A predictive and scalable health-monitoring system for handpumps

Heloise G. Marais
Balliol College
University of Oxford

Supervised by
Professor David A. Clifton
December, 2016

This Transfer Report is submitted to the Department of Engineering Science, University of Oxford.
A predictive and scalable health-monitoring system for handpumps

Heloise G. Marais

Balliol College, December, 2016

Abstract

The handpump remains a 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. Practical challenges in the supply of spare parts combined with a lack of local skills result in an estimated 30% of handpumps not working at any given time. A broken handpump in a remote village can force women or girls to walk up to 20 kilometers to find alternative sources which may be contaminated or expensive. Reliable and sustainable water supplies are important to ensure healthy communities, societies, and economies in all regions of the world. The use of predictive maintenance in handpumps (whereby potential faults can be fixed in advance of handpump failure) has the potential to limit the substantial interruptions that exist in the rural water supply network [1].

This report describes preliminary research into condition monitoring approaches for rural handpumps. The research presents a framework for detecting handpump deterioration and failure using vibration signals from an accelerometer retrofitted in the pump handle, and applying lightweight machine learning methods to flag potential failure alarms, so that a corresponding subset of flagged data can then be used to perform heavyweight processing.

Different feature vectors, extracted from the vibrations signature of handpumps operating under various conditions, were used to assess the health of the handpump. The investigation identified that the type of the pump and the weight of the water and rods inside the handpumps, which is determined by the operating depth of the water-shaft, as being important features that should be considered. The speed of the pumping operation and its consequence on the observed data are also investigated.

A logistic regressor (which is sufficiently lightweight for implementation within the embedded monitoring system) is shown to provide early warning of handpump failure for extreme breakdown events. For preliminary frequency features considered in this work, it is also shown that this method does not reduce prediction accuracy significantly when compared to more complex (and computationally heavyweight) support vector equivalent.
Finally, a proposed work plan for future research is described. Methods of improved condition monitoring will be considered, as will the construction for a system of distributed inference and communication involving a cloud-based platform, and ultimately constructing a water risk assessment tool by fusing different data streams.
## Contents

1 The Need for Rural Smart Water  
1.1 Introduction ........................................... 1  
1.2 Groundwater Status, Usage and Monitoring in Kenya ......................... 2  
1.3 Handpumps .................................................. 3  
  1.3.1 Types of Handpumps .................................. 3  
  1.3.2 Selection Criteria ..................................... 3  
  1.3.3 Piston Handpump Operation .............................. 5  
1.4 Existing Techniques for Handpump Maintenance ............................... 5  
  1.4.1 Community-based models ............................... 5  
  1.4.2 Private Models ......................................... 6  
  1.4.3 Handpump Development Projects ........................ 6  
  1.4.4 Smart Handpumps ...................................... 6  
1.5 The Future of Handpump Monitoring ...................................... 6  
1.6 Overview of this Report ..................................... 7  

2 Condition Health Monitoring for Handpumps  
2.1 Introduction ............................................... 8  
2.2 Handpump Functionality and Continuity .................................. 8  
2.3 Common Handpump Failures ................................... 10  
  2.3.1 Hydrogeology ........................................... 10  
  2.3.2 Technical ............................................... 12  
  2.3.3 Socio-Economic ......................................... 14  
  2.3.4 Most common failures .................................. 15  
2.4 Vibration-Based Condition Monitoring ................................ 15  
  2.4.1 Level of identification .................................. 15  
  2.4.2 Novelty detection ....................................... 16  
  2.4.3 Anomaly Detection in Sensor Systems ....................... 16  
2.5 Classifiers .................................................... 16  
  2.5.1 Logistic Regression ..................................... 17  
  2.5.2 Support Vector Machines ................................ 18  

3 Data and Vibration Features  
3.1 Dataset ..................................................... 20  
3.1.1 Measuring Handpump Vibrations ................................ 20  
3.1.2 Data Collection .......................................... 21  
3.2 Classifying Normality ....................................... 22  
3.3 Vectorisation .................................................. 24  
3.4 Pre-processing ............................................... 24  
  3.4.1 Data Integrity Verification ................................ 25  
  3.4.2 Fast-Fourier Transforms ................................. 26  
  3.4.3 Embedded System Considerations ......................... 26  
3.5 Visualisation .................................................. 27  
  3.5.1 Sammon’s Mapping ...................................... 27  
3.6 Analysis of Spectral Data .................................... 29  
  3.6.1 Weight in the System .................................... 29  
  3.6.2 Normality ................................................. 30  
  3.6.3 Pumping Speed .......................................... 30  
3.7 Conclusion ..................................................... 33  

4 Handpump Condition Classification  
4.1 Introduction ............................................... 34  
4.2 Method ......................................................... 34  
  4.2.1 Data set ..................................................... 34  
  4.2.2 Train-Test .................................................. 35  
  4.2.3 Principal Component Analysis ............................. 36
4.2.4 Feature Selection ........................................................................... 36
4.3 Performance of Selected Vibration Features .................................................. 40
4.4 Conclusion ......................................................................................... 42

5 Discussion and Future Work ................................................................. 43
5.1 Conclusions ......................................................................................... 43
5.2 Improved Condition Monitoring of Handpumps ............................................ 43
  5.2.1 Validating Performance ..................................................................... 43
  5.2.2 Specifying Abnormality ................................................................... 44
  5.2.3 Improved Modelling ......................................................................... 44
  5.2.4 Data Warping .................................................................................. 44
  5.2.5 Improving Sensor Hardware .............................................................. 44
  5.2.6 Risks related to improved handpump condition monitoring .............. 45
5.3 Cloud-based Platform for Data Analysis .................................................. 46
  5.3.1 Expanding into New Regions ............................................................. 47
  5.3.2 Risks of a Cloud-Based Platform ....................................................... 47
5.4 Water security risk assessment tool ....................................................... 48
  5.4.1 Risks of a water security risk assessment tool ..................................... 48
5.5 Other Risks ......................................................................................... 48
5.6 DPhil Timeline .................................................................................... 49

Bibliography ......................................................................................... 50
Chapter 1

The Need for Rural Smart Water

1.1 Introduction

For the first time in 2016, the UN Sustainable Development Goals (SDGs) specified water security as one of the most pressing universal goals of our time, acknowledging that water crosses many other social sectors [2]. 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. The perpetual water crisis in Africa is a result of both economic and physical water scarcity [3]. The increased pressure on limited natural resources due to global climate change are universal, and do not impact Africa alone. Although many African governments recognise that access to clean water is a human right, responding to the issue is not always their primary concern given the multitude of other challenges they face related to development. Improved water monitoring and management should improve both physical and economic water scarcity [4].

Since the early 1980s [5], 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. Over two million people use handpumps everyday in sub-Saharan Africa to meet their daily water needs. A typical handpump supports around 360 villagers and is normally used for up to 14 hours each day to extract water at a rate of approximately 0.7 - 1 m$^3$/hour [6].

Most community handpumps, including the Afridev pumps used in Kenya, are designed to allow village level operations and maintenance (VLOM) [7]. However, a broken supply chain for spare parts, lack of finance, and limited technical knowledge mean this level of community maintenance is largely ineffective. It is estimated that 30% of the near 1 million handpumps in sub-Saharan Africa are broken at any given time [8]. Of those pumps that break, 70% are repaired within one week [9]; however, many remain broken for longer periods, leaving entire communities without access to safe water for prolonged periods.

Functionality of handpumps is a poorly-defined parameter that remains open to the judgment of the observer,
which is normally measured at a single point in time, is susceptible to changes in seasonality, and which lacks a universally-agreed definition [10]. Therefore, functionality alone is not an informative parameter for measuring the performance and sustainability of rural water services [11, 12]. 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 [10]. Regular inspection and monitoring of individual water points is important to ensure continuity in water supply; however, this is unlikely to occur frequently in high-income countries, let alone rural villages in low- and middle-income settings [13]. Currently there is no way to collect quantitative data of this nature to inform priority-setting for improved service delivery models [12].

Advancements in communication technology are opening up a host of possibilities for the continuous monitoring of water services in remote villages that often lack infrastructure, making it difficult or unsafe to reach [14]. Since the introduction of the “Smart Handpump” in 2011 [15], a number of other organisations are working on innovations in handpump maintenance systems [16, 17, 18].

This report describes the development of a novel, lightweight condition-monitoring system, mounted within the handles of manual water pumps, to provide early warning of eventual failure such that preventive maintenance action may be taken. This work is funded by UNICEF, which along with the World Bank are one of the main stakeholders in the sector; the aim of the UNICEF programme is to scale-up the results of this research within rural regions of the countries most in need.

1.2 Groundwater Status, Usage and Monitoring in Kenya

Kenya has an estimated groundwater potential of 619 million m³ [19]. In 2012, it was estimated that the abstraction rate of 7.21 million m³/year was still within the safe abstraction rate of 193 million m³/year [19]. This is a gross average estimation and certain aquifers are over-used; this results in the decline of water levels and a deterioration in water quality [20].

