FEATURE SELECTION AND ADAPTIVE THRESHOLD
FOR AUTOMATED CAVITATION DETECTION IN
HYDROTURBINES

by

Seth W. Gregg
A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Mechanical Engineering).

Golden, Colorado
Date _______________________

Signed: _______________________
Seth W. Gregg

Signed: _______________________
John P.H. Steele
Thesis Advisor

Signed: _______________________
Douglas L. Van Bossuyt
Thesis Advisor

Golden, Colorado
Date _______________________

Signed: _______________________
Dr. Greg Jackson
Professor and Head
Department of Mechanical Engineering
ABSTRACT

Hydroturbines produce 6.3% of all electrical generation and 48% of renewable energy in the United States of America. While hydro power plants have existed for well over 100 years, cavitation damage on hydroturbine runners remains as an expensive problem that reduces power production and shortens the life of the turbine. Hydroturbine operators who wish to perform cavitation detection and collect intensity data for estimating the remaining useful life (RUL) of the turbine runner face several practical challenges related to long term cavitation detection. This thesis presents both a method for comparing and evaluating cavitation detection features as well as a method for creating adaptive cavitation thresholds and automating the cavitation detection process.

The method for cavitation feature selection can be used to quickly compare features created from cavitation survey data collected on any type of hydroturbine, sensor type, sensor location, and cavitation sensitivity parameter (CSP). Although the cavitation feature selection process is based on manual evaluation and knowledge of hydroturbine cavitation, the use of principal component analysis greatly reduces the number of plots that require evaluation. A case study based on data taken from a production hydroturbine is used to demonstrate the method and the results provide a clear ranking of the preferred sensors, sensor placements, and CSPs for the hydroturbine - thus demonstrating the usefulness of the method.

The second method presented in this thesis addresses several challenges encountered when detecting cavitation for long periods of time – a prerequisite to developing a data-driven method for estimating cavitation erosion rates. First, adaptive cavitation thresholds are generated by collecting sensor data from a hydroturbine ramp-down, then creating CSPs from the data and calculating the Mahalanobis distance (MD) to create clear separation between the healthy running state and conditions where the hydroturbine is experiencing
cavitation. Next, in order to automate the cavitation detection process, the cavitation threshold is used to create class labels for the ramp-down data which is then used to train a supervised learning algorithm for classifying cavitation from sensor data. Although domain knowledge is still required to select appropriate CSPs, the remainder of the process can be automated by applying unsupervised learning to label the training set. This method is also demonstrated utilizing data collected on production hydroturbines in a power plant environment. The results of the case studies indicate that the fully automated process for selecting cavitation thresholds and classifying cavitation performed well when compared to manually selected thresholds. Our methods provide hydroturbine operators and researchers with a clear and effective way to perform automated cavitation detection while also laying the groundwork for determining RUL in the future.
TABLE OF CONTENTS

ABSTRACT ................................................................. iii
LIST OF FIGURES .......................................................... viii
LIST OF TABLES ............................................................. x
LIST OF SYMBOLS .......................................................... xi
LIST OF ABBREVIATIONS .................................................. xiii
ACKNOWLEDGMENTS ....................................................... xiv
DEDICATION ................................................................. xv

CHAPTER 1 INTRODUCTION ............................................... 1
  1.1 Thesis Organization and Research Contribution ......................... 4

CHAPTER 2 FEATURE SELECTION FOR MONITORING EROSIVE
CAVITATION ON A HYDROTURBINE ........................................ 6
  2.1 Abstract ............................................................... 6
  2.2 Introduction ........................................................... 6
    2.2.1 Specific Contributions ........................................... 9
  2.3 Background ........................................................... 9
    2.3.1 Cavitation ......................................................... 10
    2.3.2 Cavitation in Hydroturbines .................................... 11
    2.3.3 Cavitation Detection Features for Hydroturbines .............. 14
    2.3.4 Sensor Type and Sensor Placement ............................... 14
    2.3.5 Diagnostic Methods ............................................. 15
2.3.6 Prognostics ....................................................... 19
2.4 Methodology ...................................................... 20
  2.4.1 Data Preparation ............................................. 20
  2.4.2 Feature Analysis ............................................. 24
  2.4.3 Feature Selection ............................................. 27
  2.4.4 End Result .................................................... 32
2.5 Case Study ....................................................... 33
2.6 Discussion ....................................................... 42
2.7 Future Work ..................................................... 44
2.8 Conclusion ....................................................... 45

CHAPTER 3 FROM FEATURE SELECTION TO CAVITATION DETECTION . . . 47

CHAPTER 4 A METHOD FOR AUTOMATED CAVITATION DETECTION WITH ADAPTIVE THRESHOLDS ........................................... 50
  4.1 Abstract ......................................................... 50
  4.2 Introduction ..................................................... 51
  4.3 Background ..................................................... 54
    4.3.1 Cavitation Detection in Hydroturbines ..................... 56
    4.3.2 Instrumentation for Cavitation Detection ................... 57
    4.3.3 Cavitation Intensity ....................................... 58
    4.3.4 Mahalanobis Distance ..................................... 61
    4.3.5 Prognostics and Erosion Rate Prediction ................. 62
  4.4 Methodology .................................................... 64
    4.4.1 Select Cavitation Features ................................. 65
LIST OF FIGURES

Figure 2.1 Cavitation blade damage on a hydroturbine runner (courtesy of the U.S. Bureau of Reclamation) ................. 10

Figure 2.2 Side view of a Francis style hydroturbine with major components labeled (CC BY-SA 3.0, Voth Siemens Hydro Power Generation, n.d.) . 12

Figure 2.3 Cavitation Feature Selection Process ........................................ 21

Figure 2.4 Example of principal component scores plotted versus hydroturbine running conditions. ........................................ 27

Figure 2.5 Scree plot of PCA results on the cavitation ................................. 36

Figure 2.6 The first principal component score plot represents a mode of variance related to erosive cavitation ................................. 37

Figure 2.7 The second principal component score plot represents a mode of variance related to early developing draft tube swirl .......... 38

Figure 2.8 The third principal component score plot represents a mode of variance related to draft tube vortex collapse at high power output ........ 38

Figure 2.9 Correlation coefficients between the first principal component scores, and features based on the accelerometers and acoustic emission sensors . 39

Figure 2.10 Correlation coefficients between the first principal component scores, and features that use proximity probes or a pressure transducer . .... 39

Figure 2.11 Comparison of standard deviation around the features’ minimum CSP value ..................................................... 40

Figure 2.12 Comparison of standard deviation around the features’ maximum CSP value ..................................................... 41

Figure 2.13 Ranking of remaining features by standard deviation around minimum and maximum CSP values ................................. 41

Figure 4.1 Sensor vibration amplitude comparison from a hydroturbine ramp-down as measured in RMS ........................................... 74
Figure 4.2  Sensor vibration amplitude comparison from a hydroturbine ramp-down as measured in Mahalanobis Distance .......................... 75
Figure 4.3  Hydroturbine power versus time during the ramp-down .................. 76
Figure 4.4  Variance spectrum of all four proximity probe signals ..................... 78
Figure 4.5  Variance spectrum showing the frequency ranges used for calculating the CSP values ................................................................. 79
Figure 4.6  CSP values plotted versus ramp-down sample number ..................... 80
Figure 4.7  The manually selected, single variable cavitation threshold (dashed red line) ................................................................. 81
Figure 4.8  The multivariate cavitation threshold found through k-means clustering (dashed red line) .................................................. 82
Figure 4.9  Test cavitation data shown with correct classification labels as determined using traditional, manual analysis techniques ............ 87
Figure 4.10 Test cavitation data shown with labels predicted by the naïve classifier using the manually selected threshold .............................. 87
Figure 4.11 Test cavitation data shown with labels predicted by the naïve classifier using a threshold found through k-means clustering .......... 88
Figure 4.12 Test cavitation data shown with labels predicted by a linear SVM model trained from data labeled using a manually selected, single variable threshold and the training set PP3-CSP1, PP4-CSP1, PP1-CSP3, PP4-CSP3, PP3-CSP2 ......................................... 88
Figure 4.13 Test cavitation data shown with labels predicted by a 5th order polynomial SVM model trained from data labeled using the multivariate threshold found with k-means clustering ................. 89
LIST OF TABLES

Table 2.1  Cavitation Diagnostic Methods ........................................... 16
Table 2.2  Rule of thumb for comparing correlation coefficients ............... 29
Table 2.3  Calculations used for feature CSP values .............................. 34
Table 2.4  Cavitation sensitivity parameter details for each sensor type ........ 35
Table 4.1  Cavitation thresholds for labeling training data ........................ 81
Table 4.2  Classifier test results for single variable thresholds .................. 85
Table 4.3  Classifier test results for multivariate thresholds ....................... 86
LIST OF SYMBOLS

mean ................................................................. $\mu$
standard deviation .................................................. $\sigma$
standard deviation at the minimum CSP value ................. $s_{CSP-min}$
standard deviation at the maximum CSP value ................. $s_{CSP-max}$
variance ............................................................. $\sigma^2$
acceleration ......................................................... $a$
displacement ........................................................ $d$
period ................................................................. $T$
root mean square of a block of time series data ............... $f_{rms}$
peak value of a block of data time series data ................. $f_{peak}$
crest factor of a block of time series data ...................... $f_{CF}$
kurtosis of a block of time series data ......................... $f_{kurt}$
total cavitation intensity ......................................... $I_{total}$
Mahalanobis distance .............................................. $MD$
standardized Mahalanobis distance ............................. $\hat{x}_{MD}$
baseline (healthy) data .......................................... $X_{baseline}$
orthogonal basis matrix ........................................... $P$
principal component scores matrix .............................. $Y$
covariance matrix .................................................. $\Sigma$
cavitation feature matrix ........................................ $F$
correlation coefficients between the cavitation features and the principal component scores \[ \rho(y, f) \]

blade passing frequency \[ f_b \]

vane passing frequency \[ f_v \]

sample frequency \[ f_s \]

spectrum frequency resolution \[ f_{res} \]

normalized half spectrum \[ \hat{z} \]

normalized half spectrum matrix \[ \hat{Z} \]
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RUL</td>
<td>remaining useful life</td>
</tr>
<tr>
<td>CSP</td>
<td>cavitation sensitivity parameter</td>
</tr>
<tr>
<td>RMS</td>
<td>root mean square</td>
</tr>
<tr>
<td>MW</td>
<td>Megawatts</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PC</td>
<td>Principal Component</td>
</tr>
<tr>
<td>BPML</td>
<td>Blade Pass Modulation Level</td>
</tr>
<tr>
<td>SVM</td>
<td>support vector machine</td>
</tr>
<tr>
<td>FFT</td>
<td>fast Fourier transform</td>
</tr>
<tr>
<td>DFT</td>
<td>discrete Fourier transform</td>
</tr>
<tr>
<td>DC</td>
<td>direct current</td>
</tr>
<tr>
<td>MD</td>
<td>Mahalanobis distance</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

I wish to express my gratitude to Professor Van Bossuyt for his patience and extensive writing guidance as well as Professor Steele for his inspiration, knowledge, and belief in this research. I also wish to acknowledge the contributions of John Germann and James DeHaan for collecting the cavitation survey data and their guidance with analysis of the data.

The information, data, or work presented herein was funded in part by the Office of Energy Efficiency and Renewable Energy (EERE), U.S. Department of Energy, under Award Number DE-EE0002668 and the Hydro Research Foundation.

The information, data or work presented herein was funded in part by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes and warranty, express or implied, or assumes and legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
To my wife, Christine, who has supported me from the beginning.
CHAPTER 1
INTRODUCTION

Hydropower is the largest renewable source of electricity in the world. Many nations rely heavily on energy generated from hydraulic turbines including China, Brazil, India, France, Russia, Norway and Canada [1]. In the United States, hydropower is the largest and most mature renewable energy source accounting for 48% of all renewable energy and 6.3% of all electrical energy generated in the United States [2]. In the Northwestern region of the United States, including Washington, Oregon, and Idaho, hydropower is the primary power source accounting for over half the electrical energy generated.

Like other hydraulic machinery such as pumps, ships, and valves, hydraulic turbines are susceptible to damage caused by cavitation. Cavitation is a potentially destructive and complex phenomenon involving the formation and rapid collapse of vapor bubbles in the liquid. The vapor bubbles, or cavities, form due to local pressure drops caused by sudden changes in the fluid dynamics caused by rotating blades, sharp curves or turbulence. Once formed, the cavities can gather into clouds of vapor bubbles that periodically shed portions of the cloud and violently collapse when they reach a higher pressure region in the fluid [3]. When the vapor cavities collapse, they radiate a high energy acoustic pressure wave that can lead to pit formation and aggressive material erosion in nearby surfaces.

Cavitation events in hydroturbines can lead to erosion damage on turbine runners that reduces the life of turbine runner requiring costly maintenance and loss of power production. Despite advancements in hydroturbine design, damage caused by cavitation remains one of the primary causes of turbine failure [4–6]. This problem is highlighted by recent and ongoing cavitation studies performed by the United States Bureau of Reclamation at major hydroelectric plants in northern California and eastern Washington that have recently experienced costly cavitation damage [7, 8]. These events highlight the need to develop better
methods of detecting erosive cavitation in hydroturbines.

The journal papers contained within this research present both a structured method for choosing diagnostic indicators that are sensitive to erosive cavitation – also called CSPs – as well as a method for automating long term cavitation detection in hydroturbines. The underlying motivation for this research is to work toward the goal of developing a data-driven model for estimating the remaining useful life (RUL) of hydroturbine runners. The necessary steps to make data-driven RUL predictions for hydroturbines are as follows [9–12]:

1. Select a sensor-based cavitation detection method for identifying erosive cavitation and measuring cavitation intensity.

2. Collect cavitation intensity data for a test period that is long enough for accumulative cavitation damage to be measured.

3. Measure the runner material loss over the test period and correlate the loss with the measured cavitation intensity over the same period.

4. Create an erosion rate model to use for estimating runner RUL at any future state based on accumulated cavitation intensity.

The first three steps of the RUL prediction process have been carried out in laboratory tests, but the methods used are not feasible for many hydroturbine operators or practical in a power plant environment. Complications with data quality, sensor placement, long term robustness of the data collection hardware, and the requirement of manual interaction with the detection system have thwarted attempts to carry out similar tests on production hydroturbines. At the time of this writing, there have yet to be published results that correlate cavitation erosion rates with data taken from a production hydroturbine. The lack of widespread acceptance or implementation of cavitation monitoring for estimating erosion rates suggests the existing methods are either not effective or not accessible to most hydroturbine operators.
It is important to note that cavitation detection and intensity measurements are an important part of creating a data-driven RUL model. Hydroturbine researchers generically use the term 'cavitation detection' to refer to diagnostic methods that involve sensor measurements, signal processing, and data analysis to aid in determining when cavitation is present [13–15]. This definition, however, is ambiguous about key elements of collecting long term cavitation data for studying erosion rates. For the purposes of this work, we will divide cavitation detection into three distinct actions:

1. Applying a diagnostic method to sensor measurements to create an indicator sensitive to the onset of cavitation (a CSP) as introduced in [16].

2. Establishing a cavitation threshold (when using a single CSP) or a decision boundary (when using multiple CSPs) that is used to decide when cavitation is present.

3. Measuring cavitation intensity in a way that can be used to calculate or estimate cavitation erosion rates.

Many diagnostic methods are available to hydroturbine operators for creating a CSP [13–15, 17–19]. Selecting the right diagnostic method for a given hydroturbine is difficult since no method has been shown to be effective, practical, and affordable for every hydroturbine. In addition, cavitation intensity measurements are not directly addressed in these diagnostic methods and the action of establishing a cavitation threshold is completely ignored. This is problematic because cavitation thresholds are critical for automating cavitation detection, and intensity values are needed to correlate erosion rates with sensor measurements. It would appear that outside of the work by Dorey, et al. [10], performed in collaboration with Bourdon, et al. [9] and continued by Francois [12], cavitation diagnostics studies have focused on short term data collection and manual data analysis.

The methodology of this research addresses each of the cavitation detection actions and has been broken up into two separate journal articles to be submitted for publication. The
organization of this thesis and research contribution of the journal articles is summarized in the following section.

1.1 Thesis Organization and Research Contribution

This thesis is organized into two main chapters – Chapter 2 and Chapter 4 – which are made up of two self contained journal papers. Between Chapters 2 and 4 is a brief transitional chapter and following Chapter 4 is a summary of the future work recommendations and conclusions of both journal papers.

Chapter 2 contains the article entitled "Feature Selection for Monitoring Erosive Cavitation on a Hydroturbine” which addresses the first step of developing a data-driven model for RUL prediction as well as the first action of cavitation detection. The article presents a novel method to rapidly compare cavitation detection features and select which cavitation detection features best identify when a hydroturbine runner is experiencing an erosive cavitation event. When compared to previous research aimed at comparing sensors, sensor placement, or CSPs, [20, 21] the methodology presented in this article uses a more objective, statistics-based approach to the evaluation process. It is important to note that the method presented in this article can be used to discriminate between erosive and non-erosive cavitation which is important to the future goal of determining RUL. An added benefit of using this method is determining the most useful and cost-effective sensors for cavitation detection. This method is an important step toward full automated cavitation detection and RUL calculation that will lead to more robust automated detection that can be relied on by operators.

Chapter 4 contains the article entitled ”A Method for Automated Cavitation Detection with Adaptive Thresholds” which addresses both the first and second steps of developing a data-driven model for RUL prediction as well as all three cavitation detection actions. The article outlines a method for cavitation detection based on a machine learning framework and proposes an adaptive threshold that adjusts to operational changes of the hydroturbine from a small amount (around 90 seconds) of ramp-up or ramp-down data. The article also
introduces the Mahalanobis distance (MD) to hydroturbine cavitation detection and intensity monitoring. Recently MD has been suggested as a distance metric for identifying cavitation in hydraulic pumps, in this article MD is used as a basis for both establishing cavitation detection thresholds and tracking cavitation intensity. The cavitation detection method presented in this paper is flexible, which means it can be used with many different cavitation features, it is multivariate, which allows the user to incorporate many different CSPs, and it can be fully automated. All these factors afford the hydro plant operator flexibility in deployment to suit their own specific plant conditions and greatly increasing the likelihood of successful long-term cavitation detection and cavitation intensity monitoring.
CHAPTER 2

FEATURE SELECTION FOR MONITORING EROSIVE CAVITATION ON A HYDROTURBINE

A paper submitted to the International Journal of Prognostics and Health Management

Seth W. Gregg\textsuperscript{1}, John P.H. Steele\textsuperscript{2}, and Douglas L. Van Bossuyt\textsuperscript{3}

2.1 Abstract

This paper presents a novel method for comparing and evaluating cavitation detection features - the first step towards estimating RUL of hydroturbine runners that are impacted by erosive cavitation. The method can be used to quickly compare features created from cavitation survey data collected on any type of hydroturbine, sensor type, sensor location, and CSP. Although manual evaluation and knowledge of hydroturbine cavitation is still required for our feature selection method, the use of principal component analysis greatly reduces the number of plots that require evaluation. We present a case study based on a cavitation survey data collected on a Francis hydroturbine located at a hydroelectric plant and demonstrate the selection of the most advantageous sensor type, sensor location, and CSP to use on this hydroturbine for long-term monitoring of erosive cavitation. Our method provides hydroturbine operators and researchers with a clear and effective means to determine preferred sensors, sensor placements, and CSPs while also laying the groundwork for determining RUL in the future.

