{D}_{d,1}$	Deep operating handpumps
V1.3	$\mathcal{D}_{d,2}$	Shallow operating handpumps

the same handpump accelerometry data. Figure 6.4a shows the output for a deep inter-handpump setting. In this setting, the depth-specific versions, V1.2 and V1.3, of the classifier obtain higher AUCs ($AUC_{V1.2} = 61.3$ and $AUC_{V1.3} = 55.4$) than a mixed-depth classifier V1.1 ($AUC_{V1.1} = 55.2$) which was trained using data from both shallow and deep operating handpumps. The field-based on-pump V1.2 classifier has a lower AUC ($AUC_{PumpV1.2} = 56.9$) than the lab-generated scores. From the previous comparison of the lab and real-world scores; however, it is to be expected that classifier V1.2 will slightly suffer when implemented onboard the sensor given the processing limitations. A similar result is observed for a deep intra-handpump setting, shown in Figure 6.4b. The lab-generated classifier V1.2 ($AUC_{V1.2} = 81.0$) outperforms the field-based counterpart of V1.2 ($AUC_{PumpV1.2} = 68.9$). In this scenario, the mixed-depth classifier V1.1 is able to achieve a higher AUC than the shallow classifier V1.3, $AUC_{V1.1} = 75.3$ and $AUC_{V1.3} = 69.3$, respectively.

In both inter- and intra-handpump systems, the unique deep operating classifier, V1.2, outperforms the mixed-depth classifier, V1.1, and the shallow operating classifier, V1.3. The lab-generated scores of V1.2 again offer an improvement of the sensor-based LR classifier field implementation of V1.2. For the intra-handpump setting, the mixed-depth classifier V1.1 performs better than the shallow classifier V1.3. This highlights the importance of having a balanced representation of deep- and shallow-operating handpumps in order to train a robust mixed-depth classifier - our ultimate goal given that wide-scale

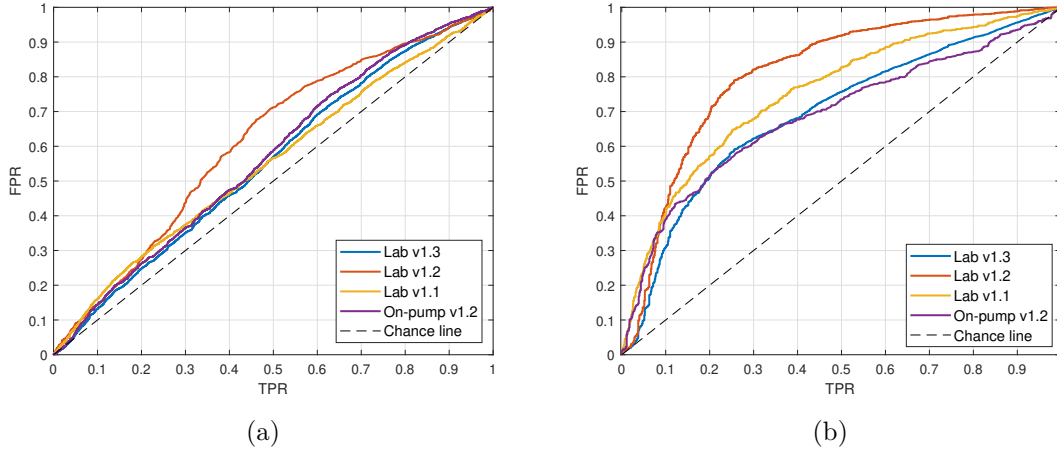


Figure 6.4: A comparison of the classifier performance for varying three different classifiers V1.1-1.3 in: (a) inter-handpump and (b) intra-handpump settings (left to right plots, respectively).

implementation of depth-specific classifiers will not be feasible. Since these versions were developed during the initial stages of investigation using the original data set, the data contained more examples of shallow operating handpumps as these were easier to access and monitor initially. However, as findings in Chapter 5 indicated the importance of the features related to the operating depth of the handpump, the data collection strategy was updated to include more examples of deep operating handpumps. The next results demonstrate that a balanced representation of deep and shallow operating handpumps in a mixed-depth classifier will perform adequately such that it can be widely deployed at a region-wide scale.

As the data collection for deep operating handpumps expanded, Figures 6.5 (a-b) compares the ROCs for inter-handpump systems, where (a) is a depth-specific classifier, $\mathcal{D}_{d,1}$, and (b) is a mixed-depth classifier, $\mathcal{D}_m$. Unlike before, when there was an under-representation of deep operating handpumps in the mixed-depth set ($AUC = 61.6$), the ROCs now look nearly identical for classifying a deep operating set ($AUC = 61.3$). In both cases, adding median filtering or taking the moving average of scores for fixed window

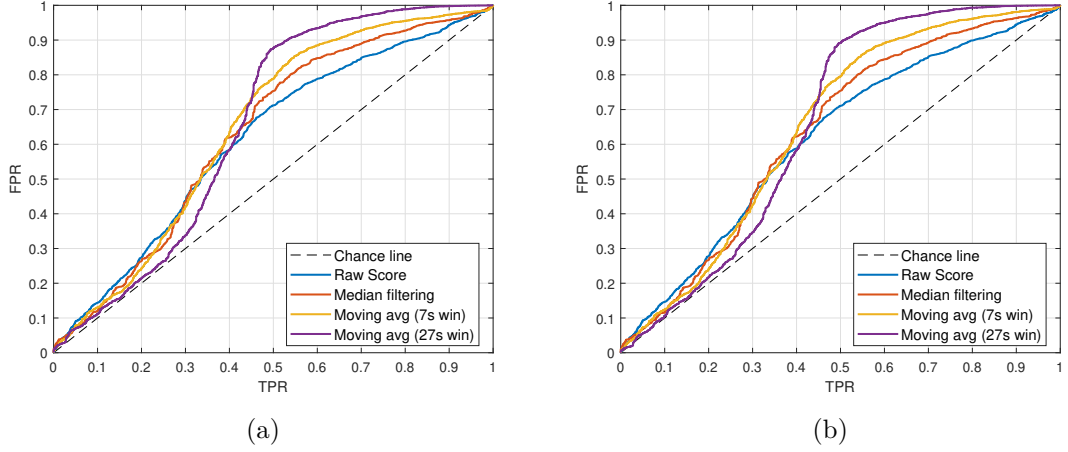


Figure 6.5: A comparison of the on-pump classifier performance for failures in inter-handpump systems: (a) a depth-specific classifier, $\mathcal{D}_{d,1}$, and (b) a mixed-depth classifier, $\mathcal{D}_m$. Novelty scores generated on the handpump (raw scores) are shown in addition to median filtered and moving averages.

sizes (7 s and 27 s) improves the output both for the deep operating set ($AUC_{raw} = 61.3$, $AUC_{med.filt.} = 63.1$, $AUC_{7sMA} = 64.8$, $AUC_{27sMA} = 65.3$) and for the mixed-depth set ($AUC_{raw} = 61.6$, $AUC_{med.filt.} = 63.4$, $AUC_{7sMA} = 65.1$, $AUC_{27sMA} = 65.7$). Given that the training and test set now contains a sufficiently large number of both deep and shallow operating handpumps, the next section will further investigate the performance of inter- and intra-handpump classifiers that can scale across large regions without the need for unique classifiers.

Inter- vs Intra-Handpump Models

Table 6.3 shows that the intra-handpump classifier, $(Q_{n,i}, \mathcal{D}_{d,2})$ pairs, performs substantially better than the inter-handpump classifiers, $(Q_{n,i}, \mathcal{D}_m)$ and $(Q_{n,i}, \mathcal{D}_{d,1})$ pairs, achieving up to 86.2 per cent AUC compared to 65.7 per cent. In all three classifiers, the performance improves when the temporal correlation is incorporated by applying moving averages. For this analysis, the performance of the general inter-handpump classifier is enough to use as a lightweight novelty filter since the large-scale implementation of pump-specific or depth-specific classifiers would be too costly and unrealistic to roll-out across

Table 6.3: Results for field-based, $Q_{n,i}$, and lab-simulated, $Q_{n,lab}$, sensor-based novelty filter classification performance, given the mean AUC of 20 iterations (one standard deviation).

	AUC $\pm$ (SD)		
	$\mathcal{D}_m$	$\mathcal{D}_{d,1}$	$\mathcal{D}_{d,2}$
$Q_{n,1}$	0.613(0.007)	0.616(0.004)	0.689(0.003)
$Q_{n,2}$	0.648(0.004)	0.651(0.002)	0.773(0.002)
$Q_{n,3}$	0.653(0.001)	0.657(0.001)	0.862(0.001)
$Q_{n,lab}$	0.689(0.003)	0.698(0.002)	0.880(0.002)

entire region-wide rural water supply networks.

The sensor-based novelty filter performance, $Q_{n,i}$, was compared to a baseline control score, $Q_{n,lab}$, generated in the lab using the same original data but assuming no processing or power constraints as would be experienced onboard the local novelty filter. In all three cases, the lab generated scores, $Q_{n,lab}$, improve those generated by the sensor-based classifier, $Q_{n,i}$, by 7.5 to 12.1 per cent. Given the limitations of the embedded system, the hypothesis was that the accuracy of the on-pump classifiers would suffer compared to the lab-simulated results. Due to the lightweight processing requirement of the onboard classifier, the temporal dependence of the accelerometer observations has not been considered. However, using the moving averages of the novelty scores indicates that the classifier performance improves when aggregating the classifier scores to incorporate temporal correlation over consecutive examples (to varying degrees as the moving average window size is increased 7 s to 27 s). This type of post-processing is lightweight and easy to implement onboard the handpump to improve the system-based novelty scores, which will bring it nearly on par with the lab-simulated results.

For the novelty scores recorded onboard the handpump, the probability density of the output score, along with the fitted gamma distributions (often used to model asymmetric distributions of speed for rotating machines [74]), for normal (blue) and abnormal (red)

conditions are shown in Figure 6.6. The distributions for the on-pump scores, both normal and abnormal, lie very close to the decision boundary $[0.5]$ for all three data sets. As such, future work can increase the classifier performance by defining a region around this boundary where the classifier could decide not to produce a score. Requesting additional data, such as features or raw waveforms, when novelty scores are close to the decision boundary will increase the overall fidelity; this analysis will be discussed in section 6.4.

Sensor-based Novelty Filtering with Real Field Data: Failure Type

The intra-handpump setting is used to investigate the performance of the classifiers to detect specific failure types. Figures 6.7 (a-b) compare the ROCs for two different failure types represented in the intra-handpump setting, $\mathcal{D}_{d,2}$: (a) a broken handpump rod ($AUC_{raw} = 68.9$) and (b) a leak in the rising main ($AUC_{raw} = 52.3$). Extreme failures (like a broken pump rod) are easier to detect than less severe failure events. Using the moving averages over the scores improve the output more for a less severe failure ($AUC_{25sMA} = 86.6$) than a severe failure ($AUC_{25sMA} = 55.5$). The misclassification of less severe failures may result due to: (1) mislabelling of the data (for example, data with restricted water flow is labelled as abnormal when in reality it is neither entirely abnormal nor entirely normal) and (2) distance between the location of the malfunction and the

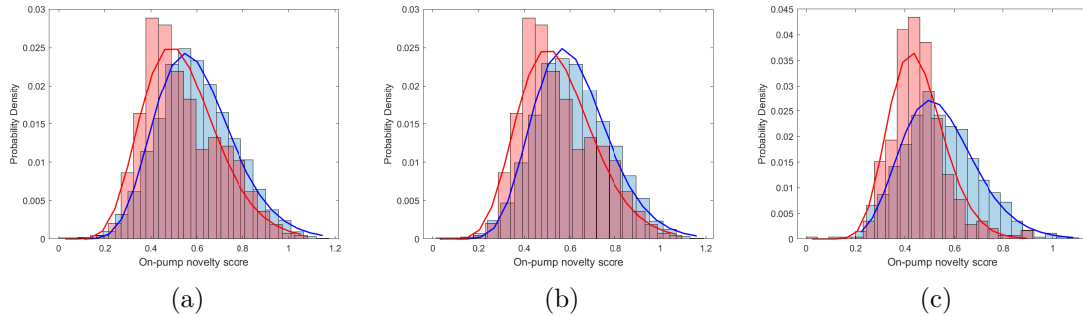


Figure 6.6: Probability density of on-pump novelty scores for normal (blue) and abnormal (red) handpump conditions in: (a) deep operating inter-handpump system, $\mathcal{D}_{d,1}$, (b) deep and shallow operating inter-handpump system, $\mathcal{D}_m$, and (c) deep operating intra-handpump system, $\mathcal{D}_{d,2}$.

sensor (for example, a leak in the rising main 50 m down the borehole is more challenging to measure than a failure on or near the pump body where the sensor is located).

For this application, the ability to detect more severe failures is important since these failures restrict water access entirely. Less severe failures, such as a leak in the rising main, tend to occur less frequently and lead only to restricted water flow initially. However, an ideal health-monitoring system should be able to identify many different failure types. Future data collection campaigns can aim to include a variety of failures, and the labelling framework used for handpump conditions can be extended to include multiple grades of failure in addition to the current binary labelling (normal and abnormal). Greater diversity in the representation of failures in the training set will assist the system to improve its ability to detect failures that are currently considered tricky.

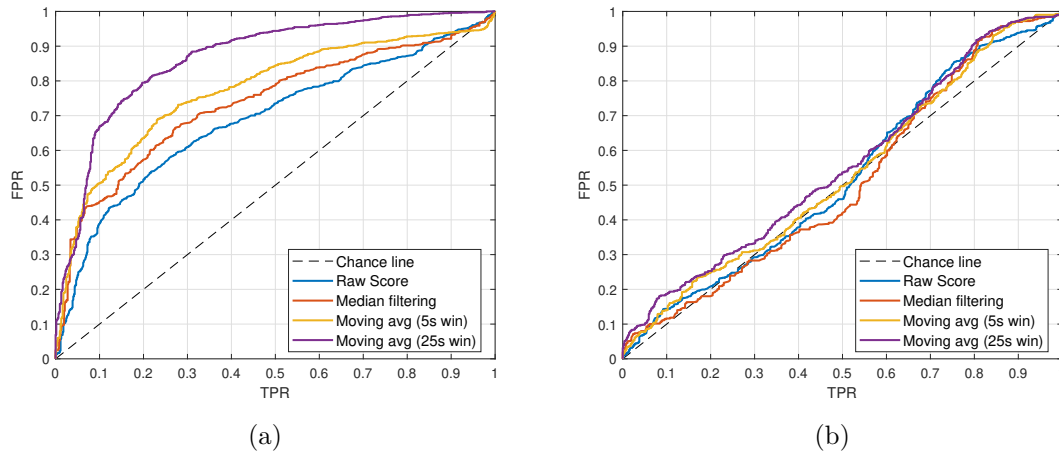


Figure 6.7: A comparison of the system-based novelty filter performance for different failures represented in the intra-handpump system: (a) broken pump rod and (b) leak in the rising main.

Cloud-based Classifier Performance using Novelty Filtered Data

The next stage of the DDI health monitoring system will involve performing heavyweight machine learning processing on the data flagged by the lightweight on-pump algorithm. For this initial investigation of the cloud-based classifier, a two-class classifier is considered

where the output of the mapping function is either normal or abnormal to specify the handpump condition. The LR classifier is compared to the RF and SVM classifiers.

The pump-based novelty filter ensures that the vast majority of data do not transmit under normal operating conditions. As the on-pump model suspects that the condition is degrading, only then will data be transmitted to the cloud. In most cases, the cloud-based system will then only receive data for novelty-filtered data related to likely abnormal conditions. To simulate the proposed operating scenario, the classifiers are trained using data which contain both normal and abnormal examples and tested using data which contain mostly abnormal examples, as would be the case for novelty-filtered data transmitted from the sensor to the cloud.

Table 6.4 compares the AUCs for the proposed classifiers using the three different data sets, $\mathcal{D}_m$, $\mathcal{D}_{d,1}$, and $\mathcal{D}_{d,2}$, where 80% of the data are used for testing and validation while the remainder is unseen test data. For both inter-handpump ($\mathcal{D}_m$, and $\mathcal{D}_{d,1}$) and intra-handpump ($\mathcal{D}_{d,2}$) cases, the LR classifier is outperformed by the RF. SVM performs better than the LR classifier only for the inter-handpump models ($\mathcal{D}_m$, and $\mathcal{D}_{d,1}$). The size of the data set in $\mathcal{D}_{d,2}$ that contains only example from a single handpump most likely limits the SVM performance. Similar to the system-based novelty filter results in section 6.3, all the cloud-based classifiers performance improve for the intra-handpump case. This was expected as an intra-handpump model would allow the algorithm to learn features unique to the specific handpump. Although it is not feasible to deploy intra-handpump models at a regional scale, this offers an important baseline from which to evaluate other models.

Figure 6.8 shows the AUC of the proposed methods when incrementally adjusting the proportion of test data or “raw data transmitted from the pump” to analyse the performance of the classifier as the amount of data being transmitted increases. In all

Table 6.4: Cloud-based condition classification using 100% of sensor-based processed data, given the mean AUC (one standard deviation) of 100 iterations.

	AUC $\pm$ (SD)		
	$\mathcal{D}_m$	$\mathcal{D}_{d,1}$	$\mathcal{D}_{d,2}$
LR	0.596(0.001)	0.600(0.001)	0.794(0.001)
RF	0.718(0.025)	0.709(0.020)	0.868(0.014)
SVM	0.608(0.002)	0.652(0.001)	0.750(0.001)

three cases, the LR classifier benefits least from the addition of extra data. Table 6.5 shows that the LR classifier reaches its optimum classification accuracy using only 6 to 15% of the raw data, compared to 89 to 98% for RFs and 97 to 100% for SVMs. However, the overall performance improvement for the RF model is marginal in all three cases, and thus the performance could be optimised by using only 15 to 20% of the overall data.

For both LR and RF, the depth-specific model ($\mathcal{D}_{d,1}$) and handpump-specific model ($\mathcal{D}_{d,2}$) offer only a marginal improvement on the performance of a more general model, which is trained using $\mathcal{D}_m$ that represents many different handpumps operating at varying depths. This shows that the practical deployment of the system could be simplified by using a model trained for both deep- and shallow-operating handpumps rather than opting for a depth-specific model. The SVM classifier performance; however, benefits significantly from the more extensive data set in $\mathcal{D}_m$ since the two-class SVMs are trained to have a low misclassification rate. It may be possible to improve the performance of an inter-handpump SVM classifier model if future data sets continue to collect more depth-specific data. The variance of the SVM classifier reduces as the proportion of data increases.

Overall, RF outperforms both the LR and SVM models for inter- and intra-handpump classifier systems and offer a 10% improvement from the system-based novelty scores. However, in contrast to LR and SVM, the variability in the prediction accuracy of the RF classifier does not decrease despite increasing the proportion of data supplied. This is a

Table 6.5: Results for lab simulated cloud-based condition classification of sensor-based processed data, given the mean AUC of 100 iterations (one standard deviation). Results show the maximum AUC achievable for a given proportion of data.

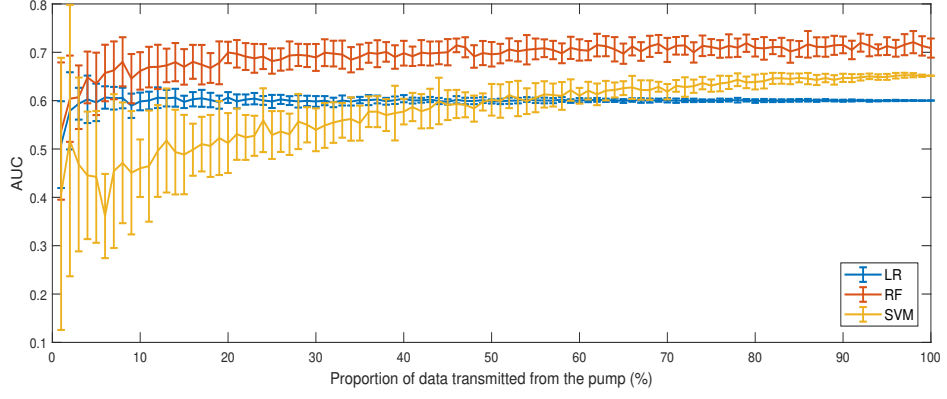
	$\mathcal{D}_m$		$\mathcal{D}_{d,1}$		$\mathcal{D}_{d,2}$	
	AUC(SD)	%Prop.	AUC(SD)	%Prop.	AUC(SD)	%Prop.
LR	0.609(0.018)	8%	0.607(0.023)	6%	0.808(0.027)	15%
RF	0.730(0.017)	89%	0.721(0.023)	98%	0.875(0.015)	91%
SVM	0.608(0.006)	99%	0.653(0.005)	97%	0.750(0.001)	100%

crucial trade-off to note as the final design will need to consider the weighting for sensitivity and specificity of the proposed DDI health-monitoring system. However, given that the input to the cloud-based classifier will be filtered data from the system-based novelty filter, the overall specificity of the DDI system may not suffer from a classifier with faster test times. A comparison of the testing time for the different classifiers and the required number of features will help to determine the trade-off when selecting parameters for the final DDI system.

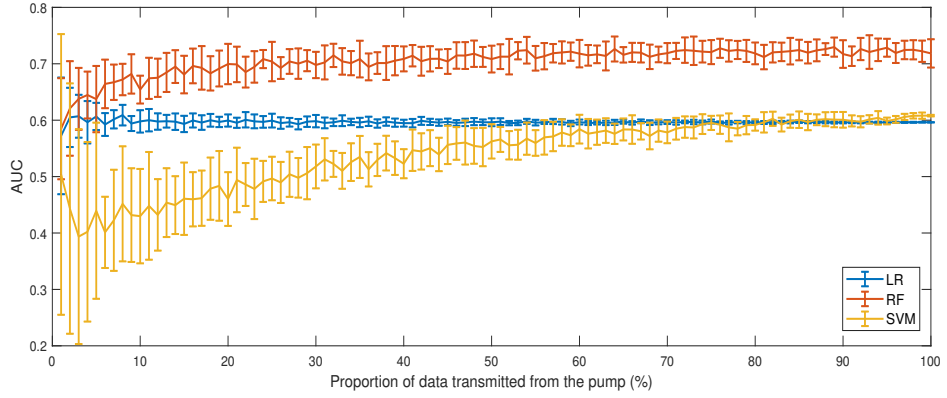
Model Training vs Prediction Time

A vital design specification to consider is the training and prediction time required for the cloud-based classifier be deployed in the final system. Prediction time is essential to ensure the DDI system can run optimally at a region-wide scale. Slow processing at run time could limit the overall number of the nodes being monitored at any given time or require additional processing power to increase the long-term operating cost of the DDI system. Similarly, faster training times will allow the DDI system to update the classifier as new data becomes available if required in future.

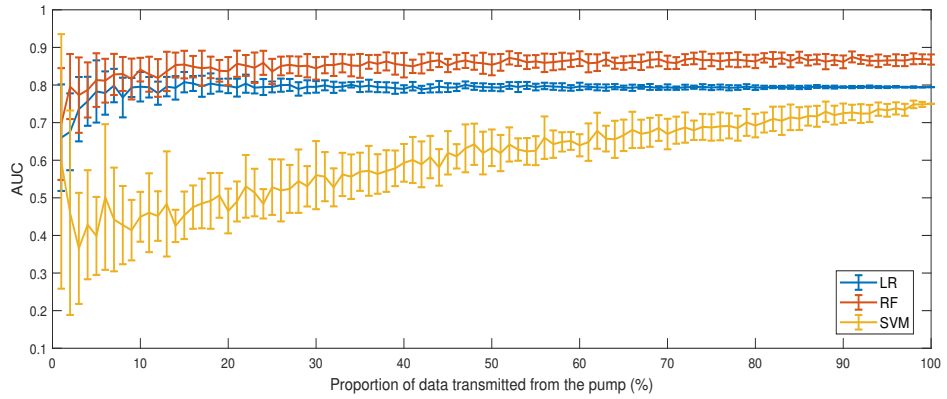
Figure 6.9 shows a comparison of the training and prediction times required by each classifier for the various proposed inter- and intra-handpump systems. In all three cases, the training time for the SVM and RF classifier increases as the proportion of data transmitted from the handpump increases, exponentially and linearly, respectively. Conversely,



(a)



(b)



(c)

Figure 6.8: A comparison of different classifier performance given varying proportions of data for: (a) non-depth-specific inter-handpump case ($\mathcal{D}_m$), (b) deep-specific inter-handpump case ($\mathcal{D}_{d,1}$) and (c) deep-specific intra-handpump case ($\mathcal{D}_{d,2}$).