As recently as 2015, it was estimated that at least 20% of the total Kenyan population of nearly 48 million people rely on a handpump for their daily water needs [21]. There are an estimated 32,000 handpumps installed across Kenya [22].

Groundwater use in Kenya is greatly limited by over-exploitation in concentrated areas, as well as saline intrusion into the water supply in coastal areas. In addition, Rift Valley aquifers in Eastern Africa are notorious for their high levels of fluoride which affects the water quality [23]. 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.

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 on a quarterly basis, and more frequently for key aquifers that have been identified as being more intensively used \[24\]. Despite the magnitude of this effort, data points remain sparse across the region, making it difficult to accurately model changes in the aquifers.

1.3 Handpumps

The handpump is a robust and affordable technology that has evolved over the past 30 years to overcome the many challenges around contamination normally associated with traditional open wells. This section will review the technical specifications of handpump, selection criteria, and discuss its operation.

1.3.1 Types of Handpumps

Over the years, many different types of handpump have been developed to meet various demands. The design of handpumps has been largely standardised across various pumps, with the most important difference being the connection between the handle and the pump rods \[25, 26\]. A comparison of different types of handpumps is shown in Figure 1.1 and highlights the evolution of the three most widely-used VLOM handpumps. The introduction of open-top cylinders made it possible for villagers to perform maintenance on the pumps. The operating depth of lift pumps far exceeds that of suction pumps, making it more suitable for use in Kenya because water is often lifted from a depth much greater than 10 meters. The limitation on the operating depth of piston pumps is determined by the physical lifting power of the human operator rather than the pump itself. A technical data sheet can be seen in Table 1.1.

1.3.2 Selection Criteria

The criteria for selecting the type of handpump to install has not yet been standardised. However, there is some agreement that the operating conditions under which the pumps operate are the most important in defining this criteria. In general, the most important factors to consider are \[30\]:

1. Lift (shallow vs deep well);
2. User-group size;
3. Groundwater conditions (e.g., factors leading to possible corrosion); and
Figure 1.1: Overview of handpump types, classified by water-lift method (displacement or direct) including VLOM capability.

Table 1.1: Technical data sheet comparing the three VLOM pumps considered in this study, for current and future work. [27][28][29]

<table>
<thead>
<tr>
<th>Technical data</th>
<th>No.6 Pump</th>
<th>Afridev</th>
<th>India MKII</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth to be used</td>
<td>0 - 6 m</td>
<td>10 - 45 m</td>
<td>45 m</td>
</tr>
<tr>
<td>Cylinder diameter</td>
<td>89.0 mm</td>
<td>50.0 mm</td>
<td>63.5 mm</td>
</tr>
<tr>
<td>Maximum stroke</td>
<td>215 mm</td>
<td>225 mm</td>
<td>125 mm</td>
</tr>
<tr>
<td>Approx. discharge at 75 watt input (m³/hour)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At 5 m head*</td>
<td>3.5</td>
<td>1.4</td>
<td>1.8</td>
</tr>
<tr>
<td>At 10 m head</td>
<td>1.1</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>At 15 m head</td>
<td>0.9</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>At 20 m head</td>
<td>0.7</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>Pumping lift</td>
<td>0 - 7 m</td>
<td>10 - 45 m</td>
<td>10 - 50 m</td>
</tr>
<tr>
<td>No. of people served</td>
<td>50 - 100</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td>No. of households served</td>
<td>5 - 10</td>
<td>30 - 50</td>
<td>30</td>
</tr>
<tr>
<td>Water consumption (litres/capita)</td>
<td>20 - 25</td>
<td>15 - 20</td>
<td>15 - 20</td>
</tr>
<tr>
<td>Type of well</td>
<td>collapsible tube</td>
<td>borehole or dug</td>
<td>borehole or dug</td>
</tr>
</tbody>
</table>

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

The criteria defined by the end-user would be different from the installer and could include: ease of repair, reliability of the handpump, user preference, availability of the spare parts, and cost.

Design parameters of the pump as well as the borehole drilling have direct cost implications when installing a small water supply system. Higher yields require deeper boreholes, bigger diameter boreholes, elaborate site investigations, and intensive well development [31]. However, the bigger rigs required to drill boreholes for these types of handpumps increase both the cost and the risk of failure of the project.

The optimum selection criteria is a trade-off between these factors which 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 m³/hour is sufficient for a handpump that can serve up to 500 people [30].
1.3.3 Piston Handpump Operation

The pump cycle of a typical handpump can be described by two distinct phases: the up-stroke and the down-stroke. Due to the coupling between the handle and the rods, the motion of the handle is in the opposite direction to that of the piston. It is the motion of the piston that is used to distinguish these phases in the cycle rather than that of the handpump handle. 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 [32].

Up-Stroke

During the up-stroke the piston assembly moves upward. In direct or chain coupled piston handpumps this occurs when the handle is dropped from a position parallel to 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 are continuously repeated, water is pushed up the rising main and flows out the spout.

1.4 Existing Techniques for Handpump Maintenance

1.4.1 Community-based models

Over the years, the lack of adequate maintenance has led to the decentralisation of handpump operation and maintenance; resulting in the wide-spread implementation of Community Based Management (CBM) models. In 1997, the World Bank OECD audit [6] identified the key role of local community organisations and water committees to sustain handpump projects post-implementation. However, the inability of CBM models to sustain revenue collection to finance handpump maintenance remains the greatest hurdle for sustainable rural water supply [33] and alternatives to the dominant CBM models, such as cross subsidisation, may be required [34].
1.4.2 Private Models

In densely populated areas of south Asia, such as Bangladesh, where handpumps are operated in much closer 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. However, the existence of high-quality services in rural areas suggest that private operator models for handpumps can be successful [35].

1.4.3 Handpump Development Projects

NGOs and donor agencies have left a legacy of installing new infrastructure in rural regions instead of implementing long-term maintenance programs. Often, a new handpump will be installed near a non-working handpump.

1.4.4 Smart Handpumps

In 2011, a proof-of-concept was demonstrated for the remote monitoring of handpumps used a simple microprocessor, accelerometer, and mobile communications (GSM) components [1]. These “Smart Handpumps” provide hourly data related to usage, which in turn provides insights into more nuanced water use patterns [15]. The use of technology to flag failure retrospectively has shown that it is possible to reduce the average down-time from 27 days to fewer than three days [36]. In 2014, a preliminary analysis used high-frequency accelerometer data to show that these patterns contain useful information. High-rate waveforms from the accelerometry data can be processed using machine learning methods for shallow aquifer monitoring at the pump [18][37].

1.5 The Future of Handpump Monitoring

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 [38]. Many of the existing models of pump service have failed because they only focus on a single dimension. The Internet of Things (IoT) is allowing economies and industries worldwide to re-imagine traditional business models and reinvent services. By using machine-to-machine (M2M) communications to connect machines, devices, and objects to the internet, the IoT is able to 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.
A large-scale predictive and scalable health monitoring system for handpumps 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. Resource-constrained devices can use lightweight, on-board machine learning approaches to perform anomaly detection in the embedded system to improve the use of its limited resources. Moreover, 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.

1.6 Overview of this Report

A preliminary analysis of the vibration signatures extracted from the handle of surface-mounted handpumps show that these data provide useful information of the overall condition of the system.

Chapter 2 presents a framework for the condition monitoring of rural handpumps using vibration data. The application of condition monitoring in other fields is described.

Chapter 3 presents an investigation of the data set and describes selection of the relevant vibration signatures. Chapter 4 shows that it is possible to provide an early warning of extreme handpump conditions within the limitation of the hardware and experimental data.

Finally, Chapter 5 draws conclusions, before discussing a plan for future work to investigate condition monitoring methods for handpumps in a cloud-based platform.
Chapter 2

Condition Health Monitoring for Hand-pumps

2.1 Introduction

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 completely. With this approach already widely used in many fields from commercial and military jet engines [39], through to patient monitoring in health systems [40], it can directly be extended to monitoring the condition of handpumps in rural villages [1].

This chapter describes the context for handpump condition monitoring by introducing different aspects of handpump functionality and different types of handpump failures. Then, building on previous condition monitoring studies using vibration analysis, a similar approach is proposed to perform lightweight condition monitoring using vibration signatures obtained from accelerometer data.