2.2 Introduction

Cavitation events in hydroturbines can lead to damage to the turbine runners and reduced RUL. Current methods of detecting cavitation events and prognosticating RUL have not been

\textsuperscript{1}Graduate student, primary researcher and Author
\textsuperscript{2}Primary Advisor
\textsuperscript{3}Co-Advisor and Corresponding Author
successful in providing hydroelectric power plant operators with meaningful information. Structured methods of data collection and feature selection as well as automated methods for cavitation detection, and RUL prediction are needed to provide plant operators with a clear view of hydroturbine health and RUL. The collection of useful data from hydroturbines is well established and the tool chain to calculate RUL is understood. However, feature selection and automated cavitation detection remain to be addressed. In this paper, we specifically examine feature selection in the larger context of calculating RUL.

Hydropower is the largest renewable source of electricity in the world. Many nations rely heavily on energy generated from hydraulic turbines including China, Brazil, India, France, Russia, Norway and Canada [1]. In the United States, hydropower is the largest and most mature renewable energy source accounting for 48% of all renewable energy and 6.3% of all electrical energy generated in the United States [2]. In the Northwestern region of the United States, including Washington, Oregon, and Idaho, hydropower is the primary power source accounting for over half the electrical energy generated.

Like other hydraulic machinery such as pumps, ships, and valves, hydraulic turbines are susceptible to damage caused by cavitation. Cavitation is a potentially destructive and complex phenomenon involving the formation and rapid collapse of vapor bubbles in the liquid. The vapor bubbles, or cavities, form due to local pressure drops caused by sudden changes in the fluid dynamics caused by rotating blades, sharp curves or turbulence. Once formed, the cavities can gather into clouds of vapor bubbles that periodically shed portions of the cloud and violently collapse when they reach a higher pressure region in the fluid [3]. When the vapor cavities collapse, they radiate a high energy acoustic pressure wave that can lead to pit formation and aggressive material erosion in nearby surfaces.

Despite advancements in runner design and cavitation resistant materials, damage caused by cavitation remains one of the primary causes of turbine failure [4–6]. This problem is highlighted by recent and ongoing cavitation surveys performed by the United States Bureau of Reclamation at major hydroelectric plants in northern California and eastern Washington.
that have recently experienced costly cavitation damage [7, 8]. These events highlight the need to develop prognostic methods for estimating the RUL of hydraulic turbine runners experiencing cavitation damage.

One starting point for estimating RUL is to calculate a cavitation erosion rate by comparing the amount of cavitation damage accumulated over a long period of time with the amount of time the turbine runner experienced cavitation over the same period. Turbine runners are inspected periodically and standard methods exist for evaluating cavitation damage [22]. Many methods of turbine cavitation event detection have been developed over the last 50 years; however, these methods are not widely used in industry for a variety of reasons including: 1) only a limited subset of cavitation events can be detected, 2) too many false positives undermine confidence in the methods, 3) methods are turbine-specific and not generalizable, and 4) installing new instrumentation to detect cavitation events is overly burdensome on hydro power plant operators especially with regards to operating budgets.

When significant cavitation damage is discovered during routine turbine runner inspections at maintenance intervals, hydro power plant operators typically perform a cavitation damage survey. The survey consists of heavily instrumenting the hydroturbine and running it through a variety of operating regimes in an attempt to understand what operating conditions lead to cavitation events that can cause turbine runner damage. After the survey is completed, the information is used to develop operating guidelines to avoid operating regions where damage can occur. While this approach works to reduce damage in the short term by avoiding operating regions that can cause damage, several problems exist with the approach including: 1) cavitation damage surveys often only examine a limited range of operating conditions available during the cavitation survey such as hydrostatic head, water temperature, and interference from sister turbines within the power plant, etc. that change seasonally or year-to-year especially due to drought conditions, 2) changes to the hydroturbine and associated equipment during repair or overhaul can change the operating regions in which cavitation occurs, 3) data is not generally collected and used beyond the cavitata-
tion damage survey to determine RUL during routine operations, and 4) due to the time consuming nature of manual comparison, a limited number of cavitation detection features are typical compared which can lead to missing cavitation events and excessive false positive identification of cavitation by not using the most effective feature.

Of specific interest to this paper is determining appropriate cavitation detection features to use on a specific hydroturbine. Many cavitation detection features have been proposed in the literature and have been used with varying degrees of success in practice; however, no single cavitation detection method is appropriate for all scenarios. The three constituent components of a cavitation detection feature include: 1) sensor type, 2) sensor placement, and 3) CSP.

2.2.1 Specific Contributions

In this paper, we present a novel method to rapidly compare cavitation detection features and select which cavitation detection features best identify when a hydroturbine runner is experiencing an erosive cavitation event. When compared to previous research aimed at comparing sensors, sensor placement, or CSPs, [20, 21] our methodology uses a more objective, statistics-based approach to the evaluation process. It is important to note here that the method presented in this paper can discriminate between erosive and non-erosive cavitation which is important in the ultimate goal of determining RUL (not addressed in this paper). An added benefit of using this method is determining the most useful and cost-effective sensors for cavitation detection. This method is an important step toward full automated cavitation detection and RUL calculation that will lead to more robust automated detection that can be relied on by operators.

2.3 Background

In this section, we present background information on cavitation damage in hydroturbines to demonstrate the need for a method to rapidly compare cavitation detection features for long term monitoring. A review of previous and current work that has attempted to address
hydroturbine cavitation damage is provided. While efforts have been made to establish reliable RUL predictions, hydro power plant operators cannot or choose not to use existing solutions. The method we present in this paper builds upon the information presented in this section toward the eventual goal of predicting RUL.

2.3.1 Cavitation

Cavitation occurs when vapor bubbles, or cavities, form in a liquid due to a local decrease in pressure below the fluid vapor pressure. In hydraulic machinery, cavitation typically develops in localized areas where a flowing liquid reaches higher than intended velocities. The liquid then becomes broken at several points and vapor cavities appear taking on different shapes depending on the structure of the flow [3]. When the vapor cavities collapse, they release a large amount of energy and can be very destructive leading to material erosion on surrounding surfaces. Consequently, cavitation and cavitation erosion is one of the most pervasive problems found in hydroturbines (see Figure 4.1), pumps, and ship propellers.

Figure 2.1: Cavitation blade damage on a hydroturbine runner (courtesy of the U.S. Bureau of Reclamation)
Cavitation damage was first noted on ship propellers in the late 1800s [23]. By the early 1900s, material research was underway to help reduce propeller damage in ocean liners caused by cavitation [24]. Soon after, Lord Raleigh published the first theoretical model analyzing the collapse of cavitation bubbles in a liquid [25] helping to explain the high pressure pulses emitted by the highly compressed bubble at the moment of collapse. Since Rayleigh’s initial models, there have been ongoing efforts to understand the bubble dynamics and wear mechanism behind cavitation in greater depth [26–30]. These studies focus primarily on the dynamics and damage caused by the collapse of single bubbles near simple, flat surfaces a situation not commonly found in hydraulic machinery.

Recent cavitation studies use experimental setups that better replicate realistic conditions of cavitation in rotating equipment [3, 31–34]. These investigations have revealed previously unseen complexity including a sheet bubble structure, periodic shedding of bubble formations, and several collapse modes that lead to varying amounts of surface damage. The complex nature of cavitation leads to difficulties in generating accurate computer models for predicting cavitation erosion [34]. Cavitation remains poorly characterized in complex flow environments which limits the ability to predict RUL of a hydroturbine runner using physics-based simulations.

2.3.2 Cavitation in Hydroturbines

Hydroturbines create energy by taking advantage of water falling between reservoirs at different elevations. The available water head and flow determine the design of the hydroturbine and play a large role in determining if cavitation will develop during turbine operation [35].

Large power plants typically have Kaplan or Francis style turbines. The major difference in these two styles of turbine is in the design of their impeller-like rotor called the runner. Kaplan turbine runners are shaped like ship propellers and are used when low water head is available. Francis turbine runners are similar to Francis vane pump impellers and are used for medium to high head applications [36]. Both turbine types are susceptible to cavitation;
however, the location and type of cavitation typically observed can vary slightly between turbine types [13]. Pump-turbines are becoming increasingly common and have a runner design similar to Francis turbines, but with the added advantage of being able to be run in reverse as a pump. Pump-turbines are susceptible to cavitation in either pump or turbine modes of operation [14, 37].

Important hydroturbine components are shown in Figure 4.2. Water flows from the inlet side of the runner into the draft tube. The amount of power produced by the hydroturbine is determined by the amount of water flowing through the impeller which is controlled by pivoting the inlet guide vanes open or closed. The area of highest concern for cavitation damage is on the blades of a turbine runner. For large turbines, the runner can be from 2 – 9 meters in diameter and is very expensive to replace or repair [38].

![Figure 2.2: Side view of a Francis style hydroturbine with major components labeled (CC BY-SA 3.0, Voth Siemens Hydro Power Generation, n.d.)](image)

Hydroturbines can be affected by several types of cavitation which are characterized by the operating conditions that cause cavitation to occur and the location where erosion damage appears. Cavitation types that lead to erosion damage on the runner include leading edge, traveling bubble, inter-blade vortex and tip vortex cavitation. Other types of cavitation including draft tube swirl can cause high vibration, loss of efficiency and fluctuations in power.
production, but typically do not lead to erosion damage [13].

Water head at the inlet and draft tube along with flow rate through the impeller dictate the operating conditions of a hydroturbine. Hydroturbines are designed to run free from cavitation; however, the complex nature of cavitation makes designing and constructing a turbine not prone to cavitation under at least some conditions very difficult. Available inlet head may also change sufficiently to lead to unexpected cavitation damage through seasonal reservoir variations or large climactic events such as drought or flood. To prevent catastrophic failure, hydroturbine runners are inspected periodically for cavitation erosion damage and repaired as necessary. When severe damage is found, a cavitation survey\textsuperscript{4} is performed to map out operating ranges where cavitation is occurring [8, 15].

During a cavitation survey, the hydroturbine is temporarily instrumented with sensors to detect vibration and acoustic emissions of the shaft and surrounding structure as well as pressure changes in the penstock. Next, the hydroturbine is run at incrementally increasing flow rates while sensor data is collected at each operating condition. The sensor data is then analyzed to identify operating conditions where cavitation is occurring so restrictions can be placed on the operating range of the turbine. In cases where permanently installed sensors and data collection equipment is installed, the cavitation survey can also be used to establish threshold values for on-line cavitation monitoring and be a basis for monitoring the condition of the turbine runner [15, 39].

Cavitation surveys provide valuable information, but the operating conditions that can be observed during a cavitation survey is limited by the available inlet and draft tube head at the time of the survey. Additionally, hydroturbines often operate in parallel with other turbines and the operating points of these units can affect the survey findings. Data from a cavitation survey is only a snapshot of current operating conditions and cavitation zones rather than a long-term operating plan, though in our experience many hydroturbine operators must treat

\textsuperscript{4}Note that the term "cavitation survey" is used internally at the Bureau of Reclamation and is used in this paper to describe a study conducted on a hydroturbine to identify operation conditions where cavitation is likely to occur.
2.3.3 Cavitation Detection Features for Hydroturbines

We define cavitation detection features to consist of three components including: 1) sensor type, 2) sensor placement, and 3) CSP. The process of extracting the appropriate information to monitor (feature selection) is a key component of both diagnostics and prognostics. Feature selection for cavitation monitoring on a hydraulic turbine involves choosing sensors, sensor placement, data collection equipment, and a CSP as well as considering location of the cavitation on the runner, influence of the turbine structure on the sensor signal, the number of turbines being operated, and the overall design of the hydroelectric plant. The sheer number of factors that influence cavitation feature selection for hydroturbines means a single cavitation detection feature is not necessarily applicable to multiple plants, turbines, and even operating conditions of the same turbine. In the below subsections, we discuss the three constituent components of cavitation detection features.

2.3.4 Sensor Type and Sensor Placement

The most common sensors used for cavitation diagnostics are accelerometers, which produce a signal proportional to acceleration, and acoustic emission sensors, which produce a signal proportional to the amplitude of small stress waves that travel through a material. Both sensors are based on piezoelectric sensing elements and are able to record high frequency events. Accelerometers used for cavitation diagnostics typically have a linear frequency response from 3 – 40,000 Hz while the acoustic emission sensors used respond well between 40 – 400 kHz. In order to take advantage of high frequency sensors, signal recording equipment must be able to record the data at a high sampling rate typically around 1 MHz. Butterworth filters are also commonly applied to recorded data in order to remove spurious signals and frequency content beyond the useful range of the sensor [8, 13–15].

Other sensors that are less frequently used for cavitation diagnostics include hydrophones and high frequency pressure sensors, sensitive to pressure events between 2 – 180,000 Hz,
and proximity probes that measure shaft movement from 0 – 10 kHz [19]. For the most part, proximity probes are used for detecting lower frequency cavitation events typical of draft tube swirl or non-cavitation related faults such as an unbalanced or misaligned hydroturbine shaft. A recent exception to this is the work by Pennacchi, et al. [40] showing the potential for cavitation detection using synchronous averaging and spectral kurtosis on low frequency proximity probe signals in a Kaplan turbine.

Typical sensor locations to monitor cavitation on hydroturbines include: 1) upper and lower turbine bearings, 2) the stem of an inlet guide vane (also called a wicket gate), and 3) the draft tube wall. In experimental setups, sensors are sometimes attached to other locations including the hydroturbine case, test stand frame, or directly to the hydroturbine shaft. Sensor placement and orientation on one of the above identified locations can significantly impact the signal response as [21] shows.

### 2.3.5 Diagnostic Methods

A summary of several options available to hydroturbine operators for cavitation diagnostics is shown in Table 4.1. To be practical for long term cavitation monitoring and RUL estimation, a diagnostic method should: 1) be effective for the turbine configuration and cavitation type, 2) produce a CSP value that correlates with cavitation erosion rates, 3) consist of sensors and hardware that are reasonable in cost and practical for installation in a power plant environment. Selecting the right diagnostic method for a given hydroturbine is difficult since no method has been shown to meet all these requirements in every situation. In addition, direct comparison of diagnostic methods in literature is rare as research instead focuses on demonstrating the efficacy of a newly proposed technique.
Table 2.1: Cavitation Diagnostic Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Application</th>
<th>Sensors</th>
<th>Signal Processing</th>
<th>CSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Varga et al. 1969</td>
<td>Laboratory turbine and pump test rig</td>
<td>1 condenser microphone, 1 accelerometer</td>
<td>Spectrum analysis 20-40,000 Hz</td>
<td>Relative average noise and overall acceleration</td>
</tr>
<tr>
<td>Bajic 2002</td>
<td>17 MW Francis turbine</td>
<td>20 acoustic emission sensors, 1 on each guide vane</td>
<td>1) normalized power spectra 0.2 kHz - 1 MHz across different turbine power output conditions and 2) polar modulation curve plots from each sensor</td>
<td>Maximum signal amplitude in RMS</td>
</tr>
<tr>
<td></td>
<td>73 MW Kaplan turbine</td>
<td>5 accelerometers, 2 on the lower guide bearing and 3 on the thrust bearing</td>
<td>1) power spectra of raw data 0 - 10,000 Hz and 2) power spectra of demodulated band-pass filtered data 5-10 kHz</td>
<td>Maximum signal amplitude in RMS</td>
</tr>
<tr>
<td>Escaler et al. 2006</td>
<td>11 MW Francis turbine</td>
<td>3 accelerometers, 2 on the lower guide bearing, 1 on the inlet guide vane. 1 acoustic emission sensor on the lower guide bearing</td>
<td>1) power spectra of raw data 0 - 20,000 Hz and 2) power spectra of demodulated band-pass filtered data 15 - 20 KHz</td>
<td>Maximum signal amplitude in RMS</td>
</tr>
<tr>
<td></td>
<td>65 MW Francis turbine</td>
<td>3 accelerometers, 2 on the inlet guide vanes, 1 on the lower guide bearing. 1 acoustic emission sensor on the lower guide bearing</td>
<td>1) Overall RMS vibration up to 49 kHz, 2) power spectra of raw data 0-50,000 Hz, and 3) power spectra of demodulated band-pass filtered data 30-50 kHz</td>
<td>Maximum signal amplitude in RMS</td>
</tr>
<tr>
<td>Method</td>
<td>Application</td>
<td>Sensors</td>
<td>Signal Processing</td>
<td>CSP</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Rus et al. 2007</td>
<td>Laboratory Kaplan turbine test rig</td>
<td>1 accelerometer, 1 acoustic emission sensor, and 1 hydrophone on the test rig suction tube</td>
<td>1) Power spectra of demodulated band-pass filtered data (several band-pass filter settings used)</td>
<td>Sum of the blade pass modulation level (BPML) normalized by the maximum value</td>
</tr>
<tr>
<td>Cencic, Hocevar, and Sirok 2014</td>
<td>185 MW Pump-Turbine</td>
<td>2 accelerometers, 1 on the lower bearing, on the inlet guide vane. 1 acoustic emission sensor on the bearing. 1 pressure sensor on draft tube wall.</td>
<td>1) normalized power spectra 2) overall RMS value of 5 different band-pass filtered frequency ranges, 3) selection of band-pass filtered value by highest coefficient of determination</td>
<td>Discharge coefficient cavitation estimator (based on band-pass RMS amplitude, water flow rate, runner discharge diameter, and rotational speed)</td>
</tr>
<tr>
<td>Escaler et al. 2015</td>
<td>26 MW Francis turbine with leading edge cavitation and draft tube swirl</td>
<td>4 accelerometers, 2 on the lower guide bearing, 1 on the guide vane, 1 on the draft tube wall. 1 acoustic emission sensor on the lower guide bearing, 1 pressure sensor on the draft tube</td>
<td>1) power spectra of raw data 0 - 45 kHz and 0 - 20,000 kHz, 2) RMS level of band-pass filtered data, 3) power spectra of demodulated band-pass filtered data</td>
<td>Power estimation of modulating frequencies</td>
</tr>
</tbody>
</table>
Bajic [18] promotes the use of a multidimensional technique that he states is effective for all hydroturbines and cavitation types; however, to be implemented it requires an acoustic emission sensor be installed on every inlet guide vane stem. Hydroturbines commonly have 20 or more inlet guide vanes and installation of this number of sensors is impractical in most hydro plants. The large quantity of data produced from this number of sensors also means analysis is a time consuming process and long term collection and storage of data is cumbersome. Escaler and Rus [13, 15, 19] show good cavitation detection results by first band-pass filtering the sensor signals, then using the power spectrum of the demodulated signal to select frequency peaks sensitive to leading edge cavitation. Escaler suggests this technique is widely applicable; however, Cencic [14] claims the methodology is not often practical because, to be effective, the sensors must be placed in largely inaccessible locations.