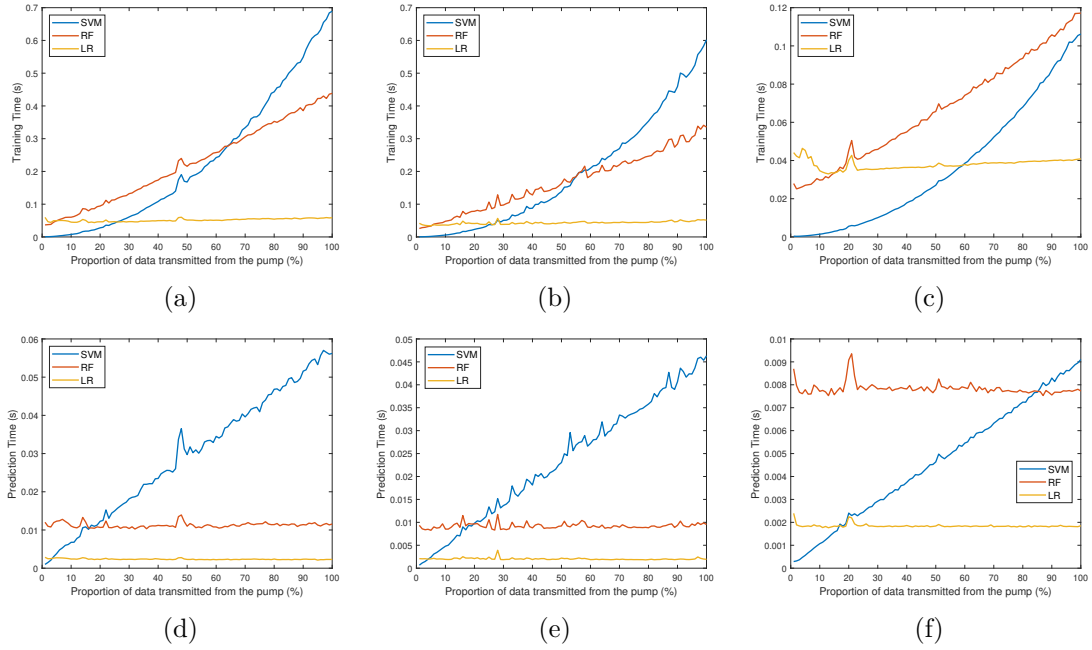


Figure 6.9: A comparison of classifier training and prediction times over a varying proportion of data for: (a & d) deep inter-handpump system, $\mathcal{D}_{d,1}$, (b & e) mixed inter-handpump system, $\mathcal{D}_m$, and (c & f) deep intra-handpump system, $\mathcal{D}_{d,2}$, (top and bottom, respectively).

the prediction time for the RF classifier remains constant as the proportion of data increases while SVM prediction time increases linearly. As the most straightforward method, the LR classifier shows the lowest training and prediction time for all three cases. The trade-offs between these times will be considered for the design of the DDI system.

Number of Features Required

Additionally, the total number of features transmitted from the sensor to the cloud will also influence the amount of data transmitted. Figure 6.10 compares the misclassification rate for the proposed methods when a varying number of features as input. For the inter-handpump cases ($\mathcal{D}_m$ and $\mathcal{D}_{d,1}$), the LR and RF classifiers have no benefit from using more than nine features, corresponding to the high-frequency components of the waveform. Similarly, the SVM is optimised by using ten features and is adversely affected by the addition of more features, which correspond to the low-frequency components. For the

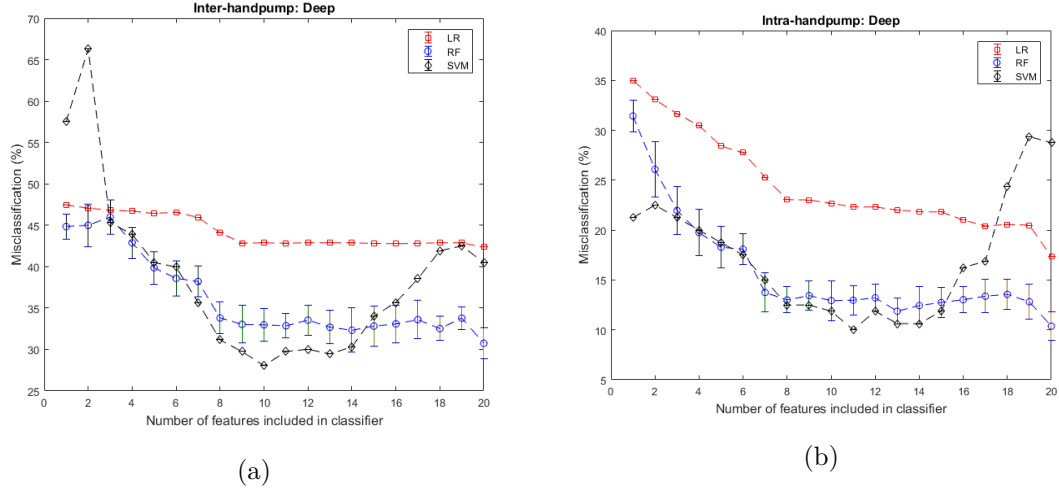


Figure 6.10: A comparison of the classifier performance for varying number of features on two deep-operating models: (a) inter-handpump, $\mathcal{D}_{d,1}$, and (b) intra-handpump, $\mathcal{D}_{d,2}$, (left to right plots, respectively).

intra-handpump case ($\mathcal{D}_{d,2}$), the LR and RF classifiers are again only marginally improved by adding additional features after eight and finally is optimised by using 20 features. Like before, the SVM is optimised by using 11 features, corresponding to high-frequency components, and adding the low-frequency components compromised performance. In both cases, it suggests that using only ten features may successfully reduce the transmission of data without severely compromising the performance of the classifier.

The results in this section suggest that an RF classifier can perform the additional analysis required once the system-based novelty filter transmits the output to the cloud. Although 20 features were used to assess the performance, ten features can be used to reduce the total amount of data transmitted. The next section compares the proposed DDI system to other standard machine learning and deep learning methods.

Comparing Proposed DDI System to Other Cloud-based Classifiers

The performance of the different proposed machine learning methods is compared using the mixed data set, $\mathcal{D}_m$, where the novelty filtered data from the previous section is the input to the respective classifiers. The standard machine learning methods used for

the comparison are: LR, RF, and SVM. The SVM is implemented with a radial basis function (RBF) kernel, and the hyperparameters are optimised using a grid search. The comparative methods also include MLP and 1D-CNN deep neural network. The MLP is a feed-forward network with two layers consisting of 14 and 10 nodes. The 1D-CNN network is constructed using a 1D convolution layer followed by a relu activation layer, and three fully connected layers (with 33, 16, and two nodes). The number of nodes for each of the layers was optimised by trial and error.

The proposed DDI method is the system-based novelty filter followed by the cloud-based RF model. Table 6.6 shows the comparison of the proposed models for the cloud-based system using the unseen test data. The results show that the proposed DDI method outperforms all the traditional standard machine learning methods. The most notable advantage of the proposed DDI method is the increase in the precision or specificity attained by the novelty filtering performed on the handpump before transmitting the data to the cloud. The proposed DDI method can provide a gain of 11.8% when compared to the next best performing model, the 1D-CNN.

Specificity is vital in low-resource settings to reduce the number of false alarms that may have a direct cost in many cases. This increase in specificity is mainly due to the

Table 6.6: Performance evaluation of the proposed distributed inference method compared to other standard classifiers methods. Results show that the DDI method offers overall improved accuracy and precision.

	CA	AUC	Precision	Recall
LR	0.688(0.015)	0.678(0.015)	0.680(0.016)	0.673(0.027)
RF	0.741(0.016)	0.741(0.016)	0.724(0.017)	0.779(0.027)
SVM	0.751(0.015)	0.751(0.015)	0.747(0.016)	0.759(0.016)
MLP	0.744(0.015)	0.744(0.015)	0.773(0.031)	0.693(0.012)
1D-CNN	0.760(0.018)	0.760(0.018)	0.754(0.031)	0.775(0.013)
Proposed DDI	0.788(0.018)	0.763(0.023)	0.872(0.015)	0.787(0.022)

DDI system’s capability to “pre-filter” data packages provided to the classifier. The unseen novelty filtered data tested this way will perform better than those not “packaged” by the system-based novelty filter embedded on the handpump. The next section looks at the implementation of the DDI platform and analyses the practical cost implication.

6.4 Dynamic Distributed Inference Platform Implementation

In the previous section, the results show that the proposed DDI method can outperform other standard machine learning methods by benefiting from the novelty filtered data transmitted from the handpump to the cloud. This section will explore the dynamic nature that can be added to the DDI system as a frugal approach to optimise the limited resources without sacrificing the fidelity of the system. Dynamic approaches will be considered by adjusting: (1) the novelty filter threshold boundary, (2) the type of data transmitted between the sensor and the cloud, and (3) the sampling rate at which the health-monitoring system “checks” the status of the rural node. The performance of the proposed method will be evaluated based on its ability to provide an early warning before the actual failure event. The DDI system will be evaluated using simulated data and validated using real cases from the field data.

6.4.1 Method

Intelligent Sampling Framework

For the logistic regressor used for novelty filtering on the sensor, the threshold is most commonly set to $P \geq 0.5$ for binary classification. Adjusting the boundary around this threshold is a quick and easy method to alter the sensitivity and specificity of the classifier. While this is a great way to increase the sensitivity of the system, it comes at the cost of a higher false-positive rate. In some instances, this may be desirable, but in others,

Algorithm 2: Dynamic Sampling Framework

Input: Sensor novelty filter score voted over the last 10 states, $P(C_{i=10})$, and last 20 states, $P(C_{i=20})$, (value between zero and one);

if $P(C_{i=10}) \leq T_{B,1}$ **then** // Assume start state is green

 | Change state to green; Send novelty filter score;

else if $P(C_{i=10}) > T_{B,1}$ **and** $P(C_{i=10}) < T_{B,2}$ **then**

 | Change to amber state and send features; Send raw waveforms if $P(C_{i=20})$ voted amber ;

else if $P(C_{i=10}) \geq T_{B,2}$ **then**

 | Change to red state and send raw waveforms ;

if $P(C_{i=20}) \geq T_{B,2}$ **then**

 | Trigger alarm to call action/help; Send novelty filter score

end

end

Result: Current system state and alarm (action required)

where false alarms have a high direct cost, it can be undesirable. False alarms have a direct cost in the case of remote monitoring, such as the cost associated with sending a technician; however, the proposed system can tolerate a higher sensitivity on the sensor-based novelty filter since the additional processing capabilities of the cloud will compensate for the specificity.

Once both the sensor-based and cloud-based classifiers are trained, the system can begin to classify the health status of the system being monitored. The handpump state and type of data transmitted to the cloud is determined based on a vote of the 10 or 20 last scores, as shown in Algorithm 2.

Setting the Threshold Boundary

The boundary around the novelty filter threshold, T , is determined arbitrarily based on the overall objective of the system, where the trade-off exists between the FPR and how early the failure is detected. In this “uncertain” boundary state, the system may not have enough information to classify the health status and instead opt to request more data to be transmitted. Our analysis described below uses three different boundaries to illustrate the effect of using different threshold boundaries. If the novelty filter threshold, T , is described

as the value midway between the minimum and maximum novelty score (usually 0.5 for a logistic regressor with an output between zero and one) the boundary, T_B , is selected based on this value of T such that $T_{B,1} = T - (\text{boundary})$ and $T_{B,2} = T + (\text{boundary})$. The first proposed boundary referred to as the large boundary, $T_{B,l}$, is the boundary where 90% of all the novelty score windows fall within this range. The second boundary referred to as the medium boundary, $T_{B,m}$, is selected such that 10% of the novelty score windows fall within the boundary range. The third boundary or zero boundary, $T_{B,z}$, is the cases where the threshold is binary. The results will be discussed in section 6.4.2.

In addition to adjusting the threshold of the novelty filter, two more methods are proposed to increase the inference while minimising the cost of classification, both in preserving battery by reducing processing requirements as well as reducing the size of data packages to be transmitted. Firstly, dynamically adjusting the type of data that is transmitted from the sensor to the cloud. Secondly, dynamically adjusting the frequency of transmitting a score from the node to the cloud or active sampling by varying the classification rate.

Dynamically Adjusting Data Type

By introducing two-way communication between the sensor node and cloud, it is possible to adjust dynamically the type of data package that is transmitted to the cloud. Under normal operating conditions, only the novelty score (a value between zero and one) would be transmitted from the sensor to the cloud. However, as the probability of the classifier deteriorates or reaches near the threshold boundary, the cloud classifier can request more data in order to perform further analysis and try to get the system out of a “state of uncertainty”. Possibly the only scenario more unwanted than a broken handpump is that of not having enough data to determine, with some reasonable level of certainty, the state

of the handpump. Therefore, sending richer data forms, such as features or raw waveforms, rather than novelty scores, will allow the cloud-based classifier to perform more advanced analysis in order to determine the system state. During periods of relative uncertainty it would also be possible for the classifier to decide not to classify until the output has reached some user-defined probability.

The classifier probability is used to define four states for the handpump condition at any instance: (1) normal, (2) unstable, (3) abnormal, and (4) alarm. The decision boundary for each state is described in Table 6.7. Since the transmission of larger data packages, such as features or raw waveforms, has a direct operating cost (considered later), these should only be sent to the cloud in periods of considerable uncertainty. Based on the probability of the classifier being in each state, $P(\text{state})$, the system may require different data types to increase the fidelity such that:

1. **Normal state** is considered a region of high certainty when the handpump is in a working (normal) condition and therefore requires only low-level, infrequent monitoring. Novelty scores can be transmitted from the sensor to the cloud to confirm that both the handpump and the health-monitoring system (sensor and battery) are in working condition. The system should be calibrated to this state after any maintenance.
2. **Unstable state** occurs when either the health-monitoring system hardware or the handpump itself deteriorates. This is the most uncertain state, and priority should be given to assist the cloud classifier to determine the condition of the handpump. High-level monitoring occurs in this state to assist the cloud classifier to increase fidelity, and the sensor will start sending features rather than novelty scores. If these additional data are still unable to provide the desired insight, the sensor will begin to transmit raw waveforms.

3. **Abnormal state** is triggered when a failure occurs or is imminent. Given that the failure may be either due to sensor noise or actual failure of the handpump, this is a high-level monitoring state and waveforms are sent in this state. The monitoring system should be able to tolerate this state for short periods before action is required from the user in order to reduce the likelihood of false alarms.
4. **Alarm state** is again a low-monitoring state. After the alarm is triggered, the system can return to a low monitoring state and transmit sensor novelty scores rather than larger waveform packages, where the latter is both expensive and depletes the battery faster. This allows the system to preserve resources to continue communication with the rural node until help arrives, likely a mechanic in the case of handpumps. This is an essential function because VLOM designs enable the local community to perform maintenance, which means that some proportion of repairs occur without assistance from the mechanic, such as a stone or stick lodged inside the pump head. Once resolved, with or without the mechanic, the novelty score may return to a normal state and reset the alarm. However, recording failure events is a crucial element to ensure the transparent allocation and management of rural resources. In an ideal remote monitoring system, a self-resetting alarm would be deprioritised for immediate resource allocation; however, routine maintenance could be brought forward or scheduled with a nearby alarm.

Given the relatively high level of noise that can occur in low-cost sensors, the system should be able to tolerate some level of fluctuation such that it can quickly move between states without false alarms.

Table 6.7: Fault development states of active sampling in examined handpump conditions. The types of data to be transmitted from the handpump sensor to the cloud are: (1) the novelty filter (NF) scores generated by the embedded LR classifier, (2) the features used by the LR classifier, or (3) the raw accelerometry waveforms. $P(C_i)$ is the probability that the state is in a given condition (C_i).

State	Output	Description	$P(C_i)$	Data Type
Normal	Green	Working	$\leq T_{B,1}$	NF score
Unstable	Amber	Condition unknown	$T_{B,1} - T_{B,2}$	Feature/Wave
Abnormal	Red	Failure likely	$\geq T_{B,2}$	Wave
Alarm	Action	Failed, action required	$\geq T_{B,2}$	NF score

Dynamically Adjusting Sampling Rate

In addition to adjusting the type of data transmitted, the system can also adjust the frequency at which it “checks” the health of the rural node. Although high-frequency measurements are required to analyse the vibrations in the handpump, it is safe to assume that the actual condition of the handpump may not change rapidly. Frequently-used handpumps (on average) break every three months for minor failures (like a broken seal) and once per year for major failures [28]. Therefore, it may be more economical to check the status of a handpump at predetermined times or intervals, rather than every 1.33 seconds (a 128 bin FFT at 96 Hz gives one novelty score every 1.33 seconds).

Furthermore, handpumps are only used during certain times of the day. The study site is near the equator, where daylight is approximately 12 hours per day without much variance throughout the year. Year-round sunrise is approximately at 6 AM and sunset at 6 PM. Given the lack of access to electricity at home, most households are predominantly active during daylight. Peak usage occurs during the early morning for a few hours after sunrise and then again in the late afternoon for a few hours before sunset, with slight variation depending on the season and region. However, some handpumps near schools or marketplaces can be used continuously throughout the day, with some even having a

Table 6.8: Proposed sampling frequency and type of data to transmit for the various monitoring levels of a health-monitoring system given the observed condition of the handpump’s health status.

Monitoring Level	Data Type	Sampling Rate
Low	Novelty score	Twice Daily
Medium	Feature/Wave	Hourly
High	Wave	15 min

long queue of users waiting their turn. Handpumps are rarely used during the night, and therefore monitoring during the 12 hours of the night is not required. It is reasonable to “check” the handpump condition periodically only during the daytime. Table 6.8 shows the proposed combination of data type and the sampling frequency for each monitoring level.

6.4.2 Results

Testing the Dynamic Framework with Simulated Data

The dynamic framework performance is evaluated using simulated data reconstructed from the existing data set, $\mathcal{D}_m$. The simulated data are created to emulate a deteriorating handpump. In the green or red state, there is a relatively high certainty of the state, with 80% of novelty scores within the boundary. The certainty of the state decreases in the amber state when only 60% of the novelty scores are correctly labelled. These probabilities are estimated given the performance of the sensor-based novelty filter on real data in Chapter 5. The simulated data are randomly generated for each iteration and repeated 1000 times to determine the mean and standard deviations. Each simulated iteration contains 2000 samples.

Figure 6.11 shows the various outputs from the proposed DDI platform for the different parameters of interest given a single iteration of 2000 samples. Figure 6.11a shows an example of the classifier output for each simulated sample. This is based on the cloud-based

output with a value between zero (abnormal) and one (normal) based on the handpump condition. The example boundary thresholds are represented by the horizontal red line, $T_{B,1} = 0.7$ and $T_{B,2} = 0.3$. Depending on the state of the system and the data type being transmitted, the classifier output is either from the system-based novelty filter (extreme green or red state) or the cloud-based classifier (amber or early red state). For each score, the cloud-based classifier performs the analysis and keeps a record of the current status of the system, which can be seen in Figure 6.11b, where the vertical red line (at sample 1400) indicates the true failure event, and the horizontal red line is the proposed threshold to trigger a failure event for this particular simulation. The failure threshold was optimised as a trade-off between how early the failure is detected and the number of false alarms (discussed later). The alarm is triggered when the system remains red longer than the threshold permits.

The DDI system uses the past individual states of the system to determine the output of the state, as shown in Figure 6.11c. In this instance, the alarm triggered before the actual event (number of red flags are more than the threshold around sample 500 and changes the system state to signal action; however, the actual failure event occurred around sample 1400). Finally, for each state, the system determines the type of data that should be transmitted from the sensor to the cloud, shown in Figure 6.11d, and records this to determine the overall data transmission required for the simulated iteration. In this iteration, mainly NF scores are transmitted. However, at sample 500 and 1600, the system requests more data as the state of the system deteriorates (which corresponds to the red counters crossing the horizontal line in Figure 6.11b).

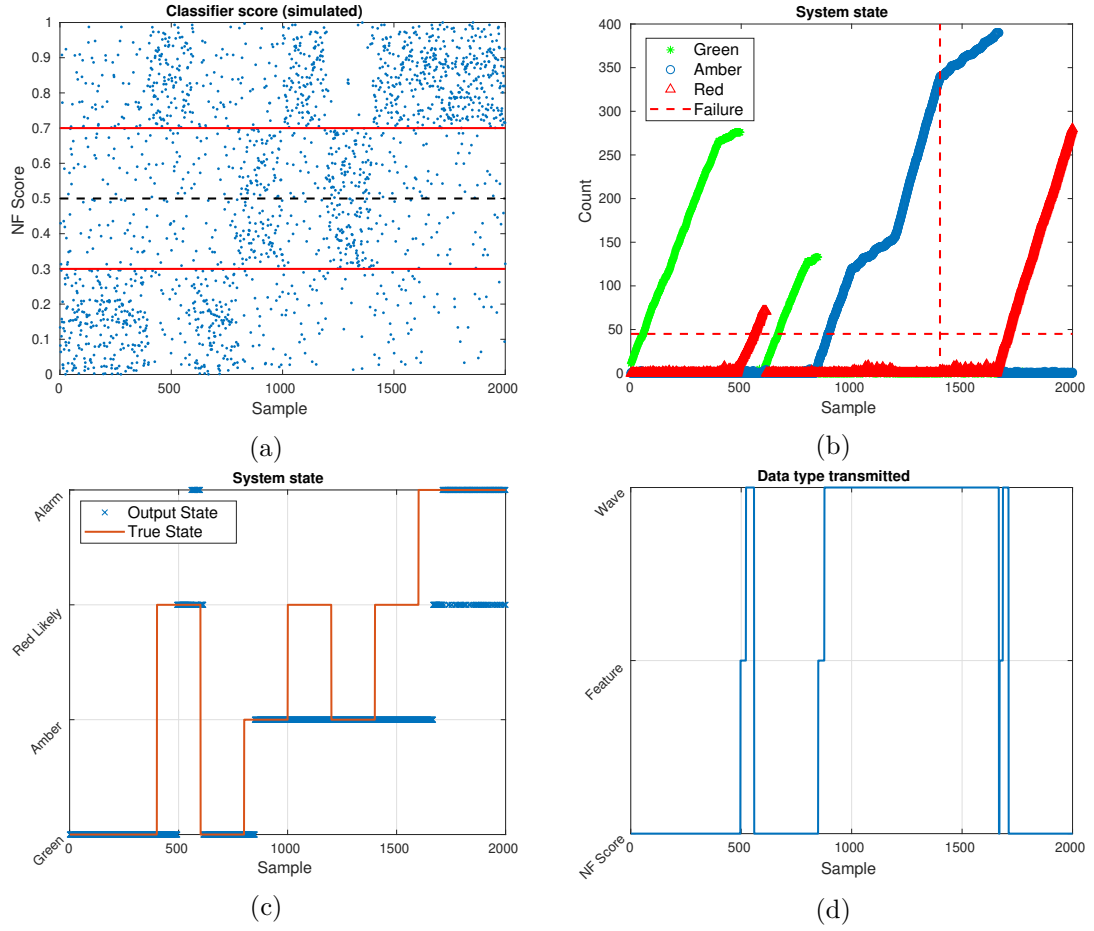


Figure 6.11: An overview of the proposed dynamic distributed inference system’s parameter outputs for a single simulated iteration with 2000 samples. The simulated novelty filter scores from the LR classifier (shown in a) feed into the DDI system. For each sample, each subplot shows the corresponding output: (a) LR classifier output, (b) count of historical handpump state for a given failure event (indicated by the vertical dashed line), (c) predicted and true condition of the handpump, and (d) the type of data requested for the next sample based on the predicted state.

Adjusting the Novelty Filter Threshold

The novelty filter threshold can be adjusted to determine the final loss function, E_{loss} .