2.2 Handpump Functionality and Continuity

Handpump functionality is a poorly defined metric in the literature. For this project, a handpump is considered functional at a given point in time $t$ if water is available at that time and non-functional otherwise. 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, 100% functionality is not expected 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 realizable target” [10].
Functionality can be modelled by the following equation [10]:

$$\text{Functionality, } f = (1 - a) - (1 - a) \frac{nd}{365}$$  \hspace{1cm} (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 breakdowns per year $n$ increase steadily over the lifespan of the pumps, from $n = 0.05$ in year 1 to 3.5 in year 10. As the severity of the failures increase, so does the duration of the 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 important parameters to improve functionality. However, neither functionality nor continuity are useful metrics in a water risk management tool as they offer little insight into the nature and the duration of the breakdowns [9]. For instance, the impact of 10 one-day breakdowns on communities will be significantly different to a single twenty-day breakdown, even though the same functionality score will be reported in both cases. Furthermore, the direct impact of non-functionality on human behavioural psychology is also not reflected in this equation and it remains unclear what the critical downtime per breakdown is which results in ultimate abandonment of the water point [11].

Condition monitoring typically considers the performance of the mechanical system; however, handpump failure is not limited to the binary classification of the performance of the mechanical system alone. Four main attributes to categorise handpump functionality are introduced [10]: 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. Understanding the range of this tolerance is an important factor to ensure the health-monitoring system is able to flag pump failure at a point that may be well 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 mechanical 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.
2.3 Common Handpump Failures

The lack of maintenance and high water demands at the pump are not the only factors contributing to non-functioning handpump parts. Other condition variables contributing to problems or failures in the system can be categorised into factors related to three key aspects: (1) hydrogeology; (2) technical; and (3) socio-economic.

A schematic summary of the interaction between multiple factors affecting handpump failures is shown in Figure 2.1.

2.3.1 Hydrogeology

Conductivity of the water

“The red water problem” is a symptom of corrosion in a handpump that is “well known by many villagers early in the morning after the pumps had not been used during the night” [42]. Moderate to aggressive groundwater with pH < 7 can rapidly accelerate the wear of pump rods and cylinders due to increased corrosion. Under these conditions, rising mains and pumping rods require more frequent replacement. It has been found that galvanisation of mild steel yields components that are not resistant to these operating conditions. It is estimated that the scale of impact of corrosion is much larger than expected; at least 66% of handpump failures in Ghana can be attributed to corrosion [42].
Furthermore, corrosion of the parts of a pump also leads to the release of iron in the water source. In some cases, working handpumps are abandoned due to excessive iron in the water [11]. In a similar study [12], it is suggested that that 30% of water points in the Niger become abandoned due to the poor taste caused by high levels of iron.

Areas with aggressive water would have to factor in the additional wear associated with corrosion. Therefore, there is strong motivation to use suitable materials, such as plastic or stainless steel, to ensure that pumps are fully corrosion resistant. Unfortunately, a design 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.

**Temperature of the water**

Pumping parts are designed to operate in water temperatures of 16 - 40°C [38]. 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 [30], 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 they are allowed to dry.

Shorter term fluctuations are caused by the drawdown in the borehole itself as well as nearby boreholes, known as borehole interference [31]. In the case of a rural water supply network, where boreholes are co-located at greater intervals of 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 overlying local soils and water table depth. Basement aquifers, with transmissivity < 1 m²/day, are more likely to experience limitations related to drawdown [13]. Precipitation and tidal-level variations are the main driving forces of recharge in coastal aquifers. Existing forecasts using machine learning methods rely exclusively on peizometric height measurements [44, 45, 46].
### 2.3.2 Technical

Worn parts and a lack of regular (or any) maintenance is the primary reason for technical failure in handpumps. However, these failures are not exclusively caused by overuse. Other causes can be due to [31]:

- Design flaws;
- Incorrect installation;
- Manufacturing defects; and
- Depth, location, and age of the borehole.

**Borehole**

The depth of the borehole determines the number of rods used in the system. It is therefore expected that the extra weight lifted in deeper boreholes may result in more frequently failures. In sub-Saharan Africa drilling depths often range from 48 to 62 metres, resulting in the pump being 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 are used to line the borehole such that it remains clear. Borehole maintenance every ten years is required to clean and re-develop the borehole [30]; however, this is rarely performed.

**Problems with parts**

Overall performance of the Afridev lies well within technical specifications with typical part failure related to frequent use. However, manufacture non-compliance with the SKAT specifications [28], such as incorrect dimensions of U-seals and poor welding craftsmanship, substantially reduce the performance of handpumps. Table 2.1 is an overview of the most common handpump breakdowns that occur as a result of failure in a single part. This table draws on the literature in addition to field notes taken by the author during the study in Kenya.

---

1 SKAT 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 2.1: Common handpump failures in deep well lift pumps (Afridev) [5, 38, 47, 48, 49].

<table>
<thead>
<tr>
<th>Part</th>
<th>Common failure</th>
<th>Possible Cause</th>
<th>Fault Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foot valve</td>
<td>Foot valve becomes stuck on the cylinder lip.</td>
<td>Foot valve oscillates when dropped down the rising main during fitting.</td>
<td>Installation</td>
</tr>
<tr>
<td>Cup Seal</td>
<td>(1) Damaged cup seal. (2) Valve forced to remain open.</td>
<td>Foot valve bobbins that break on impact with the water or, if well is dry, the foot valve receiver can cause the bobbins to break off become lodged in the valve.</td>
<td>Design Flaw</td>
</tr>
<tr>
<td>Rod Centralisers</td>
<td>Worn centralisers cause additional friction during pump operation.</td>
<td>(1) Abrasion caused by siltation; (2) Misalignment of the rod and rising main eyelet.</td>
<td>Installation</td>
</tr>
<tr>
<td>Bent Pump Rods</td>
<td>Misalignment of rods or rod joints cause additional rubbing on rising main pipes.</td>
<td>Damage to the rods can occur during: (1) transportation; (2) installation; (3) maintenance/re-installation.</td>
<td>Installation or Maintenance</td>
</tr>
<tr>
<td>Corrosion of Pump Rods</td>
<td>Worn rods have weak spots and require more frequent replacement. It affects mechanical performance when rods break but also cause damage to the cylinder.</td>
<td>Moderately to aggressive groundwater (pH &lt; 7). Galvanisation of the mild steel are not resistant to these conditions.</td>
<td>Design Flaw or Maintenance</td>
</tr>
<tr>
<td>Rising Cylinder</td>
<td>Damage at the socket joint.</td>
<td>Misalignment of the socket joint during installation.</td>
<td>Installation</td>
</tr>
<tr>
<td>Plunger</td>
<td>Plunger legs break off and become lodged in the foot valve.</td>
<td>Cause is unclear but could be during impact with the water when excessive force is applied. Possibly incorrect dimension/angle/material of the legs.</td>
<td>Design Flaw</td>
</tr>
<tr>
<td>U-Seals</td>
<td>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.</td>
<td>(1) Abrasion caused by siltation; (2) Incorrect dimensions; (3) Soft material; (4) Friction caused by withdrawal against the inside of the rising main.</td>
<td>Manufacturing defect</td>
</tr>
<tr>
<td>Rising main</td>
<td>Crack in the casing cause loss of pressure in the system restricting or preventing water flow.</td>
<td>Aggravated by bent rods, removed rod centralisers, or poor joints.</td>
<td>Design Flaw or Maintenance</td>
</tr>
<tr>
<td>Handle</td>
<td>Handle is loose or “shaky” during use.</td>
<td>(1) Worn bush bearings. (2) Loose falcrum or hanger pins caused by corrosion.</td>
<td>Design Flaw or Maintenance</td>
</tr>
<tr>
<td>Pump head and stand</td>
<td>(1) Head or stand is unstable during use. (2) Water is leaking from the pump body.</td>
<td>Loose flanges or a cracked platform cause instability during use.</td>
<td>Maintenance</td>
</tr>
</tbody>
</table>
2.3.3 Socio-Economic

Level of Use of the pump

A handpump with continuous high usage (more than 5m$^3$ or 5000 litres per day) is expected to fail every 3-6 months. The stress on the mechanical system can be described as a function of the number of users and the pumping lift in the well [30]:

$$\text{Stress, } s \approx f(u^2, x^2)$$  \hspace{1cm} (2.2)

where the stress on the pump is a function $f$ of the number of users $u$ and the total pumping lift $x$.

Thus, the number of users and pump lift are important factors affecting the pump operating conditions, long term wear, and maintenance schedules. Pumps designed for small user groups are less robust than those designed to serve large groups (>100 users).

Maintenance

Regular scheduled maintenance of VLOM pumps, such as cleaning of the rods, should be performed by the villagers to prolong the time between major breakdown events. The lack of transfer of formal “ownership” of handpumps and flow of information often prohibits community participation in practice [5].

Users Characteristics

There is no evidence in the literature to suggest that the characteristics (height, gender, age) of the user directly contribute to the failure of the pump. However, observations during field work suggest that these vary so significantly, both at the same pump as well as across pumps, that it should not be ignored during the initial investigation described by this report. For example, it was observed that the height of the user determines the range of motions in the handle. 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 [49].