Evaluation of the root mean square (RMS) amplitude of the sensor signals is the most widely used technique for cavitation diagnostics. Overall RMS calculated from raw sensor signals is sensitive to cavitation events, but also picks up unwanted contributions from other machinery faults or outside sources of noise. Two methods are suggested for reducing the effects of unwanted contributions to the sensor signals for RMS calculations: 1) apply a high-pass filter to sensor signals to remove amplitude contributions from turbine running speed and low frequency faults, and 2) apply a band-pass filter to the signals and calculate RMS amplitude from a narrow frequency range that is only sensitive to cavitation events. Some combination of high-pass and band-pass filtering is used in every cavitation diagnostic method we reviewed. Escaler et al [15] use a band-pass filter range of 15 – 20 kHz for accelerometers and 40 – 45 kHz for acoustic emission sensors to reduce the influence of outside noise. Cencc et al [14] evaluated five frequency ranges for their response to cavitation over several operating conditions, and ultimately found the frequency range between 22 – 26 kHz for accelerometers and above 50 kHz for acoustic emission sensors showed the best sensitivity. Bajic [18] suggests different frequency ranges can be used to detect different types of cavitation, but does not suggest that a single best frequency range can be assumed
before analysis of the cavitation survey data.

2.3.6 Prognostics

Prognostics is the process of using a systems state and degradation rate to predict the health of the system at a future state [41]. Prognostic methods typically utilize historical condition monitoring data combined with either physics-based or data-driven models to predict the future trend of the condition monitoring data and estimate the remaining useful life [42]. While physics-based approaches can provide accurate future health information, adequate models of cavitation erosion in complex hydraulic environments such as hydroturbines do not exist or have not been validated outside of laboratory environments [31, 34, 43]. We advocate for a data-driven prognostic method to estimate turbine runner erosion rates and RUL.

Existing attempts at data-driven hydroturbine cavitation erosion prognostics or RUL prediction have not been fully successful for a variety of reasons. Francois [12] reports on Hydro Quebec’s attempts at erosion rate estimation that have produced no published results at the time of this writing. Wolff, Jones and March [44] collected data between hydroturbine runner inspections in an attempt to establish an erosion rate model based on inspection reports but insufficient data has stymied this effort.

Several researchers have suggested that their cavitation detection features and methodologies may possibly be used for erosion estimation or RUL prediction but these researchers have yet to demonstrate a successful implementation in a hydroturbine operating at a hydroelectric plant [14, 15, 31]. To our knowledge, no one has publicly published a successful hydroturbine cavitation erosion prognostic or RUL prediction method. In addition, no research group has addressed cavitation detection feature selection, instead choosing a cavitation detection feature a priori for their studies. To date, no one has attempted to address feature selection for cavitation detection in a repeatable, objective approach appropriate for hydroturbines. The method presented in this paper attempts to provide a repeatable, structured approach for comparing and selecting cavitation detection features for hydroturbines.


2.4 Methodology

This section presents a method for determining the best cavitation detection feature(s) for a hydroturbine that experiences cavitation damage. The method is broken into three parts including: 1) Data Preparation, 2) Feature Analysis, and 3) Feature Selection. Within each part, several steps are presented that guide the practitioner through down-selecting from all of the possible cavitation detection features to the few that: 1) provide the best sensitivity to erosive cavitation, 2) the least false alarms, and 3) the most practical to implement given specific hydro plant and hydroturbine configuration. Figure 4.3 graphically shows the method.

2.4.1 Data Preparation

Data preparation is comprised of three steps that take place after a cavitation survey has been performed: 1) CSPs are calculated, 2) CSPs are organized into columns of a feature matrix, and 3) the columns of the feature matrix are normalized. The focus of this method is not on cavitation survey data collection techniques; further information can be found in [13, 19].

Step 1: Calculate Cavitation Sensitivity Parameters:

In this step, data are collected from the cavitation survey including sensor types, sensor placements, and operating conditions; matched with diagnostic methods found in Table 4.1; split into CSPs specific to each potential combination of the above listed variables, and then CSPs are calculated to feed into the matrix developed in Step 2. Most CSPs listed in Table 4.1 can be used with any combination of sensor type and sensor location. For instance, a cavitation survey that uses accelerometers located on the lower guide bearing and inlet guide vanes could use overall RMS levels [17], band-pass filtered RMS levels [14], and power estimates of modulating frequencies [15] for CSP values. It is important to note that in order for the columns of the feature matrix to have the same length, the same quantity of CSP values must be calculated for each running condition where a running condition is defined.
Figure 2.3: Cavitation Feature Selection Process
as the specific hydroturbine power output (usually denoted in Megawatts (MW))\textsuperscript{5}. However, different operating conditions can have different quantities of CSP values, if desired. If using this method to evaluate experimental features, our recommendation is to include at least one commonly accepted sensor type, location, and CSP value preferably a combination listed in Table 4.1 - to ensure useful comparative results. A rigorous feature selection process, the primary goal of this method, requires as many features as are practical to compare. For example, in the case study presented in the next section, we demonstrate the method using 61 features derived from 6 unique CSPs and 17 unique operating conditions.

*Step 2: Form the Cavitation Feature Matrix from the Cavitation Sensitivity Parameters:*

Each combination of sensor type, sensor location and CSP value (identified in Step 1) is a unique feature that will be evaluated. In this step, we combine these features into a matrix with a format conducive to the mathematical methods we use for feature analysis and selection in later steps. Features are first organized by grouping CSP values from the same operating condition (in most cases, operating condition refers to hydroturbine power output, but it could also include other variables such as head, efficiency, or number of concurrently running hydroturbines), and then creating a column vector, $f$, by concatenating the groups by increasing power output. In this way, the data can be viewed and manipulated in the operating domain of the turbine instead of in the time domain. The features are then combined into a cavitation feature matrix, $F$, where each feature becomes a vertical column of block matrices as shown in 2.1.

The column vectors of each block $f_1, \ldots, f_n$ contain the values of the features calculated from the different operating conditions of the turbine. The number of columns, $n$, is determined by the number of features being compared. The number of rows in each block, $c$, is determined by how many feature values are calculated for each operating condition. The

\textsuperscript{5}Note that while the vast majority of cavitation surveys are conducted over the same underlying operating conditions (e.g. hydrostatic head, water temperature, turbidity, other turbines active at the plant, etc.), it is possible to conduct a cavitation survey that varies more than the hydroturbine power output. In this case, multiple running conditions for each power output exist and each is treated as an individual operating condition.
number of data blocks, $s$, is determined by the number of operating conditions the turbine is run under during the cavitation survey. In this way, a column vector spanning all of the blocks contains feature values ranging across all of the operating conditions the hydroturbine experienced during the cavitation study.

$$\begin{bmatrix} f_{1,1} & \cdots & f_{1,n} \\ \vdots & \ddots & \vdots \\ f_{c,1} & \cdots & f_{c,n} \end{bmatrix}$$
Condition 1

$$\begin{bmatrix} f_{c+1,1} & \cdots & f_{c+1,n} \\ \vdots & \ddots & \vdots \\ f_{2c,1} & \cdots & f_{2c,n} \end{bmatrix}$$
Condition 2

$$\begin{bmatrix} f_{(s-1)c+1,1} & \cdots & f_{(s-1)c+1,n} \\ \vdots & \ddots & \vdots \\ f_{2sc,1} & \cdots & f_{2sc,n} \end{bmatrix}$$
Condition 3

Step 3: Normalize the Columns of the Cavitation Feature Matrix:

In the third step we normalize the columns of the feature matrix. Normalization allows CSP values with different amplitude scales and units to be directly compared without higher magnitude CSPs being given undue weighting. CSPs in the feature matrix can have different units depending on sensor type and the method used to calculate the parameter.

We use the z-score (sometimes referred to as the standard score) transformation [45] to normalize the columns of the feature matrix. The importance of normalization when comparing data is discussed in detail by Keogh and Kasetty [46] and is common in multivariate statistical analysis [47–51] as well as machinery diagnostics and prognostics [52–55]. Z-score normalization linearly transforms the data to have a mean of zero and a variance of 1. The new normalized value has no units and is a measure of the distance, in standard deviations, from the mean of the data. We recommend Z-score normalization be applied to each column independently using Equation 4.2 where $f$ is the CSP value, $\hat{f}$ is the normalized CSP value, $\mu_f$ is the column mean, and $\sigma_f$ is the column standard deviation.

$$\hat{f} = \frac{f_i - \mu_f}{\sigma_f} \quad \text{for } i = 1 \ldots n \quad (2.2)$$
For the remainder of the feature selection process, unless otherwise noted, the normalized features ($\hat{f}$) are used.

### 2.4.2 Feature Analysis

Feature analysis consists of the following steps: 4) Perform Principal Component Analysis on the feature matrix, and 5) analyze the principal component scores to select the mode of variance that is best related to erosive cavitation.

**Step 4: Perform Principal Component Analysis of the Cavitation Feature Matrix:**

In Step 4, we find the underlying modes of variance within the features in the running condition domain by applying PCA to the feature matrix. One of the modes of variance will be related to erosive cavitation and will be used during the feature selection steps to find the best sensor type, sensor location and CSP for long term cavitation monitoring.

PCA as described by [48] is one of the most important and popular methods in multivariate analysis for reducing the dimensionality of data [56]. Reducing dimensions when dealing with large data sets is helpful for both finding simplified structure within the data and removing variables or features that do not contribute significantly to patterns in the data [57]. PCA is commonly used in condition monitoring for data exploration and feature selection in diagnostics and prognostics [39, 54, 58–60].

PCA looks to re-express a data set into as few variables as possible while keeping the variance of the original data. The output of PCA is a new orthogonal basis matrix, $P$, consisting of orthogonal row vectors referred to as the principal component (PC) s of the original data, $p_1, \ldots, p_m$. The first PC, $p_1$, is the direction in the new basis that accounts for the most variance while the second PC is the orthogonal direction that accounts for the next most variance and so on for the remaining PCs.

In Step 4, we perform PCA on the correlation matrix of $F$ to obtain $P$. $F$ is transformed by $P$ to produce a new representation of the original data, $Y$. The column vectors of $Y$ are the principal component scores and are interpreted as the modes of variance of the feature matrix. Each of the PC score vectors in $Y$ is then plotted to view the mode of variance for
each principal component. The transformation is expressed as:

$$P \ast F = Y$$  \hspace{1cm} (2.3)

Step 5: Analyze Principal Component Scores and Select the Mode of Variance Related to Cavitation Erosion:

In the final feature analysis step, we view the feature scores in the running condition domain and identify the modes of variance that only represent erosive cavitation. Step 5 is needed because cavitation features pick up disturbances or events in the hydroturbine not related to erosive cavitation. These events may be related to non-erosive cavitation, bearing faults, and noise, and will vary with hydroturbine running condition in a different way than erosive cavitation. Since events other than erosive cavitation have mode of variance that are different, they will be represented by one or several principal component scores from Step 4. Selecting only the principle component scores related to erosive cavitation in Step 5 allows us to rank the features based on correlation in Step 6.

PCA on the feature matrix produces the same number of PCs and PC scores as there are columns in the feature matrix. Analyzing all the principal scores is a time consuming process; however, applying PCA shifts a majority of the important information, in terms of variance, into the first few PCs, allowing the remaining PCs to be discarded. Despite knowing that only a few PC scores need to be retained, there is no straight-forward test such as a scree plot to determine the PCs to retain.

When mining large data sets where little information is known about the data a priori, selecting the correct number of PCs that truly represent the data is a difficult task and is explored by [61] as well as [62]. In addition, [48] discusses PC selection techniques specific to time-series data similar to the feature matrix we use in this method. The task of Step 5 is not to try and fully represent the data in the feature matrix, but rather choose the data of interest for detecting erosive cavitation. When PC selection is looked at in this light, two points about the nature of the data in the cavitation matrix provide insight into picking a
selection process:

1. As suggested by [63] important PCs from time-series data will contain clear patterns when treated as time series themselves. Similarly, when PC scores from the cavitation feature matrix are plotted in the running condition domain, they show clear patterns that can be interpreted by an analyst knowledgeable about hydroturbine cavitation.

2. In our methodology, the cavitation matrix is built using features expected to be sensitive to erosive cavitation. Forming the matrix in this way builds a bias in the variance of the matrix that promotes erosive cavitation related PCs. This built in bias ensures that even when a large number of features are evaluated, only a handful of PCs will be of significance and require analysis.

Because the cavitation feature matrix has both attributes, the PC scores can be analyzed in order of decreasing overall variance until the PC relating to cavitation is found. A simple, subjective method for selecting the number of principal components to keep, such as a scree graph [48, 64], should still be used to confirm Point 2 above; however, if a scree graph indicates more than approximately 5 principal components be retained\(^6\), we suggest the practitioner re-evaluate the features or CSPs used to build the feature matrix.

Analysis of the PCs is performed by plotting the PC scores versus hydroturbine running condition and then looking for changes in amplitude that match likely changes in cavitation intensity. Figure 2.4 shows an example of a PC scores plot showing changes in amplitude versus hydroturbine running condition. Knowledge about the type(s) of cavitation the hydroturbine is experiencing (which can be gained by analyzing the erosion damage areas) is useful during analysis to help select PC scores appropriate to erosive cavitation. For additional guidance on cavitation diagnostics as well as matching types of cavitation with erosive damage location, see [13]. Further guidance is beyond the scope of this paper.

\(^6\)Note that the number of principle components to be retained is dependent upon what the cavitation feature matrix contains – specifically how sensitive the features are to erosive cavitation. For instance, if most of the features are not sensitive to erosive cavitation, more PCs will be required to find the erosive cavitation PC than if most of the features are highly sensitive to erosive cavitation.
Figure 2.4: Example of principal component scores plotted versus hydroturbine running conditions.

PCs related to erosive cavitation are retained and will be used in the next step to evaluate cavitation detection features. For the purpose of selecting a cavitation detection feature, PCs not related to erosive cavitation should be discarded. It should be noted that discarded PCs may be related to non-erosive cavitation or other hydroturbine faults and as such could be useful for selecting detection features sensitive to other faults in the hydroturbine that are of interest to hydro plant operators.

2.4.3 Feature Selection

The goal of the feature selection process is to pick the best long term cavitation detection feature for an individual hydroturbine. There is no definitive way to measure best; however, we recommend comparing the features using two statistical measurements before relying on subjective judgement. In Step 6, the correlation coefficients between the PC scores selected in Step 5 and the columns of the feature matrix are calculated. In Step 7, feature variability is compared using the estimated standard deviation at the features minimum and maximum values. The final subjective evaluation is conducted in Step 8 where features are yet again down selected based on practical considerations for long term cavitation detection.
Calculate and Compare Correlation Coefficients:

Once the principal components that represent erosive cavitation are selected (Step 5), the first feature selection step is to calculate the sample correlation coefficients between the columns of the normalized feature matrix \( f_1, \ldots, f_n \) and the columns of the PC scores \( y_1, \ldots, y_n \) related to the PCs selected in Step 5. The values of the correlation coefficients are then used as a basis for removing features that are not sensitive to erosive cavitation [65].

Sample correlation coefficients are a statistical measure of linear dependence between two population samples [66]. For our methodology, the population samples are the cavitation features and the PC scores. The correlation coefficients between these two populations are designated as \( \rho(y, f) \) and are calculated by applying z-score normalization to the principal component scores, then calculating the normalized covariance between the score vectors and each column of the feature matrix as shown in Equation 4.4.

\[
\rho(y', f') = \frac{1}{N-1} \sum_{i=1}^{N} (y_i' - \mu_{y'}) \ast (f_i' - \mu_{f'})
\] (2.4)

In Equation 4.4, \((f_i' - \mu_{f'})\) is the complex conjugate, \(N\) is the number of CSP values in each feature, \(\mu_{y'}\) is the mean of the score vector, and \(\mu_{f'}\) is the mean of the feature vector. Since the column vectors \(y'\) and \(f'\) are both real-valued and normalized to have zero mean, the equation simplifies to Equation 4.5, resulting in a scalar value between -1 and 1.

\[
\rho(y', f') = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i') \ast (f_i')
\] (2.5)

Features with correlation coefficients close to 1 or -1 are linearly dependent with the PC score, share a similar mode of variance, and are therefore sensitive to erosive cavitation. Features with coefficients closer to zero are not sensitive to erosive cavitation and can be removed from the selection process.

A rule of thumb guideline for comparing correlation coefficients [45] is shown in Table 2.2. The feature matrix is built from features meant to be sensitive to cavitation; therefore, several
features will have a high or very high degree of dependence with the PC scores associated with erosive cavitation. Based on Table 2.2 and the expectation of very high dependence, we recommend removing features with a correlation coefficient that has an absolute value less than 0.9. If a suitable feature is not found or the practitioner would like to evaluate additional options, the threshold can be relaxed to 0.7, but no lower. Below 0.7, features are expected to be of poor quality and will not be useful for cavitation monitoring.

Table 2.2: Rule of thumb for comparing correlation coefficients [45]

<table>
<thead>
<tr>
<th>Absolute Value of Correlation Coefficient $\rho$</th>
<th>Rule of Thumb</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 \leq</td>
<td>\rho</td>
</tr>
<tr>
<td>$0.3 \leq</td>
<td>\rho</td>
</tr>
<tr>
<td>$0.5 \leq</td>
<td>\rho</td>
</tr>
<tr>
<td>$0.7 \leq</td>
<td>\rho</td>
</tr>
<tr>
<td>$0.9 \leq</td>
<td>\rho</td>
</tr>
</tbody>
</table>

Step 7: Compare sample standard deviation of the features CSP values at minimum and maximum cavitation intensity:

In Step 7, we compare the dispersion (otherwise known as the spread or variation of a data set) of the cavitation features by calculating the sample standard deviation for each feature at the hydroturbine running condition with minimum and maximum erosive cavitation as depicted by the PC scores plot(s) from Step 5. Features with low standard deviation have less dispersion [66], and are less likely to produce false positive and false negative identification of erosive cavitation.

The sample standard deviation at the minimum CSP value ($s_{CSP-min}$) is calculated for each feature by first identifying the hydroturbine running condition at the minimum CSP.
value, then using only CSP values from this running condition to calculate the standard deviation. The standard deviation estimate at the maximum CSP value ($s_{CSP\text{-max}}$) is calculated in a similar way using only CSP values from the same running condition as the maximum CSP value.

The standard deviation calculation method above will result in a pair of descriptive statistics for each feature. The standard deviation pairs are used to rank the remaining features so that the best cavitation features will have the lowest values for both statistics. The following guidelines are recommended for ranking the remaining features:

1. We recommend removing features with relatively high $s_{CSP\text{-min}}$ or $s_{CSP\text{-max}}$ values from consideration. Establishing two thresholds for eliminating features by multiplying the overall lowest $s_{CSP\text{-min}}$ and $s_{CSP\text{-max}}$ values by two is the suggested method of elimination, as shown in Equation 4.6:

   \[
   \begin{align*}
   \text{threshold} 1 &= 2 \times s_{CSP\text{-min}} \\
   \text{threshold} 2 &= 2 \times s_{CSP\text{-max}}
   \end{align*}
   \tag{2.6}
   \]

   Eliminate any feature with an $s_{CSP\text{-min}}$ or an $s_{CSP\text{-max}}$ value above the respective thresholds and keep the other features for continued evaluation.