Figure 6.12 shows that the pdf of the loss function is approximately gamma distributed for all three proposed boundaries, $T_{B,l}$, $T_{B,m}$ and $T_{B,z}$. There is an increase in the value of the shape parameter of the gamma distribution when the boundary increases ($T_{B,l}$) while greatly reducing the value of the location parameter. Conversely, removing the decision boundary ($T_{B,z}$) for the novelty filter threshold leads to a reduction in the shape value of

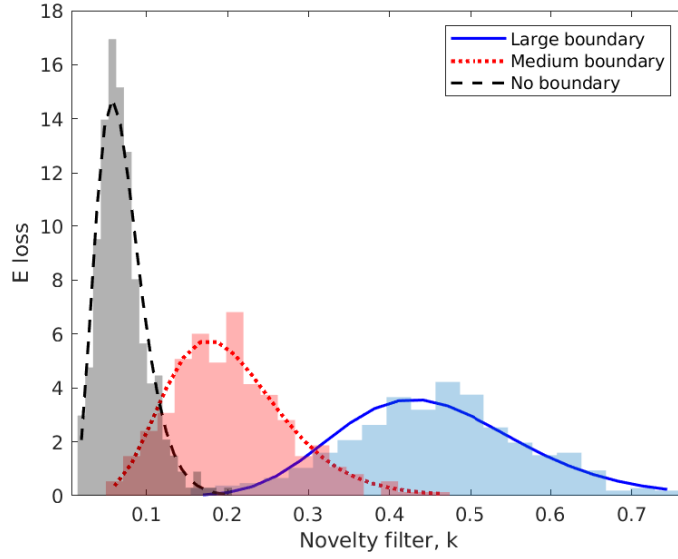


Figure 6.12: Pdf of the loss function for three different novelty filter boundaries such that $T_{B,l}$ (dashed black), $T_{B,m}$ (dotted red) and $T_{B,z}$ (solid blue), respectively.

the loss function pdf and an increase in the location value.

The boundary effect can also be seen based on the time at which the event is detected for each of the proposed boundaries. Figure 6.13 shows the individual pdfs of the time at which the failure event is detected for three different proposed thresholds, given the same data set. Zero is the point of failure, while negative time indicates detecting the event before it occurs, and positive time indicates delayed detection after the actual failure event. If the system fails to detect a failure event for the given test set, it is considered a missed event which is indicated as a large positive number and shown in red towards the right on each figure. In all three cases, the pdfs are approximately gamma distributed. In Figure 6.13a, for a high threshold when the boundary is large, the pdf is greatly reduced in magnitude, and only a small number of events missed. Of those events identified, the majority are identified before the actual event. As the threshold boundary decreases, the pdf is increased both in magnitude and location, as shown in 6.13b. A small boundary, as shown in Figure 6.13c, results in a further increase in the magnitude and location of the

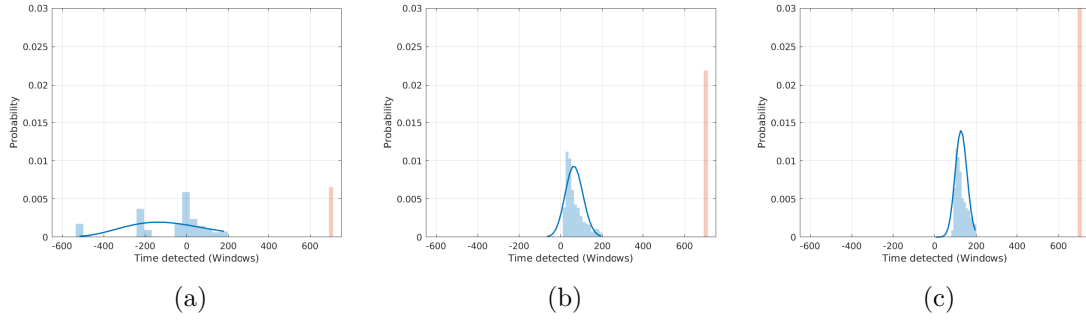


Figure 6.13: Pdf of time to detect a failure event when adjusting the novelty filter threshold for a given boundary: (a) $T_{B,l}$, $T_{B,m}$ and $T_{B,z}$, respectively. A missed event is independently counted and indicated in red on the right of each pdf figure.

pdf. However, this is at the cost of delayed detection of events and an increase in events missed entirely. This demonstrates the effect of the boundary on the sensitivity of the system, which will be further analysed in the following section.

Since it is not only crucial that the system identifies the event but also do so promptly, the next analysis considered the time at which the system identifies the event as a function of the false positive rate (FPR). Figure 6.14 shows the FPR of the system for different thresholds and different early warning scores using the simulated data set. Zero is the point of failure. Positive time windows indicate early warning and negative time windows indicate events detected after the event occurred. Figure 6.14a shows the FPR for changing threshold boundary, where the boundary is calculated as the relative percentage of the boundary around the threshold. Zero is no boundary around the threshold, and 100 is the maximum boundary over the entire novelty score range. As the threshold boundary increases, the FPR goes to zero. However, a very low FPR has a higher percentage of missed events.

Furthermore, Figure 6.14b shows that a low FPR is associated with a negative or late detection of the event. Excluding the missed events in Figure 6.14c results in the reduced CI as well as IQR boundaries. Given that the detection time for a missed event

was arbitrarily high and that the rate of missed events is less than 0.04% excluding these events in the overall calculation give a representative overview of the platform performance. Figure 6.14c shows that it is possible to achieve an early warning given an FPR of less than 0.3.

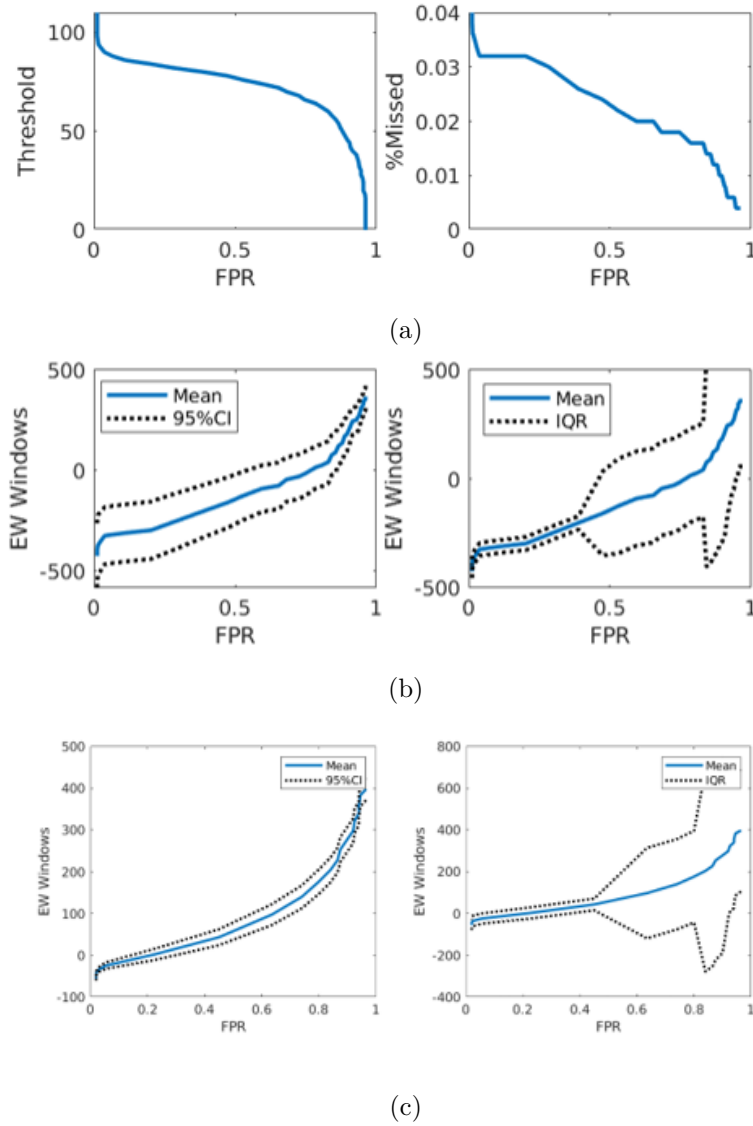


Figure 6.14: Performance of the early warning system given simulated data to investigate the change in threshold on the FPR. Mean, confidence interval (CI) and interquartile range (IQR) are calculated over 10,000 random iterations. The results show: (a) the FPR for a given threshold (as a percentage) and the subsequent proportion of missed events; (b) FPR for a given warning window including missed events, and (c) excluding missed events. In (b) and (c) the 95% CI (left) and IQR (right) are shown, respectively.

Table 6.9: Performance (AUC $\pm$ SD) of a RF classifier for different data types transmitted from the rural sensor to the cloud classifier. Median of AUC results of 1000 iterations are reported with one standard deviation.

Data Type	$\mathcal{D}_m$	AUC (SD)	
		$\mathcal{D}_{d,1}$	$\mathcal{D}_{d,2}$
NF Score	0.657(0.017)	0.648(0.023)	0.773(0.019)
Features	0.719(0.012)	0.720(0.019)	0.896(0.014)
Waves	0.809(0.008)	0.813(0.009)	0.913(0.007)

Adjusting the Data Type

The next step is to analyse the influence of the data type being transmitted from the rural node to the cloud, where more advanced analysis can be performed on the latter than the former. Table 6.9 compares the AUC of an RF classifier receiving only one data type at a time for the same period. The comparison shows that providing richer data types allows the system to increase the overall fidelity by 15.2% for $\mathcal{D}_m$, 16.5% for $\mathcal{D}_{d,1}$ and 14.0% for $\mathcal{D}_{d,2}$ when transmitting waves instead of novelty filter scores. For the general case, $\mathcal{D}_m$, there is an increase in the AUC of 6.2% when sending features over NF scores and another increase of 9.0% when sending raw waveforms over the feature packages. There is a similar increase for $\mathcal{D}_{d,1}$ and $\mathcal{D}_{d,2}$. The cloud classifier gains most from the additional data provided by the raw waveforms. Given that a mix of data types will be transmitted, the actual gain may be lower. However, these results confirm the hypothesis that providing more rich data types (such as raw waveforms) rather than those packaged by the system-based novelty filter allows the cloud-based classifier to make a more accurate estimation of the health status of the rural node.

Economics of Dynamic Distributed Inference System

After installation, the amount of data transmitted from the sensor to the cloud determines the main variable cost of operating the system. There is a vast difference in the bandwidth required to transmit each data type that would ultimately influence the overall operational

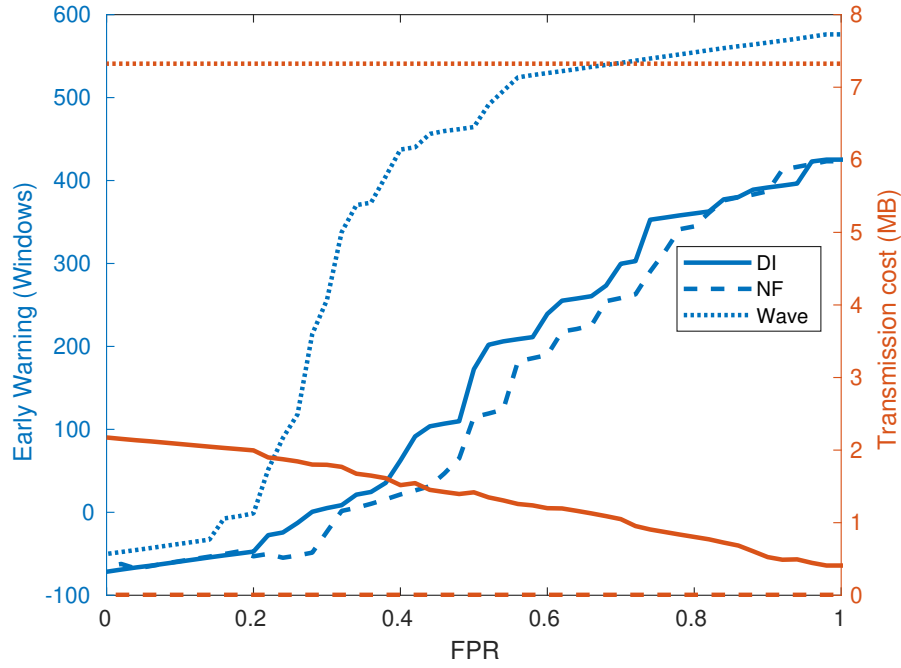


Figure 6.15: Comparison of the performance and cost of operating the proposed DDI system and traditional non-dynamic system transmitting only the novelty filter scores (dashed) or raw waveforms (dotted) from the handpump the central cloud.

cost of the system. Figure 6.15 considers the economics of transmission by comparing the data bandwidth (in MB) to the FPR, which in turns relates to whether or not the event is detected early. The proposed DDI platform, being dynamic in the type of data that is transmitted, is compared with the case of transmitting raw waveforms or novelty filter scores, exclusively. Since the size of the data package being transmitted for the NF score and waveform is constant, the total amount of data transmitted for these two instances are fixed. Conversely, the amount of data transmitted for the DDI platform is inversely linear to the FPR. This is a direct result of the dynamic nature of the platform, adjusting the type of data being transmitted based on the assumed condition of the handpump.

The results show an apparent reduction in the amount of data, and thus operating cost, required to detect the failure event before occurring for a low FPR (< 0.3). For FPR = 0.3, the DDI platform can detect the failure shortly before it occurs at the cost

of about 2 MB. For the same FPR, the NF score alone is delayed in detecting the failure albeit at a minimal cost (0.1 MB) while the raw waveform can detect the failure very early at the cost of approximately 7.5 MB. A 24-hour lead time is required to mobilise the service network responsible for repairing the handpumps in the study area. An earlier warning earlier may be ignored until the last moment in order to maximise the capacity of the current parts. Thus, the additional cost of transmitting the raw waveform would not lead to any additional gain.

Furthermore, the saving in cost of the DDI over the raw waveform is even more considerable given that for most of the time, the handpump would be in a normal state and the DDI would therefore only be transmitting the NF score. Thus, over a more extensive network of rural handpumps and an extended monitoring period, a seemingly small saving will quickly become rather significant.

Testing the DDI with Real Handpump Field Data

Once the appropriate thresholds were established and the platform validated using the simulated data, the proposed DDI framework was evaluated using real failure examples from the $\mathcal{D}_l$ data set. The failures in this data set are accurately labelled and well-documented for validating a system of this complexity. Figure 6.16 shows the output of the system for real data recorded from a single handpump in $\mathcal{D}_l$ using the platform trained using $\mathcal{D}_m$. The trend of a moving average of 45 windows (equivalent to one minute) visually changes around the failure event for this given case. Although this experiment was performed retrospectively, the platform had access to the true novelty scores generated on the sensor as well as the accompanying raw waveforms and features to simulate real-time operation if two-way communication is available.

Figure 6.16 shows the window at which the failure is detected as a function of the FPR

for two failure events at different handpumps in $\mathcal{D}_l$. Both cases show that it is possible to attain an early warning of imminent failure using the proposed DDI framework for some low FPR. The right-hand vertical axis shows the transmission cost as the total megabytes of data transmitted from the sensor to the cloud for the monitoring period leading up to the failure. Both the early warning and cost of transmission is inversely related to the FPR, such that a lower FPR result in a late warning, and high transmission cost and a higher FPR leads to earlier warning and reduced transmission cost. In the first example, the event can be detect early (before occurring) for an $\text{FPR} = 0.42$, which requires only 2.6 MB. For the second example, early detection is possible given an $\text{FPR} = 0.5$ using 4 MB. Given the trade-off between these parameters, the ideal threshold boundary can be selected such that the amount of early warning and overall cost of transmission is acceptable for the given FPR that can be tolerated. Results from the different handpumps showed that there is variability between individual pumps. Given that some failures are harder to detect, these may lead to an increased number of false alarms. Unfortunately, since these failures are from the long-term data set, $\mathcal{D}_l$, the labels and timestamps of failures are acquired from the maintenance service company responsible for tending to failures. Since the service obligation of this company is to attend to a repair within 48-hours, it is unclear whether the recorded failure time is the date of failure or repair. For standardisation, the time was taken as the repair date, and thus the actual warning could be flagged earlier than noted in these results.

For the third case, the failure in the long-term samples occurred during a field visit and so the exact failure time is known and labelled accurately. Figure 6.17 shows the results for the optimised threshold. The DDI platform tracks the health status of the handpump being monitored by sampling a five-second clip at the start of each day leading up to

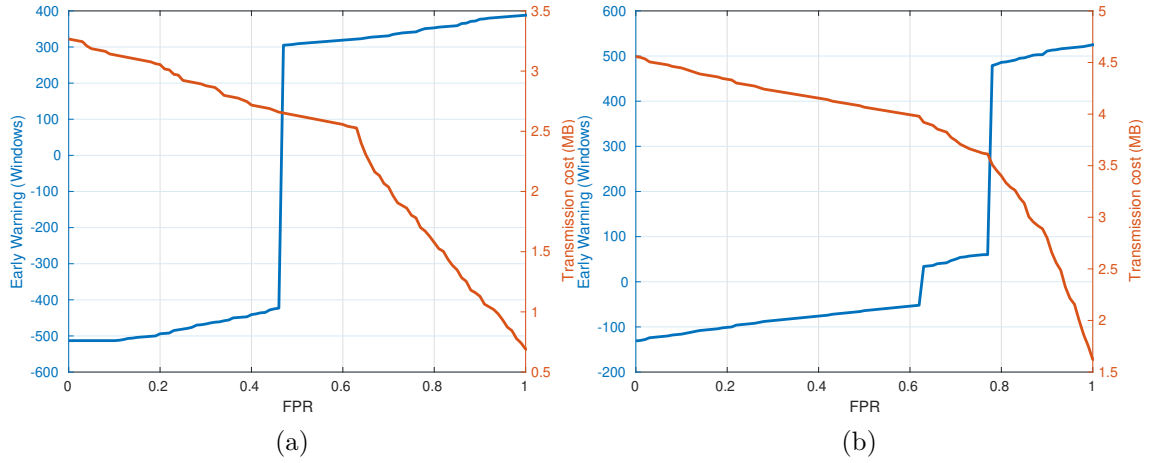


Figure 6.16: Performance of the early warning alert by the DDI system along with the associated cost (measure in amount of data being transmitted) using real failure events from $\mathcal{D}_l$, given two different failure types from the same handpump.

the failure, which occurs on Day 0. The normalised novelty scores for each monitoring period in the days leading up to the failure event is shown in Figure 6.17a along with the moving average (red line). Figure 6.17b shows that for a low FPR (less than 0.2) the DDI platform could detect the failure 418 windows before the event, which corresponds to two days before the actual failure. The results from these three failures also demonstrate that it is reasonable to sample the handpump condition only for a few seconds every day given that changes occur over a long period. While this result is purely observational, it is a promising finding given that there are limited examples of correctly-labelled failure events represented in the data set.

In the case of handpumps in our study region, the service delivery provider who is responsible for repairing the handpumps once the failure is reported typically requires a 24-hour notice. As such, this would be the minimum early warning time that would allow for handpump failures to be eliminated. For the example that was accurately labelled, the DDI platform could provide a sufficiently early warning for a low FPR. For other examples, an early warning was achievable if a higher FPR could be tolerated.

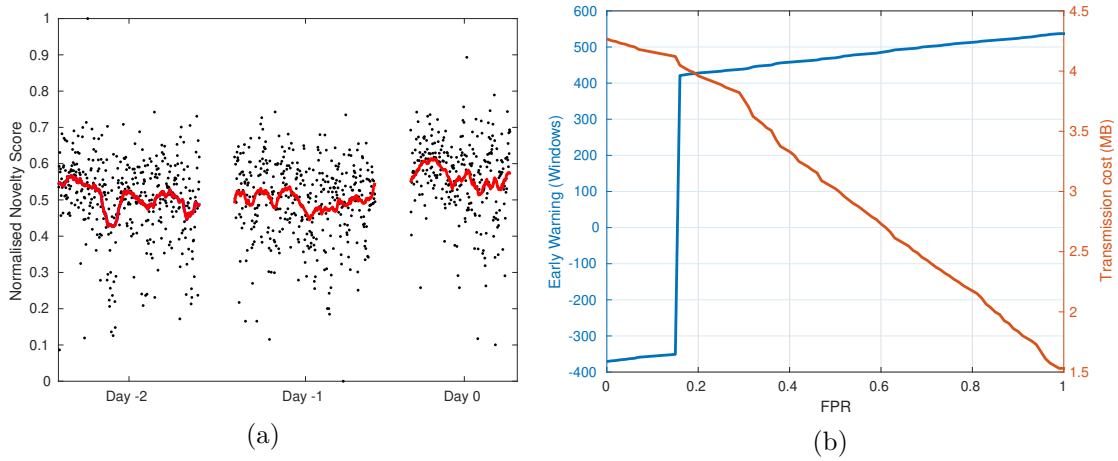


Figure 6.17: Case using real data of true failure event: (a) scores generated by the system-based novelty filter and (b) performance of the early warning alert by the DDI system along with the associated cost (measure in amount of data being transmitted).

6.5 Discussion

The aim of this chapter was threefold. Firstly, to analyse the real-world scores generated by the system-based novelty filter embedded in the handpumps. The method uses a lightweight LR classifier to generate these scores, and different algorithms were tested on the handpumps over 12 months. The second aim was to train the cloud-based health monitoring algorithm, which receives intelligent data packages produced by the sensor-based novelty filter. The approach relies on the novelty filtered data to be further processed using the additional computational capabilities of the cloud to increase the overall fidelity of the health-monitoring system. Finally, the DDI framework was tested using both simulated and real data to identify an adverse event before occurring. This was done to validate the robustness of the dynamic distributed inference system and illustrate that it is a useful approach to remote health monitoring of systems in limited-resource areas.

The first analysis, validating the real-world scores, compared different sensor-based novelty filter algorithms. This included comparing the lab-generated scores with real-world scores, as well as algorithms that were unique to the operating depth of the handpump

being either “deep” or “shallow”. The proposed algorithms that were tested in the field are listed in Table 6.2. Results shown in Figure 6.3 demonstrate that the performance of the system-based novelty filter is only somewhat outperformed by the lab-based filter operating within fewer constraints. This was a critical validation as the ongoing steps for designing the proposed DDI system would rely on these inputs. Figure 6.4 compared the three different versions of algorithms and found that a depth-specific algorithm was able to increase the fidelity. However, limitations in the representation of shallow-operating handpumps in $\mathcal{D}_o$ meant that the algorithm was biased to handpumps operating at greater depths. Fortunately, these results were obtained early in the study and could be used to inform subsequent data collection campaigns to ensure more balanced representation in the handpumps included in the data sets, $\mathcal{D}_f$ and $\mathcal{D}_l$, used in this study.

These results also demonstrated that some failures types were easier to distinguish than others, as shown in Figure 6.7. The novelty filter relies on the vibrations in the handpump to inform the health-status of the handpump. However, due to the location of the sensor relative to the entire handpump system, where most of it is often up to 80m underground compared to the 1.5m tall surface-based pedestal, vibrations far away are less distinct than those that originate near the sensor. This means that failures, such as a leak in the rising main, that occur far away from the sensor are harder to identify using the vibration features alone. However, more extreme failures that lead to immediately lack of water, such as a broken pump rod or broken u-seal, have more distinctive features that can be identified. The practical impact of this on the health monitoring system in the case of handpumps are not detrimental as the severe failures are those that are most important to detect and prevent as this could lead to immediate loss of access to water. In comparison, a leak in the rising main might be harder to detect but would likely occur slowly over time

and as such lead to a reduction in the pressure and the water flow but not necessarily immediately result in a lack of water flow entirely.

Finally, the addition of temporal correlation is incorporated by aggregating consecutive novelty scores - a simple step that could easily be added to the sensor-based novelty filter. The improvement introduced by this additional step is shown in Table 6.3. Marginal increases are seen in the inter-handpump data sets, $\mathcal{D}_m$ and $\mathcal{D}_{d,1}$, while the intra-handpump data set, $\mathcal{D}_{d,2}$, achieve additional improvement in the AUC.

Overall, the field-generated novelty scores collected from the proposed system-based novelty filter were shown to be sufficiently robust when compared to the equivalent lab-generated scores which is achievable without the limitations of an 8-bit microprocessor. This was demonstrated for a variety of different failure types, data sets, and operating considerations. The sensor-based novelty scores from the field were typically within two to three per cent of the “ideal” lab-generated scores.

The promising results from the real-world system-based novelty filter scores ensured a solid base for the second analysis, which aimed to develop the cloud-based health monitoring algorithm receiving intelligent data packages from the sensor as input. To demonstrate the robustness of the proposed DDI system, the analysis was done using real-world data from the three data sets, $\mathcal{D}_m$, $\mathcal{D}_{d,1}$ and $\mathcal{D}_{d,2}$. The analysis found that incorporating more advanced condition monitoring methods on the cloud-based platform could increase the system’s positive predictive value by around 10%, shown in Table 6.4. For the cloud-based heavyweight classifiers, the RF was shown to outperform both LR and SVM for both inter-handpump data sets, $\mathcal{D}_m$ and $\mathcal{D}_{d,1}$ as well as the intra-handpump data set, $\mathcal{D}_{d,2}$.