Furthermore, it was noted that gender, often but not always, determines the “pumping style”. Women often use both hands on one side of the 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.
2.3.4 Most common failures

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, in Table 2.1 it was shown that failure is often due to malfunctioning of more than one part. Early detection of wear in a single part may therefore prevent many other problems from occurring in the pump.

2.4 Vibration-Based Condition Monitoring

Vibration-based condition monitoring is the analysis of specific system characteristics using pattern recognition to detect changes that may indicate extensive wear, degradation, or damage. 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 [32]. 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 [50].

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 certain systems. Novelty detection, which is described in section 2.5 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 categories are based on the damage that is measured and include methods based on: natural frequency, mode shape, mode shape curvature/strain mode shape, dynamically measured flexibility, matrix update, non-linear methods, neural network methods, and other methods [51].

2.4.1 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 that is required for a specific application. A classification method based on the level of identification required offers four levels [52]:
Level 1: Determine that damage is present in the structure
Level 2: Determine the geometric location of the damage
Level 3: Quantify the severity of the damage
Level 4: Predict the remaining service life of the structure

2.5 Novelty detection

Since enough examples of different types of handpump failures are unavailable it is unlikely for a single machine learning to model to be applicable to all possible object classes. Thus, non-model based statistical pattern recognition approaches are required to complement existing model-based techniques [53]. 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 [54], such as typically occur with high-integrity systems [55].

2.5.1 Anomaly Detection in Sensor Systems

Lightweight, on-line (i.e., real-time) processing is emerging in the maturing field of wireless sensor networks (WSN). Optimal use of their limited available system resources, such as battery life and data-transmission bandwidth, is important for resource-constrained embedded devices used for remote monitoring in locations that are unreachable or dangerous.

Anomalies in sensor data can broadly be categorised by type such as spikes (transients), constant, noise, and drift [56]. These manifestations can occur as a result of the sensor itself or the environment in which the sensor is operating. It is therefore important to distinguish between the causation of these anomalies in the data. Malfunctioning can occur within the system being monitored due to mechanical degradation associated with regular use or hydrogeological factors, as discussed in section 2.3. However, malfunction can also occur due to a failure of the sensor itself; for instance, a malfunctioning hardware connection, a depleted battery, a sensor becoming “stuck” (physically or in software), or a change in sensor calibration [57].

2.6 Classifiers

Classification is an important tool used in statistics and machine learning to estimate the class label of an object from a given set of input variables. This can be supervised learning where labels are known for a set of training examples, which will be the focus of the investigation described by this report.
Two supervised classification algorithms were considered for the prediction of handpump failure. Since the computational power would be limited by the processing power and battery life, a lightweight algorithm is used that would be able to operate within the design constraints.

Although novelty detection approaches are better suited given the lack of failure examples, for this preliminary study binary classification is performed, using a data set for which both normal and abnormal examples are available, in order to investigate vibration data across various pump types and failure conditions. This is done to identify the characteristic features that are the most informative of handpump failures, which in turn will be used to inform novelty detection approaches in future.

2.6.1 Logistic Regression

Logistic regression is a standard regression model used in statistics to categorise the dependent variable and predict the probability (0,1) of membership of one class (e.g., “failure”) in a two-class setting. 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 the 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.

Models are used to describe the patterns of interactions and associations. The model parameters provide an indication of the strength of these associations. Generalised linear models (GLM) \([58]\) are a class of models that assume the response variable \( y_i \) follows an exponential family distribution with mean \( \mu_i \), where \( \mu_i \) is assumed to be a function of \( x_i^T \beta \).

Logistic regression is special class of GLMs that models how the binary response variable \( Y \) depends on a set of \( k \) explanatory variables, \( X = (X_1, X_2, X_3) \), where the random component, the probability distribution of the response variable \( Y \), has a binomial distribution for \( Y \) (in the binary log regression). The systematic component specifying the explanatory variables \( (X_1, X_2, X_3) \) for logistic regression is mixed (continuous/categorical). Finally, the link function, describing the distribution of the residuals, for logistic regression can be described by:

\[
\eta = \logit(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) \tag{2.3}
\]

In machine learning, logistic regression can be used to model the posterior probability of input variables \( X \) being associated with a class by fitting a linear model to the feature space. This approach is quick to train and
very fast to classify unknown records. The natural probabilistic views of class predictions are restricted by the use of a linear decision boundary; however, this does help to prevent over-fitting of the model.

The state of a handpump can be determined by binary classification as the pump is either functional \((Y = 1)\) or non-functional \((Y = 0)\). Thus, given the 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:

\[
P(X) \triangleq P(Y = 1|X) \tag{2.4}
\]

Although logistic regression is a linear method, the predictions are transformed using the logistic function and as a result can no longer be interpreted as linear combinations of the inputs. Thus, equation 2.4 becomes:

\[
P(X) = \frac{\exp(b_0 + b_1X)}{1 + \exp(b_0 + b_1X)} \tag{2.5}
\]

The coefficients of the logistic regression algorithm, \(b_0\) and \(b_1\), are estimated from the training data using maximum-likelihood estimation.

For training data with \(m\) data points and \(n\) features, \(((x^1, y^1), (x^2, y^2), \ldots, (x^m, y^m))\) where \(x \in \mathbb{R}^{n+1}\) with \(x_0 := 1\), \(y \in \{0, 1\}\), the cost function is optimised by:

\[
J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} (h_\theta(x^i) - y^i)^2 \tag{2.6}
\]

where \(h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}\). The coefficients should be optimised to minimise the error in the predicted probabilities.

**Applications of LR**

Logistic regression was first used in medical science to develop the Trauma and Injury Severity Score (TRISS), which is used to predict the mortality in injured patients [59]. In engineering, this technique is frequently used to predict the probability of failure in mechanical systems. Finally, logistic distributions are frequently used in hydrology to illustrate the distribution of rainfall or river discharge, which are thought to have normal distributions [60].

2.6.2 Support Vector Machines

The support vector machine (SVM) is used to form decision boundaries and recognise patterns for classification analysis and can be extended for regression [61][62]. For this initial investigation, a two-class SVM classifier is
considered where the output of the mapping function is either \( Y = 1 \) (normal) or \( Y = 0 \) (abnormal) to specify handpump condition. Assuming to processing constraints, this will be the "gold standard" used to compare the performance of the (much more lightweight) LR classifier.

**Classifying with SVMs**

The so-called kernel trick enables the classification of nonlinear inputs of a maximum-margin hyperplane in some feature space, \( F \) \(^{[63, 64]} \). The hyperplane is optimised to maximise the margin between the data points near the boundary in the high-dimensional feature space in which the decision boundary is linear \(^{[65]} \).

The classifying SVM \(^{[66, 67]} \) estimates the value of the function \( f(x) \) at the unsampled training point \( x_i \) such that the linear hyperplane is of the form:

\[
f(x) = x_i \cdot w + b \tag{2.7}
\]

where \( w \) is the support vector weight and \( b \) is the bias. For the case where the classes of data are linearly separable in \( X \), this decision boundary is defined by \( f(x) = 0 \).

Given that the “real-world” data used in this study are not noise-free or linearly separable, a *soft-margin* SVM is used to allow data mislabelling while still maximising the margin. In the non-separable case, a convex penalty term is imposed to allow violations for misclassifications on \( y_i(w^T x_i - b) \geq 1 \) in order to determine the optimised values of \( w \) and \( b \). The slack variable \( \xi_i \) is introduced such that \( y_i(w^T x_i - b) \geq 1 - \xi_i \) for \( \xi \geq 0 \). The hyperparameters of the soft margin SVM are optimised by minimising:

\[
E = \frac{1}{2} \| w \|^2 + C \sum_{n=1}^{N} \xi_i \tag{2.8}
\]

where \( C \) is a trade-off parameter of the model between the margin size and the amount of error in training tolerated. The value of \( C \) as a penalty factor plays an important role in the fitting of the model. A large value for non-separable points will over-fit while a low penalty will result in under-fitting of the model. The slack variable measures the degree of misclassification. K-fold cross-validation is used to find the optimal combination of the slack variable, \( \xi \), and kernel parameter, \( C \), which will result in high accuracy without over- or under-fitting of the model. Implementation was performed in Matlab using the public open-source LIBSVM environment \(^{[68]} \).
Chapter 3

Data and Vibration Features

3.1 Dataset

3.1.1 Measuring Handpump Vibrations

To measure the vibrations in a handpump during operation, a sensor, containing a consumer-grade accelerometer with a sampling frequency of 96 Hz, is attached to the handle of the handpump at a position closest to the pump body, as shown in Figure 3.1, without interfering with the range of motion of the handle. The sensor used in the logger contains an accelerometer that is very similar to the Water Data Transmitters (WDT) used in our long-term data loggers and other similar Smart Water infrastructure [18, 15], to ensure data consistency. The vibration sensor was connected to a nearby laptop via Bluetooth to record 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 working and non-working condition is shown in Figure 3.2. To minimise the risk of variations in the placement position and orientation of the sensor between different pumps and different sessions for the same pump, visual standardisation was always undertaken by two team members.