2. We have found ranking the remaining features based on the combination of their standard deviation values is a useful means of comparing and ranking features:

   \[
   s_{CSP\text{-combined}} = s_{CSP\text{-min}} + s_{CSP\text{-max}}
   \tag{2.7}
   \]

3. Features with lower standard deviation around their minimum value are given preference to features with lower standard deviation around their maximum value. This evaluation is made because identifying individual cavitation events is not critical and, in our opinion, erring toward reducing the number of false positives is beneficial to long term acceptance of cavitation prognostics [67].
An additional and roughly equivalent method for evaluating feature dispersion is to calculate interquartile range around the minimum and maximum CSP values. In this method, box plots or quantile-quantile plots give the practitioner additional insight into the structure of CSP variability [66]. In the our experience with calculating dispersion of cavitation sensitivity parameters, ranking features using either interquartile range or standard deviation shows similar results.

Step 8: Evaluate remaining features based on practical considerations for long term cavitation detection:

In Step 8, we evaluate the remaining cavitation features and make a final selection based on the following practical considerations:

1. Are the sensor locations specific to the features practical for long term usage in the hydroelectric plant?

2. Are the hardware and installation costs required to generate a feature drastically different than other features?

3. Are the features sensitive enough to erosive cavitation to be used alone or is more than one feature required?

4. Is the hardware and software specific to each feature reliable and can it be maintained by plant personnel?

For the first practical consideration, each remaining feature can be evaluated by how practical the sensor location is for permanent installation. Cavitation survey data may have been collected at sensor locations that work well for cavitation detection, but due to access restrictions, safety regulations, or the need for equipment modifications, the sensor locations may not be deemed acceptable for permanent installation. Cencic et al [14] notes that sealed bearing housings prevented the consideration of diagnostic techniques that require a clean transmission path between the runner and the sensor, such as demodulation, from being used for long term cavitation monitoring.
The second consideration is to evaluate the hardware and installation costs required to generate each feature. Cavitation diagnostics based on data acquired from a large number of sensors [18] or using custom-built wireless technology [8, 68] have significantly higher equipment cost and complexity when compared to the other methods shown in Table 4.1 based on fewer, commercially available sensors. The cost of handling, storing and maintaining the data generated by each feature is also a part of this consideration. The number of sensors, required sample rate for recording the data, and duration of the recorded signals all affect data storage requirements and must be considered when evaluating feature costs.

The third consideration requires evaluating whether each feature is sensitive enough to erosive cavitation to be used on its own for long term cavitation monitoring. Features that generate noisy data or are sensitive to hydroturbine faults not related to erosive cavitation are unreliable on their own, but combining multiple features may lead to robust results. When trying to monitor cavitation to determine erosion rates in a hydroturbine, Wolff et al [44] reported problems due to noisy data and noted that additional sensors would have been helpful for making more accurate erosion rate estimations.

The final evaluation for cavitation feature selection is to consider reliability of the hardware and software system required to generate the feature. If the system requires maintenance or troubleshooting, consider whether the hydroelectric plant personnel have the resources to keep the system reliable. Bourdon et al. [6, 11] developed a sophisticated monitoring system for cavitation detection meant for making long term erosion rate estimates. Francois et al. [12] report however, that lack of reliability in the monitoring system led to incomplete data over an 8 year period preventing erosion rates from being estimated. The monitoring system was recently upgraded; however, cavitation erosion estimates from the system have yet to be published.

2.4.4 End Result

The output of the feature selection method is the best feature or group of features to use for long term erosive cavitation monitoring in a specific hydroturbine. The selected
feature(s) specifies the sensor type, and sensor location for permanent installation as well as the cavitation sensitivity parameter to be monitored over time.

2.5 Case Study

We present here a case study using a real cavitation survey conducted on a Francis turbine at a hydro power plant located in the American West. The data was collected using the following sensors and sensor placement. An accelerometer and acoustic emission sensor (Acc3 and AE3) were located directly on the hydroturbine shaft and data was collected at a sample rate of 1,330,000 S/s. One accelerometer and one acoustic emission sensor were located on both the lower guide bearing (Acc1 and AE1) and inlet guide vane stem (Acc2 AE2), and signals from these sensors were sampled at a rate of 1,000,000 S/s. A total of four proximity probes were mounted 90 degrees apart facing the shaft, two near the lower bearing (PP1 and PP2) and two near the upper bearing (PP3 and PP4) of the turbine. In addition, a pressure sensor was located in the wall of the draft tube (PR1). Signals from the proximity probes and the pressure sensor were sampled at a rate of 10,000 S/s. The turbine operating conditions captured in the cavitation survey ranged from 5 MW to 85 MW in 5 MW increments resulting in 17 unique operating conditions. Other running condition variables such as hydrostatic head, other turbines in the plant operating, and other factors were held effectively constant throughout the survey.

Step 1: Calculate Cavitation Sensitivity Parameters:

In the first step of the feature selection process, we chose to calculate six CSP values for each sensor where high frequency data was collected and five CSP values for each sensor where medium frequency data was collected. The number of calculated CSP values was selected to demonstrate the method without added confusion from many tens or hundreds of calculated CSP values. The practitioner can decide to use more or less CSP values depending upon the situation and desired results. In our experience, between five and ten CSPs per sensor is effective in identifying desirable features.
Every CSP listed in Table 4.1 uses band-pass filters on raw sensor data and RMS amplitude calculations as either the primary CSP or as a step to calculating the CSP. Since RMS amplitude of band-pass filtered data is so common in cavitation detection, we use it as the basis for a majority of CSPs calculated in this case study. Alternative CSPs were also calculated using peak amplitude, crest factor, and kurtosis, which are common calculations used for condition monitoring outside of cavitation detection [69]. The alternative CSPs were included for experimental purposes to compare methods other than RMS that are very rarely if ever found in hydroturbine cavitation studies. Practitioners may wish to include other experimental CSPs to determine if borrowing a CSP from a different field may provide better results as compared to CSPs traditionally used with hydroturbines.

Table 2.3 shows the formulas used for calculating RMS, peak, crest factor, and kurtosis values. Table 2.4 lists the specific CSPs calculated for each sensor type used in this case study.

<table>
<thead>
<tr>
<th>Calculation</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMS</td>
<td>( f_{\text{rms}} = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}} )</td>
</tr>
<tr>
<td>Peak</td>
<td>( f_{\text{peak}} = \max(x) )</td>
</tr>
<tr>
<td>Crest Factor</td>
<td>( f_{\text{CF}} = \frac{f_{\text{peak}}}{f_{\text{rms}}} )</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>( f_{\text{kurt}} = \frac{\frac{1}{N} \sum_{i=1}^{N} [x_i - \mu_x]^4}{\left(\frac{1}{N} \sum_{i=1}^{N} [x_i - \mu_x]^2\right)^2} )</td>
</tr>
</tbody>
</table>

Step 2: Form the Cavitation Feature Matrix from the Cavitation Sensitivity Parameters:

Next, we formed the cavitation feature matrix by calculating the CSP values listed in Table 2.4 for the three acoustic emission sensors, three accelerometers, four proximity probes
and one pressure transducer used in the cavitation survey resulting in 61 total features. 32 CSP values were calculated for each of the 17 operating conditions resulting in 544 CSP values for each feature. The cavitation feature matrix is therefore a 544 x 61 matrix organized as described in Step 2 of the methodology. Throughout the rest of this document, we have adopted the feature naming convention from the combination of the abbreviation of the sensor type and the CSP number shown in Table 2.4.

Table 2.4: Cavitation sensitivity parameter details for each sensor type

<table>
<thead>
<tr>
<th>Sensor Type</th>
<th>Cavitation Sensitivity Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accelerometers</td>
<td>1) RMS amplitude 1,000 - 20,000 Hz</td>
</tr>
<tr>
<td></td>
<td>2) RMS amplitude 20,000 - 30,000 Hz</td>
</tr>
<tr>
<td></td>
<td>3) RMS amplitude 30,000 - 100,000 Hz</td>
</tr>
<tr>
<td></td>
<td>4) Peak amplitude 1,000 - 20,000 Hz</td>
</tr>
<tr>
<td></td>
<td>5) Crest factor 1,000 - 20,000 Hz</td>
</tr>
<tr>
<td></td>
<td>6) Kurtosis 1,000 - 20,000 Hz</td>
</tr>
<tr>
<td>Acoustic Emission Sensors</td>
<td>1) RMS amplitude 1,000 - 400,000 Hz</td>
</tr>
<tr>
<td></td>
<td>2) RMS amplitude 50,000 - 400,000 Hz</td>
</tr>
<tr>
<td></td>
<td>3) RMS amplitude 1,000 - 50,000 Hz</td>
</tr>
<tr>
<td></td>
<td>4) Peak amplitude 1,000 - 400,000 Hz</td>
</tr>
<tr>
<td></td>
<td>5) Crest factor 1,000 - 400,000 Hz</td>
</tr>
<tr>
<td></td>
<td>6) Kurtosis 1,000 - 400,000 Hz</td>
</tr>
<tr>
<td>Proximity Probes</td>
<td>1) RMS amplitude 40 - 1,000 Hz</td>
</tr>
<tr>
<td></td>
<td>2) RMS amplitude 1 - 40 Hz</td>
</tr>
<tr>
<td></td>
<td>3) Peak amplitude 40 - 1,000 Hz</td>
</tr>
<tr>
<td></td>
<td>4) Crest factor 40 - 1,000 Hz</td>
</tr>
<tr>
<td></td>
<td>5) Kurtosis 40 - 1,000 Hz</td>
</tr>
<tr>
<td>Pressure Transducer</td>
<td>1) RMS amplitude 40 - 1,000 Hz</td>
</tr>
<tr>
<td></td>
<td>2) RMS amplitude 1 - 40 Hz</td>
</tr>
<tr>
<td></td>
<td>3) Peak amplitude 40 - 1,000 Hz</td>
</tr>
<tr>
<td></td>
<td>4) Crest factor 40 - 1,000 Hz</td>
</tr>
<tr>
<td></td>
<td>5) Kurtosis 40 - 1,000 Hz</td>
</tr>
</tbody>
</table>
Steps 3 and 4: Normalize the Feature Matrix and Perform PCA:

Steps 3 and 4 were performed using MATLAB Software and resulted in a principal component scores matrix \( \mathbf{Y} \). The matrix \( \mathbf{Y} \) is not reproduced here due to the size of the matrix.

Step 5: Analyze Principal Component Scores and Select the Mode of Variance Related to Cavitation Erosion:

In Step 5, we first created a scree plot from the PCA results to determine the number of principal component scores to analyze. Figure 2.5 shows the results of the scree plot which clearly indicate the first PC represents a large majority of the total variance.

![Scree plot of PCA results on the cavitation](image)

Figure 2.5: Scree plot of PCA results on the cavitation

The scree plot also indicates a slight drop off in variance after the fourth principal component. Based on the scree plot, we evaluated the first four PC scores to capture the vast majority of the variance. The first PC score plot (Figure 4.6) shows a steady increase in normalized amplitude values from 35 - 45MW, peak amplitude from 50 – 60MW, then an amplitude decrease from 60 – 70MW. Based on previous cavitation diagnostics performed on this hydroturbine by personnel at the Bureau of Reclamation, using techniques and resources similar to those discussed by [13], the first PC score plot best matches the operating
The first principal component score plot represents a mode of variance related to erosive cavitation conditions and mode of variance associated with erosive cavitation. Additionally, the second and third score plots (Figure 4.7 and Figure 4.8) represent modes of variance associated with draft tube swirl and draft tube vortex collapse. Draft tube swirl can be damaging to hydroturbines; however, for this specific hydroturbine it does not cause erosive damage and is therefore not of interest for this case study. The fourth PC score plot was a twin to the third PC score plot, but at a slightly different running condition.

**Step 6: Calculate and Compare Correlation Coefficients:**

In Step 6, the correlation coefficients were calculated between the first principal component scores and each feature. Correlation coefficients for the accelerometers and acoustic emission sensors are shown in Figure 4.9. Correlation coefficients for the proximity probes and pressure transducer are shown in Figure 4.10. Based on the large number of features with a very high degree of dependence with the first principal component scores, features with a correlation coefficient less than 0.9 were removed from consideration for the remainder of the selection process.
Figure 2.7: The second principal component score plot represents a mode of variance related to early developing draft tube swirl.

Figure 2.8: The third principal component score plot represents a mode of variance related to draft tube vortex collapse at high power output.
Figure 2.9: Correlation coefficients between the first principal component scores, and features based on the accelerometers and acoustic emission sensors

Figure 2.10: Correlation coefficients between the first principal component scores, and features that use proximity probes or a pressure transducer
Step 7: Compare sample standard deviation at the features minimum and maximum CSP values:

In the next step, the standard deviation at minimum ($s_{CSP-min}$) and maximum ($s_{CSP-max}$) CSP values were calculated to compare dispersion within each of the remaining features. As described in Step 7 of our methodology, features with high standard deviations were removed from consideration and the remaining features were ranked in order of their combined $s_{CSP-max}$ and $s_{CSP-min}$ values. Figure 4.11 and Figure 4.12 show the features compared by $s_{CSP-min}$ and $s_{CSP-max}$ values, respectively. Both figures also show the threshold line of two times the minimum standard deviation used for determining which features to remove from consideration. Based on the threshold line, 12 additional features were removed from consideration. The remaining 12 features were ranked from smallest to largest by their combined standard deviation values as shown in Figure Figure 4.13.
Figure 2.12: Comparison of standard deviation around the features’ maximum CSP value

Figure 2.13: Ranking of remaining features by standard deviation around minimum and maximum CSP values
Step 8: Evaluate remaining features based on practical considerations for long term cavitation detection:

In the final step of the feature selection process, we evaluated the remaining 12 features from Step 7 based on the practical considerations outlined in our methodology. Features Acc3.1, Acc3.2, Acc3.3, AE3.1, AE3.2, and AE3.3 are all shaft mounted sensors that require higher cost and complexity to install and maintain. Given that there are 6 additional features that have a similar sensitivity to erosive cavitation, we eliminated these features based on their higher cost and complexity.

The remaining 6 features were based on RMS amplitude in different frequency ranges and come from sensors mounted to the hydroturbines lower guide bearing. Based on the similarity between their $s_{CSP-min}$ values and the low cost of installation and maintenance, any of the remaining 6 features are adequate for erosive cavitation monitoring. In addition to having low $s_{CSP-min}$ values, the two highest ranking features, AE1.2, and AE1.1 have the lowest $s_{CSP-combined}$ value and as such, we consider them the best features for long term monitoring of erosive cavitation.

2.6 Discussion

The methodology outlined in this paper provides several benefits to a researcher or hydroturbine operator wishing to estimate RUL on a hydroturbine runner through long term monitoring of erosive cavitation. In this section we discuss the benefits of using our cavitation feature selection process, and issues that the practitioner must keep in mind. While the method presented here does not yet provide RUL calculations, it is a step in the direction of a full RUL method for hydroturbines—a long-sought goal in the industry.

The feature selection process method described in this paper was demonstrated on cavitation survey data taken on a Francis hydroturbine experiencing leading edge erosive cavitation. While we demonstrated the method on a Francis hydroturbine, the method can be used on any hydroturbine regardless of type. This also holds true for common sensors used in hydroturbine plants to monitor cavitation and other health monitoring applications.
(e.g.: bearing monitoring, etc.), for sensor locations, and multiple cavitation types (e.g.: draft tube swirl actually seen in the case study's data, trailing edge cavitation, etc.). The cavitation survey in the case study was performed using sensors and sensor locations that are commonly found in hydroturbine plant industry cavitation studies. Practitioners must note that the feature selection process, specifically analysis of the principal component scores, is difficult if variance of all the features is dominated by noise or events not related to erosive cavitation. Thus, it is important that the initial features investigated be at least in part known useful features used on other hydroturbines such as overall RMS values taken from an acoustic emission sensor on a lower guide bearing.

One benefit of the presented method is that several aspects of the feature comparisons are automated which allows many different cavitation detection features to be compared quickly. Increasing the number of different features being compared increases the likelihood of finding the best all-around feature. The sensitivity and precision of the features being compared are ranked based on statistical values versus purely subjective evaluation. This combined with the quickness of the process also allows new or experimental features to be evaluated and compared. However, it should be noted, accuracy of the features is not addressed in this paper due to the lack of visual confirmation of cavitation intensity. In industrial settings, it is very rare to have visual confirmation of cavitation intensity.

A few points to keep in mind when using this method include that the methodology compares signal dispersion in order to reduce the likelihood of false positives and false negatives, but the methodology does not directly evaluate the false positives or false negatives associated with each feature. Doing this evaluation requires establishing cavitation thresholds, which is beyond the scope of the work presented in this paper. Another important point is that cavitation intensity is not addressed by the methodology presented here. This prevents the method from being directly used to determine RUL. However, it is expected that future efforts with establishing cavitation thresholds can help to adapt the method presented here to be more useful in calculating RUL. A final point to note is that two large sources of error
exist including data collection and data organization. Poor data collection and organization methods, sometimes seen in cavitation studies, can lead to results that are not accurate or relevant to hydroturbine operation.

The method presented in this paper is a good starting point for researchers and hydroturbine operators to better understand how to monitor hydroturbines for cavitation during operation. The method can be used to identify the most appropriate sensors, sensor placements, and CSPs that provide the most insight into erosive cavitation. Previously, operators and researchers did not have a direct method of comparison for sensors, sensor placements, and CSPs. The method presented here is already showing great promise with some hydroturbine operators and is expected to be deployed in the field soon.

2.7 Future Work

We are actively pursuing several areas of future work and propose the hydroturbine prognostics community pursue several larger goals. One area requiring further study is to better understand why different RMS frequency bands do not distinguish themselves from one another. We discovered this issue using F-tests. A potential direction of research is an in-depth investigation of spectral data produced from RMS frequency bands.

Another area that the community needs to investigate is the evaluation of feature plots viewed in the z-score normalized domain that may be useful for establishing thresholds for long term cavitation detection or for training supervised machine learning algorithms. Establishing cavitation detection thresholds will lead to a better understanding of cavitation intensity that can then be used to develop a RUL method for hydroturbine operators.

Spectrum-based methods such as demodulation and spectral kurtosis were not explored in this paper; however, the foundation has been laid here for evaluating spectrum-based features against traditional RMS-based features. It is possible that spectrum-based methods may be more sensitive to different types of erosive cavitation on the same hydroturbine. While we have demonstrated in this paper that we can detect leading edge cavitation and we also have seen this method work to detect draft tube swirl cavitation on the same dataset, there are
several other types of cavitation that can be important depending upon the hydroturbine design and operating conditions. Multiple erosive cavitation events can occur at the same time and this should be captured for a complete understanding of RUL.

Proximity probes did not show as high a degree of dependence as the accelerometers and acoustic emission sensors to the mode of variance related to erosive cavitation; however, proximity probes did show high dependence and were also sensitive to the modes of variance associated with draft tube swirl. Due to their low cost and higher likelihood of already being installed on a hydroturbine to monitor common low speed faults such as bent shafts, additional investigation of using these sensors for erosive cavitation detection and broader condition monitoring is warranted. It is possible that using proximity probes may make detecting erosive cavitation significantly less expensive and intrusive for hydroturbine operators.

Finally, experimental CSPs including peak, crest factor, and kurtosis did not measure well against RMS for erosive cavitation. These features do however show stark differences between different sensor locations specifically between sensors mounted on the shaft versus sensors mounted off the shaft. We do not yet understand why this is the case. It is possible that a deeper understanding of the physics of the situation may help to develop significantly improved CSPs.