Furthermore, the results from this analysis, shown in Table 6.5 and Figure 6.8, found that these improvements could be achieved by using only a small fraction (less than 30%) of the original data. In the entire $\mathcal{D}_m$ data set, the RF achieved an average AUC of 70.1 ± 1.9 by using only 20% of the windows. However, the RF model, when compared to the other classifiers, does not offer the same level of robustness in the prediction accuracy as some amount of variability persists despite the addition of more data. This suggests that in the event of imminent failure, when all the raw data are transmitted from the handpump to the base-station, RFs will continue to have a higher rate of uncertainty in the prediction accuracy. LR and SVM variability reduces to almost zero as the amount of data transmitted increases. This is quite an important trade-off to consider when selecting a suitable cloud-based classifier as the increased cost of transmitting more data should ideally lead to higher sensitivity and specificity of the prediction. There are, however, other ways of improving the overall prediction accuracy, such as changing the threshold of the sensor-based novelty filter to create a dynamic distributed monitoring system, which was analysed in the next phase.

An RF offers the ability for much faster training and testing times should the proportion of data increase (as shown in Figure 6.9); a feature which outweighs the previous limitation. The analysis in Figure 6.10 showed that further optimisation of the system could be achieved by sending a reduced number of features (10 rather than 20) without compromising the performance. Given these results, it was recommended that the RF classifier be used for the cloud-based classifier of the DDI system as it is sufficiently robust with the additional benefit of fast test times, which could future-proof the system to perform at region-wide scale.

Finally, the proposed DDI system was compared to other standard machine learning

and deep learning methods. The results in Table 6.6 show that the proposed DDI system was able to outperform all the standard machine learning methods, including LR, SVM, MLP, and 1D-CNN. Importantly, the proposed method also improved precision by more than 10% compared to any of the other methods. Specificity is an essential consideration in low-resource settings where scarcity of resources requires careful prioritisation of the allocation. Low specificity can also be costly in the case of remote monitoring, such as those in rural nodes, where tending to alarm have a direct cost.

By considering the optimal means of representing the data packages, the results have shown that it is possible to increase fidelity as the risk of deterioration of the handpump is deemed to increase by dynamically adjusting specific parameters of the platform. This interaction between the system-based LR novelty filter and more advanced machine learning, such as RF, SVM, MLP, and CNN, processes in the cloud, is a novel research field and which applies to other cognate fields.

6.6 Conclusion

The results in this chapter show that the proposed DDI platform offers an essential tool in the remote monitoring of rural nodes. It presented the development and validation of a DDI system that takes as input the data packages created by the sensor-based novelty filter developed in Chapter 5 to reduce the amount of data being transmitted from the sensor to the cloud. It demonstrated that the inputs from real sensor data are valid when compared to those generated in a lab with no limitation. The results show that a robust DDI platform incorporating the input from the system-based novelty filter and cloud-based classifier improves the condition estimation compared to traditional machine learning and deep learning methods. These results further show that the addition of dynamically adjusted parameters, such as the LR threshold, the type of data transmitted,

and the sampling rates of health checks, greatly reduce the amount of data transmitted and subsequent operating cost compared to static monitoring.

The cloud-based classifiers that were considered are logistic regressor, random forest, support vector machine, multilayer perceptron and convolutional neural network. The objective of creating the DDI system was to improve the overall accuracy and robustness compared to the system-based novelty filter while optimising the limited resource by being frugal about the use of processing power and data transmission. The proposed DDI model met this standard while outperforming all the other standard machine learning methods. Notably, the DDI method demonstrated superior specificity over the other methods, an essential requirement for a remote health-monitoring system. Finally, the ability of the DDI system to perform early warnings on both simulated and real cases were validated as a trade-off between how early the warning is triggered (time before event) and the cost of achieving the warning (amount of data to be transmitted). The platform showed high variability in the early detection of critical events in individual subjects which is due to the location of the failure type.

Chapter 7

From Handpump to Human

This chapter aims to demonstrate the potential wide-scale application of the proposed DDI system that was developed and validated in Chapter 5 and 6 using handpumps as real examples of rural nodes or “patients”. The analysis will evaluate the ease with which the system adjusts to a different application, specifically the monitoring of patient vital signs in a step-down unit (SDU). The first part will introduce the patient data set used in this analysis before it will discuss the changes required to adapt the system for a different application. Finally, the analysis will validate the ability of the previously-defined DDI platform presented in Figure 6.1 to predict physiological deterioration in a clinical setting. Consideration will be given to how this type of rural “node” is different from a rural handpump as well as discuss the similarities that make it possible to transfer the DDI system.

7.1 Remote Patient Monitoring

Remote patient monitoring (RPM) offers a solution to meet healthcare needs in limited-resource settings. As the global demand on healthcare systems continue to increase, resources can be considered “limited” even in developed country settings. In 2016, more than 7.1 million patients worldwide were remotely monitored by health-monitoring devices [217]. In the same year, the National Audit Office in the United Kingdom found that “bed blocking” cost the NHS over £820m [218]. Bed blocking is a phenomenon ob-

served in developed countries with an ageing population as the social care system can often delay the discharge of elderly patients. RPM could save the NHS nearly £1bn over five years [219]. Furthermore, overstretched services of “unavoidably small” rural sites in the United Kingdom place an enormous financial burden on the NHS as many end in annual deficits. Underperforming sites account for nearly a quarter (23%) of the overall fiscal deficit; however, represent only three per cent of all trusts [220].

In low- and middle-income countries (LMICs), the large-scale shortage of healthcare workers decreases access to healthcare services. However, there is often a great disparity between the population concentration and the subsequent access to public services. Table 7.1 shows an inverse relationship in the patient-to-nurse ratio and rural populations for a subset of countries. This shortage in the workforce affects rural populations more than their urban counterparts. Rural populations typically also account for a significantly larger proportion of the overall population in LMICs. More than half the population in LMICs live in rural areas, compared to 19% in high-income countries [221]. In India, for example, 76% of the available medical facilities are in urban areas. However, 66% of the total population live in rural areas, creating a considerable disease burden [221]. This disparity between practitioners and population is not unique to India but widespread in many other LMICs.

A recent review found that in LMICs where primary healthcare improved between 2010 and 2017, strategies tend to focus on finance, policy and workforce as the core system functionalities while overlooking the importance of social accountability and improvements in management and facilities [223]. This is a major limitation for RPM or mobile health (mHealth) adoption, which requires insights into socio-cultural dynamics and local business practices in order to succeed [224]. Overall, RPM and mHealth have the potential to

Table 7.1: Density of nurses and rural population in selected countries. (Source: [221, 222])

Year	Country	Nurses*	% Rural Pop.
2016	Canada	9.9	19
2015	USA	8.6	18
2017	United Kingdom	8.3	17
2016	Saudi Arabia	5.6	16
2017	South Africa	3.5	34
2010	Philippines	3.3	53
2015	China	2.3	41
2017	India	2.1	66
2014	Kenya	1.5	73
2016	Malawi	0.3	83

*per 1000 population

help alleviate many of the ongoing pressures of health services in both LMICs and high-income countries. Along with addressing some of the system issues, RPM also offers new avenues for patient care not possible with existing healthcare technologies. Continuous patient monitoring offers data streams required to evolve empirical methods that define the relationships between early predictors and potential patient outcomes.

7.2 Early Detection of Deterioration

Clinicians typically use two types of parameters to assess patients. Firstly, ‘vital signs’ refer to the most commonly used physiological parameters related to the human body’s primary function. These include heart rate (HR), breathing rate (BR), temperature, blood pressure (BP), and blood oxygen saturation (SpO₂). Technology facilitates the continuous measurement of these parameters non-invasively. The second set of parameters that clinicians use to monitor patients are the biochemical parameters. The measurement of these are usually invasive, such as via withdrawing a blood sample, and which is necessarily measured intermittently.

Studies dating as far back as the early 1990s suggest that premonitory signs of physiological deterioration occurring before a cardiopulmonary arrest could be used to prevent

these adverse events if appropriate action is taken in the hours leading up to the event [225, 226, 227]. More recent studies have attempted to detail the threshold of change, rate of change and time before the event that acceptable action should be taken. Many of these adverse events occur when in-hospital patients move out of the intensive care unit (ICU) to locations that have a higher patient to nurse ratio and less frequent monitoring. All of these studies were conducted in a developed world setting where the patient to nurse ratio is closely regulated and in theory, should be 1:1 or 1:2 in United Kingdom and USA to allow for continuous observations.

Heuristic patient monitoring approaches originally used univariate alarm thresholds to generate warnings about abnormalities in single vital sign measurements [228]. The National Early Warning Score (NEWS) intended for use in all NHS hospitals addressed shortcomings in single vital sign monitoring by proposing an aggregated score across multiple vital signs for patients at risk of deteriorating [229, 230]. However, it was found that the additional workload on healthcare staff may be disproportionate to the proposed benefits [231] and only offered clinicians a statistical summary and risk score at a single point in time [232].

Probabilistic patient monitoring approaches use computational inference by continuously comparing the patient vital sign measurements to a healthy individual in the global population. As one common example, Kernel density estimates (KDE) using Parzen windows identify parameters that indicate deterioration [233]. Extreme value theory distinguishes extreme observations due to actual deterioration from those caused by measurement artefacts resulting from long-term continuous monitoring [234, 235]. To overcome the limitations of approaches taking only a summary at a single point in time, Gaussian process regression (GPR) has been suggested as a robust method to forecast the time-

series vital signs of patients [236] and is becoming increasingly popular in the field. The intense exploration required to identify preferable kernels and priors for patient-specific models often limits this latter approach to applications using single vital signs, such as HR.

In the case of handpumps, 24-48 hours was sufficient lead-time to ensure timely repair and allow the service repair company adequate time to respond to any imminent failures. In the case of SDU patients, the aim is to detect any patient deterioration as early as possible, but at least before the actual event. The DDI platform goals in this chapter remain the same as those laid out in Chapter 6. However, for the application of the proposed system for patient monitoring in an SDU, the clinical aims would be to:

1. have a low false alarm rate so as not to add any additional burden to the clinical staff,
2. provide continuous patient monitoring in addition to existing clinical observations (which may occur several hours apart), and
3. offer sufficient warning that can allow clinical staff to respond according to existing protocols.

At a minimum, the system should facilitate an increase in the quality of care that patients receive by offering clinical staff an additional data stream to inform decision-making.

7.3 Data Description

The data set used for this analysis was acquired as part of a three-phase trial of a patient monitoring system at the Presbyterian Hospital at the University of Pittsburgh Medical Centre (UPMC) SDU. Typically, the SDU provides intermediary care to support patients recently discharged from the ICU before being transferred to the general wards. Although

the precise guideline for SDUs varies by country, a typical SDU will have a nurse-to-patient ratio of 1 : 4-6 in comparison with more staff-intensive ICUs that have a nurse-to-patient ratio of 1 : 1-2 [237]. The intermediary nature of the SDU means that staff levels can support data collection but are more limited than an ICU as to resemble settings with lower monitoring levels.

The data set considered in this chapter is from the UPMC surgical-trauma SDU and contains data from 333 adult patients collected as part of a three-phase study to validate the patient monitoring system in [233]. We use data from November 2006 and which lasted for eight weeks. All the patients in the data set were older than 18 years of age. Male and female representation in the data set is both approximately 50%. A total of 18,692 hours of continuous monitoring data was collected over eight weeks containing only seven activation alerts for the Medical Emergency Team (MET). The MET calling criteria is shown in Table 7.2 as defined in [238]. Patients in this data set were labelled for having a ‘clinically significant critical event’ when triggering the MET via “Condition A” (cardiac or respiratory arrest requiring cardiopulmonary resuscitation) or “Condition C” (defined by the MET calling criteria).

7.3.1 Labelling Critical Events

The activation criteria use a univariate threshold lasting for at least four of the previous five minutes. This meant that a large number of events, often those related to BR, were triggered and merged alarms that occurred close to one another. A clinician validated 407 such events in the data set. Of these, 237 events from 83 patients were identified as non-artefactual since the clinician confirmed that the time-series data does exceed the MET limits, shown in Table 7.3. These events were labelled as C’-events. Of those events that exceeded the MET threshold for a sustained time, clinicians labelled 112 C”-events

Table 7.2: Clinical criteria for MET activation at UPMC Presbyterian Hospital [238].

Parameter
<i>Heart rate:</i>
• < 40 or > 140 beats/min with symptoms, or any rate > 160.
<i>Breathing rate:</i>
• < 8 or > 36 breaths/min; • new onset of breathing difficulty; • new $\text{SpO}_2 < 85\%$ for > 5 min unless known chronic hypoxaemia.
<i>Blood pressure:</i>
• < 80 or > 200 mmHg systolic, or > 110 mmHg diastolic with symptoms.
<i>Acute neurological change:</i>
• acute loss of consciousness; • new onset lethargy or narcan (naloxone) use without immediate response; • seizure (outside seizure monitoring unit); • sudden loss of movement (or weakness) of face, arm, or leg.
<i>Other:</i>
• chest pain unresponsive to nitroglycerine or doctor unavailable; • colour change (of patient or extremity): pale, dusky, gray, or blue; • unexplained agitation > 10 min; suicide attempt; uncontrolled bleeding.

in 59 patients that would require a MET activation alert or emergency intervention. The continuous monitoring of vital signs was able to identify 105 MET activation alerts previously missed. Labelled C"-events record the relevant start- and end-time as well as the primary cause of the event.

7.3.2 Vital Sign Measurements

Physiological patient data in the UPMC data set contains the time-series data of five vital signs: HR, BR, SpO_2 , systolic blood pressure (SBP) and diastolic blood pressure (DBP). HR was acquired using a portable computer connected to a multi-parameter patient monitor to measure the electrocardiogram (ECG) and detect the intervals of the QRS peaks, typically between 60 to 120 beats per minute (bpm), resulting in a typical measurement frequency of 1 to 2 Hz. BR was determined by applying peak detection to the impedance pneumography signal to calculate the time between respective peaks. BR, typically ranges between 12 to 30 breaths per minute, requiring a frequency of measure-

Table 7.3: Threshold for MET alarm and artefact rejection thresholds applied to patient data.

Vital Sign	MET Alarm Threshold		Artefact Threshold	
	Lower	Upper	Lower	Upper
HR (bmp)	40	140	30	300
BR (bpm)	8	36	3	45
SpO ₂ (%)	85	-	10	101
SBP (mmHg)	80	200	0	1000
DBP (mmHg)	85	-	0	1000

ment of 0.2 to 0.5 Hz. Blood pressures (both SBP and DBP) were measured manually with an inflatable cuff around the patient’s arms. Readings were taken every 30 minutes while awake and hourly while sleeping. A sensor was used to measure SpO₂ with sampling frequency set to 1 Hz.

Given that different devices with different sampling frequencies were used to monitor each of the vital signs, the measurement timestamps are often misaligned for the same patient. As such, all signals were down-sampled to a sampling rate equivalent to the minimum sampling rate (that of BR) such that $f_s = \frac{1}{60}$ Hz. Different starting times and probes becoming detached also means that some vital signs have a higher collection rate than others. Artefacts for each vital sign were removed for those exceeding the thresholds in Table 7.3 [233].

Missing data are quite common in patient monitoring studies where up to 20% of the data can be lost [239]. This is most commonly caused by sensors disconnecting from the patients or patients removing the sensors to wash. At least 11 patients were missing data entirely or had less than one hour of data. Of the remaining patients, some had at least one vital sign with less than one hour of monitoring data. The breakdown in Table 7.4 shows that BR is the most likely to be missing and non-C”-patients are more likely to be missing data than their closely monitored unwell counterparts, C”-patients.

The non-C"-patient in Figure 7.1, for example, has the first 10 hours of data missing while the C"-patient is missing measurements between 51 and 58 hours and again between 75 to 79 hours. Shorter gaps exist individually during different times for both patients. The C"-patient in this example experienced three C"-events due to increased BR and low SpO₂. Even though the BR remains elevated after these events, the clinician did not mark another C"-event as the other vital signs of the patients appears to have stabilised.

Figure 7.2 shows the distributions for the five parameters broken down firstly for all patient data and then for C"- and non-C"-patient data. All the vital sign measures are normally distributed, except for SpO₂ which has a one-side distribution given the maximum value is 100%. A reading of 100% is possible given that the patients were in the SDU and receiving oxygen through a mask.

Individually, the distribution of vital sign measurements for C"- and non-C"-patients look nearly identical. Both have a median HR of around 83 bpm and SpO₂ levels of 97%. The C"-patients have only slightly higher median BR of 19, compared to non-C" cohort. Similarly, SBP and DBP are slightly different at 128 and 65 for C"-patients, compared to 127 and 67 for non-C"-patients. Independently these vital signs would not distinguish deterioration; however, the importance of continuous monitoring and the aggregate monitoring of different vital signs is evident for early detection.

Table 7.4: Count of patient files with less than one hour of data broken down by vital sign.

	HR	BR	SpO ₂	SBP	DBP
C"-Patient	0	3	0	0	0
Non-C"-Patient	5	21	12	2	2

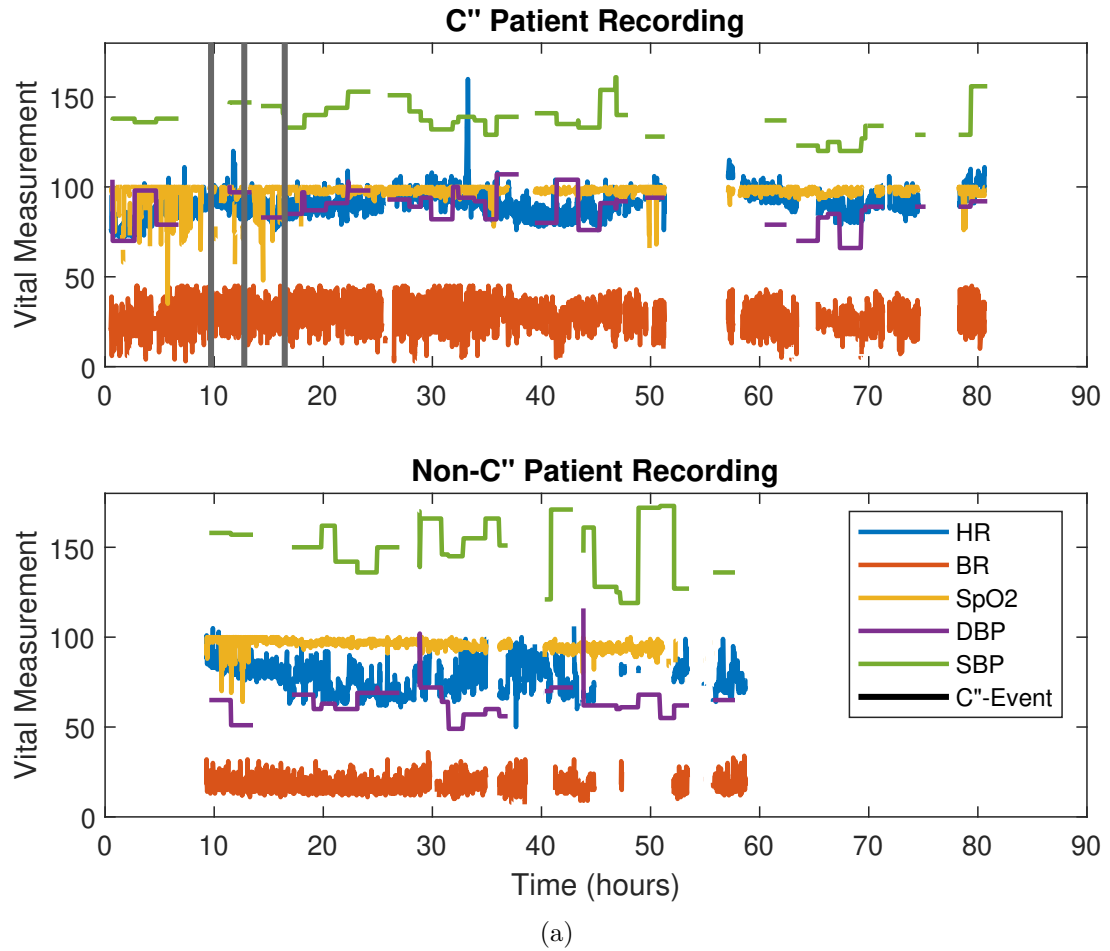


Figure 7.1: Vital sign measurements for a patient with C''-events and a patient without, top and bottom respectively.

Table 7.5: Count of the primary cause of the first C''-event.

Cause	Low HR	High HR	Low BR	High BR	Low SpO ₂	Low SBP	High SBP	High DBP
Count	2	10	4	6	20	11	3	3

7.3.3 Description of Critical Event

Although some patients may experience multiple events, we assume that the first event would be the most critical indicator of deterioration (following accepted practice in the literature). Figure 7.3 shows the distribution for the length of stay for patients with C''-events and those without, respectively. On average, patients who experienced a C''-event are observed for 141.4 hours (5.9 days), which is twice as long as non-C''-patients, who are

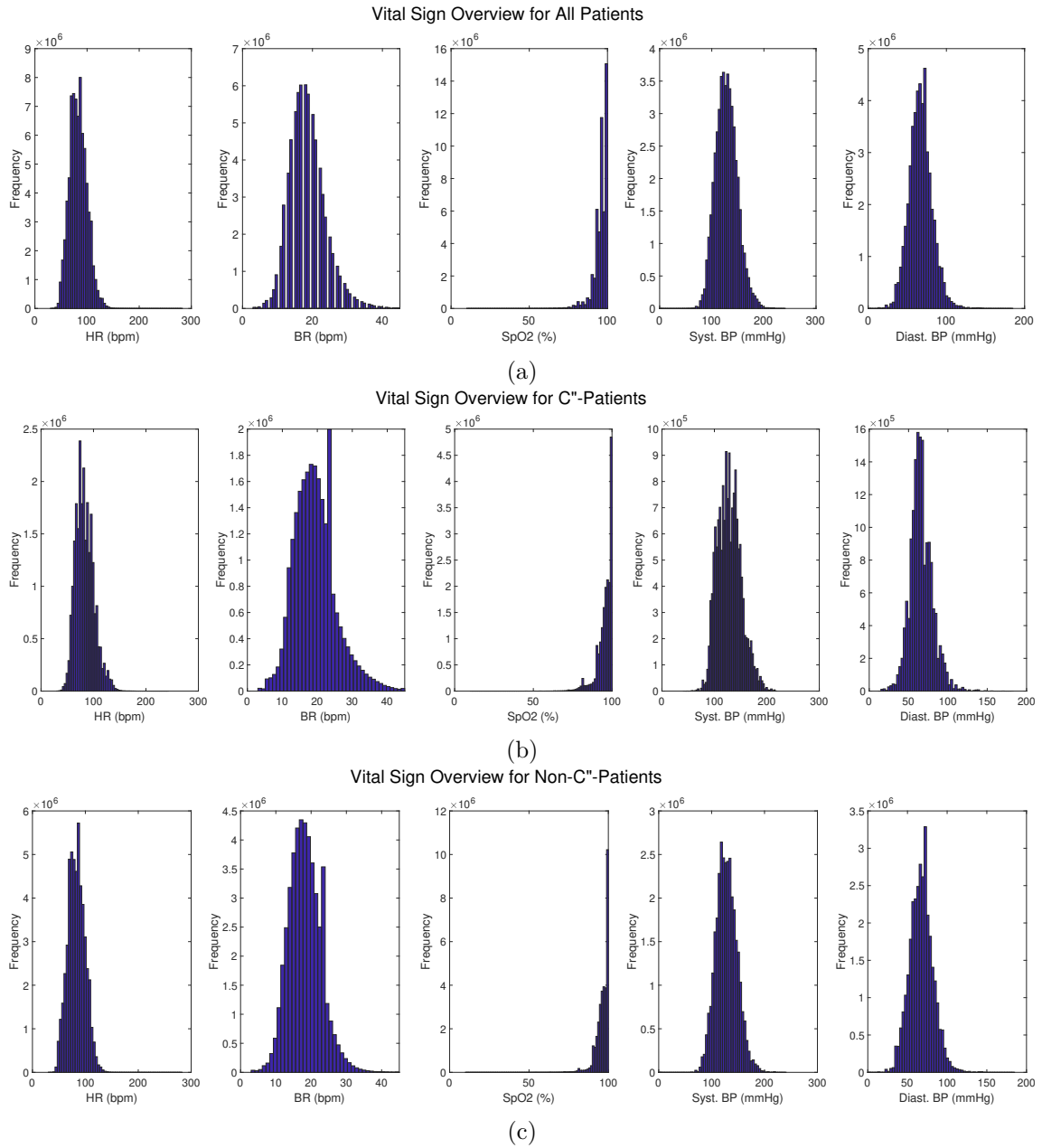


Figure 7.2: Distribution of all the vital sign measurements for (a) all patients, (b) non-C''-patients, and (c) C''-patients.