Figure 3.1: Diagram showing the experimental set up of a vibration sensor attached to a surface-mounted handpump and the relevant orthogonal axes of accelerometry.
Figure 3.2: A five-second interval of unprocessed accelerometer data of the same pump in a (a) normal and (b) abnormal condition in the X, Y, and Z dimensions (upper to lower plots, respectively) for the same user.

3.1.2 Data Collection

In March 2016, we acquired short-term accelerometry data from handpumps in Kwale County, Kenya, over a two-week period. The Afridev is the most commonly-used VLOM pump in Kenya and although one India MKIII was used for recordings, it is not included in the research described by this report. Each recording is for the duration required to fill a predefined water container, normally 20 litres, and with durations from 30 to 180 seconds.

Variability Experiments

To ensure sufficient representation of variability in the data set, recording from different pumps and users are included. The data set contains recordings collected for both members of the project team as well as volunteers from the local community at 28 different Afridev pumps across the county (where their geographical locations may be mapped here: https://goo.gl/VoOeFx).

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 data set, 50.4% of the recordings are from deep wells, operating at depth of more than 25 m below ground level, and the remaining from shallow wells, operating at a depth of less 25 m below ground level.
Long-term Deterioration Experiments

In addition to the 28 pumps sampled during the field visit, long-term data loggers containing vibration sensors were installed at five handpumps to track the deterioration of handpump mechanics leading up to the breakdown event over time. No failures were recorded at one of these sites by the time this report was published.

Seeded Fault Experiments

During the field work, we worked closely with a local Kenyan maintenance company, FundiFix, which is responsible for handpump repairs in the area. 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 abnormality.

3.2 Classifying Normality

Following the discussion of handpump functionality in section 2.2, two main attributes to classify normal handpump conditions are introduced:

1. **Short-Term Water Quantity:** a pump is either classed as normal (1) or abnormal (0). A pump is considered normal when water flows from the spout while pumping and abnormal when no water flows from the spout while pumping. In Figure 3.3a it can be see that there are significantly fewer examples of abnormal handpump conditions compared to the relative frequency of data for normal conditions. This is the first level of identification to “determine that damage is present in the structure”, as described in section 2.4.1.

2. **Mechanical Performance:** ten sub-categories are used to classify the mechanical attributes that describe the underlying condition of each pump. These conditions correspond to the most common failures identified in table 2.1. A histogram of the relative frequency of each mechanical condition across all recordings is shown in Figure 3.3b.

The affect of mechanical performance on normality is shown in Table 3.1. Recordings were retrospectively labeled according to pump performance.

Identifying Multiple Conditions

In the second level of identification, to determine the geometric location of the damage, it is possible that a single pump could be assigned more than one label to describe its mechanical performance. However, the
Figure 3.3: Histogram of the mechanical condition labels across all recordings.

Table 3.1: Description of the mechanical condition and short-term water quantity classification labels assigned to the recordings of each pump.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Description</th>
<th>Label</th>
<th>Water Flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Excellent working condition</td>
<td>1</td>
<td>Unrestricted</td>
</tr>
<tr>
<td>2</td>
<td>Noisy but working</td>
<td>1</td>
<td>Varied</td>
</tr>
<tr>
<td>3</td>
<td>Dry borehole</td>
<td>0</td>
<td>None</td>
</tr>
<tr>
<td>4</td>
<td>Rising main leak</td>
<td>0</td>
<td>None</td>
</tr>
<tr>
<td>5</td>
<td>Broken U-seal</td>
<td>0</td>
<td>None</td>
</tr>
<tr>
<td>6</td>
<td>Worn U-seal</td>
<td>1</td>
<td>Restricted</td>
</tr>
<tr>
<td>7</td>
<td>Pump body leaking</td>
<td>1</td>
<td>Restricted</td>
</tr>
<tr>
<td>8</td>
<td>Worn bush bearing</td>
<td>1</td>
<td>Unrestricted</td>
</tr>
<tr>
<td>9</td>
<td>Stiff handle</td>
<td>1</td>
<td>Unrestricted</td>
</tr>
<tr>
<td>10</td>
<td>Other</td>
<td>1</td>
<td>Varied</td>
</tr>
</tbody>
</table>

primary condition was considered as the true cause of the condition. For example, a pump with a broken U-seal could also have a worn bush bearing; however, the broken U-seal is most likely the cause of no water flow and thus labeled as the primary condition of the pump.

**Limitations of this classification of normality**

Condition 3 (dry borehole) should be treated with caution given that the lack of water flow is an attribute of seasonality (long term) rather than mechanical performance of the handpump. However, since the condition of the pump is being classified using a short-term snapshot of data (from a single field visit), this is considered abnormal handpump performance related to the mechanical condition, where the health of the aquifer forms part of the overall handpump health. Once one full cycle of long-term data (one year with all seasons) is collected, classification could be updated to contain a seasonal aspect to the analysis.
3.3 Vectorisation

Each recording is summarised by a shape vector $x$ to describe the the underlying sinusoidal waveform. Each vector consists of a set of points $(t_i, x_i)_{i=1}^N$, which represent the measured force $x_i$ at time $t_i$. The signal waveform contains vibration noise such that $x = f(t) + \eta$, where $f(t)$ is the function describing the underlying periodic waveform of the pumping motion and $\eta$ 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 3.2, 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, when compared to rest of the signal.

3.4 Pre-processing

In order to perform frequency-domain approaches to modelling handpump conditions, where the mechanical failures are most likely to be obvious in the frequency domain [69, 70], transformation of the time-domain recordings is required. In the following section, the steps taken to convert the time signal to the frequency domain are described.

Pre-processing includes a number of steps to prepare the data for feature extraction and ensure hardware compatibility. 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 subsequent assessment of data integrity. For waveforms on the Y- and Z-axes, we chose to consider periods with a difference of greater than 0.25 s between the peaks and troughs and 0.015 s for the signal measured in the X-direction.

The second pre-processing step is to filter each unique vibration waveform. 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 is then removed from the original signal. This process acts as a high-pass filtering (HPF) to remove the low-frequency components associated with the pumping tempo that are not indicative of handpump failure, and which retains an estimate of the fast-moving trend components of the noisy vibrations. Our hypothesis is that changes in pump conditions are not affected by the component of the signal cause by the relative low-frequency motion of the user moving the pump handle.
A moving average filter suitable for use on resource-constrained hardware applications. The filter calculates the average of a number of points from the input signal such that each point of the output signal $y$ is calculated as [71]:

$$y[i] = \frac{1}{M} \sum_{j=0}^{M-1} x[i+j]$$  \hspace{1cm} (3.1)

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 will also be used to filter the test data in addition to the training data. This is to ensure that the testing process closely simulates the on-board processing to be performed on the handpump logger. A simple FIR filter will allow continuous filtering in real-time. An example of the output of the filtering step is shown in Figure 3.4 with the unprocessed pumping waveform, the low-passed moving average, and the final high-passed output of the vibrations.

![Figure 3.4: A six-second interval of the final output of the pre-processing steps for the Y-axis accelerometry data.](image)

The final pre-processing step is to create 1.3 s windows from each recording with 50% overlap. This creates 128 samples per window, equivalent to 64 frequency components with a resolution of 0.75 Hz per component for a sampling frequency of 96 Hz.

### 3.4.1 Data Integrity Verification

In addition to pre-processing, key steps were implemented to ensure that windows which were not of suitable quality or unusable are discarded. A waveform that varies significantly could, for example, indicate unusual use of the pumps, such as children playing with the handle.

The first data verification step was to verify that data vectors contained non-zero values. The second step was to clip the recordings by removing the first and last periods of each recording.
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 being the time between
subsequent troughs, was used to evaluate pumping rate during normal use. Waveform periods that were 1.5
times larger or smaller than the median period were removed.

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

### 3.4.2 Fast-Fourier Transforms

The Fourier transform decomposes a 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 \[72\]. The discrete Fourier
transform $X_k$ for a discrete time-signal $x_m$ can be defined as:

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

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

#### Windowing

In Figure 3.4 it is shown 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 \[73\] \[74\]. Windowing can be used to 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 feature vector with 64 frequency component per window, up to the Nyquist
rate.

### 3.4.3 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) two additional steps were taken.
Handling floating point units

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.

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 as a way to optimise resources on the embedded system and to ensure compatibility with the hardware constraints of the embedded microprocessor that will be used. However, 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 originally sampled 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.

3.5 Visualisation

Visualisation is a powerful tool that can be used 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 dimensionality of the data into a “latent space” while aiming to preserve the structure of the data in the original (higher-dimensional) data space.