2.8 Conclusion

This paper presents a novel method for comparing and evaluating cavitation detection features - the first step toward estimating RUL of hydroturbine runners. The method can be used to quickly compare features created from cavitation survey data collected on any type of hydroturbine, sensor type, sensor location, and CSP. Although manual evaluation and knowledge of hydroturbine cavitation is still required for our feature selection method, the use of principal component analysis greatly reduces the number of plots that require evaluation. We are not aware of anyone in academia or industry taking this approach with hydroturbines. We applied the method presented in this paper to cavitation survey data collected on a
Francis Hydroturbine and were able to select the best sensor type, sensor location, and CSP to use on this hydroturbine for long term monitoring of erosive cavitation, thus demonstrating the usefulness of the method. Our method provides hydroturbine operators and researchers with a clear and effective way to determine preferred sensors, sensor placements, and CSPs while also laying the groundwork for determining RUL in the future.
CHAPTER 3
FROM FEATURE SELECTION TO CAVITATION DETECTION

The feature selection process presented in Chapter 2 helps hydroturbine operators choose the best combination of sensor, sensor location and CSP for detecting erosive cavitation and monitoring cavitation intensity. The process may also be used by researchers who wish to compare well established CSPs with new and more advanced signal processing methods for detecting erosive cavitation. One of the original research goals for this thesis was to explore and compare CSPs proposed by Escaler et al. [15] and Pennacchi et al. [40] to look for significant improvements or differences between these methods and more established methods suggested by Varga et al. [17] and Hammitt et al. [70]. However, initial application of the demodulation method suggested by Escaler et al. to the cavitation survey data available did not produce results that differentiated themselves from the results detailed in Chapter 2. Additionally, the method for creating CSPs suggested by Pennacchi et al. and based on spectral kurtosis required knowledge about the natural frequency of the hydroturbine shaft that was not available. Based on the initial modulation results and the lack of critical information about the hydroturbine, the research focus of the second part of this thesis changed to addressing issues associated with long term cavitation detection in production environments.

The complexities involved with tracking cavitation detection and intensity data for long periods in industrial environments have historically been a barrier to creating a prognostic model. For instance, many CSPs have specific hardware requirements such as using many specialty sensors or high speed acquisition hardware that is not commonly found in hydro plants and is difficult to maintain [12, 18, 44]. Collecting and evaluating data through cavitation surveys to develop CSPs is disruptive to hydro plant operations and data-intensive, especially when developing a cavitation threshold. In addition, although a cavitation detec-
tion threshold is needed for automated cavitation detection, many diagnostic methods found in the literature do not suggest a way to establish a threshold, leaving the decision up to the hydro plant operator.

Once a cavitation threshold is established, maintaining the threshold over a long period poses its own set of challenges. Static threshold can quickly become invalid due to changing hydro plant operating conditions such as variation in flow rate, hydrostatic head changes (e.g.: the reservoir’s height changes due to drought or flooding), the number of hydroturbines operating in the hydro plant simultaneously, and disturbances to the inlet or outlet flow. Sensor signals can also be affected by internal changes, causing detection errors from a variety of sources including: repairs made to the hydroturbine runner, worsening of cavitation damage to the runner, faults related to the hydroturbine shaft or bearings, changes in detection instrumentation (intentional or otherwise), and sensor drift.

Determining the root cause of a static cavitation detection threshold becoming invalid is difficult; a stationary threshold cannot determine if plant operating conditions or hydroturbine conditions are the source of the error. Cavitation intensity measurements are affected by the same problems that impact a static cavitation detection threshold. In summary, the existing methods used in industry and in the literature to detect cavitation in a hydroturbine are based on single source measurements, require manual analysis of many different CSPs, or combine many of the same CSPs while maintaining the issues noted above. The research presented in Chapter 4 address many of the long term issues with collecting sufficient cavitation detection and intensity data to advance the industry toward the goal of accurate RUL prediction.

An additional note about Chapter 4 is the specific data used to create the case study presented in this chapter is motivated by the results from the case study in Chapter 2. It is evident from the results presented in Chapter 2 that at least for the specific Francis hydroturbine operated by the United States Bureau of Reclamation, there are several sensor and CSP options available for cavitation detection. An accelerometer or an acoustic emission
sensor combined with a fairly simple CSP - RMS of a band pass filtered high frequency signal - offer the the best features for cavitation detection and intensity measurements. Something that is not evident in the result shown; however, is that features based on proximity probes are also sensitive to erosive cavitation. The case study for Chapter 4 is based purely on CSPs created from proximity probe measurements. This result is relevant to hydroturbine operators who already have proximity probes installed for condition monitoring – a common scenario in older hydroturbines – and may not have the choice to install the additional sensors and data collection equipment needed to perform high frequency vibration or acoustic emission measurements.
CHAPTER 4

A METHOD FOR AUTOMATED CAVITATION DETECTION WITH ADAPTIVE THRESHOLDS

A paper to be submitted to the International Journal of Prognostics and Health Management

Seth W. Gregg\textsuperscript{7}, John P.H. Steele\textsuperscript{8}, and Douglas L. Van Bossuyt\textsuperscript{9}

4.1 Abstract

Hydroturbine operators who wish to collect cavitation intensity data to estimate cavitation erosion rates and calculate remaining useful life (RUL) of the turbine runner face several practical challenges related to long term cavitation detection. In this paper, we present a novel method that addresses several of these challenges including a method to create an adaptive cavitation threshold and automate the cavitation detection process – two strategies to aid in collecting consistent cavitation intensity data. Although domain knowledge and manual interpretation are needed to choose an appropriate cavitation sensitivity parameter (CSP), the remainder of the process is automated through the use of both unsupervised and supervised learning methods, and enhanced by the use of Mahalanobis distance (MD) when generating the threshold. We present a case study based on ramp-down taken from a production hydroturbine and verify the accuracy of our cavitation detection process on cavitation survey data from the same hydroturbine. Our results indicate that our fully automated process for selecting cavitation thresholds and classifying cavitation performed well when compared to manually selected thresholds. Our methods provide hydroturbine operators and researchers with a clear and effective way to perform automated cavitation detection while also laying the groundwork for determining RUL in the future.

\textsuperscript{7}Graduate student, primary researcher and Author
\textsuperscript{8}Primary Advisor
\textsuperscript{9}Co-Advisor and Corresponding Author
4.2 Introduction

Hydroturbines produce 6.3% of all electrical generation and 48% of renewable energy in the United States of America [2]. While hydro power plants have existed for well over 100 years, issues such as cavitation damage to hydroturbine runners remain problematic for plant operators. In this paper, we present a method to automatically detect damaging cavitation events using existing installed sensors whose data is used to recalibrate the cavitation detection algorithm on hydroturbine ramp-down or ramp-up. Of particular interest to hydro plant operators is the reduction in required user input and hydroturbine downtime; the method we present here automatically calibrates at hydroturbine ramp-down or ramp-up.

Our underlying motivation for this work is the goal of estimating remaining useful life (RUL) of hydroturbines. Currently, hydro plant operators service hydroturbines on a fixed schedule based on operating experience to repair cavitation damage to the hydroturbine runners. If RUL can be accurately estimated, then condition-based maintenance of hydroturbines can be implemented – a significant advancement for the industry. The necessary steps to develop RUL predictions for hydroturbines are as follows:

1. Select a sensor-based cavitation detection method for identifying erosive cavitation and measuring its intensity.

2. Collect cavitation intensity data for a test period that is long enough for accumulative cavitation damage to be measured.

3. Measure the runner material loss over the test period and correlate the loss with the measured cavitation intensity over the same period.

4. Create an erosion rate model to use for estimating runner RUL at any future state based on accumulated cavitation intensity.

It is important to note that a significant amount of data is required including: 1) cavitation detection data, 2) cavitation intensity data, and 3) runner material loss data. These
data would then be correlated to develop an erosion rate model to estimate runner RUL. The complexities involved with tracking cavitation detection and intensity data for long periods in industrial environments have historically been a barrier to creating a prognostic model. For instance, many indicators sensitive to the onset of cavitation (a cavitation sensitivity parameter (CSP) as first introduced in [16]) have specific hardware requirements such as using many specialty sensors or high speed acquisition hardware that is not commonly found in hydro plants and is difficult to maintain. Collecting and evaluating data through cavitation surveys to develop CSPs is disruptive to hydro plant operations and data-intensive, especially when developing a cavitation threshold. Many diagnostic methods found in the literature do not suggest a way to establish a cavitation detection threshold, leaving the decision up to the hydro plant operator. A static cavitation detection threshold can quickly become invalid due to changing hydro plant operating conditions such as changes in flow rate, hydrostatic head changes (e.g.: the reservoir’s height changes due to drought or flooding), the number of hydroturbines operating in the hydro plant simultaneously, and disturbances to the inlet or outlet flow. The vibrations that sensors monitor can also be affected by internal changes, causing detection errors from a variety of sources including: repairs made to the hydroturbine runner, worsening of cavitation damage to the runner, faults related to the hydroturbine shaft or bearings, changes in detection instrumentation (intentional or otherwise), and sensor drift. Determining the root cause of a static cavitation detection threshold becoming invalid is difficult; a stationary threshold cannot determine if plant operating conditions or hydroturbine conditions are the source of the error. Cavitation intensity measurements are affected by the same problems that impact a static cavitation detection threshold. In summary, the existing methods used in industry and in the literature to detect cavitation in a hydroturbine are based on single source measurements, require manual analysis of many different CSPs, or combine many of the same CSPs while maintaining the issues noted above. In this paper, we address collecting sufficient cavitation detection and intensity data to advance the industry toward the goal of accurate RUL prediction.
The first three steps of the RUL prediction process have been carried out in laboratory tests, but the methods used are not practical for monitoring a hydroturbine in a production power plant environment. Complications with data quality, sensor placement, long term robustness of the data collection hardware, and the requirement of manual interaction with the detection system have thwarted attempts to carry out similar tests on production hydroturbines. To our knowledge, results have yet to be published that correlate cavitation erosion rates with data taken from a production hydroturbine. The lack of widespread acceptance or implementation of cavitation monitoring for estimating erosion rates suggests the existing methods are either not effective or not accessible to most hydroturbine operators.

The issues with establishing a RUL prediction process described above suggest that an adaptive approach that is easily automated may be more successful for long term RUL prediction on a production hydroturbine. In this paper, we address the first two steps in developing a RUL prediction method: 1) detecting erosive cavitation and, 2) collecting cavitation intensity data. Here we view cavitation detection as both a supervised and unsupervised learning problem. Cavitation detection is a simple classification problem with two classes: cavitation exists (class 1) or it does not (class -1). With a properly labeled set of training data, many different supervised classification methods can be used to solve this problem. Supervised learning provides a more sophisticated approach to cavitation detection when compared to setting linear thresholds; however, even these algorithms will still become inaccurate as sensor data and operating conditions change over time. To solve problems with drift in the data and operating conditions, a classification algorithm (classifier) can be re-trained over time (tantamount to re-calibrating); however, labeled training data must be re-generated under the new hydroturbine conditions. The need to manually generate labeled training data reduces the automation of the process and increases the likelihood of mis-classification due to sensor failure, changing operating conditions, or neglect. A more robust approach is to view the creation of training data as an unsupervised learning problem that can be automated after initial parameters are set using domain knowledge. We follow this approach.
to identify operating regions where the hydroturbine is experiencing cavitation through an initial manual process that is then automated to re-calibrate the classifier during ramp-up or ramp-down of the hydroturbine. We determine the intensity of cavitation through calculation of the Mahalanobis distance (MD) from a set of baseline data. The baseline data is generated from the ramp-down or ramp-up data, with the initial ramp-down or ramp-up requiring manual selection of cavitation and cavitation-free operating zones. After initial manual selection of the operating zones, the process is automated and auto-updates based on current hydroturbine running conditions and sensor data.

This paper specifically contributes to the literature a process that addresses the first two steps of developing a RUL prediction for hydroturbines. While we demonstrate the process using proximity probes, it is important to note that this process will work with any sensor commonly used to monitor hydroturbines and that can detect cavitation events. Below we demonstrate a feature selection method that is simple and can be generalized to many different sensors and CSPs. We use a feature selection process that can be performed on a small amount of data with minimal intrusion to the hydroturbine and hydro plant. After an appropriate CSP is selected, our method can be fully automated, greatly increasing the likelihood of successful long-term cavitation detection and cavitation intensity monitoring. In this paper we propose using an adaptive threshold that automatically learns the new conditions by collecting a small amount of ramp-up or ramp-down data. We introduce the MD to hydroturbine cavitation detection and intensity monitoring from the field of cavitation detection in hydraulic pumps where we use the MD as a basis for both establishing cavitation detection thresholds and tracking cavitation intensity. Our method is flexible and multivariate, allowing for the incorporation of many different CSPs which affords hydro plant operators flexibility in deployment to suit their own specific plant conditions.

4.3 Background

Hydroturbines create energy by taking advantage of water falling between reservoirs at different elevations. The available water head and flow between the reservoirs determines
the amount of power that can be produced and affects the design and type of hydroturbine that will be installed [35]. The most common types of hydroturbines used for commercial power generation are the Kaplan, Francis, and pump-turbine designs. The primary difference between Francis and Kaplan hydroturbines is the design of the runner – the impeller-shaped rotor that captures energy from flowing water. Kaplan turbine runners are designed to be most effective in low head applications while Francis turbine runners are common in medium and high head use cases [36]. Pump-turbines are similar to Francis turbines, but have the added advantage of being able to be used as a pump, such as in pump storage hydro projects [14, 37].

Cavitation is one of the most common faults that occurs in hydroturbines [4, 5] and the damage caused by cavitation can be very costly to repair [6, 38]. Cavitation in Hydroturbines is the formation of vapor bubbles in the water flowing through the hydroturbine and occurs when abrupt changes in water velocity cause local pressures to fall below the fluid vapor pressure [3]. Vapor bubbles typically develop on or near the hydroturbine runner, but can form in any area where the flowing water reaches higher than intended velocities. When cavitation bubbles collapse, they release a large amount of energy that is destructive to nearby surfaces.

The available water head and flow play a significant role in determining if cavitation will develop during turbine operation [35]. Hydroturbines are designed to prevent cavitation from forming under normal running conditions; however, several factors outside of the control of designers make eliminating cavitation, and damage caused by cavitation, a difficult task including: 1) available head may change outside of design conditions due to seasonal reservoir variations, floods, or drought; 2) turbulent flow caused by damage or obstructions at the inlet of the hydroturbine; 3) erosion damage on the runner can encourage the formation of cavitation; and 4) the complexity of cavitation formation and collapse makes the amount of damage caused by cavitation difficult to predict in Hydroturbines.
4.3.1 Cavitation Detection in Hydroturbines

Hydroturbine researchers generically use the term ‘cavitation detection’ to refer to diagnostic methods that involve sensor measurements, signal processing, and data analysis to aid in determining when cavitation is present [13–15]. This definition, however, is ambiguous about key elements of collecting long term cavitation data for studying erosion rates. For the purposes of this paper, we will divide cavitation detection into three distinct actions:

- Applying a diagnostic method to sensor measurements to create an indicator sensitive to the onset of cavitation (a CSP) as introduced in [16].

- Establishing a cavitation threshold (when using a single CSP) or a decision boundary (when using multiple CSPs) that is used to decide when cavitation is present.

- Measuring cavitation intensity in a way that can be used to calculate or estimate cavitation erosion rates.

Many diagnostic methods are available to hydroturbine operators for creating a CSP [13–15, 17–19]. Unfortunately, cavitation intensity measurements are not directly addressed in these diagnostic methods and the action of establishing a cavitation threshold is completely ignored. This is problematic because cavitation thresholds are critical for automating cavitation detection, and intensity values are needed to correlate erosion rates with sensor measurements. It would appear that outside of the work by Dorey, et al. [10], performed in collaboration with Bourdon, et al. [9] and continued by Francois [12], cavitation diagnostics studies have focused on short term data collection and manual data analysis.

Selecting the right diagnostic method for a given hydroturbine is difficult since no method has been shown to be effective, practical, and affordable for every hydroturbine. Although we present a method for selecting spectral-based CSPs in this paper, selecting the best diagnostic method is beyond the scope of this paper. Our previous work on this topic provides hydro plant operators with a guide to select the most appropriate diagnostic method for specific
plant configurations [71]. We do not suggest the CSPs created in our case study are superior in all scenarios, but rather focus on the larger cavitation detection process as defined above.

4.3.2 Instrumentation for Cavitation Detection

When a cavitation bubble collapses on the surface of the hydroturbine runner, the shock wave it creates propagates through the hydroturbine and surrounding water. Cavitation creates significant erosive damage when many thousands of bubbles collapse over a short period of time which produces a vibration response between 3000 and 400,000 Hz [13, 14]. Detecting the high frequency response of cavitation directly requires sophisticated sensors and equipment meant for high frequency applications. Accelerometers and acoustic emission sensors are frequently used in cavitation diagnostics due to their high frequency response. Accelerometers capture vibratory response by producing a signal proportional to absolute acceleration at the sensor location. Acoustic emission sensors produce a signal proportional to the amplitude of small stress waves that travel through the hydroturbine and surrounding water [13, 38, 72]. The signals from these sensors are collected using data acquisition equipment capable of very fast sample rates that has large storage capabilities.

Since hydroturbines have relatively low shaft speeds (typically well below 20 Hz), high frequency monitoring equipment is specific to cavitation detection. Other fault conditions such as balance and alignment problems occur well below 500 Hz and are monitored with low sample rate data acquisition equipment and proximity probes, which produce a signal proportional to the relative movement between the sensor and the hydroturbine shaft. The added cost of a separate, more sophisticated cavitation detection system means many hydroturbines are not constantly monitored for cavitation.

It has been shown by Pennacchi, et al. [40] that proximity probes can also be used for diagnosing cavitation. Instead of measuring cavitation events directly, Pennacchi used synchronous averaging and spectral kurtosis to monitor the hydroturbine shaft’s natural frequency response fluid instability. For their method to be implemented however, the signal is filtered around the natural frequency of the shaft. The shaft natural frequency, especially
while surrounded by water, is often not known for older hydroturbines and cannot easily be obtained.

4.3.3 Cavitation Intensity

Dular et al. [31] developed a cavitation damage model that relates cavitation damage, $A_{rel}$ (damage area), to the amount of time a surface is exposed to cavitation, $\tau$, the cavitation shedding frequency, $f$, the probability of a cavitation event (referred to as a micro-jet by Dular et al.), $P(mj)$, and the velocity characteristics of the flowing water, $v_{ref}$ and $v$ with Equation 4.1

$$A_{rel} = A_{pit} \tau f P(mj) \left( \frac{v}{v_{ref}} \right)^2 \left( \frac{v}{v_{ref}} \right)$$  

(4.1)

In Equation 4.1, $A_{pit}$ is the pit area and $A_{ref}$ is the total reference area. The damage model was verified on a radial pump with $f$, and $v$ being measured during the experiment and $P(mj)$ being held constant. The significance of this model is that cavitation damage was related to a cavitation intensity based on local fluid velocity, exposure time, and the frequency of cavitation events. Although local fluid velocity and cavitation event frequency cannot practically be measured in a production hydroturbine, in a later laboratory experiment on a model turbine [19], it was verified that measurements from acoustic emission sensors and accelerometers correlated well with cavitation intensity as predicted by the model and verified with a high speed camera. In summary, cavitation intensity can be measured through vibration or acoustic emission sensors and combined with cavitation exposure time to estimate cavitation damage.