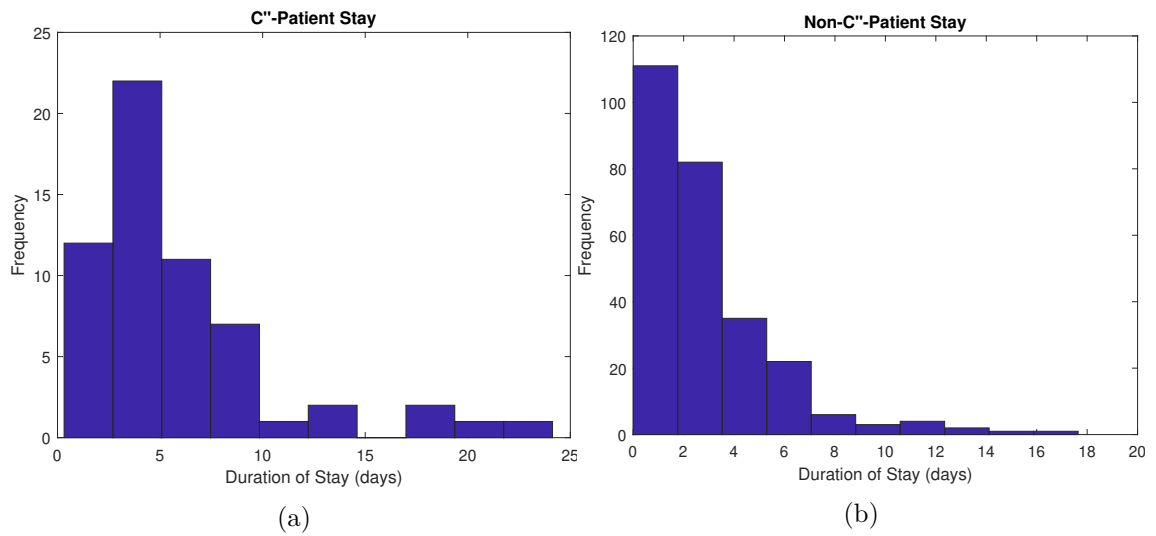


Figure 7.3: Distribution of the total monitoring duration for (a) C''-patients, and (b) Non-C''-patients as a factor of the length of their stay in the SDU.

observed for 70.1 hours (2.9 days).

A critical event can last between a few minutes and up to three hours; however, Figure 7.4a shows that even though the average event in the UPMC data lasted for 29.7 minutes, the majority of events only lasted for 7.6 minutes. The first C''-event typically occurred most frequently within the first hour but on average within the first 2.6 days of the patient's stay in SDU. On average, most patients who experience an event leave the SDU after 3.2 days.

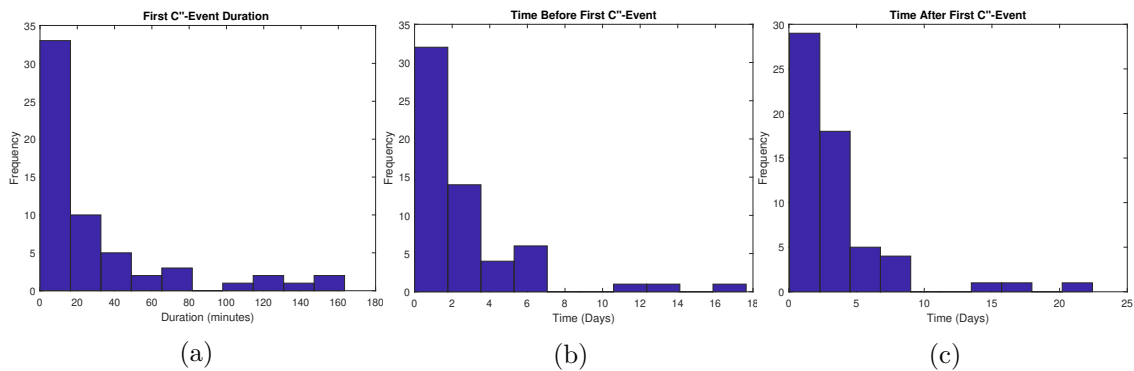


Figure 7.4: Distribution of the duration of C''-events for (a) total duration of the event, (b) amount of monitoring before the first event, and (c) the amount of monitoring after the first event.

7.4 Techniques

Previous chapters presented the investigation and technique for the frequency analysis of time-series accelerometer data. The current chapter describes two techniques that are common in patient monitoring. Previous investigations considered the KDE [233] and GP [240]. For comparison, these two methods will also be considered for the sensor-based novelty filter in addition to the five classifier models (LR, KNN, NBC, CART, NN) that were previously considered in Chapter 5.

7.4.1 Kernel Density Estimator

The kernel density estimator (KDE), or “Parzen windows”, approach is a popular non-parametric approach used for patient vital sign monitoring [233]. This approach is particularly useful for application with no *a priori* knowledge [241, 242]. Thus, to determine the probability density, $p(\mathbf{x})$, in some D-dimensional space (here $D = 5$) for N observations in the training set, we consider some small region, R , with a volume, V , and K training points falling within this region such that [205]:

$$p(\mathbf{x}) = \frac{K}{NV} \quad (7.1)$$

For the kernel approach, we fix the volume V in order to determine the number of samples K . To do this we assume that the region, R , is a sufficiently small hypercube on the origin $\mathbf{x}$ described by the unit cube function:

$$k(\mathbf{u}) = \begin{cases} 1 & |u_i| \leq 1/2; i = 1, \dots, D \\ 0 & \text{otherwise} \end{cases} \quad (7.2)$$

where $k((\mathbf{x} - \mathbf{x}_n)/h)$ is unity when $\mathbf{x}_n$ falls within the hypercube of side h centred on $\mathbf{x}$, and zero otherwise. The total number of samples in the hypercube is:

$$K = \sum_{n=1}^N k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right) \quad (7.3)$$

If we substitute this equation for K into Equation 7.1, the estimated density at $\mathbf{x}$ becomes:

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N \frac{1}{h^D} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right) \quad (7.4)$$

However, the sum over N cubes results in artificial discontinuities at the boundaries of the cubes. To overcome this, a Gaussian kernel function is commonly used as smoother kernel function to obtain a smoother density model. The kernel density model can be denoted by placing a Gaussian over each point instead of a hypercube such that:

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N \frac{1}{(2\pi h^2)^{D/2}} \exp\left\{-\frac{\|\mathbf{x} - \mathbf{x}_n\|^2}{2h^2}\right\} \quad (7.5)$$

where h is equivalent to the standard deviation of the Gaussian distribution, σ . To avoid over-fitting due to a small h , the standard deviation is derived analytically for the sample $\sigma^2 = \frac{1}{m} \sum_{n=1}^m (y_n - \bar{y})^2$, where $m = 10$ is commonly used [205].

Previous designs of the KDE [233, 240] found the location of 400 cluster-centroids in the training which is a standard approach to KDE computation using hundreds of kernels [205, 243]. Thus, the kernel density estimate of the 5-D space for each vital sign can be described by:

$$p(\mathbf{x}) = \frac{1}{400(2\pi)^{5/2}\sigma^5} \sum_{n=1}^{400} \exp\left\{-\frac{\|\mathbf{x} - \mathbf{x}_n\|^2}{2\sigma^2}\right\} \quad (7.6)$$

where σ is the mean of the variances of the distributions around each cluster-centroid (as described).

In this non-parametric approach, both the density function and model parameters are determined by the training data. An essential advantage of this approach is that the training phase is computationally light since the training set is stored. However, the trade-off is that it can lead to slower run-time during the test phase, particularly as the size of the data set increases, because many comparisons must occur for every point.

7.4.2 Gaussian Processes Classification

In supervised machine learning, Gaussian processes (GPs) are a non-parametric stochastic process used to govern the properties of functions as a generalisation of the Gaussian probability distribution over the entire function rather than individual data points [244]. As such, any finite subset of the original range will follow a multivariate Gaussian distribution, $f(\mathbf{x}) \sim GP(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$, where the mean function, $m(\mathbf{x})$, is commonly assumed to be zero and $\kappa(\mathbf{x}, \mathbf{x}')$ is the kernel or covariance function which defines how smoothly the latent function varies from a given $\mathbf{x}$. The kernel is governed by a vector of hyperparameters θ .

Linear models for classification use the same Gaussian prior $\mathbf{w} \sim \mathcal{N}(0, \Sigma_p)$ as regression to obtain the log of the un-normalized posterior that is defined by [245]:

$$\log p(\mathbf{w}|X, \mathbf{y}) = -\frac{1}{2}\mathbf{w}^T \Sigma_p^{-1} \mathbf{w} + \sum_{i=1}^n \log \sigma(y_i f_i) \quad (7.7)$$

where $f_i \triangleq f(\mathbf{x}_i) = \mathbf{x}_i^T \mathbf{w}$.

For the classification of patient condition in this project, the outputs of the GP are discrete. In the binary GP classification problem, the posterior probabilities for the input variable $\mathbf{x}$ must on the interval of the target variable $t \in (0, 1)$. As GP is a regression

method that predicts on the entire real axis, the output can be transformed by using a nonlinear activation function, such as a logistic sigmoid $y = \sigma(f) = \frac{1}{1+\exp(-f)}$ over some latent function $f(\mathbf{x})$.

For the classification case, the posterior $P(f|\mathcal{D}, \theta)$ is intractable since the marginal likelihood integral $p(f|D, w) = \int \sigma(f(x_i))P(\mathbf{f}|\mathbf{X}, \mathbf{w})df$ is a product of sigmoids multiplied by a Gaussian prior. However, this can be overcome by using an analytical approximation, such as Laplace Approximation.

Laplace Approximation for Binary GP Classifier

For the Laplace approximation (LA), we obtain a Gaussian approximation to the posterior distribution by making a second order Taylor expansion of the logarithm around the maximum of the posterior such that the Gaussian approximation to the posterior is described by:

$$p(f|\mathcal{D}, \theta) \sim q(f|\mathcal{D}, \theta) = \mathcal{N}(f|\mathbf{m}, \mathbf{A}) \quad (7.8)$$

where $\mathbf{m}$ is the mean and $\mathbf{A}$ is the covariance. LA is used to find these two parameters by placing $\mathbf{m}$ at the mode (MAP) and $\mathbf{A}$ is equivalent to the negative inverse Hessian of the log posterior density at $\mathbf{m}$.

The log marginal likelihood (LML) of the observed points is calculated by evaluating the Gaussian integral for a given value of the covariance function hyperparameters, θ :

$$\log q(\mathbf{y}|X, \theta) = -\frac{1}{2}\hat{\mathbf{f}}^T K^{-1}\hat{\mathbf{f}} + \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2} \log |B| \quad (7.9)$$

where $|B| = |K| \cdot |K^{-1} + W| = |I_n + W^{\frac{1}{2}}KW^{\frac{1}{2}}|$ and K is the covariance matrix. In this case $W \triangleq -\nabla^2 \log p(\mathbf{y}|\mathbf{f})$ is a diagonal matrix because the likelihood factorizes over cases.

For the covariance kernel, we consider the commonly used squared-exponential or Gaus-

sian kernel described by:

$$k_{SE}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right) \quad (7.10)$$

where ℓ is the lengthscale and σ^2 is the output variance. The kernel hyperparameters, $\theta_{SE} = [\sigma, \ell]$, is fit according to a MAP estimate. GPs offer the advantage over SVM that the kernel parameters are learned from the training data without the need for cross-validation.

7.5 Method

7.5.1 Pre-processing

Before extracting the features, several steps ensure that standard data pre-processing occurs. Firstly, we apply artefact removal to each non-empty vital sign channel using the limits described in Table 7.3. A plausible physiological range determined by clinical knowledge defines these limits. We reject any values outside these limits.

Secondly, we remove the monitor artefact from the BR channel. Monitor artefact occurs when a single value appears for an abnormally long time, and as such, we remove these segments of artefact measurements. Once removed, we confirm whether the patient has valid BR measurements for at least 60% of the recording to be included in the database. The inclusion criteria of a valid reading is done by calculating the mean, μ_{BR} , and standard deviation, σ_{BR} of each five-minute segment in the recording and determining if each segment has a mean BR value in the range of $8 \text{ bpm} > \mu_{BR} < 40 \text{ bpm}$ OR the standard deviation $\sigma_{BR} < 4 \text{ bpm}$.

Next, the time scales are shifted and scaled uniformly. A timestamp, usually the HR start time, is selected around which all other timestamps shifts and each recording scales

to minutes. Finally, we linearly interpolate the values of each vital sign channel to account for any missing data or artefact segments removed in previous steps.

7.5.2 Parameter Normalisation

Each of the five vital sign is normalised to ensure each parameter shares a common scale and contribute equal weight to the trained model without distorting the variations contributed by each. This is done by applying the zero-mean unit-variance transformation to each vital sign measurement, x_i , for $[i = 1, 2, 3, 4, 5]$ for each of the five channels such that:

$$x_{n,i} = \frac{x - \mu_t}{\sigma_t} \quad (7.11)$$

where x_n is the normalised vital sign value. The mean, μ_t , and standard deviation, σ_t , are calculated using the training data. Table 7.6 shows the mean and standard deviation for a typical training set. Both the training and test data are normalised using the mean, μ_t , and standard deviation, σ_t , of the training set.

7.5.3 Labelling

For patients with C"-events, we assume that there is some distinct point in time when the patient's condition changes from normal to abnormal. The timestamp of the start of the first C"-event was retrospectively marked by the clinician. Thus, there is some window, W_1 , before the event starts labelled as abnormal and a second window, W_2 , after the event. Any vital sign measurements obtained before W_1 is considered normal, and those after W_2 are removed. Even though some patients may experience a second and third C"-

Table 7.6: Mean, μ_t , and standard deviation, σ_t , for each patient vital sign in the training data.

	HR	BR	SpO ₂	SBP	DBP
μ_t	84.58	18.69	94.67	124.49	69.15
σ_t	16.95	5.37	5.56	16.51	13.24

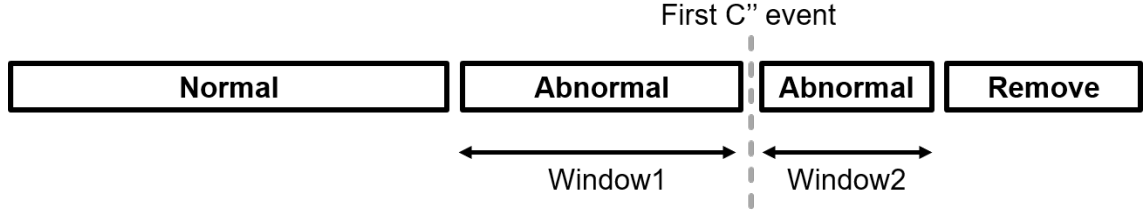


Figure 7.5: Timeline for labelling the data as normal and abnormal for a given window size, W_1 and W_2 , given the patient's first C''-event.

event while others only experience a single event, the first event is the primary indicator of deterioration. To ensure that patients with multiple events are not over-represented in the data, we only consider the first event for all C''-patients and discard the remainder of the recordings.

7.5.4 Feature Extraction

The features for all five vital sign channels were considered in the time-domain only. We determine the standard statistical parameters commonly used in patient monitoring for the time-domain features. Features were created by taking a fixed window, W_f , and calculating the mean, median, minimum, maximum and standard deviation of each of the five vital signs resulting in 25-D feature vector per window. The window then slides forward for some step size, S_f , which is smaller than W_f such that some amount of overlap occurs. The final result is a $N \times 25$ feature matrix, where N is determined by the final selection of W_f and S_f . The selection of W_f and S_f is discussed in the results section.

7.5.5 Train-Test

For each iteration, we allocated the patients into two mutually exclusive groups, for C''-patients, $\mathcal{D}_{C''}$, and non-C''-patients, $\mathcal{D}_C$. The patients are then randomly divided into the 80-20 training and test set, and each is balanced to include the same number of C''- and non-C''-patients. As such, the training set would contain approximately 92 patients, 46 each from $\mathcal{D}_{C''}$ and $\mathcal{D}_C$, and the test set would contain approximately 24 patients, 12

each from $\mathcal{D}_{C_{pp}}$ and $\mathcal{D}_C$.

7.5.6 Novelty Filter Score Estimation

Following the framework from Figure 6.1, the sensor nodes (previously represented by the accelerometry data from rural handpumps) now represent the physiological patient data collected from the bedside monitors. Instead of a single sensor, multiple sensors collect the physiological data. In a real-world set-up, the sensor node for collecting physiological data from patients may be a wearable sensor, such as an electrocardiogram (ECG) patch or photoplethysmogram (PPG) sensor. Figure 7.6 shows two different PPG sensors with ear clip and finger probe connections for practical implementation in future. The sensor on the patient produces an in-situ novelty score which filters the data such that only those examples deemed as abnormal transmits to the cloud. As in Chapter 5, this will initially be done retrospectively in a simulated lab environment, but unlike the handpump case, it will not be deployed for testing in real-time as part of this project.

The proposed classifiers (LR, KNN, NBC, CART, NN, KDE, GP) will be used to simulate the novelty filter score and a comparison discussed in the results. The training of the proposed models will follow the same framework used for the handpump application



Figure 7.6: Low-cost wearable sensors used for collecting mobile physiological data from patients in limited-resource settings. PPG sensor with ear clip and storage-battery unit (left) and PPG sensor with a finger probe and display unit (connected to tablet or smart-phone via Bluetooth) (right).

in Chapter 5. Since the KDE stores the training data to evaluate new points, there is no computation involved in training this classifier. During testing, we use the training data to determine the probability density estimate for C"- and non-C"-Patients to evaluate the unseen data points.

7.5.7 Central- or Remote-Monitoring Training Framework

These novelty filtered data produce a novelty score of the patient health for the cloud-based classifiers to evaluate. In the ideal system design, the novelty filtered data can be transmitted from the remote wearable sensor to a central monitoring system for patients in an ICU or hospital ward or a cloud monitoring system when the patient is not near a clinic or hospital. In this case, visits to a district hospital are reserved for extreme cases since it involves costly travel and time. The central-monitoring platform will consider the same classifier models that were proposed for the handpump DI platform: LR, SVM, RF, MLP and 1D-CNN. Five-fold cross-validation was used in the same training framework described in Algorithm 1 in Chapter 6 to optimise the hyperparameters of each classifier.

7.5.8 Evaluation

The DDI system in a different application will be evaluated as before, considering both measurement accuracy and robustness. We compare the proposed DDI method to the five cloud-based classifiers. In this comparison, we used the same techniques to assess the performance for monitoring patients as we those used for assessing the performance for handpump monitoring: prediction accuracy, sensitivity, precision and recall. We evaluate the timeliness of the alerts as a trade-off of the FPR. Similarly, we evaluate the robustness by considering the standard deviation of these parameters for at least 20 iterations. Robustness is an important parameter because as we saw in the case for handpump monitoring, the DDI method is designed to improve robustness, which is essential since false

alarms can have a high cost in remote monitoring or even be fatal when ignored in the case of patient monitoring.

7.6 Results

7.6.1 Labelling and Feature Selection

The size of the pre-event window, W_1 , is set to two hours following other studies [233, 240]. To determine the ideal window size after the start of the first C"-event, W_2 , we consider a range of values between 15 min and 24 hours as recommended by the clinician that we consulted. Figure 7.7 compares the misclassification rate of nine proposed window sizes, W_2 . From this, we see that $W_2 = 4$ hours or 24 hours minimises the misclassification rate. Some subset of patients may experience second and third C"-events in the 24 hours following the first event and choosing such as window size would increase the total number of abnormal examples. Since not all patients with a C"-event experience multiple events, such labelling could lead to learning incorrect features.

The post-C" window size W_2 was selected to be 4 hours to ensure that all abnormal data from C"-patients are likely to represent only one event per patient. Any subsequent

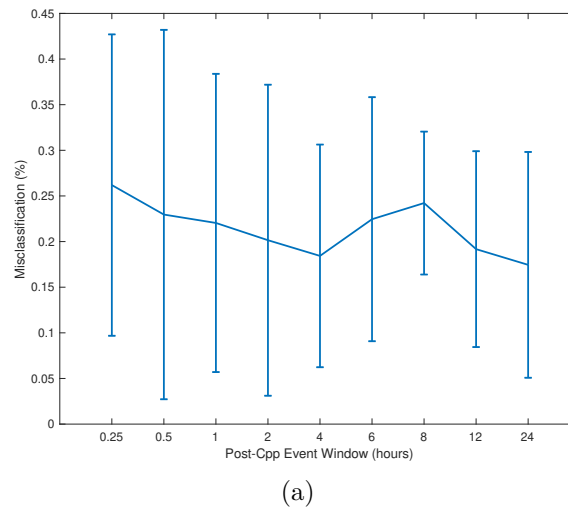


Figure 7.7: Results for misclassification of the test set when varying the size of the post-C" window, W_2 .

events are linked to relapse and features may be altered since treatment has already been administered following the first C"-event. Thus, if the first C"-event for a C"-patient was to start at $T_{C_{pp1}}$, then all vital sign measurements from the beginning of the recording until time $(T_{C_{pp1}} - W_1)$ were labelled as normal, recordings from $(T_{C_{pp1}} - W_1)$ to $(T_{C_{pp1}} + W_2)$ were labelled as abnormal, and all recordings after $(T_{C_{pp1}} + W_2)$ were discarded. All non-C"-patient data were labelled as normal since there was no C"-event.

Once the vital sign recordings for each patient are appropriately labelled, we extract the features to train the respective models for both the patient- and cloud-based classifiers. Figure 7.8 compares the misclassification rate and standard deviation for the proposed feature vectors, FV_1 to FV_7 . This shows that the misclassification rate is minimised for FV_5 where $W_f = 30$ minutes and $S_f = 5$ min. Window sizes smaller than 30 minutes are likely unable to capture sufficient variability in the vital signs of the patients as deterioration is gradual rather than sudden. Similarly, window sizes greater than 30 minutes are too long and unable to capture variability. The selected window and step size combination of FV_5 results in a feature set of 66517×25 , which is balanced and split to an 80-20 train-test set of size 2904×25 and 881×25 , respectively.

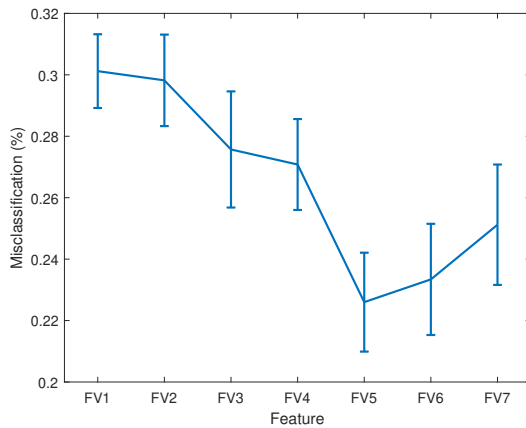


Figure 7.8: Comparison of the misclassification rate for different feature vectors, FV_1 to FV_7 .

	Win. (min)	Step (min)
FV_1	5	1
FV_2	5	2.5
FV_3	10	5
FV_4	15	5
FV_5	30	5
FV_6	30	10
FV_7	60	20

Table 7.7: Proposed feature vectors, FV_1 to FV_7 , determined by the window, W_f , and step, S_f , combinations.

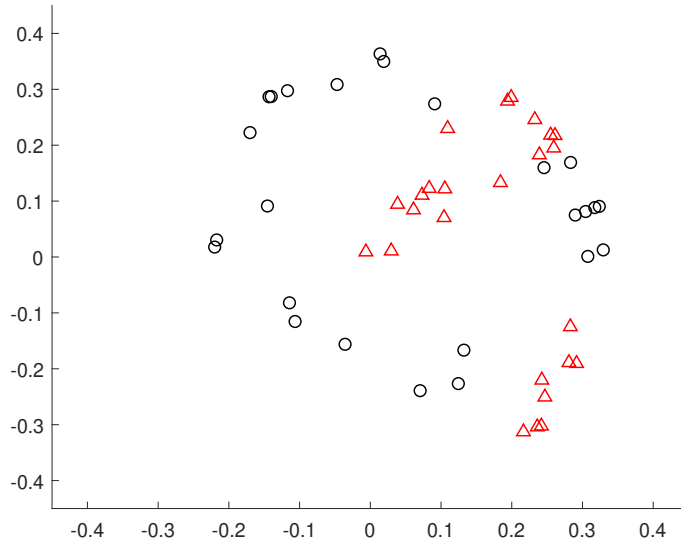


Figure 7.9: Projection of 25-D feature vector, FV_5 , for vital sign measurements for all C''- (red triangle) and non-C''-patients (black circle), respectively.