3.5.1 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, typically Euclidean, between each pair of points in the original multi-dimensional data space in the two dimensional latent space [75]. 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:

$$E_{sam} = \frac{1}{\sum_{i<j} |d_{ij}^*|} \sum_{i<j} \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*}$$  \hspace{1cm} (3.3)
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. From the resulting projection in Figure 3.5 it can be seen that abnormal classes 2 to 4 internix with the normal class 1. However, classes of individual abnormal conditions form separate clusters. The projected points of class 2 and class 4 form tight clusters, respectively, indicating that their shape vectors are very similar.

Figure 3.6 shows the patterns projected for individual pumps, where each one of the 28 sub-plots highlights a single pump in the data set. The spread of tight clusters for individual pumps show the clear variability that occurs across individual pumps of the same type.

![Figure 3.5: Projection of 128-D feature vectors for handpump vibrations on the Y-axis. Labels are described in Table 3.2.](image-url)

Table 3.2: Class labels within visualisation of normal and abnormal handpump conditions.

<table>
<thead>
<tr>
<th>Class</th>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Green □</td>
<td>Normal</td>
</tr>
<tr>
<td>2</td>
<td>Red △</td>
<td>Abnormal: Dry borehole</td>
</tr>
<tr>
<td>3</td>
<td>Black *</td>
<td>Abnormal: Leak in rising main</td>
</tr>
<tr>
<td>4</td>
<td>Blue ○</td>
<td>Abnormal: Broken U-seal</td>
</tr>
</tbody>
</table>
3.6 Analysis of Spectral Data

As part of the preliminary analysis described by this report, the FFT is used to represent approximations of the power spectra of the handpump vibration for the accelerometer sensor output.

3.6.1 Weight in the System

Two factors determine the weight inside the handpump system being lifting with each pump stroke: (1) water and (2) rods in the pump. From Figure 3.5 it is clear that the presence or absence of water in the system for normal and abnormal classes separate into distinct clusters when projecting the feature vectors. The number of rods in the pump, connecting the handle and the plunger just above the foot valve, is determined by the depth of the borehole during the handpump installation. Each rod is 3 metres in length and made of either mild or stainless steel. 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. Visualisation of the feature vectors grouped by the depth of the well, as determined by the number of rods in the pump, is shown in Figure 3.7. For deep wells (blue), operating at depth of more than 25 m or more than 7 rods, and shallow wells (red), operating at a depth of less 25 m or fewer than 7 rods, there is some sub-set of feature vectors that is unique to deep wells. Feature vectors from shallow pumps vary more.
3.6.2 Normality

For normal handpump conditions, shown in Figure 3.8a and 3.8b, distinct high-frequency components appear to take high energy. 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.

For abnormal handpump conditions at the same pump, these bands of high energy are significantly less obvious than under normal operating conditions. This may be due to the lack of water (weight) in the system.

3.6.3 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 affect of the speed at which is handpump is operated, 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 maximum recorded rotation of the handle above the horizontal is 61.3 degrees and -34.5 degrees 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 degrees and the median arc on the up stroke (handle moves below the horizontal) is -15.5 degrees. 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 of 28 pumps, the probability of recorded speeds, along with fitted gamma distributions (often used to model asymmetric distributions of speed for rotating machines [39]), for normal (blue) and abnormal (red) conditions are shown in Figure 3.9. It can be seen that a speed sub-range of [70 160] degrees/second is appropriate to cover the speeds examples contained in the data set for both normal and abnormal conditions. The change observed in the probability density function of speed for normal (blue) and abnormal (red) handpump conditions indicates that there is an increase in pumping speed under abnormal handpump conditions. This may be due to the lack of weight in the system, making it easier to move the handle at greater, or an exerting attempt by the user trying to extract water.

![Figure 3.9: Probability of pumping speed for normal (blue) and abnormal (red) handpump conditions, respectively.](image_url)

Figure 3.9 considers a speed-based representation of the spectral data shown in Figure 3.8 based on work modelling jet engine vibration spectra [39]. The speed-frequency space is quantised into \( N_s \times N_f = 10 \times 10 \) bins. The horizontal axis shows speed bins ranging from [70 160] degrees/second. 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. We conclude from this preliminary representation of the data that the effect of speed on vibration of the pump could be considered in future, but there is little evidence for disregarding the effect of speed.
Figure 3.10: 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.

3.7 Conclusion

This chapter described the method used to prepare the data set for feature extraction. It was shown that separate clustering of normal and abnormal data occur while significant variability exists across individual pumps. The affect of normality, weight in the system, and pumping speed were considered for the selection of significant vibration features. While there was insufficient evidence to suggest that pumping speed should be eliminated from the analysis, the weight in the system and subsequently the operating depth of the handpump was shown to be factors that significantly effect the distribution of feature vectors.
Chapter 4

Handpump Condition Classification

4.1 Introduction

This section describes an analysis of the performance of the accuracy scores for the classification of the handpump condition using different vibration features. Support vector machines (SVMs) are used as the gold standard to assess the performance of the logistic regressor (LR) classifier performance, assuming no hardware processing constraints. These binary classifiers were used to map the outcomes of the function as either one (normal) or zero (abnormal).

4.2 Method

4.2.1 Data set

From the original data set, containing examples of all the labeled mechanical conditions across 28 pumps, a subset of the data was used to construct a training set for the classifier. The training set, $D_1$, contains only cases for normal pumps in excellent working condition (condition 1) and abnormal cases, including dry borehole (condition 3), leak in the rising main (condition 4) and broken U-seal (condition 5), as specified in section 3.1. 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 $D_1$.

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. In Figure 3.3a it was shown that the abnormal class is significantly under-represented in the data set. Thus, it was important to ensure that for each failure case the same number of normal and abnormal points are included in the final data set. As a result, the final data set, $D_1$, contained $N = 2530$ samples with equal-sized representation of normal and abnormal conditions, after sub-sampling the larger class.
4.2.2 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 confirm 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 case, the data set is balanced such that it contains the same number examples for normal and abnormal conditions. A flow diagram of the train-test method for the LR classifier is shown in Figure 4.1.

Each feature vector in the training data is 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.

A LR classifier has no hyperparameters to control model flexibility and as such no validation step (e.g., 10-fold cross-validation) is required to optimise its parameters. However, to ensure the robustness of the model, the accuracy score of the trained LR model using the unseen test data was 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) of the accuracy scores and found to be negligible (less than 0.02).
Hyperparameters of the SVM classifier were optimised using cross-validation (CV). K-fold CV, for $K = 5$, is used to partition the available balanced data set, $D_1$, into $K$ subset “folds” [76]. 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 SVM is reported as the average of $K$ outcomes using the unseen data.

### 4.2.3 Principal Component Analysis

Principal component analysis (PCA) is another statistical analysis tool used to reduce the dimensionality of a data set by using the principal components that describe the underlying structure in the data. By simplifying complex data in this way, it is possible to observe the hidden structure that underlies it in the parsimonious form when considering the dominant principal components. This is done by finding the eigenvector of the data covariance matrix with the highest eigenvalue such that the vectors of weights map each row vector $x_t$ to a new linearly-uncorrelated principal component score [77]. The first principal component accounts for as much of the variability in the data as possible such that the weight is described by:

$$w_{(1)} = \arg\max \left\{ \sum_i (x_{(i)}w)^2 \right\}$$ (4.1)

PCA was used to analyse the selected feature vectors in a visual manner.

### 4.2.4 Feature Selection

Feature extraction and selection are 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 in Chapter 3. 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 gold standard to evaluate the performance of selecting a subset of these features. For feature vectors $FV_1 - FV_3$, the frequency components with indices in 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, mid, and upper range, respectively. Feature vector $FV_4$ is constructed by uniformly sampling over the entire range. $FV_5$ selects the 20 frequency components with the greatest absolute difference between the median power spectra for normal and abnormal handpump conditions. Finally, $FV_6$ randomly selects 20 frequency components from the entire frequency range from bin 3 to 64. Definitions of the feature vectors are described in Table 4.1.
The first three frequency components, equivalent to 0 - 2.25 Hz, were not considered for feature construction to avoid dominant vibration features within the median pumping frequency (associated with the manual operation of the pump) of 1.1 Hz. In Figure 3.8 it can be seen that the moving average filter is not successful in removing all the frequency components in the lower range.

Table 4.1: 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 $[j, j + i, j + 2i, ..., j + m \times i]$, where $m = 20 = \frac{k - j}{i}$.