These results are encouraging; however, in a practical implementation one must choose sensor types and locations as well as CSPs that give reliable intensity measurements. Variation in the structure and layout of different hydroturbines combined with different sensor types and placement make amplitude measurements inconsistent when compared to one another. The measurement scale (or unit) of a CSP is dependent on the sensor type and the measured value is affected by the sensor location [21]. Cavitation tests on production hy-
droturbines are usually performed with accelerometers and acoustic emission sensors placed on the upper and lower hydroturbine bearings as well as the stems of the guide vanes that control water flow rate into the turbine runner [13–15, 20]. Proximity probes are typically located in or near the hydroturbine’s bearings. Each accelerometer, acoustic emission sensor, and proximity probe will produce a signal with a different amplitude. There is no single best location for detecting cavitation events since signal strength will depend on the location of the cavitation as well as the structure of the hydroturbine and response of the shaft. In our experience (and from conversation with hydroturbine experts), the best sensor location should be determined on a case-by-case basis [71].

Unfortunately, this means cavitation intensity measurements performed directly from the sensor’s native measurement scale can only be performed once the sensor’s response to cavitation excitation is known. In the past, response to cavitation has been estimated using the coherence between each sensor and a known input excitation [9, 15]; however, this approach requires specialized data collection equipment, manually collecting test data while the hydroturbine is out of commission, and specialized analysis of the test results. The inconvenience, added cost, and loss in production required to perform a coherence test prevents it from being practical for most production hydroturbines. Additionally, since these tests cannot be performed while the hydroturbine is running, they do not estimate the rotor dynamics, flowing water, and additional sources of noise that affect the sensor response. An alternative way to measure and compare cavitation intensity is to instead normalize the sensor signals using Z-score standardization.

Z-score standardization is a popular method of normalization when comparing and analyzing multivariate data with different amplitude scales [46, 47, 55, 73]. Z-score standardization - more commonly called 'standardization' - linearly transforms the data to have a mean of zero and a variance of 1. A data set \(X = [x_1, x_2, \ldots, x_n]\) is standardized by normalizing the difference between the set mean \(\mu_x\) and each set value by the set standard deviation, \(\sigma_x\),
as shown in Equations 4.2.

\[ \hat{X} = \frac{x_i - \mu_x}{\sigma_x} \quad \text{for } i = 1 \ldots n \]  \hfill (4.2)

The standardized amplitude values are unit-less and measure the distance, in standard deviations, from the mean of the data. In vibration analysis, standardization prevents high amplitude signals from dominating the analysis and obscuring important low amplitude features.

Standardization is frequently used as a data preparation step for machinery diagnostics and prognostics [52–55]; however, we were unable to find it as a step in any published hydroturbine cavitation diagnostic research. Instead of standardization, researchers apply other methods of normalization such as dividing a set of frequency spectra by the first spectrum collected [14, 18] or do not normalize at all. Presumably, normalization is not deemed necessary because researchers and practitioners often compare vibration signals that have the same magnitude scale or are following a collection and analysis process specified in an international standard [74]. We disagree with this notion and choose to standardize our vibration signals for two reasons: 1) vibration amplitude has a non-linear relationship with respect to frequency, and 2) vibration amplitude is affected by the transmissibility between the vibration source and the sensor location.

Commonly used vibration amplitude scales, such as displacement and acceleration, have a non-linear relationship with respect to frequency. Equation 4.3 shows the relationship between acceleration, \(a\), and displacement, \(d\), with respect to a single vibration cycle of frequency \(f\) (in Hz).

\[ a = 2df^2 \]  \hfill (4.3)

As an example, imagine a simply supported steel I-beam, several meters long, vibrating vertically 1 cm at its center. The vibration amplitude in displacement is 0.01 meters. If the vibration frequency is 1 Hz, the vibration amplitude in acceleration is \(0.02 \frac{m}{s^2}\). If the vibration frequency is 10 Hz, the amplitude in displacement remains 0.01 meters; however, the
acceleration amplitude experiences a 100 fold increase to $2 \frac{m}{s^2}$. Vibration signals recorded for cavitation detection contain vibration at many different frequencies that require comparison to each other. This relationship shows how comparing vibration signals with the same units, but with different frequency content, is similar to comparing signals with different scales.

The second reason we chose to standardize vibration signals with the same scale is because vibration amplitude is path dependent. The hydroturbine’s structure changes the transmissibility between vibration at the runner and different sensor locations commonly chosen for cavitation detection. Sensors installed at different locations will observe different amplitudes for the same vibration event [21]. Even sensors placed close to each other, but in different orientations can show significant differences in the observed vibration amplitude. For vibration detection we are not concerned with the local structural or directional response to cavitation, but rather estimating the amplitude of the cavitation intensity at its source. We have found that standardizing signals between different types of sensors, sensor locations, and frequency ranges allows for a consistent comparison of vibration amplitude (discussed in later sections in this paper and also in our previous work [71]).

4.3.4 Mahalanobis Distance

Cavitation detection can be viewed as an on-line process that examines new vibration signal observations as they become available to determine if cavitation is present. When viewed this way, previously examined vibration observations can be represented as the set $X = [x_1, x_2, \ldots, x_n]$ and $x_{n+1}$ becomes the next observation available for examination. If we limit $X$ to observations collected during a known healthy state of the hydroturbine (i.e. the baseline set when cavitation is not present), we can extend the principals of standardization by finding the difference between $x_{n+1}$ and the mean of the baseline set, $\mu_{base}$, then dividing by the standard deviation of the baseline set, $\sigma_{base}$, to compare the distance between $X$ and the new observation. The Mahalanobis distance (Equation 4.4) is a multivariate extension of this concept that is useful for outlier detection, structural health monitoring, clustering,
and detecting cavitation in pumps [16, 75–77].

$$MD = \left[(x_{n+1} - \mu_{base})\Sigma^{-1}(x_{n+1} - \mu_{base})^T\right]^{\frac{1}{2}}$$ (4.4)

In the multivariate case, $X$ now becomes a set of variables, such as observations from multiple sensors while the hydroturbine is in a healthy state, and $x_{n+1}$ contains the next observation from every sensor. The covariance matrix, $\Sigma$, is calculated with the formula:

$$\Sigma_x = \frac{1}{n-1}(X)^T(X)$$ (4.5)

The MD is useful for cavitation detection because it takes into account the correlation of the sensor data and allows us to describe and compare the distribution of several sensors using a single metric. In terms of establishing a threshold for identifying cavitation, instead of creating a threshold for each available sensor, we can now use a single threshold that incorporates all the signals.

It is important to note that when $X$ contains observations from a single sensor, Equation 4.4 reduces down to an equation closely related to Equation 2 used to standardize a data set:

$$MD = \left[\frac{(x_{n+1} - \mu_{base})^2}{\sigma_{base}^2}\right]^{\frac{1}{2}}$$ (4.6)

This single variable form no longer contains a covariance matrix, but still takes into account the distribution of the healthy baseline data for its distance metric. Equation 4.6 should be used when only one sensor is available for cavitation measurements or sensor signals are considered as completely independent observations.

### 4.3.5 Prognostics and Erosion Rate Prediction

The definition we use here for prognostics is the process of forecasting the remaining useful life RUL, probability of failure, or future condition of a component or system [39, 42, 55]. Prognostic models are categorized into physics-based, data-driven, or combination approaches. Physics-based models require a mathematical understanding of the degradation phenomenon affecting the system of interest while data-driven models rely on condition
monitoring or training data collected from the system. Under the right circumstances, both models are effective. In practice, both strategies are needed since mathematical models require experimental validation, which is fundamentally data drive. Similarly, data-driven methods require an understanding of the underlying physics to collect meaningful data. Both data-driven and physics-based approaches are being pursued to developed a prognostic model for estimating hydroturbine RUL.

Current physics-based approaches for cavitation prognostics focus on predicting erosion rates. The underlying mechanisms of cavitation have been shown to be quite complex [3], yet numerical methods developed for erosion rate prediction have been experimentally verified in simplified systems [34, 78]. Though progressing, numerical methods for predicting erosion rates have yet to be verified under conditions and geometries as complex as an operating hydroturbine. Physics-based prognostic models require knowledge of very complex environments and mechanisms that make them hard to build for practical applications [41, 55].

Researchers developing data-driven prognostic models also focus on estimating erosion rates. As previously mentioned, laboratory experiments have verified that damage caused by cavitation is related to cavitation intensity, which in turn can be measured through vibration and acoustic emission. Producing similar results outside of the controlled environment of the laboratory has proven to be much more complex. Hammitt and De discussed predicting erosion rates from sensor measurements as early as 1979 [70], but focused primarily on cavitation erosion on simple shapes in laboratory environments. Francois [12] has written about a major power producers’ attempts at erosion rate estimation; however, no results have been published as of yet. Wolff, Jones and March [44] attempted a similar endeavour at another major power plant in an attempt to establish an erosion rate model, but insufficient data stymied this effort. Similar research in other fields has shown that data-driven prognostic models are often plagued by problems with data quality and data quantity. It is for this reason that we focus our research in this paper on improving long-term cavitation detection and intensity monitoring for production hydro plants.
4.4 Methodology

In this section, we present a methodology for collecting the sensor data needed to create remaining useful life models for hydroturbine runners. The underlying idea of our methodology is that sensor signals collected from a hydroturbine ramp-down and ramp-up – a small data set that requires minimal disruption to power production – can be used to select a CSP, create a threshold for identifying cavitation, and to create a baseline for measuring cavitation intensity. When automated means are used for creating training sets (an unsupervised learning problem) and for classifying cavitation (a supervised learning problem), our method can be used to create a fully automated cavitation detection strategy that can adjust for sensor drift and changes in operating conditions of the hydroturbine.

To setup our methodology, we suggest viewing cavitation detection in a hydroturbine from a machine learning framework by breaking it into four steps: 1) Select Cavitation Features, 2) Create Training Sets, 3) Train a Classifier, and 4) Measure Intensity. By categorizing hydroturbine cavitation detection in this way, it is viewed from the perspective of fundamental areas of machine learning research, making it easier to identify connections between the two fields.

Our methodology was developed using vibration data collected from four proximity probes mounted on an 85 MW hydroturbine. Our feature selection process can easily be used with other sensor types more commonly selected for cavitation detection including accelerometers or acoustic emission sensors; however, an advantage to using proximity probes for cavitation detection is that many older hydroturbine units are permanently instrumented with proximity probes and the associated collection hardware. This is often not the case with accelerometers and acoustic emission sensors that have higher frequency response ranges, but require hardware capable of faster sampling rates. Additionally, the use of four sensors demonstrates how the method has multi-dimensional capability which both improves the classification accuracy and is more robust for long term usage since it doesn’t rely on a single signal source which can be corrupted by noise.
4.4.1 Select Cavitation Features

In the field of machine learning, feature selection is the process of selecting or creating from raw input data a subset of variables that are used as predictors. For the purposes of this paper, the feature being selected is the frequency range used within the CSP calculations used to predict when a hydroturbine is experiencing cavitation. This definition of feature could easily be expanded to include the sensor type and sensor location when these additional options exist [71].

Many sophisticated methods exist for selecting machine learning features. When available, domain knowledge is often an effective way to construct features that are efficient for making predictions [79]. In this section, we present a method to create and select features based on knowledge about vibration signal processing, and the nature of cavitation in hydroturbines.

Step 1: Collect Ramp-Down Data

The features used in our method are created from raw data collected from the hydroturbine as it linearly ramps between its maximum and minimum power output running conditions\(^\ref{foot1}\). When using proximity probes for cavitation detection, the minimum sampling rate used to collect the data should be roughly based on the higher of either the blade passing frequency, \(f_b\), or the guide vane passing frequency, \(f_v\). For a given hydroturbine running speed, \(N\), \(f_b\) and \(f_v\) are defined as follows:

\[
\begin{align*}
  f_b &= N \times (\# \ of \ runner \ blades) \\
  f_v &= N \times (\# \ of \ guide \ vanes)
\end{align*}
\] (4.7)

Based on the typical values of running speed, the number of guide vanes, the number of runner blades on hydroturbines found in literature [13, 14], and taking into account the Nyquist Theorem, a sample rate of at least 1,000 Hz is recommended.

\(^{10}\)The direction of the ramp – up from minimum to maximum or down from maximum to minimum power output – is not important to the research presented here, although in other applications the differences in ramp – up and ramp – down are important. Throughout this text, unless otherwise noted, we generically use the term ramp-down to signify a ramp in either direction.
The amount of time in seconds the hydroturbine takes to go through the ramp-down will affect the amount of data collected, its frequency resolution, and total number of points available to create training data. We have observed in our research that a 60 - 90 second ramp-down produces sufficient data, however, these lengths were based on data available for analysis. The minimum practical ramp-down time could be estimated through progressive decimation of the existing data; however, obtaining this estimate would produce information that is only useful for the hydroturbines included in our study and is beyond the scope of our methodology.

**Step 2: Calculate the Variance of Each Frequency**

In Step 2, we search for vibration frequency ranges in the ramp-down data that significantly change in amplitude over time. During the hydroturbine ramp-down, the speed of the turbine remains constant and the only variables that change are generation load and water flow through the turbine. Vibration frequencies dependent on water flow can be further analyzed to determine if they are related to cavitation. The following process, when applied to the ramp-down data collected in Step 1, allows us to identify frequencies dependent on water flow: 1) The ramp-down data is divided into 1 second blocks, 2) The direct current (DC) (zero frequency) trend is removed in each block resulting in data centered around zero, 3) The discrete Fourier transform (DFT) of each block is computed, and 4) The sample variance of each frequency value across all blocks is calculated.

The first two parts of the process, dividing the ramp-down data into intervals then removing the DC trend, are performed to prepare the data for calculating the DFT. The DFT is a powerful tool used widely in engineering [80] and condition monitoring [81] that allows us to convert time-domain vibration data – data where amplitude varies with time – into the frequency-domain – data with amplitude that varies with frequency. The DFT calculation is performed on a block of time-series data and for vibration analysis, the output of interest is a plot of frequency versus amplitude called a frequency spectrum. The frequency resolution of a spectrum, \(f_{res}\), is dependent on the period of the data collected, \(T\). Frequency
resolution and the data collection are related to sample frequency, \( f_s \), and the number of data samples, \( N \), by Equation 4.8.

\[
\frac{f_{res}}{T} = \frac{f_s}{N} \tag{4.8}
\]

By selecting a ramp-down data block length of 1 second, the resulting DFT calculation will produce a spectrum with a resolution of 1 Hz, which is sufficient to differentiate between cavitation related frequencies ranges within the ramp-down data. The total number of 1 second blocks of data that will be created, \( t \), is dependent on the total length of ramp-down data collected. Selecting block lengths of 1 second not only provides sufficient frequency resolution, but also provides plenty of training data while keeping the ramp-down length reasonable.

When used to detect shaft vibration on a hydroturbine, proximity probes produce a signal proportional to the distance between the tip of the proximity probe and the surface of the turbine shaft. The vibration signal from a proximity probe will therefore oscillate around the average distance between the proximity probe and the shaft which adds a DC offset above zero to the signal. In addition to the added offset, each vibration block is likely to have a slight linear trend in the DC portion of the signal which will cause the DFT to have a large zero frequency amplitude that obscures the amplitude of higher frequencies of interest. The DC offset and linear trend should be calculated and subtracted from each data block.

The DFT (Equation 4.9) of each block of ramp-down data is now calculated using the fast Fourier transform algorithm [82]. Each block of ramp down data, represented as vector \( \mathbf{x} \), has \( n \) data points as does the resulting output vector, \( \mathbf{z} \).

\[
z_k = \sum_{j=1}^{n} x_j e^{\frac{-2\pi i}{n} (j-1)(k-1)} \tag{4.9}
\]

The DFT vector output, \( \mathbf{z} \), is mirrored around its center point and contains both real and complex information. For convenience and ease of interpretation \( \mathbf{z} \) is converted into a normalized half-spectrum [83] using Equation 4.10. The result, \( \hat{\mathbf{z}} \), has \( \frac{n}{2} + 1 \) values, each one representing the amplitude of a \( f_{res} \) Hz wide frequency range within the original time series.
data.

\[ \hat{z} = \frac{2}{n} \left| z_1 \ldots z_{\frac{n}{2} + 1} \right| \quad (4.10) \]

The result of using Equations 4.9 and 4.10 on each data block is a total of \( t \) half spectrum vectors, \( \hat{z} \), which are then used as row vectors to form a \( t \times n \) half spectrum matrix, \( \hat{Z} \), as shown in Equation 4.11.

\[ \hat{Z} = \begin{bmatrix} \hat{z}_1 \\ \vdots \\ \hat{z}_t \end{bmatrix} \quad (4.11) \]

Each row of \( \hat{Z} \) is a frequency spectrum created from a 1 second block of the ramp-down data. Each column of \( \hat{Z} \) contains the amplitude of consecutive 1 Hz frequencies from 0 Hz to \( \frac{n}{2} \) Hz.

Recall that the flow rate of water through the turbine runner is the only running condition variable that changes in the Hydroturbine during ramp-down. As noted multiple times by Escaler et al. [13, 15, 68], cavitation is related to flow rate and causes vibration at multiple frequencies including running speed, \( f_b \), and \( f_v \), as well as through broad-band high frequency noise. As such, vibration frequencies with significant change in amplitude throughout the ramp-down data are likely to be related to cavitation. We recommend finding the change in amplitude of vibration frequencies through the ramp-down by taking the variance of each column of the \( \hat{Y} \) matrix. Variance is a statistical method for measuring the dispersion or variability of sampled data and is calculated from the mean, \( \mu \), using Equation 4.12 [84].

\[ \sigma^2 = \frac{\sum_{0}^{n} [x_i - \mu]^2}{n - 1} \quad (4.12) \]

The result of applying Equation 4.12 to the columns of \( \hat{Z} \) is a single vector that is plotted to form a variance frequency spectrum. The variance frequency spectrum is used to quickly identify frequencies that change during ramp-down and are subsequently related to changes in water flow rate through the hydroturbine.
Step 3: Select CSP Frequency Ranges

The CSP chosen in our methodology for cavitation detection is calculated from the root mean square (RMS) amplitude of proximity probe vibration within one or multiple frequency bands. CSPs based on RMS calculations and frequency filters has been shown to be effective for cavitation detection in both hydroturbines and pumps, [14, 16] and is practical to implement since it can be easily derived using either digital or analog methods. For accelerometers or acoustic emission sensors, demodulation methods [13, 15, 85] can also be used as a basis for the CSP; however, demodulation relies on the use of frequency bands beyond the sensing capabilities of proximity probes.

The frequency bands to use for RMS calculations are based on the variance frequency spectrum created in Step 2 (above). More generally, when using proximity probes for cavitation detection in hydroturbines, three frequency regions are of interest:

1. Vibration frequencies below running speed are affected by draft tube swirl, and Von Karmen vortex shedding, or other hydraulic instabilities [13].