Figure 7.9 shows the resulting PCA projection of the first two principal components from the proposed 25-D feature vector, FV_5 . The projection shows that the selected feature vector, FV_5 , captures the variability in the data sufficiently well such that visible clusters appear. Despite some inter-mixing, apparent clustering of abnormal conditions for C''-patients (red triangles) can be seen confirming the separability of the conditions.

Finally, the top of Figure 7.10 shows an example of the raw data for each vital sign for a single patient along with a marking (vertical black line) to indicate the start time of the first C''-event. The extracted features for this patient can be seen below the original raw data.

7.6.2 Performance of the Patient-based Novelty Filter

We used the proposed feature vector, FV_5 , to train and evaluate the seven simple classifiers to embed in a wearable device. Table 7.8 shows a comparison of the performance metrics for the trained classifiers using the unseen test data containing the five vital signs to classify whether a patient experienced a C''-event or not. In this instance, as with the handpumps in Chapter 5, KNN is the best performing model, while CART and LR are the worst-



(a)

Figure 7.10: Example of the raw vital sign measurements (top) and associated features (bottom) for a C⁺-patient.

performing classifiers. As with the handpump case, when selecting the onboard classifier, there is a 15% difference in accuracy between the best and worst-performing classifier. The performance of the LR for this application is very similar to before ($AUC = 0.724 \pm 0.01$). KDE and GPs both offer suitable performance, along with the original five simple classifiers that we considered for the handpumps. For a first stage in-situ novelty detector embedded in an 8-bit microprocessor of a wearable sensor, any one of these model would offer suitable performance and robustness to meet the design criteria. Therefore, in the interest of evaluating the transferability of the proposed DDI platform, the LR model is used as the sensor-based novelty filter.

7.6.3 DDI Platform Performance

Once we have the in-situ novelty scores, we analyse the DDI platform introduced in the previous chapters using the proposed data set from UPMC. As with the handpump results

Table 7.8: Performance metrics for the seven different classification models proposed to be embedded onboard the wearable device. Results are presented with mean and standard deviation for 20 iterations.

Model	CA	Sens.	Spes.	AUC
LR	0.705(0.01)	0.685(0.02)	0.725(0.02)	0.704(0.01)
KNN	0.862(0.01)	0.893(0.02)	0.832(0.02)	0.863(0.01)
NBC	0.859(0.02)	0.926(0.04)	0.792(0.02)	0.859(0.02)
CART	0.708(0.01)	0.651(0.02)	0.765(0.02)	0.708(0.01)
NN	0.769(0.05)	0.738(0.17)	0.799(0.08)	0.769(0.05)
KDE	0.789(0.02)	0.677(0.03)	0.732 (0.03)	0.786 (0.02)
GP	0.786(0.01)	0.678(0.03)	0.899(0.01)	0.806(0.02)

Table 7.9: Performance evaluation of the proposed distributed inference method compared to other standard methods. Standard deviation calculated for 100 iterations.

	CA	AUC	Precision	Recall
LR	0.716(0.010)	0.718(0.012)	0.721(0.013)	0.641(0.023)
RF	0.759(0.010)	0.752(0.011)	0.783(0.013)	0.644(0.022)
SVM	0.733(0.012)	0.734(0.012)	0.886(0.025)	0.369(0.025)
MLP	0.750(0.013)	0.735(0.016)	0.876(0.024)	0.425(0.024)
1D-CNN	0.721(0.018)	0.725(0.013)	0.797(0.020)	0.429(0.014)
Proposed DDI	0.817(0.019)	0.781(0.009)	0.778(0.016)	0.664(0.022)

in Chapter 5, the SVM is implemented with an RBF kernel, and grid search is used to optimise the hyperparameters. The number of layers and the number of nodes per layer for the MLP were selected using Bayesian optimisation. The 1D-CNN network uses a 1D convolution layer followed by a relu activation function and three fully connected layers. The number of nodes for each layer is optimised through experimentation.

Table 7.9 shows the comparison of the performance for the proposed models for the cloud-based system using test data from unseen patients. The results show that the classification accuracy of the proposed DDI method is also able to outperform other standard machine learning methods for physiological patient data. Moreover, the DDI method offers both excellent recall and precision when compared to the other standard methods, which sacrifice recall. The MLP and SVM both have better precision than the proposed DDI method; however, they also share the worst recall.

Finally, one of the key design criteria of the DDI platform is not only to be a reliable health-monitoring system but also cost-effective to operate. The long term maintenance and financial support of any healthcare infrastructure in limited-resource settings remain the greatest challenges to ensure optimal and effective use over its lifespan. As such, we also consider the overall transmission cost (in MB) of the DDI method to provide a timely warning of an imminent C"-event for the SDU patients. Figure 7.11 compares how early (in windows where 1 window is 30 minutes) the C"-event is detected for specific FPR. The DDI platform is compared to the alternative of transmitting only the novelty scores or only the features to the central monitoring system. In both alternatives, the transmission cost is fixed since the data type is predetermined. However, the novelty threshold of the DDI system is dynamically adjusted based on the predictive fidelity and determines the amount of data transmitted, which inversely proportional to the FPR.

Similar to the handpump results, the results show that the DDI platform can achieve a low FPR (< 0.3) by using only a quarter of the data required to transmit the features

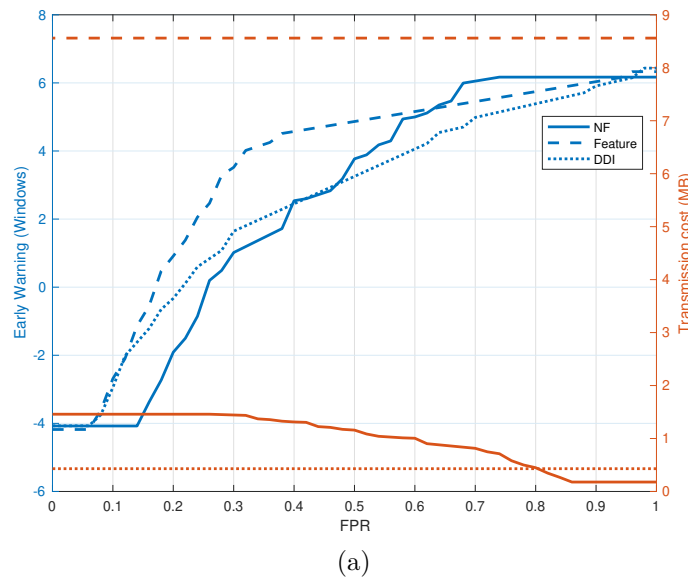


Figure 7.11: Comparison of the performance and cost of operating the DDI platform (dotted) to detect C"-events in patients over non-dynamic systems transmitting only novelty filter scores (solid) and features (dashed), respectively.

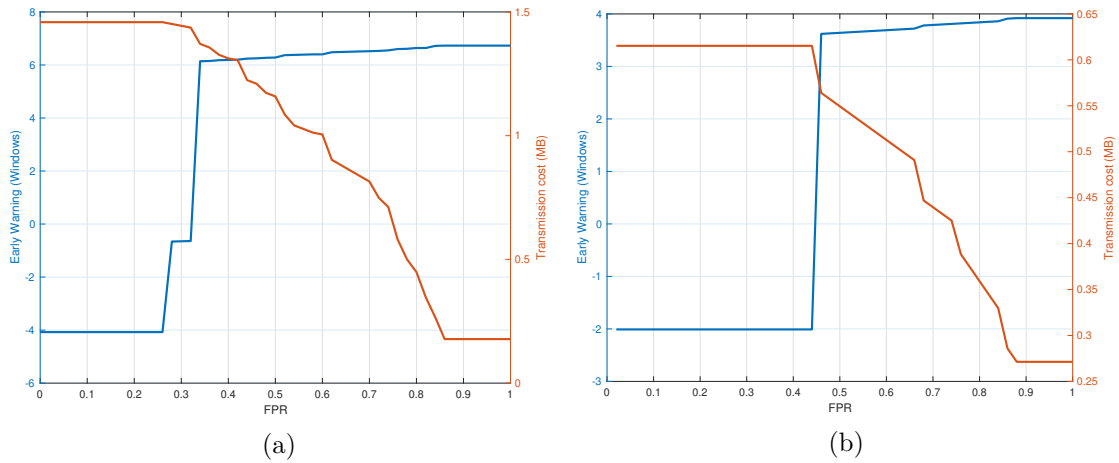


Figure 7.12: Two cases using individual patient data with C''-event to perform the early warning alert by the DDI system along with the associated cost (measured in amount of data being transmitted).

in order to detect the C''-event before it occurs. At an FPR of 0.3, a system transmitting features would be able to alert 2 hours before the event while the DDI platform will only alert 1 hour before the event. Depending on the context, this may or may not be sufficiently early. For instance, if the patient was being monitored remotely in a rural village, it may be that they are unable to access appropriate help in time.

Figure 7.12 shows the outcome of the DDI platform for two individual patients. In the first example, the patient will benefit from the DDI platform since this event could be alerted ahead of time. Within a low FPR, approximately 0.3, the DDI method could alert the C''-event up to three hours early. However, in the second example, Figure 7.12b shows that the system may miss alerting the event entirely for a low FPR. On closer inspection, this recording contains many segments with artefact noise, and so additional optimisation of the novelty filter threshold would ensure that the DDI platform is more specific for the use in patient monitoring.

7.7 Conclusion

The results in this chapter presented the application and analysis of the proposed DDI platform to patient data using vital sign measurements. It was demonstrated that the DDI platform could be applied to patient data with minimal adjustment. The DDI offers at least 5% and up to 10% improvement in prediction accuracy over other standard machine learning methods. The preliminary results show that the DDI platform may be suitable for monitor patients, although optimisation of the novelty threshold may be required to ensure sufficient specificity. Without adjusting the threshold, the system can alert one hour before the event for a low FPR (< 0.3).

However, given that the alert could be triggered two hours before the C"-event when sending only features for the same FPR, it may be necessary to adjust the boundaries of the threshold which are optimised for handpumps that tend to deteriorate much slower. Individual cases show that motion artefact can lead to missed events. Thus the DDI platform should be further adapted for patient monitoring, especially in mobile monitoring using wearable sensors. One hour might be a sufficiently long lead time for a patient in an ICU or clinic in a limited resource-setting where a nurse or doctor would be available within the hour. However, additional adjustments are required to ensure that the DDI platform can detect failures sufficiently early before it is applied to remote monitoring, where the lead time for patients to access care following an alert is often longer.

The proposed DDI approach should be validated on new data sets within a related clinical setting, specifically from wearable sensors given the anticipated clinical setting of the DDI platform would be most likely out-of-hospital patients or those in clinics in limited-resource settings. The motivation for applying the DDI platform developed for

handpumps to patient data was to demonstrate the versatility of the platform for real-world application. This proof-of-concept makes a case for large-scale data collection using low-cost wearable devices that can be used by clinics in limited-resource settings as the DDI platform easily transfers to other applications.

Chapter 8

Conclusion and Future Work

This thesis addressed how to use machine learning techniques in a frugal context to monitor the condition of equipment remotely. The work presented a novel approach to remotely monitor the health of a rural water handpump given limited-resources in an extreme operating setting. We demonstrated that it is possible to distinguish normal handpumps from abnormal handpumps, with more than 80% precision, using machine learning techniques combined with suitable hardware. We addressed the four critical components of a predictive and scalable health-monitoring system for handpumps by showing that:

1. Low-cost, embedded sensors can acquire high-rate accelerometry data from routine daily handpump usage. We successfully collected, curated and published three data sets (Chapter 4), including a longitudinal data set containing associated novelty filter scores produced by the sensor embedded in the handpump handle. The in-situ signal condition monitoring and pre-processing of the acquired data required additional processing steps (Chapter 5) to account for the embedded system constraints.
2. In-situ machine learning analysis of the resulting data can produce an informative assessment of the equipment health condition with more than 70% sensitivity. The validated and implemented novelty filter using LR (Chapter 5) can extract and assess features from the accelerometry vibration spectra onboard the rural handpump handle such that “intelligent” data summaries, containing almost exclusively “ab-

normal” data, may be transmitted to a cloud-based system for further analysis.

3. Communication with a cloud-based monitoring centre via the domestic telecommunications network can perform more advanced machine learning techniques on the cloud-based system to increase the accuracy and certainty of the classified health condition of the equipment. We designed and evaluated a dynamic distributed inference system (Chapter 6) in which the novelty filtered data from the handpump are classified without the processing and power constraints of an in-situ embedded sensor to increase the precision of the proposed health-monitoring system to more than 80%.
4. The DDI proposed system can identify imminent equipment failure in advance and is cost-effective (Figure 6.15). The proposed DDI approach was assessed using physiological patient data to ensure the approach is suitably scalable and transferable to other applications. The results (Chapter 7) demonstrate that the proposed system can easily be adapted and scaled to monitor the health of other rural nodes by evaluating the performance of the approach to predict deterioration in patients.

Contribution to Knowledge

This thesis made a contribution to the current state of knowledge in the field of machine learning applications by its ability to:

1. Create a low-cost data stream to monitor and assess the condition of rural hand-pumps remotely.
2. Provide an innovative theoretical method to increase the fidelity of a classifier while reducing the operational efficiency.
3. Offer a frugal practical and methodological approach in the application of machine

learning in limited-resource settings.

8.1 Summary of the Thesis

8.1.1 Remote Health Monitoring

Remote health monitoring of rural infrastructure is becoming more critical as global development, and tracking of SDGs demand more transparent metrics for progress. Despite many years of development efforts to advance the reliability of rural water handpumps, there is still no universal definition for the “functionality” of a rural water point, and thus multiple measures of monitoring are used. Broadly, the device should be reliably producing sufficient yield to meet the needs of the water users who rely on it [56]. Chapter 2 presented a comprehensive list of the most common handpump failures, shown in Table 2.3, which formed the bases for the standardised categories proposed for labelling the handpump accelerometry data in this thesis. Furthermore, the critical relationship of the handpump within rural communities was explored concerning the existing conventional handpump maintenance systems. Understanding the value of ownership as a critical driver of handpump sustainability is an essential factor that any successful health-monitoring system should take into account. These novel systems would play an important role to address the lack of accountability that currently results in maintenance service delays of rural infrastructure.

In Chapters 3, we describe the urgent need for objective and reliable monitoring approaches due to the lack of existing methods to perform remote health monitoring in rural settings, especially in water supply management. Most monitoring projects rely on manual user inputs which can be unreliable and inaccurate. Projects that collect automated data streams use flow sensors and generally report the raw data; missing out on valuable insights about usage patterns that can be obtained through minimal analysis. To understand the

potential of machine learning approaches for health-monitoring systems, we present cases using vibration-based monitoring, shown in Table 3.1. Combined with the visualisation of handpump maintenance, shown in Figure 3.2, this offers a valuable resource to underpin the design of a scalable health-monitoring system for limited-resource settings.

The first design step involved the accelerometer sensors embedded inside the handpump handles to collect the high-rate vibration data. The bespoke sensors used to collect the accelerometry data are introduced in Chapter 4. These same sensors were later adapted to complete the in-situ processing using a logistic regressor to produce a novelty score onboard the handpump. The extreme operating conditions, such as high temperature and humidity, tested the limits of the electronic components not designed to operate in such conditions for extended periods. Improvements to the casing and seals enhanced both performance and battery life of the embedded sensors.

Three unique databases were constructed over three years and contain data from 48 handpumps, where 28 handpumps contain only high-rate vibration waveforms with condition labels. The remaining 20 handpumps have high-rate waveforms, condition labels, and the novelty scores produced onboard the handpumps. The final database is a carefully selected cluster of ten handpumps that contain continuous, longitudinal high-rate waveforms and in-situ novelty scores of closely located wells for at least 18 months. The construction of the three databases was curated along with the proposed standardised condition labels. These databases, along with additional aquifer level data collected alongside the vibration data [89], has been archived as open access with the British Geological Survey to encourage broader work in this field [86].

8.1.2 Distributed Inference

The design of the sensor-based novelty filter and the construction of suitable feature vectors are discussed in Chapter 5. The time-series accelerometry data from the handpump handle were analysed in the frequency domain to identify relevant high energy bands. The PDFs for “normal” and “abnormal” operating conditions showed that these conditions have distinctively different high-frequency components for which high energy appears. Unique features clusters were also shown for handpump operating at different pumps. The depth of the handpump well is an important influence in the pumping characteristics since the number of internal rods directly influence the weight lifted with each stroke. As such, both handpump condition and operating depth are considered as important factors that influence the pump characteristic. Speed-based representations of the spectral data could not confirm whether the pumping speed influenced the pumping characteristics.

An LR classifier was used to perform condition monitoring inference onboard the handpumps ($\text{AUROC} = 0.725$). The feature vectors contain 20 points uniformly sampled across the entire frequency spectrum and shown to be optimal. The LR is compared to other standard machine learning methods suitable for use on embedded sensors (KNN, NBC, CART, NN) and deemed to be a sufficiently good performer given that the probabilistic framework would allow the threshold of the novelty filter to be adjusted dynamically by the cloud-based classifier.

The novel framework for the proposed dynamic distributed inference platform, shown in Figure 6.1, is presented in Chapter 6. The real novelty scores collected from the field are validated and compared to those generated in the lab, which only had a 3% higher AUROC than the field-generated novelty scores. Despite the lower accuracy, the novelty scores produced by the embedded sensors in the field handpumps were used for the ongoing

analysis. The field-generated novelty scores also revealed that certain failure types are more difficult to detect than others. This was mostly the case for failures that occur on a handpump section, such as the rising main, that is physically further from the sensor or in cases where the failure was not extreme, such as a broken pump rod.

An RF classifier was identified as a suitable cloud-based classifier to perform further processing once the novelty scores are received from the handpump sensors. The distributed inference system was compared to standard machine learning methods (RF, SVM, MLP, CNN) and found to offer improvement in the classification accuracy but more importantly, delivered at least 10% improvement in the precision over any other method. It is shown that the novelty filter can be adjusted to change the type of data being transmitted dynamically. Transmitting features or raw waveforms offer improved fidelity compared to novelty scores but increase the transmission cost. The DDI system is shown to perform early warning on both simulated and real data while reducing the overall transmission cost.

Finally, the transferability of the proposed DDI method was demonstrated in Chapter 7, where the system was used to monitor physiological data from patients in a step-down unit. The DDI method offers an improvement of at least 6% in the classifier accuracy over other standard machine learning methods (RF, SVM, MLP, CNN, KDE, GP). Similarly, the DDI can produce an early warning for patient deterioration for a low FPR (< 0.3) at a reduced cost of transmission.

8.2 Future Work

The use of machine learning methods combined with low-cost technologies to develop frugal, reliable remote health-monitoring systems present multiple opportunities for ongoing research related to low-cost health monitoring in addition to ongoing improvements for

the proposed dynamic distributed inference framework.

8.2.1 Deployment for Impact

It was always the intention that the work produced by this thesis should have an impact beyond academic and literary output. The deployment will enable the design of a health-monitoring system for handpumps and other rural infrastructures which is universally applicable to maximise the potential impact for the rural poor. As such, setting up a production pipeline for the cloud-based platform would enable “plug-and-play” options for data to be transmitted from different locations and subsequently different pump types (other than the Afridev), as described in section 2.2. Figure 8.1 shows the projected clusters of recorded vibration signatures from three different pump types, commonly used in Bangladesh (Samrat) and Ghana/Uganda (India MK II), represented in previous data sets of this study. Transfer learning is a technique that may be applied to reduce the training time by providing prior knowledge (such as model parameters or weights). This approach can ensure that models from the proposed method are easily adapted for other handpump types without the need for large data sets, which can be expensive and time consuming to collect.

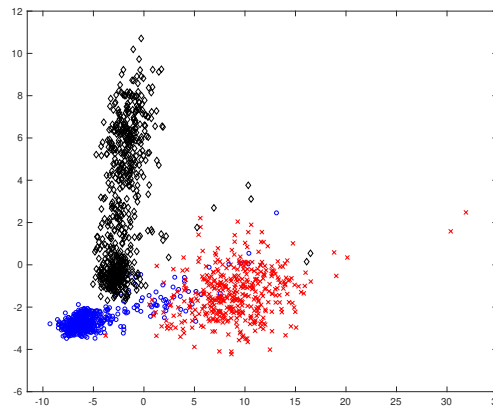


Figure 8.1: PC projection of different pump types: (1) Afridev (blue $\circ$), (2) Samrat (red $\times$), and (3) India MKII (black $\diamond$).

The prominent spectral bands that appear at different frequencies based on the oper-

ating depth of the well introduces the need for including *a priori* knowledge about the system. Depth-specific models are unpractical and expensive to implement. The field of data-space warping has been used in other fields of machine learning, in which a Gaussian process regressor is learned jointly with the warping transformation of the data-space. The use of warping techniques to scale the Y-axis of the power spectra for deep and shallow cases should be investigated, such that a standardised classifier can be applied independently of the well depth. For large-scale deployment, this would greatly simplify the roll-out of condition loggers in the field.

The implementation of the cloud-based platform will also enable two-way communication between the rural node and the cloud classifier such that the dynamic nature of the proposed method could be verified in real-time (this was simulated in this thesis using the existing one-way communication) by adjusting the novelty threshold, the type of data being transmitted, and the sampling frequency of the health-monitoring system. Such an implementation may be useful to provide evidence of further reductions in the amount of data transmission and thus associated operating cost. The onboard novelty filter may benefit from the use of Discrete Wavelet Transform (DWT) [246] to extract features as a more suitable input (than FFT) for Gaussian process, which is commonly used in monitoring applications.

Finally, such a platform may facilitate the collection of additional raw data to refine the existing models and allow further investigations to consider more data-intense methods (such as unsupervised deep learning). Recurrent neural networks (RNNs), such as long short-term memory (LSTM), have proven useful to detect the presence of damage in vibration-based monitoring of structures [247]. Although this format of data collection will come at the additional cost of transmitting raw waveforms (initially to obtain a more

extensive data set), it is less costly and time-consuming than it has been in the past, which required numerous field visits.

8.2.2 Prognosis for Planning

The DDI method described in this thesis can identify “abnormal” handpump conditions by raising a timely alarm such that the maintenance service company can respond and repair the handpump before failure occurs. This output is crucial for the water users who rely on the handpumps for their daily water needs. However, the practical operations of the maintenance service company will be enhanced by the system’s ability to perform multi-class prognosis. Knowing, ahead of time, precisely what the cause of the failure is, allows the maintenance team to carry the appropriate parts and supplies to avoid the unnecessary trip, especially when failures occur in remote locations that are difficult or dangerous to reach. This is particularly important for crucial failures that cause total loss of water, such as a broken pump rod or rising main. These replacement parts are bulky and difficult to carry on motorbikes, the primary means of travel between handpumps and supply parts. Furthermore, objective, continuous records of the type of failure and spare part requirements go well beyond impacting the final node in the supply chain. These records will be a valuable resource for local governments, NGOs, and business to adjust the part supply chain to seasonal requirements and support future work to improve parts that have a high failure rate.

The output of Chapter 6 offers the proof-of-principle to confirm that different “abnormalities” show different characteristics. However, the existing databases are limited in the type and number of failure types represented. Since obtaining this data was time-consuming, costly and often sporadic, seeding faults experiments may offer a way to obtain such data quickly. Determining the type and extent of the damage requires prior knowledge

about the “normal” behaviour parameters which often requires long-term observations to monitor and evaluate the operational performance.

Notably, the application of CNNs has recently proven useful for multi-class damage identification problems [248], which can identify the particular failure and facilitate more efficient planning as well as make a prediction regarding the remaining life of the system should the problems not be addressed immediately. Additionally, unsupervised waveform clustering (UWC) could be considered for the cloud-based classifier as structural health analysis has found it a suitable tool for reporting both the presence and location of damage [249].

8.2.3 Clinical Decision Support with Wearables for Limited-Resource Settings

The successful translation from handpump to human in Chapter 7 shows the robustness of the proposed DDI approach in monitoring various rural systems. Wearable health devices (WHD) continue to increase in popularity and offer a low-cost solution to monitoring physiological patient data in limited-resource settings. Wearable devices are no longer only considered for fitness monitoring, but large-scale use has seen a rise in demand. Mobile health (mHealth) applications for health status monitoring are essential for both patients outside the hospital to track recovery following medical interventions and inside the hospital, especially in ICU of clinics and hospitals in developing countries where limited resources often necessitate clinicians to share monitors amongst patients.