<table>
<thead>
<tr>
<th>Feature Vector</th>
<th>Description</th>
<th>Frequency bin index</th>
<th>Equivalent Frequency (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$FV_0$</td>
<td>Full range</td>
<td>[1:1:64]</td>
<td>[0.75:0.75:48]</td>
</tr>
<tr>
<td>$FV_1$</td>
<td>Lower range</td>
<td>[4:1:23]</td>
<td>[3:0.75:17.25]</td>
</tr>
<tr>
<td>$FV_2$</td>
<td>Mid range</td>
<td>[24:1:43]</td>
<td>[18:0.75:32.25]</td>
</tr>
<tr>
<td>$FV_3$</td>
<td>Upper range</td>
<td>[44:1:63]</td>
<td>[33:0.75:47.25]</td>
</tr>
<tr>
<td>$FV_4$</td>
<td>Uniformly sampled</td>
<td>[3:3:60]</td>
<td>[2.25:2.25:45]</td>
</tr>
<tr>
<td>$FV_5$</td>
<td>Hand-picked</td>
<td>[11,26:1:36, 45:1:51,53]</td>
<td>[8.25,19.5:0.75:27, 33.75:0.75:38.25,39.75]</td>
</tr>
<tr>
<td>$FV_6$</td>
<td>Randomly selected</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The resulting 20-D projections from each feature vector are shown in Figure 4.2(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 proposed feature vectors. Some clustering of specific conditions can be seen in Figure 4.2d and 4.2f. A similar projection from each feature grouped by the operating depth of the handpump is shown in Figure 4.3. For the case of shallow operating pumps, separate clusters of blue and green appears in Figure 4.3d, 4.3e and 4.3f. For the case of deep operating pumps, similar clusters of black and red points are seen in Figure 4.3c, 4.3i and 4.3d. In both cases it appears that FFT bins across the entire frequency range are important to capture variability in the data. Further investigation will verify this work with larger data sets.
Figure 4.2: 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 4.2.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>Normal Train</td>
</tr>
<tr>
<td>Red</td>
<td>Abnormal Train</td>
</tr>
<tr>
<td>Black</td>
<td>Normal Test</td>
</tr>
<tr>
<td>Red</td>
<td>Abnormal Test</td>
</tr>
</tbody>
</table>

Table 4.2: PCA labels within visualisation of normal and abnormal handpump conditions.
Figure 4.3: 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 4.3.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black *</td>
<td>Deep Normal</td>
</tr>
<tr>
<td>Red ○</td>
<td>Deep Abnormal</td>
</tr>
<tr>
<td>Green △</td>
<td>Shallow Normal</td>
</tr>
<tr>
<td>Blue +</td>
<td>Shallow Abnormal</td>
</tr>
</tbody>
</table>

Table 4.3: PCA labels within visualisation of normal and abnormal handpump conditions for different operating depth: (a) deep (46 m) and (b) shallow (15 m).
4.3 Performance of Selected Vibration Features

The receiver operating characteristic (ROC) is used to compare the performance of the LR classifier for different feature vectors to verify CM reliability. This metric compares the actual and predicted outputs for each class. For the case of a two-class classifier, the outcomes are either labeled as positive (normal), $p$, or negative (abnormal), $n$. Negative outcomes incorrectly classed as positive are called false positives. Conversely, positive outcomes incorrectly classed as negative are called false negatives. Positive outcomes are those that are predicted correctly.

The true positive rate (TPR), or sensitivity, of a classifier is defined to be the probability of detection such that:

$$TPR = \frac{\sum TruePositive}{\sum ConditionPositive}$$ \hspace{1cm} (4.2)

The false positive rate (FPR), or fall-out, is defined to be the probability of a false alarm such that:

$$FPR = \frac{\sum FalsePositive}{\sum ConditionNegative}$$ \hspace{1cm} (4.3)

Optimising the area under the ROC (AUC) will maximise pump failure detection while simultaneously minimising false alarms, which can be costly real-life. In the ideal case, the classifier would be very sensitive (TPR = 1) with no false alarms (FPR = 0). The ROC plot for the LR classifier trained using the different feature vectors can be seen in Figure 4.4.

![Figure 4.4: ROC plot comparing the performance of the LR classifier for different features, $FV_1 - FV_6$.](image)

In Table 4.4 it is shown that for all the proposed feature vectors, the LR classifier does not sacrifice performance significantly compared to the more complex SVM. Most notable is the advantage of minimal computational resources required by the LR classifier which is far more suitable for an embedded application than the
Table 4.4: Comparison of LR and SVC classifier accuracy scores for different feature vectors, $FV_1 - FV_6$, respectively.

<table>
<thead>
<tr>
<th>Feature Vector</th>
<th>Logistic Regression</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy (%)</td>
<td>AUC</td>
</tr>
<tr>
<td>$FV_0$</td>
<td>91.9</td>
<td>0.75</td>
</tr>
<tr>
<td>$FV_1$</td>
<td>73.9</td>
<td>0.74</td>
</tr>
<tr>
<td>$FV_2$</td>
<td>72.6</td>
<td>0.67</td>
</tr>
<tr>
<td>$FV_3$</td>
<td>72.5</td>
<td>0.62</td>
</tr>
<tr>
<td>$FV_4$</td>
<td>75.3</td>
<td>0.77</td>
</tr>
<tr>
<td>$FV_5$</td>
<td>72.8</td>
<td>0.75</td>
</tr>
<tr>
<td>$FV_6$</td>
<td>74.0</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Table 4.5: Comparison of classifier accuracy scores for deep and shallow.

<table>
<thead>
<tr>
<th>Feature Vector</th>
<th>Shallow</th>
<th>Deep</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy (%)</td>
<td>AUC</td>
</tr>
<tr>
<td>$FV_0$</td>
<td>87.0</td>
<td>0.93</td>
</tr>
<tr>
<td>$FV_1$</td>
<td>76.4</td>
<td>0.84</td>
</tr>
<tr>
<td>$FV_2$</td>
<td>77.3</td>
<td>0.85</td>
</tr>
<tr>
<td>$FV_3$</td>
<td>65.8</td>
<td>0.71</td>
</tr>
<tr>
<td>$FV_4$</td>
<td>80.6</td>
<td>0.88</td>
</tr>
<tr>
<td>$FV_5$</td>
<td>78.3</td>
<td>0.86</td>
</tr>
<tr>
<td>$FV_6$</td>
<td>79.3</td>
<td>0.88</td>
</tr>
</tbody>
</table>

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.

The change in classifier performance for grouping according to depth is shown in Table 4.5 for feature vectors $FV_0 - FV_6$. The values in boxes show that the best performing feature vector in the case of shallow and deep wells, respectively, improve on the highest accuracy for case of mixed wells in Table 4.4. The classifier performance improves for both shallow and deep well grouping, however, the improvement is much greater for shallow wells. This may be because feature vector $FV_4$, uniform sampled across the entire frequency range, is more favourable for handpump condition monitoring at shallow wells. It could be that another feature vector, such as $FV_5$, is preferred for deeper wells. Similarly, it can be seen in Figure 4.5 that the sensitivity of the classifier improves for data sets grouped by depth compared to the mixed data set.

The final outcomes of the trained LR classifier can be seen in Figure 4.6. Figure 4.6b shows the outcomes for a model considering only data from handpumps operating at a similar depth (>25 m) which shows higher sensitivity than the outcomes in Figure 4.6a when features of deep and shallow pumps are mixed. In both cases, however, there is a clear change in trend of predictions for normal and abnormal conditions, showing that a LR classifier can be used to classify the condition of a handpump given the selected feature set.
Figure 4.5: ROC plot comparing the performance of the LR classifier for $FV_4$ grouped by depth.

Note that there is no ordering of data in figure 4.6. The horizontal axis shows the index of feature vectors, which are independent and which are not consecutive in time.

Figure 4.6: Predicted (black :) and actual (blue) outputs of the LR classifier for $FV_4$ for (a) mixed depth and (b) grouped by depth, respectively.

4.4 Conclusion

This chapter validated the accuracy of a LR classifier to distinguish between normal and abnormal handpump conditions for the proposed feature vectors. A SVM classifier was used as gold standard to compare the performance of the LR classifier. Given the data set and features considered, the LR classifier was able to perform condition monitoring with a reasonable level of sensitivity, AUC = 0.77, using unseen data with mixed depth characteristics. It was shown that eliminating deterministic features, such as operating depth of the handpump, that can be learned by the model improve the performance of the trained classifier. Finally, the LR classifier performed condition monitoring with a satisfactory level of sensitivity, AUC = 0.84 and 0.87, for deep and shallow operating handpumps, respectively.
Chapter 5

Discussion and Future Work

5.1 Conclusions

The results of the investigation described by this report have shown that it is possible to determine, without a priori knowledge, the condition of a rural handpump based on the vibrations signatures collected from a sensor retrofitted inside the handle. However, sensitivity and accuracy could be improved by including knowledge of the operating depth of the pump. This research used Fourier analysis to test six proposed sets of feature vectors that describe the characteristics of the data. It was determined that selecting 20 points, uniformly sampled across the entire frequency range, was able to describe the variability in the data set, better than any of the other representations that we considered.