2. Increased vibration frequencies at running speed can also be an indicator of hydraulic instability; however, running speed vibration may also be influenced by other types of faults including unbalance, misalignment, and bearing wear. [86].

3. High frequency vibration at $f_v$, $f_b$, as well as general broadband vibration is associated with cavitation that causes erosion on runner blades.

Vibration amplitude measurement in the third frequency range is of most importance for estimating the intensity of erosive cavitation. We recommend using the third frequency range for the primary CSP and using the first and second frequency ranges for classification features, but not as indicators of cavitation intensity. In our experience, creating training
data using all three frequency ranges improves classifier accuracy and provides the option for using multi-class learning algorithms to identify additional faults within the turbine [87].

As suggested by McKee et al. [16], an alternative approach to using the above method to select frequency bands is to divide the spectrum into octave bands with the second band centered on running speed. This method is particularly effective when many different hydroturbines or pumps are being evaluated; however, the method still requires knowledge about the running speed of the machine and evaluating which bands are most appropriate for monitoring cavitation.

### 4.4.2 Create Training Sets

A supervised machine learning algorithm infers a prediction function based on examples of labeled data. Labeled examples of data are referred to as training sets (or more generally as training data), and consist of feature values, often in the form of a vector, and labels that identify the category assigned to each feature [88]. The size and diversity of a training set has a direct impact on the accuracy and generalisation ability of a machine learning algorithm. Training sets must have relevant features from each category of data the machine learning algorithm is expected to classify, and the number of relevant features is important to predicting the accuracy of the classification predictions. Understanding the number of features and amount of data needed for a classifier to be effective form the basis of statistical learning theory [88, 89]; however, inconsistent data and situational variability make it unpractical to define a generalized method for determining the amount of data needed when working with sensor based prognostics in an industrial environment. First, sensor data is highly susceptible to being influenced by variables not being measured such as operator input, mechanical repairs, and environmental changes that aren’t measured. Second, the amount of training data available and number of features available varies widely based on the type of asset being analyzed, the number of factors that reduce RUL, and the technology used for monitoring [41]. Third, inconsistency of both the source and quality of the sensor data make estimating prediction accuracy very time consuming [55, 90, 91]. For cavitation
detection in hydroturbines, the ability to create training data through a ramp-down process means problems with data quality and variation can be addressed by frequently collecting new training data sets, and then re-training the classification algorithms to account for the most recent conditions.

The challenge in this cavitation detection scenario now becomes creating accurate labels for the training data. Often, training sets are created manually or generated from a process that requires human judgement to determine the correct labels. In our cavitation detection scenario, manual labeling is possible (we suggest manually labeling data for the first training set); however, minimizing human interaction in the process makes it easier for training data to be captured and therefore more likely for the classification algorithm to stay accurate for long periods. For this reason, we propose that labeling training data be viewed as an unsupervised machine learning problem where the goal is to infer the training labels with as little outside input as possible.

In our methodology, we treat erosive cavitation detection as a binary classification problem with two categories: Cavitation and No-Cavitation, and numerically represent them as 1 and -1 respectively. For the reasons described earlier, we recommend MD be used to establish labels for the initial set of training data and find standardizing MD helps with separation of data and interpreting the results. Each point in the training set can be categorized manually, or in an automated fashion using an unsupervised learning algorithm using the following steps:

1. Band pass filter the previously collected sensor data ramp-down signals around each frequency range of interest determined from the Cavitation Feature Selection step.

2. Divide the filtered signals into 1 second blocks and calculate the RMS of each block. The result will be a ramp-down data set for each frequency range of interest ($x_{f_1}...x_{f_2}$).

3. Select the baseline data for calculating MD by plotting the standardized RMS amplitude of each ramp-down data set versus sample number and identifying a continuous
sample range free from cavitation or other faults. This sample range, \( X_{\text{baseline}} \), is the baseline data and is meant to be representative of the fault free distribution of the data for each frequency range and sensor. As a general rule of thumb, the baseline data should contain at least 30 samples\(^{11}\).[84]

4. Combine the ramp down sets, \((x_{f1}...x_{fn})\), into a single matrix \(X\) and calculate the MD of the values in the ramp-down data sets by first calculating the covariance matrix of \(X_{\text{baseline}}\) with Equation 4.5 then applying Equation 4.4 to the remaining values in \(X\). Values for \(\mu\) are calculated from \(X_{\text{baseline}}\). The MD values can then be standardized by applying Equation 4.2.

5. To categorize the data manually, use ramp-down data from the frequency range(s) most representative of erosive cavitation and plot the standardized MD of the ramp-down data versus sample number. Select a cavitation threshold value that is visually above the points in the \(X_{\text{baseline}}\) sample range. When using standardized MD, a conservative threshold, corresponding to fewer false positives, will be close to 1 and a more aggressive threshold, corresponding to more false negatives, will be close to or below 0. All points with a MD larger than the threshold belong in the Cavitation category and all the other points belong in the No-Cavitation category.

6. To automate data categorization, instead of visually selecting a threshold, use an unsupervised learning algorithm such as k-means clustering [93] to separate the ramp-down data from Step 5 into two clusters. The No-Cavitation cluster should minimally contain all the samples in the \(X_{\text{baseline}}\) range.

4.4.3 Train a Classifier

Once cavitation features are selected and training sets are created, cavitation detection is automated by applying a classification algorithm to new cavitation features that are gen-

\(^{11}\)If the RMS values of the vibration data is assumed to be normally distributed, the number of sample points can be reduced[92].
erated to predict if the hydroturbine is experiencing erosive cavitation. When classifying cavitation with a supervised machine learning algorithm, an additional training step is required that allows the algorithm to generate its own cavitation threshold (more generally, this is called a decision boundary or hyperplane) from the training sets created in the previous step.

As a method for evaluating classifiers, we suggest comparing the classifier predictions to a naïve, single variable algorithm that calculates the standardized MD, $\hat{x}_{MD}$, of each new value based on $X_{baseline}$ and compares this value to the threshold established to create the training set. Given a threshold, pseudo code for this classifier is as follows:

```plaintext
FOR new RMS value $x$
    calculate $\hat{x}_{MD}$
    IF $\hat{x}_{MD} > threshold$
        classify $x$ as 1 (Cavitation)
        Calculate and save cavitation intensity based on MD
    ELSE
        classify $x$ as $-1$ (No Cavitation)
END
```

The accuracy obtained by applying the naïve classification algorithm can be used as a baseline for comparing more sophisticated classification algorithms. The advantages of using a naïve classifier are ease of implementation, low computing cost which makes it feasible to use in either an on-line or batch mode, and good accuracy. The disadvantage of such a simple classifier is that it is based on a single variable that is not sensitive to other, non-cavitation related faults so it cannot be used for more generalized fault detection. A multi-dimensional classification algorithm such as a support vector machine (SVM) may be used to take advantage of features created from other frequency ranges to both enhance
cavitation detection and classify other fault states such as non-erosive cavitation.

4.4.4 Measuring Cavitation Intensity

We use the MD of the CSP most representative of erosive cavitation as our cavitation intensity measurement. MD is suited well for measuring cavitation intensity because it automatically accounts for variability in the sensor signal. The benefit of this is best shown graphically using real hydroturbine data. Figure 4.1 shows RMS vibration amplitude with respect to time of a hydroturbine going through a ramp-down as measured by two sensors mounted at different locations. Sensor 1 clearly records a higher maximum amplitude as well as accumulated amplitude (area under the curve) from the erosive cavitation zone in sample range 11 - 38. It is also evident that Sensor 1 increases in amplitude more than Sensor 2 over the baseline range from sample 55 - 100. By contrast, Figure 4.2 shows the same sensor data, but with amplitude measured as MD. Sensor 2 now clearly shows a higher total, and accumulated amplitude since the MD calculation takes into account the lower variance (as measured by standard deviation) of Sensor 2 through the base-line range. In

![Figure 4.1: Sensor vibration amplitude comparison from a hydroturbine ramp-down as measured in RMS](image)

Figure 4.1: Sensor vibration amplitude comparison from a hydroturbine ramp-down as measured in RMS
this way, signals that are more stable when cavitation is not present can contribute more to the intensity measurement.

In summary, the first half of our cavitation detection method takes hydroturbine ramp-up or ramp-down data as an input and returns a training set of CSPs that are used to establish erosive cavitation thresholds. The second half of our cavitation detection strategy creates cavitation thresholds based on the CSPs. The final output is a set of cavitation CSPs created from the sensors being used for cavitation detection, and cavitation thresholds that can adapt to changes in running condition whenever the hydroturbine goes through a ramp-up or ramp-down.

4.5 Case Study

We present here a case study using vibration data collected from an 85Megawatts (MW) hydroturbine known to be experiencing erosive cavitation and located at a hydro power plant in the American West\textsuperscript{12}. Vibration data were collected from four proximity probes mounted 90 degrees apart facing the hydroturbine’s main shaft. Proximity Probes 1 and

\textsuperscript{12}Our data source has asked for the exact location and details of the hydro plant to remain confidential.
2 were located near the hyroturbine’s lower bearing while Proximity Probes 3 and 4 were located near the upper bearing. Signals from the proximity probes were sampled at a rate of 10,000 Hz. The data used for feature selection and to create training sets were collected while the hydroturbine ran through a continuous ramp-down from 85 MW to 0 MW over a 100 second period, which was divided into 1 second blocks. The power produced versus time by the hydroturbine during the ramp-down is shown in Figure 4.3.

![Figure 4.3: Hydroturbine power versus time during the ramp-down](image)

The goal of this case study is to both demonstrate the methodology presented in this paper and compare hydroturbine cavitation classification accuracy using the following four approaches: 1) Classify cavitation with a naïve threshold classifier and a manually selected cavitation threshold, 2) Classify cavitation with a naïve threshold classifier and a cavitation threshold found by applying an unsupervised learning algorithm, 3) Classify cavitation with a supervised learning algorithm and training data that is manually labeled, and 4) Classify cavitation with a supervised learning algorithm and training data that is labeled by applying an unsupervised learning algorithm. A SVM was selected as the supervised learning algorithm to use for predicting cavitation classes and a K-Means was selected as the unsupervised algorithm for labeling training data. SVM and K-means algorithms used for this case study
are based on the corresponding built-in functions of Matlab (v2015a) with the Statistics and Machine Learning Toolbox.

The SVM, as described by Cortes and Vapnik [94], is a machine learning algorithm for binary classification problems that is frequently used to detect machine faults in the field of condition monitoring [95]. The concept behind a SVM is that input data, derived from a training set, is mapped to a high-dimensional feature space where a decision surface – often called the hyperplane, but for our purposes it can also be thought of as a threshold – is constructed. The SVM maximizes the distance between the classes and creates a hyperplane based on points near the decision surface, called support vectors, to create a globally optimal hyperplane for the training set. SVMs are capable of finding both linear and non-linear decision surfaces using a kernel function. Details about the kernel function and how the SVM finds the optimum hyperplane are beyond the scope of this paper and the reader is referred to [89, 96] for additional information. SVMs were selected for this case study due to their high accuracy, low computational burden, ease of use, and popularity in the machine learning community [97–99].

K-Means clustering, as described by Hartigan [100], is a heuristic algorithm that aims to divide $M$ data points into $K$ clusters so that the sum of squares is minimized within each cluster. The K-means algorithm used in this case study [101, 102] is iterative and requires the practitioner to choose a value for $K$ as well as $K$ data points, called seeds, that are initially assigned to their own cluster. Next, the point to cluster centroid distance of each data point is calculated and all points included in the cluster analysis are assigned to the cluster with the closest centroid. The new cluster centroid is then calculated and the data points are then re-assigned based on the new centroid. This repeats until clusters are no longer re-assigned after the new centroids are calculated. The final cluster results are dependent on the value of the $K$ seeds selected for the first centroid calculation. To obtain consistent results for establishing a cavitation threshold, a segmentation technique similar to bi-level thresholding [103] was used where the input value of $K$ was always equal to 2, and the minimum and
maximum CSP values in the training set were used as seeds.

4.5.1 Step 1: Select Cavitation Features

The fast Fourier transform (FFT) was calculated for each block of ramp-down data, then the variance spectrum was created to determine which vibration frequencies responded during the hydroturbine ramp-down. Each proximity probe had a similar response to the ramp down as can be seen in Figure 4.4.

![Figure 4.4: Variance spectrum of all four proximity probe signals](image)

Based on the variance spectrum, three frequency ranges (Figure 4.5) were identified as features to use for calculating CSPs:

- **Frequency Range 1** = 1 – 3 Hz
- **Frequency Range 2** = 3 – 30 Hz
- **Frequency Range 3** = 50 – 90 Hz

Frequency Range 1 is made up of frequencies below running speed while Frequency Range 2 includes the shaft rotating frequency and its first several harmonics. Frequency Range 3 includes the hydroturbine blade-pass and vane-pass frequencies.
4.5.2 Step 2: Create Training Sets

Next, the sample range to use as baseline data for the MD calculation, \( X_{\text{baseline}} \), was chosen by analyzing the standardized amplitude of the three CSPs in the time domain over the hydroturbine ramp-down as shown in Figure 4.6. The CSP’s amplitudes from sample 55 to 100 are relatively low and steady, so this sample range was chosen for \( X_{\text{baseline}} \).

As previously explained, Frequency Range 3 is expected to be the most sensitive to erosive cavitation; however, when a multi-dimensional classifier is used for prediction, all three ranges can be used to improve accuracy. One reason for the improvement in accuracy is that each frequency range has an independent response to flow during the ramp-down. The independence of each CSP is evident when comparing their standardized amplitude in the time domain during the ramp-down, as shown in Figure 4.6.

Once \( X_{\text{baseline}} \) was selected, the standardized MD distance was calculated for all of \( X \). It is important to note that MD can be calculated in its multivariate form (Equation 4.4) where \( X \) is a combination of CSPs from all the proximity probes, or the single variable MD can be calculated independently for each sensor with Equation 4.6. When performing the multivariate calculation, there will be a single set of MD values which means only a single
threshold will be needed for all the sensor measurements. The single variable calculation; however, will produce 4 sets of MD values and 4 thresholds, one of which will need to be selected for labeling training data. Results from both methods are presented in our case study.

Based on the Frequency Range 3 CSP values, cavitation thresholds were first selected manually – using both the multivariate MD calculation and the single variable method – then by automating the method utilizing a K-means clustering algorithm. The selected cavitation thresholds are shown in Table 4.1.

These cavitation thresholds are used for labeling training sets as well as for classifying cavitation when applying the naïve classifier. Training sets for binary classification can only have one label; however, a unique set of labels will be produced for each proximity probe due to slight variations in amplitude between each sensor. For example, Figure 4.7 shows several CSP values between sample 1 and sample 10 are above the cavitation threshold for Proximity Probe 3, but below the cavitation threshold for the other proximity probes. In our analysis, the classification labels established by applying the thresholds to data from Proximity Probe 2 were used for labeling the training sets.
Table 4.1: Cavitation thresholds for labeling training data

<table>
<thead>
<tr>
<th></th>
<th>Single Variable Thresholds</th>
<th>Multivariate Thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual 1</td>
<td>0.018</td>
<td>Manual</td>
</tr>
<tr>
<td>Manual 2</td>
<td>-0.110</td>
<td>K-means</td>
</tr>
<tr>
<td>Manual 3</td>
<td>-0.386</td>
<td>K-means 2</td>
</tr>
<tr>
<td>Manual 4</td>
<td>-0.392</td>
<td>K-means 3</td>
</tr>
<tr>
<td>K-means 1</td>
<td>0.338</td>
<td>K-means 4</td>
</tr>
<tr>
<td>K-means 2</td>
<td>0.162</td>
<td></td>
</tr>
<tr>
<td>K-means 3</td>
<td>0.494</td>
<td></td>
</tr>
<tr>
<td>K-means 4</td>
<td>0.622</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.7: The manually selected, single variable cavitation threshold (dashed red line)

Figure 4.8 shows the multivariate threshold and resulting classification labels found by applying a K-Means clustering algorithm to the training set containing all the proximity...
probe CSPs.

Figure 4.8: The multivariate cavitation threshold found through k-means clustering (dashed red line)

4.5.3 Step 3: Train Classifiers

The naïve classifier does not require additional training beyond establishing a cavitation threshold. A label for each new observation is generated by directly comparing its standardized MD to the cavitation threshold then labeling the observation ”1” if the value is above the threshold, or ”-1” if it is not. The observations the naïve classifier uses for comparison are data points from one proximity probe, for the single variable case, or all the proximity probes, for the multivariate case, with a single value based on the CSP calculated from Frequency Range 3. In other words, the naïve classifier uses a one-dimensional cavitation threshold and acts on one-dimensional data. Accuracy testing for the naïve classifier included one test for each proximity probe, and one for the multivariate threshold, with both the manually and K-Means selected thresholds resulting in 10 accuracy values.

The SVMs also relies on the labeled training sets to construct a decision boundary; however, the boundary can be multi-dimensional, which means the training set and testing
set can simultaneously include any or all of the proximity probes and CSPs. The benefits of a multi-dimensional decision boundary include more accurate classification predictions on data that is not linearly separable as well as the ability to extend the capabilities of a classifier to recognize more than just two categories of data. The multi-dimensional capability of a SVM also means a decision must be made about which proximity probes and CSPs to include in the training. For our analysis, we decided to train and test a SVM for every unique combination proximity probe and CSP, and compare the combinations with the highest accuracy. There are 4 proximity probes, and 3 CSPs for each of the proximity probes, which means that there are 12 individual training sets and \(2^{12} - 1 = 4095\) unique combinations of these 12 training sets.

We also looked at the multivariate threshold case where there is only one CSP for each frequency range for a total of 7 unique combinations. A potential advantage of using an SVM is its capability to find non-linear thresholds. The correctly classified test data shows that a non-linear cavitation threshold may be appropriate, to test this hypothesis we trained SVM models with polynomial kernels with orders 1 to 8 to test how a non-linear boundary affected classification accuracy. Non-linear SVM models were only trained and tested for the multivariate threshold case.

### 4.5.4 Classification Test Results

Data used for testing accuracy of the SVM and naïve classifiers as well as calculating cavitation intensity were collected while the hydroturbine ran for prolonged periods in 17 unique flow rates ranging from 5 MW to 85 MW in 5 MW increments. 24 seconds of data was collected for each flow rate which was then divided into 1 second blocks resulting in 408 total blocks of vibration data used to create the test data. Other running condition variables such as hydrostatic head, other turbines in the plant operating, and other factors were held effectively constant throughout the data collection period. The correct class labels for the training set were created manually using more traditional cavitation detection methods as well as sensor data from accelerometers and acoustic emission sensors. Additional informa-
tion on the full analysis and general cavitation detection methods used to create the class labels can be found in [13, 68? ? ].