One of the biggest challenges in mobile monitoring is to obtain a reliable BR from the PPG waveform. Of all the vital signs that wearables sensors have focused on (HR, BR, BP, temperature, SpO₂), BR is of great importance since it is widely used as one of the earliest and most accurate predictions of deterioration. The conventional analytical methods used

to process PPG waveforms are computationally intensive and have considerable power requirements. Efforts to overcome this problem is ongoing, and the proposed DDI solution described in this thesis may utilise some of the processing and power capacity available from a mobile platform connected to the sensor, either via cable or Bluetooth, to increase the fidelity without requiring additional resources. Hence, the mobile platform will interact with the central cloud-server to adjust the threshold to decrease the FPR as required and perform higher-order computational integration not currently possible with the wearable sensor or mobile device alone.

Finally, this set-up will allow the development of clinical decision support systems (CDSS) that offers an important tool to address the challenge of medical staff shortages in LMICs. Early warning scores developed for high-income countries typically do not translate successfully to low-resource clinical settings. Active learning is a useful approach to address fairly sizeable unlabelled data sets, such as incoming patient data, which in turn may be used to develop robust machine learning models for the early prediction of patient deterioration. CDSS, like remote health-monitoring systems for rural handpumps, can support multiple stakeholders as clinicians have better information from continuous monitoring to enhance treatment plans while patients can take more ownership of their condition.

8.3 Conclusion

A lack of reliable data streams in rural settings is currently one of the most significant barriers that prevent transparent monitoring of vital infrastructure and patient health in limited-resource settings. This thesis aimed to demonstrate that novel data streams in a limited-resource setting, coupled with advanced machine learning methods can transform rural infrastructure to reliable resources through continuous monitoring.

This thesis presents a novel approach to remote health-monitoring by distributing the inference between the sensor and the cloud to minimise in-situ computational and power requirements as well as data transmission while improving the overall fidelity of the system. This approach was demonstrated by retrofitting sensors, consisting of an accelerometer, 8-bit microprocessor, and battery, inside the handles of traditional surface-mounted hand-pumps in rural Kenya. The approach was easily adapted to other applications, such as monitoring of patient vital signs to detect deterioration.

Further work and validation are required to test the real-time performance with two-way communication; however, the work presented in this thesis is a pivotal contribution to ensure that positive impacts of machine learning research impact the lives of even the most rural communities in limited-resource settings.

Bibliography

- [1] William F. Fox and Salena Porca. Investing in Rural Infrastructure. *International Regional Science Review*, 24(1):103–133, 2001.
- [2] Saul Arlosoroff, Gerhard Tschannerl, David Grey, William Journey, Andrew Karp, Otto Langeneffer, and Robert Roche. *Community water supply: The handpump option*. World Bank, Washington, D.C., US, 1987.
- [3] WHO/UNICEF Joint Water Supply and Sanitation Monitoring Programme. *Progress on drinking water and sanitation: 2014 update*. World Health Organisation, 2014.
- [4] Rob Hope, Tim Foster, and Patrick Thomson. Reducing risks to rural water security in Africa. *Ambio*, 41(7):773–776, 2012.
- [5] Jess MacArthur. Handpump standardisation in sub-Saharan Africa. *RWSN, St. Gallen, Switzerland*, 2015.
- [6] Michael Fisher, Katherine Shields, Terence Chan, Elizabeth Christenson, Ryan Cronk, Hannah Leker, Destina Samani, Patrick Apoya, Alexandra Lutz, and Jamie Bartram. Understanding hand-pump sustainability: Determinants of rural water source functionality in the Greater AframPlains region of Ghana. *Water Resource Research*, 51:8431–8449, 2015.
- [7] Erich Baumann. May-day! May-day! Our handpumps are not working! *RWSN Perspectives*, RWSN/A/200(1), 2009.
- [8] WHO and UNICEF. Progress on Drinking Water, Sanitation and Hygiene: 2017 update and SDG baselines. Technical Report February, WHO, UNICEF, Geneva, Switzerland, 2017.
- [9] John. Briscoe and David M De Ferranti. *Water for rural communities : helping people help themselves*. World Bank,, Washington, D.C., 1988.
- [10] R C Carter, S F Tyrrel, and P Howsam. The Impact and Sustainability of Community Water Supply and Sanitation Programmes in Developing Countries. *Water and Environment Journal*, 13(4):292–296, 8 1999.
- [11] Anthony Churchill, David de Ferranti, Robert Roche, Carolyn Tager, Alan Walters, and Anthony Yazer. *Rural water supply and sanitation: Time for a change*, volume 18. 1987.
- [12] Peter Harvey and R Reed. Rural water supply in Africa: Building blocks for handpump sustainability. *Water, Engineering, and Development Centre (WEDC)*, pages 1–316, 2004.
- [13] WHO. How does safe water impact global health? *Safe water, Better health*, 2008.
- [14] H.C. Bonsor, N. Oates, P.J. Chilton, Carter R.C., V. Casey, A.M. MacDonald, R. Calow, R. Alowo, P. Wilson, M. Tumutungire, and M. Bennie. UPGRO A Hidden Crisis: strengthening the evidence base on the sustainability of rural groundwater supplies – results from a pilot study in Uganda. Technical report, 2015.
- [15] REACH. Country diagnostic report: Kenya. Technical report, REACH, University of Oxford, Oxford, UK, 2015.
- [16] Pradeep K. Naik. Water crisis in Africa: myth or reality? *International Journal of Water Resources Development*, 0627(November):1–14, 2016.
- [17] SDSN TReNDS. Handpump Data Improves Water Access. Technical report, TReNDS, 2018.
- [18] Heloise Greeff. Can a rural handpump tell you it’s not well? *UNICEF Innovation Stories*, 2016.
- [19] Johanna Koehler, Patrick Thomson, and Rob Hope. Pump-priming payments for sustainable water services in rural Africa. *Water Development*, 74:397–411, 2015.

- [20] UN Water. Post-2015 global goal for water. Technical report, 2014.
- [21] UN Water. Integrated Monitoring Guide for SDG 6: Targets and global indicators. Technical Report April, UN Water, 2016.
- [22] UN Water. Transboundary Waters: Sharing Waters, Sharing Responsibilities. Technical report, 2008.
- [23] S Parry-Jones, R Reed, and B H Skinner. Sustainable Handpump Projects in Africa: A Literature Review. Technical report, DFID, UK, 2001.
- [24] World Bank. Hardware Elements of Handpump Technology : Field Experience and views on the future. Technical report, 1994.
- [25] Tim Foster and Rob Hope. A multi-decadal and social-ecological systems analysis of community waterpoint payment behaviours in rural Kenya. *Journal of Rural Studies*, 47:85–96, 2016.
- [26] Anthony Oyo. Spare Part Supplies for Handpumps in Africa Success Factors for Sustainability Rural Water Supply Series Creating Sustainable Supply Chains. (October), 2006.
- [27] PA Harvey, BH Skinner, and R Reed. Sustainable Handpump Projects in Africa: Report on Fieldwork in Zambia, April 18–May 4 2002. Technical report, 2002.
- [28] Richard Carter and Ian Ross. Beyond “Functionality” of Handpump-Supplied Rural Water Services in Developing Countries. *Waterlines*, 35(1), 2016.
- [29] M Adank, T C Kumasi, T L Chimbar, J Atengdem, B D Agbemor, N Dickinson, and E Abbey. The State of Handpump Water Services in Ghana: Findings from Three Districts. *Sustainable Water and Sanitation Services for All in a Fast Changing World*, pages 1–7, 2014.
- [30] Vida Affum Duti. Tracking Functionality for Sustainability - A Key to Sustainable Services. In *Annual Review Conference of the Community Water and Sanitation Agency*, pages 1–11, 2012.
- [31] J. Crocker and J. Bartram. Comparison and cost analysis of drinking water quality monitoring requirements versus practice in seven developing countries. *Int. J. Environ. Res. Public Health*, 11(7):7333–7346, 2014.
- [32] Jenny T Grönwall, Martin Mulenga, and Gordon Mcgranahan. *Groundwater, self-supply and poor urban dwellers A review with case studies of Bangalore and Lusaka Human Settlements Working Paper Series Water and Sanitation -26*. Number November. 2010.
- [33] Jennifer Sara and Travis Katz. Making Rural Water Supply Sustainable: Report on the Impact of Project Rules. Technical report, UNDP-World Bank Water Sanitation Programme, 2005.
- [34] Jess Macarthur. Handpump Standardisation in Sub-Saharan Africa Seeking a champion. *RWSN*, 1, 2015.
- [35] P.A. Harvey, P.N. Ikumi, and D.K. Mutethia. Sustainable Handpump Projects in Africa: Report on Fieldwork in Kenya. Technical report, WEDC, Loughborough University, UK, 2003.
- [36] J. Cook, P. Kimuyu, and D. Whittington. The costs of coping with poor water supply in rural Kenya. *Water Resources Research*, 52(2):841–859, 2016.
- [37] Patrick Thomson, David Bradley, Adamson Katilu, Jacob Katuva, Michelle Lanzoni, Johanna Koehler, and Rob Hope. Rainfall and groundwater use in rural Kenya. *Science of the Total Environment*, 649:722–730, 2019.
- [38] P. Pavelic, M. Giordano, B. Keraita, V. Ramesh, and T. Rao. Groundwater availability and use in Sub-Saharan Africa: A review of 15 countries. Technical report, International Water Management Institute, 2012.
- [39] Maxwell Barasa, Brighid Dochartaigh, Emily Crane, and Kirsty Upton. Africa Groundwater Atlas: Hydrogeology of Kenya, 2020.
- [40] Albert Mumma, Michael Lane, Edward Kairu, Albert Tuinhof, and Rafik Hirji. Kenya Groundwater Governance: Case Study. Technical Report June, Water Partnership Programme, 2011.

- [41] S.G. Furey and K. Danert. Sustainable Groundwater Development: use , protect and enhance. *Rural Water Supply Network*, 1:1–12, 2012.
- [42] Tim Foster, Juliet Willetts, Mike Lane, Patrick Thomson, Jacob Katuva, and Rob Hope. Risk factors associated with rural water supply failure: A 30-year retrospective study of handpumps on the south coast of Kenya. *Science of the Total Environment*, 626:156–164, 2018.
- [43] Eric Baumann and John Keen. Afridev Handpump Specification. Technical report, SKAT - RWSN, RWSN/SKAT, St. Gallen, CH, 2007.
- [44] K. Erpf. India Mark Handpump Specifications (Revision 2-2007). Technical report, RWSN/SKAT, St. Gallen, CH, 2007.
- [45] UNICEF. WASH technology information packages. *Unicef*, 2010.
- [46] World Bank. Sustaining Rural Water Systems: The Case of Mali. *World Bank Operations Evaluation Department*, 154, 1997.
- [47] K. Erpf. Afridev: Installation and Maintenance Manual for the Afridev Handpump (Revision 2-2007) v.2. Technical report, RWSN/SKAT, St. Gallen, CH, 2007.
- [48] RWSN. No.6 pump specifications. Technical report, RWSN/Skat , St Gallen, Switzerland, 2009.
- [49] E. Baumann. *Water Lifting*, volume 7. SKAT, St. Gallen: CH, 2000.
- [50] P Wurzel. Drilling Boreholes for Handpumps. 2001.
- [51] Marije van den Broek and Julia Brown. Blueprint for breakdown? Community Based Management of rural groundwater in Uganda. *Geoforum*, 67:51–63, 2015.
- [52] E. Kleemeier and J. Narkevic. Private Operator Models for Community Water Supply. Technical report, World Bank Water and Sanitation Program, 2010.
- [53] P.A. Harvey and S.M. Kayaga. Report on Fieldwork in South Africa. Technical report, DFID, UK, 2003.
- [54] Emily Kumpel, Rachel Peletz, Mateyo Bonham, Annette Fay, Alicea Cock-Esteb, and Ranjiv Khush. When are mobile phones useful for water quality data collection? An analysis of data flows and ICT applications among regulated monitoring institutions in Sub-Saharan Africa. *International Journal of Environmental Research and Public Health*, 12(9):10846–10860, 2015.
- [55] Vincent Casey, Lawrence Brown, Jacob D Carpenter, Jacinta Nekesa, and Bonny Etti. The role of handpump corrosion in the contamination and failure of rural water supplies. *Waterlines*, 35(1):59–77, 2016.
- [56] Paul Wilson, Helen Bonsor, Alan MacDonald, Luke Whaley, Richard Carter, and Vincent Casey. UPGRO Hidden Crisis Research Consortium: Initial project approach for assessing rural water supply functionality and levels of performance. *British Geological Survey Open Report*, OR/16/044:35, 2017.
- [57] Heloise Greeff. Transfer Report: A predictive and scalable health-monitoring system for handpumps. Technical report, University of Oxford, 2016.
- [58] O. Langenegger. Groundwater Quality: An Important Factor for Selecting Handpumps. Technical report, World Bank, World Bank, Washington, DC, 1987.
- [59] A M MacDonald and H C Bonsor. An initial estimate of depth to groundwater across Africa. Technical report, British Geological Survey, 2011.
- [60] Tirusew Asefa, Mariush W. Kemblowski, Gilberto Urroz, Mac McKee, and Abedalrazq Khalil. Support vectors-based groundwater head observation networks design. *Water Resources Research*, 40(11), 11 2004.
- [61] Heesung Yoon, Seong-Chun Jun, Yunjung Hyun, Gwang-Ok Bae, and Kang-Kun Lee. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. *Journal of Hydrology*, 396(1-2):128–138, 1 2011.

- [62] Lily Tapak, Ali Reza Rahmani, and Abbas Moghimbeigi. Prediction the groundwater level of hamadan-bahar plain, west of iran using support vector machines. *Journal of research in health sciences*, 14(1):81–6, 1 2014.
- [63] Mohammed Ahmed, Abdel Bagi, Mahmoud Hassan Onsa, and Yasir Ismail Adam. Investigation on the Problems of Handpumps in Sudan. *Sudan Engineering Society Journal*, 56(55):1–8, 2010.
- [64] Raphael Nampusuor and Stein Mathisen. COWAP Field Paper 27: Report on The Performance of Afridev and Nira Handpumps on the Upper Regions Community Water Project (COWAP) Ghana. Technical report, GWSC/CIDA: Canada, 2001.
- [65] Philip Hankin. The afridev handpump: problems and solutions. *27th WEDC Conference*, pages 121–124, 2001.
- [66] GSMA. The Mobile Economy: Sub-Saharan Africa. Technical report, 2019.
- [67] K. Fogelberg. Filling the knowledge gap: Monitoring post-construction water and sanitation sustainability. *Waterlines*, 29(3):220–235, 2010.
- [68] Evan A. Thomas, Christina K. Barstow, Ghislaine Rosa, Fiona Majorin, and Thomas Clasen. Use of remotely reporting electronic sensors for assessing use of water filters and cookstoves in Rwanda. *Environmental Science and Technology*, 47(23):13602–13610, 2013.
- [69] Rob Hope, Tim Foster, Alex Money, and Michael Rouse. Harnessing mobile communications innovations for water security. *Global Policy*, 3(4):433–442, 2012.
- [70] Avri Doria, Maria Kristina Udén, and Durga Prasad Pandey. Providing connectivity to the Saami nomadic community. In *2nd International Conference on Open Collaborative Design for Sustainable Innovation*, pages 1–8, 2002.
- [71] Alex Pentland, Richard Fletcher, and Amir Hasson. Daknet: Rethinking connectivity in developing nations. *Computer*, 37(1):78–83, 2004.
- [72] Christian Raffelsberger and Hermann Hellwagner. A hybrid MANET-DTN routing scheme for emergency response scenarios. In *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 505–510, San Diego, CA, 2013.
- [73] John Whitbeck and Vania Conan. HYMAD: Hybrid DTN-MANET routing for dense and highly dynamic wireless networks. *Computer communications*, 33(12):1483–1492, 2010.
- [74] David A Clifton and Lionel Tarassenko. Novelty Detection in Jet Engine Vibration Spectra. *The International Journal of Condition Monitoring*, 5(2):2–7, 2015.
- [75] Lei Clifton, David A Clifton, Marco A F Pimentel, Peter J Watkinson, and Lionel Tarassenko. Predictive Monitoring of Mobile Patients by Combining Observations With Data From Wearable Sensors. *IEEE Journal of Biomedical and Health Informatics*, 18(3):722–730, 2014.
- [76] Gracon H.E.L. De Lima, Lenardo C. E Silva, and Pedro F.R. Neto. WSN as a tool for supporting agriculture in the precision irrigation. *6th International Conference on Networking and Services, ICNS 2010, Includes LMPCNA 2010; INTENSIVE 2010*, pages 137–142, 2010.
- [77] Ning Wang, Naiqian Zhang, and Maohua Wang. Wireless sensors in agriculture and food industry - Recent development and future perspective. *Computers and Electronics in Agriculture*, 50(1):1–14, 2006.
- [78] William Merrill. Where is the return on investment on investment in wireless sensor networks? *IEEE Wireless Communications*, (February):4–6, 2010.
- [79] J Behar, A Guazzi, Jorge J., M A Maraci, S Laranjeira, T Papastylianou, P Thomson, G D Clifford, and R A Hope. Software Architecture to Monitor Handpump Performance in Rural Kenya. *Social Implications of Computers in Developing Countries: Proceedings of the 12th International Conference of IFIP working group 9.4, Ocho Rios, Jamaica*, pages 978–991, 2013.
- [80] Patrick Thomson, Rob Hope, and Tim Foster. GSM-enabled remote monitoring of rural handpumps: a proof-of-concept study. *Journal of Hydroinformatics*, 14(4):829–839, 2012.

- [81] P. Thomson, R. Hope, and T. Foster. Is silence golden? Of mobiles, monitoring, and rural water supplies. *Waterlines*, 31(4):280–292, 2012.
- [82] T. Papastylianou, J. Behar, A. Guazzi, J. Jorge, S. Laranjeira, M.A. Maraci, G.D. Clifford, R.A. Hope, and P. Thomson. Smart handpumps: technical aspects of a one-year field trial in rural Kenya. *Appropriate Healthcare Technologies for Low Resource Settings (AHT 2014)*, (1):9–9, 2014.
- [83] F.E. Colchester, H. Greeff, P. Thomson, R. Hope, and D.A. Clifton. Smart Handpumps: A Preliminary Data Analysis. In *Appropriate Healthcare Technologies for Low Resource Settings*, pages 1–4, 2014.
- [84] Patrick Thomson, Heloise Greeff, Farah Colchester, David Clifton, and Rob Hope. Distributed Monitoring of Shallow Aquifer Level using Community Handpumps. In *World Water Week*, number August, pages 1–3, Stockholm, Sweden, 2015.
- [85] F.E. Colchester, H. Greeff (Marais), P. Thomson, R. Hope, and D.A. Clifton. Accidental Infrastructure for Groundwater Monitoring in Africa. *Environmental Modelling and Software*, 91:241–250, 2017.
- [86] H. Greeff, A. Manandhar, and P. Thomson. Daily handpump accelerometer data and borehole water level data, Kwale County, Kenya (Dataset), British Geological Survey, 2020.
- [87] P. Thomson, A. Manandhar, and H. Greeff. Machine Learning for Aquifer Monitoring using Handpump Vibration Data. In *45th International Association of Hydrologist Congress*, Seoul, South Korea, 2018.
- [88] A. Manandhar, H. Greeff, P. Thomson, R. Hope, and D.A. Clifton. Groundwater Monitoring Using Handpump Vibration Data for Rural Africa. In *2019 Conference on Neural Information Processing Systems, AI for Social Good Workshop*, pages 1–4, 2019.
- [89] Achut Manandhar, Heloise Greeff, Patrick Thomson, Rob Hope, and David A. Clifton. Shallow aquifer monitoring using handpump vibration data. *Journal of Hydrology X*, 8(May), 2020.
- [90] P. Sharma, A. Manandhar, P. Thomson, J. Katuva, R. Hope, and D.A. Clifton. Combining Multi-Modal Statistics for Welfare Prediction Using Deep Learning. *Sustainability*, 11(22):6312, 2019.
- [91] Akvo Data Platform. Akvo Flow.
- [92] E. H. Jürrens, A. Bröring, and S. Jirka. A human sensor web for water availability monitoring. In *Proceedings of OneSpace*, 2009.
- [93] mWater. mWater Project Highlights.
- [94] Mtega. Maji Matone, 2013.
- [95] IRC-Wash. Monitoring the Functionality of Rural Water Sources Using Mobile Phone Technology (M4W). Technical report, Triple-S, 2013.
- [96] Welldone. MoMo Technology.
- [97] Akvo. Prepaid Handpumps Kenya, 2013.
- [98] Corey Nagel, Jack Beach, Chantal Iribagiza, and Evan A. Thomas. Evaluating Cellular Instrumentation on Rural Handpumps to Improve Service Delivery-A Longitudinal Study in Rural Rwanda. *Environmental Science and Technology*, 49(24):14292–14300, 2015.
- [99] EWaterPay. Smart Technology.
- [100] Charity:Water. Our Work.
- [101] Appropriate Technology for Water & Sanitation. Recommendations for the sustainability and scalability of the metered hand pump in Kamwenge district , Uganda. Technical Report October, TAF Assessment Summary: Technology Recommendations, 2015.
- [102] MSABI. Pump for Life Program.

- [103] J. Pearce, N. Dickinson, and K. Welle. Technology, data, and people: opportunities and pitfalls of using ICT to monitor sustainable WASH service delivery. In T. Shouten and S. Smits, editors, *Infrastructure to Services: Trends in Monitoring Sustainable Water, Sanitation and Hygiene Services*, pages 85–108. Practical Action Publishing, Rugby, UK, 2014.
- [104] N. Dickinson and K. Bostoen. Using ICT for monitoring rural water services: from data to action. Technical report, , Triple-S Working Paper 4. Geneva, Switzerland: IRC Water and Sanitation, 2013.
- [105] Evan Thomas, Zdenek Zumr, Jodi Graf, Carson Wick, James McCellan, Zachary Imam, Christina Barstow, Kelly Spiller, and Michael Fleming. Remotely Accessible Instrumented Monitoring of Global Development Programs: Technology Development and Validation. *Sustainability*, 5(8):3288–3301, 2013.
- [106] Evan Alexander Thomas. Improving global impact : How the integration of remotely reporting sensors in water projects may demonstrate and enhance positive change. *GWF Discussion Paper 1349, Global Water Forum*, 12, 2013.
- [107] Daniel L. Wilson, Jeremy R. Coyle, and Evan A. Thomas. Ensemble machine learning and forecasting can achieve 99% uptime for rural handpumps. *PLoS ONE*, 12(11):1–13, 2017.
- [108] Zaid Tashman, Christoph Gorder, Sonali Parthasarathy, Mohamad M. Nasr-Azadani, and Rachel Webre. Anomaly detection system for water networks in northern ethiopia using bayesian inference. *Sustainability*, 12(7):1–16, 2020.
- [109] P. J. C. Nel, Marthinus J. Booysen, and B. Van Der Merwe. ICT-enabled solutions for smart management of water supply in Africa. In *Proceeding or the 1st International Conference on the Use of Mobile ICT in Africa 2014*, pages 40–44, 2014.
- [110] D.J. MacAllister, A.M. MacDonald, S. Kebede, S. Godfrey, and R. Calow. Comparative performance of rural water supplies during drought. *Nature Communications*, 11(1099):1–13, 2020.
- [111] H. Sohn, C. R. Farrar, F. M. Hemez, D. D. Shunk, D. W. Stinemat, and B. R. Nadler. *A Review of Structural Health Monitoring Literature: 1996-2001*. Los Alamos National Laboratory, USA, 2003.
- [112] S.W. Doebling, C.R. Farrar, M.B. Prime, and D.W. Shevitz. Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: a literature review. Technical report, Technical Report LA-13070-MS, Los Alamos National Laboratory, Los Alamos, NM, 1996.
- [113] M. E.Beth Smith, Joseph C. Chiovaro, Maya O’Neil, Devan Kansagara, Ana R. Quiñones, Michele Freeman, Makalapua L. Motu’apuaka, and Christopher G. Slatore. Early warning system scores for clinical deterioration in hospitalized patients: A systematic review. *Annals of the American Thoracic Society*, 11(9):1454–1465, 2014.
- [114] Charles R. Farrar and Keith Worden. An introduction to structural health monitoring. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1851):303–315, 2007.
- [115] Nathan Cooper, Andrew Swan, and David Townend. A confluence of new technology and the right to water: Experience and potential from South Africa’s constitution and commons. *Ethics and Information Technology*, 16(2):119–134, 2014.
- [116] Charles R. Farrar and Nick A.J. Lieven. Damage prognosis: The future of structural health monitoring. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1851):623–632, 2007.
- [117] D. E. Bently and C. T. Hatch. *Fundamentals of rotating machinery diagnostics*. New York, NY: ASME Press, 2003.
- [118] Iman Talebinejad, Chad Fischer, and Farhad Ansari. Numerical evaluation of vibration-based methods for damage assessment of cable-stayed bridges. *Computer-Aided Civil and Infrastructure Engineering*, 26(3):239–251, 2011.
- [119] Heloise Greeff. *MSc. Thesis: Smart Water Systems*. PhD thesis, University of Oxford, 2014.