Furthermore, this research has shown that a lightweight LR classifier can be used to classify vibration signatures as being associated with either normal or abnormal handpump conditions. Together with the background research in Chapter 2, the research presented in this report demonstrated the importance of condition monitoring to ensuring stable systems in rural water supply networks, which directly impact the health, economics, education, and well-being of large communities across rural sub-Saharan Africa. The initial investigation presented in this report provides sufficient evidence, based on our positive results, to support further research and ultimately provide a large-scale implementation of a health-monitoring system for handpumps.

5.2 Improved Condition Monitoring of Handpumps

5.2.1 Validating Performance

A limited data set was used for this initial investigation. Although we concluded that the general framework for condition monitoring of handpumps has been successfully illustrated, the values of the weights associated with the LR classifier should be validated using a much larger data set, containing examples of all the failures types proposed in this study.
In the immediate future, this can be overcome by incorporating the data from five long-term loggers which were installed in Kenya earlier this year. However, a planned field trip to install a further 20 loggers is scheduled for February 2017. During this trip, short-term data for seeded faults will be collected to ensure that we have sufficient representation of abnormal conditions according to the specified classes described by this report.

5.2.2 Specifying Abnormality

Research to date has focused on the first level of identification to distinguish between normal or abnormal conditions. Additional examples of abnormality, obtained through seeded fault experiments for examples of failures that have already occurred and like examples from the long-term loggers leading to failures, may be exploited for the diagnosis of handpump abnormality at the second level to specify the type of failure. This would not only enable quantified insights to inform handpump development strategies, but also enable local service delivery to apply strategic resource allocation to increase handpump maintenance efficacy.

5.2.3 Improved Modelling

The LR classifier can be extended to incorporate the feature selection step by applying a regularisation term to the likelihood function used to construct the model. For example, the LASSO framework [78] attempts to shrink weights for uninformative input variables to zero, effectively removing them from the model. Additionally, we will evaluate other computationally-efficient classifiers, such as linear state-space approximations of Gaussian processes [79].

5.2.4 Data Warping

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 prominent spectral bands that appear at different frequency ranges depending on the depth of the well, introduces the need for including a priori knowledge about the system. 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 independent of well depth. For large scale deployment, this would greatly simplify the roll-out of condition loggers in the field.

5.2.5 Improving Sensor Hardware

This investigation and proposed future work surpass the original purpose of the Water Point Transmitters, which were designed to collect hourly estimates of the volume of water that has been pumped (by counting the strokes of the handle). Even though the most basic processing methods were selected for each step of
this investigation, it may prove too computationally heavy for the 8-bit processor used in the existing loggers. Ongoing hardware improvements involve upgrading the existing hardware with a 32-bit microprocessor.

The proposed large-scale roll out of the sensors across multiple geographical regions requires the use of standardised parts, such as replacing the existing lithium-ion batteries with more readily-available “AA” batteries. The use of off-the-shelf components, particularly those readily available in rural regions, will ensure the longevity and future maintenance of the project. However, smaller batteries, in combination with a larger microprocessor, are expected to reduce the battery-life of the system; making an ever stronger case for the use of extremely light-weight processing methods.

Finally, the several instances of failure that have occurred with the current loggers suggests that the current enclosure design is not suited for frequent modifications. Most notably, the collection of the data on a monthly, rather than yearly, basis (as is now being performed for those pumps with continuous-monitoring loggers) means that faster wear of the logger enclosure seals will occur, resulting in water-damage of the hardware. This enclosure is now being used beyond the original design specification. Re-design is currently under way.

5.2.6 Risks related to improved handpump condition monitoring

Obtaining access to a large data set that contains a spread of failure types is the most pressing requirement to ensure improved condition monitoring. This also involves the greatest area of technical risk. This is mitigated by the fact that we are installing a long-term logger at pumps with high usage levels to ensure a higher probability of capturing potential failure events.

Furthermore, working in close collaboration with a local mechanic to respond to failure reports will mitigate the risk of missing failure events; this also gives us the opportunity for rapid collection of relevant data. This was previously only undertaken for two days. During the next field trip (scheduled for February, 2017), recording reported faults will be performed for at least 5 to 7 days.

Validating prediction performance could prove to be a challenge. This can be overcome by using of existing data sets and seeded failures can be used to test the models. Moreover, the recent installation of a handpump in Oxford offers a test-bed for testing between field work, which will substantially contribute towards de-risking the considerations described above.
5.3 Cloud-based Platform for Data Analysis

Without a platform to host the data (and to store, and post-process, early warning flags generated by the on-board condition monitoring loggers) it is very difficult to reach the potential and scale that we require. Subsequent research will focus on developing a dynamic framework for measuring, processing, and transmitting data between sensors 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 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 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. An overview of the proposed dynamic cloud-based platform is shown in Figure 5.1.

![Figure 5.1: Cloud-based platform overview. [Clifton, Howey 2016, unpublished]](image)

In this dynamic platform, the sensor, as first point of contact with the system, will continue to perform the lightweight condition monitoring introduced in this study. Given the likelihood that the estimate is accurate, the sensor could adjust the amount of data transmitted to the central servers accordingly. For higher likelihoods, it would only need to transmit accuracy scores, but as the likelihood of the estimated accuracy decreases, sending either features or raw data to a central server via the cloud-based system will allow more heavyweight estimation algorithms to be performed.
This will require research into the “optimal” means of representing the data, which should increase in fidelity as the risk of deterioration of the handpump is deemed to be increasing. Research into the “optimal” means of 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 Bayesian Gaussian processes in the cloud, is a novel research field and which is applicable to other cognate fields.

5.3.1 Expanding into New Regions

A cloud-based platform would enable “plug-and-play” options for data to be transmitted from different locations and subsequently pump types other than the Afridev, as described in section 1.3.1. This will enable the design of a health-monitoring system for handpumps which is universally applicable to maximise the potential impact for the rural poor. Figure 5.2 shows projected clusters of recorded vibration signatures from three different pump types, commonly used in different countries, that are represented in previous data sets of this study.

Figure 5.2: PC projection of different pump types: (1) Afridev (blue ◦), (2) Samrat (red x), and (3) India MKII (black ◦).

5.3.2 Risks of a Cloud-Based Platform

Setting up a central server and establishing the relevant device-cloud communications require expert skills and can introduce the potential for delay. However, by collaborating with relevant partners to use an existing M2M embedded system will mitigate the risk of establishing initial prototypes. Expanding into new regions requires expert local knowledge and trusted partners. Our close collaboration with the Oxford School of Geography ensures that, as in Kenya, there is an opportunity to build on existing networks.
5.4 Water security risk assessment tool

Water security is central to many development goals directly related to poverty reduction in Africa and Asia. As such, measuring poverty from water-related changes introduces a multi-dimensional framework for poverty assessment that could overcome the limitations of existing methods, such as absolute (food basket) or relative (USD1.25 per day) poverty metrics.

Long-term, continuous monitoring of handpump conditions introduces a novel indicator for access to water. Incorporating other dynamic trend data, such as hydrological or social indicators, could result in more robust water security and poverty metrics to stimulate evidence-based practitioners and policy action. This will build on previous work which showed that this high-frequency data from the handpump sensor could be used to estimate the aquifer levels at the handpump [18, 37].

This final aspect of the work can examine data-fusion methods for combining socio-economic data and the output of existing hydrogeological models, with the aim of predicting risk of water insecurity. Conventional algorithms (e.g., SVMs) will be compared with Bayesian Gaussian processes.

5.4.1 Risks of a water security risk assessment tool

Data related to sensitive socio-economic and health indicators are often privately owned and difficult to access. Working in close collaboration with the Oxford School of Geography will allow access to the in-house expertise and data sets already available within this department.

5.5 Other Risks

Political instability and unrest in Kenya pose a continuous threat to preventing in-country field work and hamper subsequent data collection. We frequently monitor FCO advice to ensure track any changes that might impact the proposed research plan. Regular contact and close collaborations with local partners could overcome short-term travel restrictions. Fortunately, handpumps are widely used across sub-Saharan Africa and identifying an alternative test site could overcome long-term instability. Finally, the set-up of a handpump in Oxford will ensure that research can continue unhindered.
## 5.6 DPhil Timeline

<table>
<thead>
<tr>
<th>Project Area</th>
<th>2017</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Condition Monitoring (Afridev)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop Optimal Binary Classifier</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Build loggers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implementation and field testing (Kenya)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data collection for failure cases (Kenya)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improve condition monitoring</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cloud-Based Platform</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Enclosure Design/Prototype</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop algorithms for predictive condition monitoring</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloud-Based Software Field Testing</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Water Security Risk Assessment Tool</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aquifer Estimation Long Term Data Collection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop Optimal Aquifer Estimation algorithm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aquifer Estimation Implementation (Oxford)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Field Testing (Bangladesh/Kenya)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confirmation of Status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writing Thesis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.3: Proposed Gantt Chart of DPhil Project Timeline.
Bibliography