- [120] L Tarassenko, D A Clifton, P R Bannister, S King, and D King. Novelty Detection. *Encyclopaedia of Structural Health Monitoring*, pages 653–675, 2009.
- [121] Yu Jin Hong, Ig Jae Kim, Sang Chul Ahn, and Hyoung Gon Kim. Mobile health monitoring system based on activity recognition using accelerometer. *Simulation Modelling Practice and Theory*, 18(4):446–455, 2010.
- [122] A Rytter. Vibrational Based Inspection of Civil Engineering Structures. Technical report, Department of Building Technology and Structural Engineering, Aalborg Universitet, Denmark, 1993.
- [123] A.H. Christer. Innovative decision making. In K.C. Brown and D.J. White, editors, *Proceedings of the NATO Conference on Role and Effectiveness of Theory of Decision Practice*, pages 368–377. Hodder and Stoughton, UK, 1976.
- [124] A H Christer and W M Waller. Reducing Production Downtime Using Delay-Time Analysis. *Journal of the Operational Research Society*, 35(6):499–512, 1984.
- [125] R. D. Baker and W. Wang. Estimating the delay-time distribution of faults in repairable machinery from failure data. *IMA Journal of Management Mathematics*, 3(4):259–281, 1991.
- [126] R D Baker and W Wang. Developing and Testing the Delay-Time Model. *Journal of the Operational Research Society*, 44(4):361–374, 4 1993.
- [127] A H Christer, W Wang, R D Baker, and J Sharp. Modelling maintenance practice of production plant using the delay-time concept. *IMA Journal of Management Mathematics*, 6(1):67–83, 1 1995.
- [128] A. H. Christer. Developments in delay time analysis for modelling plant maintenance. *Journal of the Operational Research Society*, 50(11):1120–1137, 1999.
- [129] Wenbin Wang. A two-stage prognosis model in condition based maintenance. *European Journal of Operational Research*, 182(3):1177–1187, 2007.
- [130] Robert M. Vandawaker, David R. Jacques, and Jason K. Freels. Impact of prognostic uncertainty in system health monitoring. *International Journal of Prognostics and Health Management*, 6(Lcc):1–13, 2015.
- [131] Iman Izadi, Sirish L. Shah, David S. Shook, and Tongwen Chen. An Introduction to Alarm Analysis and Design. In *Proceedings of the 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes*, volume 42, pages 645–650. IFAC, 2009.
- [132] Wayne B. Nelson. *Accelerated testing: statistical models, test plans, and data analysis*. John Wiley & Sons, 2009.
- [133] Neville Stanton. Alarm Information in Fault Diagnosis. *Condition Monitoring for Fault Diagnosis, IEE Colloquium on*, pages 1–9, 1987.
- [134] Antonio Ginart, Irtaza Barlas, Jonathan Goldin, and Jordan Lewis Dorrity. Automated feature selection for embeddable Prognostic and Health Monitoring (PHM) architectures. *AUTOTESTCON (Proceedings)*, (404):195–201, 2006.
- [135] Gang Niu and Bo Suk Yang. Intelligent condition monitoring and prognostics system based on data-fusion strategy. *Expert Systems with Applications*, 37(12):8831–8840, 2010.
- [136] C. S. Byington, M. Watson, P. Kalgren, and R. Safa-Bakhsh. Metrics evaluation and tool development for health and usage monitoring system technology. In *Third International Conference on Health Usage Monitoring*, number Forum 59, pages 35–44, Phoenix, Arizona, 2003.
- [137] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(1):1157–1182, 2003.
- [138] Nikolai Helwig, Eliseo Pignanelli, and Andreas Schutze. Condition monitoring of a complex hydraulic system using multivariate statistics. In *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, pages 210–215, 2015.
- [139] K. Altun, B. Barshan, and O. Tunçel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10):3605–3620, 2010.

- [140] P. Reimann and A. Schütze. Fire detection in coal mines based on semiconductor gas sensors. *Sensor Review*, 32(1):47–58, 2012.
- [141] Robert T Olszewski. Generalized feature extraction for structural pattern recognition in time-series data. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 2001.
- [142] Fabian Morchen. Time series feature extraction for data mining using DWT and DFT. Technical report, 2003.
- [143] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.
- [144] Ruqiang Yan, Robert X. Gao, and Xuefeng Chen. Wavelets for fault diagnosis of rotary machines: A review with applications. *Signal Processing*, 96(PART A):1–15, 2014.
- [145] T. I. El-Wardany, D. Gao, and M. A. Elbestawi. Tool condition monitoring in drilling using vibration signature analysis. *International Journal of Machine Tools and Manufacture*, 36(6):687–711, 1996.
- [146] Chuangwen Xu, Zhe Liu, and Wencui Luo. A frequency band energy analysis of vibration signals for tool condition monitoring. *2009 International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2009*, 1:385–388, 2009.
- [147] P. D. O. Conceicao Junior, M. Marchi, C. H. Rossinoli Martins, D. D’addona, P. R. de Aguiar, and E. C. Bianchi. Spectral estimation of vibration signal for monitoring the wear of single-point dresser. *Rev. Mater.*, 4:827–840, 21.
- [148] Seetaram Maurya, Vikas Singh, and Nishchal Verma. Condition Monitoring of Machines using Fused Features from EMD based Local Energy with DNN. *IEEE Sensors Journal*, PP(7):1–1, 2019.
- [149] F. Al-Badour, M. Sunar, and L. Cheded. Vibration analysis of rotating machinery using time–frequency analysis and wavelet techniques. *Mechanical Systems and Signal Processing*, 25(6):2083–2101, 2011.
- [150] Jihong Yan and Jay Lee. Degradation Assessment and Fault Modes Classification Using Logistic Regression. *Journal of Manufacturing Science and Engineering*, 127(4):912–914, 7 2004.
- [151] M Straczekiewicz, P Czop, and T Barszcz. Supervised and unsupervised learning process in damage classification of rolling element bearings. *Diagnostyka*, 17(2):71–80, 2016.
- [152] A Malhi and R X Gao. PCA-based feature selection scheme for machine defect classification. *IEEE Transactions on Instrumentation and Measurement*, 53(6):1517–1525, 2004.
- [153] Seunghye Park, Jong-Jae Lee, Chung-Bang Yun, and Daniel J Inman. Electro-Mechanical Impedance-Based Wireless Structural Health Monitoring Using PCA-Data Compression and k-means Clustering Algorithms. *Journal of Intelligent Material Systems and Structures*, 19(4):509–520, 5 2007.
- [154] F Tong, S K Tso, and M Y Y Hung. Impact-acoustics-based health monitoring of tile-wall bonding integrity using principal component analysis. *Journal of Sound and Vibration*, 294(1):329–340, 2006.
- [155] J.-C. Golival, P. De Boe, A. Yan, and G. Kerschen. Structural damage detection based on PCA of vibration measurements. In *58th Meeting of the Society for Machinery, Failure Prevention Tech*, 2004.
- [156] N Saravanan, V N S Kumar Siddabattuni, and K I Ramachandran. A comparative study on classification of features by SVM and PSVM extracted using Morlet wavelet for fault diagnosis of spur bevel gear box. *Expert Systems with Applications*, 35(3):1351–1366, 2008.
- [157] P Konar and P Chattopadhyay. Bearing fault detection of induction motor using wavelet and Support Vector Machines (SVMs). *Applied Soft Computing*, 11(6):4203–4211, 2011.
- [158] A Soualhi, K Medjaher, and N Zerhouni. Bearing Health Monitoring Based on Hilbert–Huang Transform, Support Vector Machine, and Regression. *IEEE Transactions on Instrumentation and Measurement*, 64(1):52–62, 1 2015.
- [159] A. Bulut, A.K. Singh, P. Shin, T. Fountain, H. Jasso, L. Yan, and A. Elgamal. Real-time nondestructive structural health monitoring using support vector machines and wavelets. In *Advanced Sensor*

- Technologies for Nondestructive Evaluation and Structural Health Monitoring*, pages 180–189, San Diego, CA, USA, 2006.
- [160] F.L. Greitzer, L.J. Kangas, K.M. Terrones, M.A. Maynard, B.W. Wilson, R.A. Pawlowski, D.R. Sisk, and N.B. Brown. Gas turbine engine health monitoring and prognostics. In *International Society of Logistics (SOLE) 1999 Symposium*, pages 1–7, Las Vegas, NA, USA, 1999.
 - [161] D H Wang and S L Huang. Health Monitoring and Diagnosis for Flexible Structures with PVDF Piezoelectric Film Sensor Array. *Journal of Intelligent Material Systems and Structures*, 11(6):482–491, 2000.
 - [162] P. Jayaswal, S. Verma, and A. Wadhwani. Application of ANN, fuzzy logic and wavelet transform in machine fault diagnosis using vibration signal analysis. *Journal of Quality in Maintenance Engineering*, 16(2):190–213, 2010.
 - [163] Roque A Osornio-Rios, Juan Pablo Amezcuita-Sanchez, Rene J Romero-Troncoso, and Arturo Garcia-Perez. MUSIC-ANN Analysis for Locating Structural Damages in a Truss-Type Structure by Means of Vibrations. *Computer-Aided Civil and Infrastructure Engineering*, 27(9):687–698, 2012.
 - [164] A C Neves, I González, J Leander, and R Karoumi. Structural health monitoring of bridges: a model-free ANN-based approach to damage detection. *Journal of Civil Structural Health Monitoring*, 7(5):689–702, 2017.
 - [165] C.R. Farrar and S.W. Doebling. Damage detection II: field applications to large structures. In J.M.M. Silva and N.M.M. Maia, editors, *Modal Analysis and Testing, Nato Science Series*. Dordrecht, Netherlands: Kluwer Academic Publishers, 1999.
 - [166] Marco A F Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.
 - [167] David A Clifton. *Novelty Detection with Extreme Value Theory in Jet Engine Vibration Data*. PhD thesis, University of Oxford, 2009.
 - [168] T. Heyns, P. S. Heyns, and J. P. De Villiers. Combining synchronous averaging with a Gaussian mixture model novelty detection scheme for vibration-based condition monitoring of a gearbox. *Mechanical Systems and Signal Processing*, 32:200–215, 2012.
 - [169] C T Yiakopoulos, K C Gryllias, and I A Antoniadis. Rolling element bearing fault detection in industrial environments based on a K-means clustering approach. *Expert Systems with Applications*, 38(3):2888–2911, 2011.
 - [170] Alberto Diez, Nguyen Lu Dang Khoa, Mehrisadat Makki Alamdari, Yang Wang, Fang Chen, and Peter Runcie. A clustering approach for structural health monitoring on bridges. *Journal of Civil Structural Health Monitoring*, 6(3):429–445, 2016.
 - [171] Mehrisadat Makki Alamdari, Thierry Rakotoarivelo, and Nguyen Lu Dang Khoa. A spectral-based clustering for structural health monitoring of the Sydney Harbour Bridge. *Mechanical Systems and Signal Processing*, 87:384–400, 2017.
 - [172] J. Tian, M.H. Azarian, and M. Pecht. Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm. In *European Conference of the Prognostics and Health Management Society*, Turin, Italy, 2020.
 - [173] Varun Mehra, Rajeev Ram, and Claudio Vergara. A novel application of machine learning techniques for activity-based load disaggregation in rural off-grid, isolated solar systems. *GHTC 2016 - IEEE Global Humanitarian Technology Conference: Technology for the Benefit of Humanity, Conference Proceedings*, pages 372–378, 2016.
 - [174] Amin Jahromi Torabi, Meng Joo Er, Xiang Li, Beng Siong Lim, and Gan Oon Peen. Application of clustering methods for online tool condition monitoring and fault diagnosis in high-speed milling processes. *IEEE Systems Journal*, 10(2):721–732, 2016.
 - [175] E. Pazouki and S. Choi. Fault diagnosis and condition monitoring of bearing using multisensory approach based fuzzy-logic clustering. *Proceedings - 2015 IEEE International Electric Machines and Drives Conference, IEMDC 2015*, pages 1412–1418, 2016.

- [176] Mikko Perttunen, Oleksiy Mazhelis, Fengyu Cong, Mikko Kauppila, Teemu Leppänen, Jouni Kantola, Jussi Collin, Susanna Pirttikangas, Janne Haverinen, Tapani Ristaniemi, and Jukka Riekk. Distributed Road Surface Condition Monitoring Using Mobile Phones. In Ching-Hsien Hsu, Laurence T Yang, Jianhua Ma, and Chunsheng Zhu, editors, *Ubiquitous Intelligence and Computing*, pages 64–78, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.
- [177] Kush R. Varshney, George H. Chen, Brian Abelson, Kendall Nowocin, Vivek Sakhrani, Ling Xu, and Brian L. Spatocco. Targeting Villages for Rural Development Using Satellite Image Analysis. *Big Data*, 3(1):41–53, 2015.
- [178] Felan Carlo C. Garcia, Alvin E. Retamar, and Joven C. Javier. A real time urban flood monitoring system for metro Manila. In *TENCON 2015 - IEEE Region 10 Conference*, pages 1–5, Macao, 2015. IEEE.
- [179] M.R. Borchers, Y.M. Chang, K.L. Proudfoot, B.A. Wadsworth, A.E. Stone, and J.M. Bewley. Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. *Journal of Dairy Science*, 100(7):5664–5674, 2017.
- [180] Alexander G. Logan, Warren J. McIsaac, Andras Tisler, M. Jane Irvine, Allison Saunders, Andrea Dunai, Carlos A. Rizo, Denice S. Feig, Melinda Hamill, Mathieu Trudel, and Joseph A. Cafazzo. Mobile Phone-Based Remote Patient Monitoring System for Management of Hypertension in Diabetic Patients. *American Journal of Hypertension*, 20(9):942–948, 2007.
- [181] Kevin A. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press: Cambridge, US, 2012.
- [182] Irina Rish. An empirical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 3(22):41–46, 2001.
- [183] Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall, 1984.
- [184] Safavian Rasoul and Landgrebe David. A Survey of Decision Tree Classifier Methodology. *IEEE Transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.
- [185] Ian Goodfellow, Yoshua Bengio, and Aarib Courville. *Deep Learning*. MIT Press: Cambridge US, 2016.
- [186] Andrew Ng. CS229 Lecture notes Deep Learning. Technical report, Stanford, 2000.
- [187] Padraig Cunningham and Sarah Jane Delany. K-Nearest Neighbour Classifiers. *Multiple Classifier Systems*, 34(8):1–17, 2007.
- [188] Min Ling Zhang and Zhi Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.
- [189] E Boser, N Vapnik, Isabelle M Guyon, and T Bell Laboratories. Training Algorithm Margin for Optimal Classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.
- [190] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.
- [191] N Cristianini and J Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.
- [192] M.A. Aizerman, E.M. Braverman, and L.I. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–835, 1964.
- [193] N.J. Nilsson. *Learning machines: Foundations of Trainable Pattern Classifying Systems*. McGraw-Hill, 1965.
- [194] V Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.
- [195] V. Vapnik. *Statistical Learning Theory*. John Wiley, Hoboken, NJ, 1998.

- [196] Chih-Chung Chang and Chih-Jen Lin. LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.
- [197] E. P. Carden. Vibration Based Condition Monitoring: A Review. *Structural Health Monitoring*, 3(4):355–377, 12 2004.
- [198] Subhasis Nandi, Hamid A Toliyat, and Xiaodong Li. Condition Monitoring and Fault Diagnosis of Electrical Motors — A Review. *IEEE Transactions on Energy Conversion*, 20(4):719–729, 2005.
- [199] Steven W. Smith. Moving Average Filters. In *The Scientist and Engineer’s Guide to Digital Signal Processing*, pages 277–284. California Technical Publishing, US, 2nd edition, 1999.
- [200] Richard G. Lyons. *Understanding Digital Signal Processing*. Prentice Hall, Upper Saddle River, NJ, 3rd edition, 2011.
- [201] F.J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.
- [202] Alfredo Testa, Daniele Gallo, and Roberto Langella. On the Processing of Harmonics and Interharmonics: Using HanningWindow in Standard Framework. *IEEE Transactions on Power Delivery*, 19(1):28–34, 2004.
- [203] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2008.
- [204] J Sammon. A non-linear mapping for data structure analysis. *IEEE Transactions on Computers*, 18(5):401–409, 1969.
- [205] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag: Berlin, Germany, 2006.
- [206] R. R. Tenney and N. R. Sandell. Detection with distributed sensors. *IEEE Trans. Aerosp. Electron. Sys.*, 17:501–510, 1981.
- [207] Stanislav Funiak, Carlos Guestrin, and Mark Paskin. Distributed Inference in Dynamical Systems. In *NIPS*, 2006.
- [208] V.V. Veeravalli and P.K. Varshney. Distributed inference in wireless sensor. *Phil. Trans. R. Soc.*, A(370):100–117, 2012.
- [209] Pramod K. Varshney. *Distributed detection and data fusion*. Berlin, Germany: Springer, 1996.
- [210] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [211] Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. MIT Press: Cambridge US, 2012.
- [212] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
- [213] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. *Advances in Neural Information Processing Systems 26*, pages 431–439, 2013.
- [214] Y. Le Cun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.
- [215] Y. Le Cun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.
- [216] K Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(11):559–572, 1901.
- [217] Johan Fagerberg and Lars Kurkinen. mHealth and Home Monitoring. *Berg Insight: M2M Research Series*, 2016.
- [218] Comptroller and Audit General. Discharging older patients from hospital. *National Audit Office*, HC 18(May), 2016.

- [219] IMECHE. Remote Health Management: Reducing Bed Blocking in the NHS. 2017.
- [220] Billy Palmer, John Appleby, and Jonathan Spencer. Rural health care: A rapid review of the impact of rurality on the costs of delivering health care. Technical Report January, 2019.
- [221] World Bank. United Nations Population Division’s World Urbanization Prospects: Rural Population (% of total population). Technical report, 2018.
- [222] WHO. Density of nursing and midwifery personnel, 2019.
- [223] Asaf Bitton, Jocelyn Fifield, Hannah Ratcliffe, Ami Karlage, Hong Wang, Jeremy H Veillard, Dan Schwarz, and Lisa R Hirschhorn. Primary healthcare system performance in low-income and middle-income countries: a scoping review of the evidence from 2010 to 2017. *BMJ Global Health*, 4(8):1–13, 2019.
- [224] G.D. Clifford. E-Health in Low to Middle Income Countries. *J Med Eng Technol*, 40(7-8):336–341, 2016.
- [225] R. M.H. Schein, N. Hazday, M. Pena, B. H. Ruben, and C. L. Sprung. Clinical antecedents to in-hospital cardiopulmonary arrest. *Chest*, 98(6):1388–1392, 1990.
- [226] C. Franklin and J. Mathew. Developing strategies to prevent inhospital cardiac arrest: Analysing responses of physicians and nurses in the hours before the event. *Critical Care Medicine*, 22(2):244–247, 1994.
- [227] Andrew F. Smith and Jeremy Wood. Can some in-hospital cardio-respiratory arrests be prevented? A prospective survey. *Resuscitation*, 37(3):133–137, 1998.
- [228] L. Clifton, D. A. Clifton, Marco A. F. Pimentel, P. J. Watkinson, and L. Tarassenko. Gaussian processes for personalized e-health monitoring with wearable sensors. *IEEE Transactions on Biomedical Engineering*, 60(1):193–197, 2013.
- [229] D.R. Prytherch, G.B. Smith, P.E. Schmidt, and P.I. Featherstone. ViEWS–Towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation*, 81(8):932–7, 2010.
- [230] Gary B. Smith, David R. Prytherch, Paul Meredith, Paul E. Schmidt, and Peter I. Featherstone. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, 84(4):465–470, 2013.
- [231] Stuart Jarvis, Caroline Kovacs, Jim Briggs, Paul Meredith, Paul E. Schmidt, Peter I. Featherstone, David R. Prytherch, and Gary B. Smith. Aggregate National Early Warning Score (NEWS) values are more important than high scores for a single vital signs parameter for discriminating the risk of adverse outcomes. *Resuscitation*, 87(2):75–80, 2015.
- [232] C. P. Subbe, R. G. Davies, E. Williams, P. Rutherford, and L. Gemmell. Effect of introducing the modified early warning score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions. *Anaesthesia*, 58(8):797–802, 2003.
- [233] A. Hann. *Multi-parameter monitoring for early warning of patient deterioration*. PhD thesis, University of Oxford, 2008.
- [234] D. A. Clifton, S. Hugueny, and L. Tarassenko. Novelty detection with multivariate Extreme Value Theory, part I: A numerical approach to multimodal estimation. In *2009 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2009.
- [235] David Andrew Clifton, Samuel Hugueny, and Lionel Tarassenko. Novelty detection with multivariate extreme value statistics. *Journal of Signal Processing Systems*, 65(3):371–389, 2011.
- [236] Glen Wright Colopy, Stephen J. Roberts, and David A. Clifton. Bayesian Optimization of Personalized Models for Patient Vital-Sign Monitoring. *IEEE Journal of Biomedical and Health Informatics*, 22(2):301–310, 2018.
- [237] K. Yousef, M. R. Pinsky, M. A. DeVita, S. Sereika, and M. Hravnak. Characteristics of Patients With Cardiorespiratory Instability in a Step-down Unit. *American Journal of Critical Care*, 21(5):344–350, 2012.

- [238] MA DeVita, RS Braithwaite, R Mahidhara, S Stuart, M Foraida, RL Simmons, and MERIT. Use of medical emergency team responses to reduce hospital cardiopulmonary arrests. *Qual Saf Health Care*, 13(4):251–4, 2004.
- [239] D. A. Clifton, D. Wong, L. Clifton, S. Wilson, R. Way, R. Pullinger, and L. Tarassenko. A large-scale clinical validation of an integrated monitoring system in the emergency department. *IEEE Journal of Biomedical and Health Informatics*, 17(4):835–842, 2013.
- [240] Glen Wright Colopy, Marco A F Pimentel, Steven J Roberts, and David A Clifton. Bayesian Gaussian Processes for Identifying the Deteriorating Patient. *IEEE EMBC*, pages 5311–5314, 2016.
- [241] C.M. Bishop. Novelty detection and neural network validation. *IEE Proceedings - Vision, Image and Signal processing*, 141(4):217–222, 1994.
- [242] B.D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, 1996.
- [243] M.J. Desforges, P.J. Jacob, and J.E. Cooper. Applications of probability density estimation to the detection of abnormal conditions in engineering. *Proc Inst Mech Eng*, 212:687–703, 1998.
- [244] C E Rasmussen and C K I Williams. *Gaussian Processes for Machine Learning*. US: MIT Press, 2006.
- [245] C E Rasmussen and C K I Williams. *Regression*. 2006.
- [246] M.R. Taha, A. Noureldin, J. Lucero, and T. Baca. Wavelet transform for structural health monitoring: A compendium of uses and features. *Structural Health Monitoring*, 5(267-295), 2006.
- [247] Mohsen Azimi, Armin Dadras Eslamlou, and Gokhan Pekcan. *Data-driven structural health monitoring and damage detection through deep learning: State of the art review*, volume 20. 2020.
- [248] Panagiotis Seventekidis, Dimitrios Giagopoulos, Alexandros Arailopoulos, and Olga Markogiannaki. Structural Health Monitoring using deep learning with optimal finite element model generated data. *Mechanical Systems and Signal Processing*, 145(Nov-Dec):106972, 2020.
- [249] John P McCrory, Safaa Kh. Al-Jumaili, Davide Crivelli, Matthew R Pearson, Mark J Eaton, Carol A Featherston, Mario Guagliano, Karen M Holford, and Rhys Pullin. Damage classification in carbon fibre composites using acoustic emission: A comparison of three techniques. *Composites Part B: Engineering*, 68:424–430, 2015.