The naïve and SVM classifier algorithms were applied to the test data and the resulting class predictions were compared to the correct class labels to determine the prediction accuracy. Equation 4.13 was used to calculate accuracy.

\[
\text{Accuracy} = \frac{\text{total \# correct class labels}}{\text{total \# of class labels}} \times 100 \tag{4.13}
\]

Cavitation intensity was calculated directly from the MD of the test data. The accumulated cavitation intensity over the whole data set, \( I_{total} \), is calculated by taking the MD of each CSP identified by the classifier as being in the cavitation class, \( X_{MD-cavitation} \) and multiplying it by the time block length used to create the CSP, \( t_{block} \) as shown in Equation 4.14. For the training and testing data, the time block length is 1 second and only CSPs created from Frequency Range 3 are used for intensity measurements.

\[
I_{total} = \sum (X_{MD-cavitation}) (t_{block}) \tag{4.14}
\]

Classifier accuracy results for the top performing training set combinations based on single variable thresholds are shown in Table 4.2. For the SVM results, proximity probe/CSP pairs are abbreviated with the proximity probe number first, "-", then 'CSP' followed by the frequency range used to create the CSP. For example, a training set created with data collected from Proximity Probe 1 that uses Frequency Range 1 for the CSP calculation would be abbreviated "PP1-CSP1".

Classifier accuracy results based on multivariate thresholds are shown in Table 4.3. Since all the proximity probe data is combined in the multivariate case, only the frequency range used for training and the order of the non-linear polynomial threshold are noted.

Figure 4.9 graphically shows the correct classification labels for the test data. Labels predicted by the naïve classifiers are shown in Figure 4.10 and Figure 4.11. Labels predicted by the SVM classifiers are shown in Figure 4.12 and Figure 4.13.
Table 4.2: Classifier test results for single variable thresholds

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Training and Testing Set</th>
<th>%</th>
<th>$I_{total}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>naïve, with manually selected threshold</td>
<td>Proximity Probe 1</td>
<td>95.3</td>
<td>2457</td>
</tr>
<tr>
<td></td>
<td>Proximity Probe 2</td>
<td>95.3</td>
<td>2473</td>
</tr>
<tr>
<td></td>
<td>Proximity Probe 3</td>
<td>93.6</td>
<td>2696</td>
</tr>
<tr>
<td></td>
<td>Proximity Probe 4</td>
<td>92.2</td>
<td>2395</td>
</tr>
<tr>
<td>K-Means selected threshold</td>
<td>Proximity Probe 1</td>
<td>91.2</td>
<td>2433</td>
</tr>
<tr>
<td></td>
<td>Proximity Probe 2</td>
<td>94.6</td>
<td>2473</td>
</tr>
<tr>
<td></td>
<td>Proximity Probe 3</td>
<td>89.0</td>
<td>2661</td>
</tr>
<tr>
<td></td>
<td>Proximity Probe 4</td>
<td>85.1</td>
<td>2395</td>
</tr>
<tr>
<td>SVM, trained manually selected threshold</td>
<td>PP2-CSP1, PP2-CSP3, PP3-CSP3, PP2-CSP2</td>
<td>95.6</td>
<td>2446</td>
</tr>
<tr>
<td></td>
<td>PP2-CSP1, PP2-CSP3, PP3-CSP3, PP2-CSP2, PP3-CSP2</td>
<td>95.3</td>
<td>2433</td>
</tr>
<tr>
<td></td>
<td>PP1-CSP3, PP2-CSP3, PP3-CSP3, PP1-CSP2, PP2-CSP2</td>
<td>95.1</td>
<td>2520</td>
</tr>
<tr>
<td>SVM, K-Means selected threshold</td>
<td>PP3-CSP1, PP4-CSP1, PP1-CSP3, PP4-CSP3, PP3-CSP2</td>
<td>99.0</td>
<td>2638</td>
</tr>
<tr>
<td></td>
<td>PP3-CSP1, PP4-CSP1, PP1-CSP3, PP2-CSP3, PP4-CSP2</td>
<td>98.7</td>
<td>2633</td>
</tr>
<tr>
<td></td>
<td>PP3-CSP1, PP4-CSP1, PP1-CSP3, PP2-CSP3, PP3-CSP2</td>
<td>98.7</td>
<td>2640</td>
</tr>
<tr>
<td></td>
<td>PP3-CSP1, PP4-CSP1, PP1-CSP3, PP1-CSP3, PP4-CSP2</td>
<td>98.5</td>
<td>2625</td>
</tr>
<tr>
<td>Classifier</td>
<td>Training and Testing Set</td>
<td>%</td>
<td>$I_{total}$</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-----------------------------------------------</td>
<td>------</td>
<td>------------</td>
</tr>
<tr>
<td>non-linear SVM trained with</td>
<td>CSP2, CSP3, 1st order</td>
<td>94.1</td>
<td>2711</td>
</tr>
<tr>
<td>manually selected multivariate</td>
<td>CSP2, CSP3, 2nd order</td>
<td>94.1</td>
<td>2711</td>
</tr>
<tr>
<td>threshold</td>
<td>CSP1, CSP2, CSP3, 3rd order</td>
<td>94.4</td>
<td>2720</td>
</tr>
<tr>
<td></td>
<td>CSP1, CSP2, CSP3, 4th order</td>
<td>96.8</td>
<td>2829</td>
</tr>
<tr>
<td></td>
<td>CSP1, CSP2, CSP3, 5th order</td>
<td>97.8</td>
<td>2866</td>
</tr>
<tr>
<td></td>
<td>CSP1, CSP2, CSP3, 6th order</td>
<td>97.3</td>
<td>2833</td>
</tr>
<tr>
<td></td>
<td>CSP1, CSP2, CSP3, 7th order</td>
<td>96.8</td>
<td>2798</td>
</tr>
<tr>
<td></td>
<td>CSP1, CSP2, CSP3, 8th order</td>
<td>96.1</td>
<td>2740</td>
</tr>
<tr>
<td>K-means selected multivariate</td>
<td>CSP2, CSP3, 1st order</td>
<td>94.1</td>
<td>2711</td>
</tr>
<tr>
<td>threshold</td>
<td>CSP3, 2nd order</td>
<td>93.6</td>
<td>2777</td>
</tr>
<tr>
<td></td>
<td>CSP3, 3rd order</td>
<td>94.0</td>
<td>2696</td>
</tr>
<tr>
<td></td>
<td>CSP1, CSP2, CSP3, 4th order</td>
<td>95.1</td>
<td>2916</td>
</tr>
<tr>
<td></td>
<td>CSP1, CSP2, CSP3, 5th order</td>
<td>97.5</td>
<td>2891</td>
</tr>
<tr>
<td></td>
<td>CSP1, CSP2, CSP3, 6th order</td>
<td>97.3</td>
<td>2891</td>
</tr>
<tr>
<td></td>
<td>CSP1, CSP2, CSP3, 7th order</td>
<td>96.8</td>
<td>2951</td>
</tr>
<tr>
<td></td>
<td>CSP1, CSP2, CSP3, 8th order</td>
<td>96.1</td>
<td>2891</td>
</tr>
<tr>
<td>naïve, with multivariate,</td>
<td>CSP3</td>
<td>94.1</td>
<td>2711</td>
</tr>
<tr>
<td>manually selected threshold</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>naïve, with K-means selected</td>
<td>CSP3</td>
<td>94.1</td>
<td>2711</td>
</tr>
<tr>
<td>threshold</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.9: Test cavitation data shown with correct classification labels as determined using traditional, manual analysis techniques.

Figure 4.10: Test cavitation data shown with labels predicted by the naïve classifier using the manually selected threshold.
Figure 4.11: Test cavitation data shown with labels predicted by the naïve classifier using a threshold found through k-means clustering.

Figure 4.12: Test cavitation data shown with labels predicted by a linear SVM model trained from data labeled using a manually selected, single variable threshold and the training set PP3-CSP1, PP4-CSP1, PP1-CSP3, PP4-CSP3, PP3-CSP2.
4.6 Discussion

The methodology outlined in this paper provides several benefits when compared to other cavitation detection strategies. Additionally, our method addresses common problems associated with cavitation thresholds and intensity measurements. In this section we discuss the benefits of using our cavitation detection process, as well as issues that the practitioner must keep in mind. While the method presented here does not yet provide cavitation erosion rate calculations, it provides the tools necessary to automate the collect of cavitation intensity data, a crucial step toward creating an erosion rate model for production hydroturbines.

The cavitation detection process described in this paper was demonstrated on ramp-down data collected from proximity probes on a hydroturbine experiencing erosive cavitation. We chose to demonstrate our method with proximity probe data because these types of sensors do not require data acquisition equipment capable of high sample rates and they are more frequently already installed on older hydroturbines. A benefit of our method, however, is that it can also be applied to data collected from other types of sensors including accelerometers,
acoustic emission sensors, or pressure transducers – any sensor type that can be used to create a CSP sensitive to erosive cavitation.

Another benefit of our cavitation detection process is that we address issues unique to long term data collection by establishing a cavitation threshold from hydroturbine ramp-down data and showing how the process can be automated with an unsupervised learning algorithm. These strategies allow the thresholds to adapt to changes in running condition with minimal disruption to power production, and without human intervention. For example, in our case study we established our thresholds from a 90 second ramp-down, while cavitation surveys traditionally used to collect data for cavitation detection require stepping the hydroturbine through a series of running conditions which can take several hours or even days to perform and many more hours of manual analysis. Practitioners should note that the frequency range used to calculate the CSP and ramp-down rate will change the amount of data required to establish a threshold.

The classifier accuracy results reported in our case study show that our methodology is flexible enough to be applied in several different ways to successfully classify cavitation. Accuracy results from cavitation thresholds found when combining single variable MD calculations with K-means clustering were quite promising. Optimism about these results should be tempered; however, because of potential sources of variability that are difficult to quantify in the classifier testing process. First, since there is no way to visually confirm the presence and intensity of cavitation in a production hydroturbine, the only way to create ground truth labels for the test data is through manual analysis of the data. The analysis techniques used are well-established and based on years of field experience, but still involve human judgement and cannot be more rigorously verified. In addition, the most accurate SVM-based test results may be misleading. By testing out all 4095 combinations of the 12 different sensors and CSPs, we are in a sense over fitting the data. A more general approach would be to note sensors that show up frequently in the top performing combinations and use those as an estimate of best case accuracy. For instance, combinations PP3-CSP1, PP4-CSP1,
PP1-CSP3, PP4-CSP3 show up in several of the top results and the accuracy result for this combination is just above 98% (not shown in the table). Another more obvious source of overfitting the data is in selecting the polynomial order for the non-linear thresholds. The most accurate non-linear threshold results are from using a 5th order polynomial; however, when the results are examined closely (Figure 4.13) there are a few points well above the visual cavitation threshold that are classified as non-cavitation.

The method presented in this paper is a good starting point for researchers and hydroturbine operators to better understand how to collect cavitation intensity data on a hydroturbines for an extended period of time. The method can be used to identify a CSP, automate the training and classification process, and keep thresholds relevant through changes in operating conditions. Previous researches did not directly address the concerns of outdated thresholds and automation of the detection process. The method presented here is already showing great promise with some hydroturbine operators and is expected to be deployed in the field soon.

4.7 Future Work

We are actively pursuing several areas of future work and propose the hydroturbine prognostics community pursue several larger goals. The most obvious next step is to verify the methods for cavitation detection and intensity measurements in a long term study on a production hydroturbine environment. The larger data sets collected from such a study could be used to verify the accuracy and adaptability of process we propose and ultimately lead to enough results to start correlating cavitation intensity measurements with erosion damage on the turbine runner.

Another area that should be investigated in parallel with collecting more data is to continue investigating supervised and unsupervised methods for labeling training sets and classifying cavitation. One suggestion would be to investigate relevance vector machines for classification [104]. The relevance vector machine is a classifier that provides a probabilistic interpretation of its output. When combined with logistic regression, this may serve as a
better tool for estimating erosion rates.

Finally, the use of multiple class SVMs for more general condition monitoring was not explored in this paper. SVMs could be used to classify faults associated with misalignment, bearing wear, and generator stator interaction to provide a better overall view of the hydroturbine’s health.

4.8 Conclusion

This paper presents both a novel method for creating adaptive cavitation thresholds as well as a machine learning framework for automated cavitation detection for hydroturbines. Adaptive thresholds can be used to address issues encountered during long term cavitation detection caused by variability in the hydroturbine’s operating conditions – a critical part of collecting consistent intensity data for estimating erosion rates on hydroturbine runners. The framework outlined in this paper for automated cavitation detection provides a guideline for making data collection more practical and accessible for hydroturbine operators and researcher who wish to estimate cavitation erosion rates and runner remaining useful life (RUL).

Adaptive cavitation thresholds are generated by first collecting sensor data from a hydroturbine ramp-down, creating cavitation sensitivity parameter (CSP)s from the data and calculating the Mahalanobis distance (MD) to create clear separation between the healthy running state and conditions where the hydroturbine is experiencing cavitation. This method allows a new cavitation threshold to be generated quickly and with minimal impact to the power production of the hydroturbine which allows them to be easily adapted to variation of the turbine’s running conditions. To automate the cavitation detection process, the cavitation threshold is used to create class labels for the ramp-down data which is then used to train a supervised learning algorithm for classifying cavitation from sensor data. Although domain knowledge is still required to select appropriate CSPs, the remainder of the process can be automated by applying unsupervised learning to label the training set.
To verify our methodology and demonstrate several ways it can be applied, we created single variable and multivariate thresholds from production hydroturbine ramp-down data then use them to create multiple training sets. The training sets were used to train a series of support vector machine (SVM) models and their accuracy was compared to a simple single variable naïve classifier based on the same training sets. For comparison purposes, cavitation thresholds were found through both a manual selection process as well as by applying K-Means clustering to the ramp-down data. Our results indicated that the fully automated process that utilized K-Means and SVMs for cavitation detection generally performed better than a process based on manually selected thresholds, thus demonstrating the usefulness of the machine learning framework. Our method provides hydroturbine operators and researchers with a clear and effective way to perform automated cavitation detection while also laying the groundwork for determining RUL in the future.
CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Future Work

The research contained within this thesis is thorough, but by no means complete. There exist several areas of future work that should be pursued by the hydroturbine prognostics community moving forward. The most obvious future work is to verify the methods for cavitation detection and intensity measurements in a long term study on a production hydroturbine environment. The larger data sets collected from such a study could be used to verify the accuracy and adaptability of process we propose and ultimately lead to enough results to start correlating cavitation intensity measurements with erosion damage on the turbine runner. A study of this magnitude would not be complete without collaboration with researchers from the cavitation modeling and experimental communities to help validate the findings and incorporate experimental erosion rate models into the remaining useful life (RUL) estimates.

Future investigation that can take place on a smaller scale should include additional development of supervised and unsupervised methods for labeling training sets and classifying cavitation. One suggestion would be to investigate relevance vector machines for classification [104]. The relevance vector machine is a classifier that provides a probabilistic interpretation of its output and, when combined with logistic regression, may serve as a better tool for quantifying the probability that cavitation is occurring and give a more direct way of estimating the risk associated with RUL calculations.

For cavitation feature selection, one area requiring further study is to better understand why different root mean square (RMS) frequency bands do not distinguish themselves from one another when being used to detect erosive cavitation. We discovered this issue when attempting to using F-tests to help rate cavitation detection features. A potential direction
of research is an in-depth investigation of spectral data produced from RMS frequency bands.

Finally, the use of multiple class support vector machine (SVM)s for more general condition monitoring was not explored in this paper. SVMs could be used to classify faults associated with misalignment, bearing wear, and generator stator interaction to provide a better overall view of the hydroturbine’s health.

5.2 Conclusion

The papers contained within this thesis present novel methods for comparing and evaluating cavitation detection features as well as creating adaptive cavitation thresholds to be used for automated cavitation detection. The goal of this research is to advance the hydroturbine community towards estimating RUL of hydroturbine runners. Additionally, the methodologies contained in Chapter 2 and 4 provide a structured path forward for hydroturbine operators wishing to implement or improve an erosive cavitation detection strategy at their plant.

The methodology outlined in Chapter 2 can be used to quickly compare features created from cavitation survey data collected on any type of hydroturbine, sensor type, sensor location, and cavitation sensitivity parameter (CSP). Although manual evaluation and knowledge of hydroturbine cavitation is still required for our feature selection method, the use of principal component analysis greatly reduces the number of plots that require evaluation. The method was applied to cavitation survey data collected on a Francis Hydroturbine resulting in a ranked list of the best sensor type, sensor location, and CSP to use on this hydroturbine for long term monitoring of erosive cavitation, thus demonstrating the usefulness of the method.

Chapter 4 presents both a novel method for creating adaptive cavitation thresholds as well as a machine learning framework for automated cavitation detection for hydroturbines. Adaptive cavitation thresholds are generated by first collecting sensor data from a hydroturbine ramp-down, creating CSPs from the data and calculating the Mahalanobis distance (MD) to create clear separation between the healthy running state and conditions where the
hydroturbine is experiencing cavitation. This method allows a new cavitation threshold to be generated quickly and with minimal impact to the power production of the hydroturbine which allows them to be easily adapted to variation of the turbine’s running conditions. To automate the cavitation detection process, the cavitation threshold is used to create class labels for the ramp-down data which is then used to train a supervised learning algorithm for classifying cavitation from sensor data. Although domain knowledge is still required to select appropriate CSPs, the remainder of the process can be automated by applying unsupervised learning to label the training set.

To prove out the methodology contained in Chapter 4 and demonstrate several ways it can be applied, single variable and multivariate thresholds were created from production hydroturbine ramp-down data then use them to create multiple training sets. The training sets were used to train a series of SVM models and their accuracy was compared to a simple single variable naïve classifier based on the same training sets. For comparison purposes, cavitation thresholds were found through both a manual selection process as well as by applying K-Means clustering to the ramp-down data. Our results indicated that the fully automated process that utilized K-Means and SVMs for cavitation detection generally performed better than a process based on manually selected thresholds, thus demonstrating the usefulness of the machine learning framework. This method provides hydroturbine operators and researchers with a clear and effective way to perform automated cavitation detection while also laying the groundwork for determining RUL in the future.
REFERENCES CITED


[38] The Knowledge Stream - Detecting Cavitation to Protect and Maintain Hydraulic Turbines. 2014.


APPENDIX A - MATLAB CODE FOR CAVITATION FEATURE SELECTION

MATLAB scripts used for analysis in Chapter 2. (Listing A.1)

Listing A.1: Code used for Feature Selection

```matlab
% Hydroturbine Cavitation Feature Selection:
% After initial loading and arranging of the cavitation survey data,
% the
% following steps are taken to select hydroturbine cavitation features:

% Step 1: Generate cavitation sensitivity parameters
% Step 2: Build the cavitation feature matrix
% Step 3: Normalize the columns of the cavitation feature matrix
% Step 4: Perform PCA on the cavitation feature matrix
% Step 5: Analyze PCA score plots
% Step 6: Calculate Correlation Coefficients
% Step 7: Calculate Min/Max Value Standard Deviations

% Load time series data and structure data matrices

% for raw twf signals load the following:
load 'on_off lf_twf_data.mat';
load 'off shaft raw U1.mat';
AEdata = [off shaft raw(:,1) off shaft raw(:,3)];
ACCdata = [off shaft raw(:,2) off shaft raw(:,4)];
clear('off shaft raw');
load 'on shaft raw U1.mat';
% BE CAREFULL: on shaft variable name is called 'off shaft raw'
AEdatanon = [off shaft raw(:,1)];
ACCdataon = [off shaft raw(:,2)];
clear('off shaft raw');
load 'low f raw U1.mat';
% BE CAREFULL: low f variable name is still called 'off shaft raw'
% Variable Order: LGB North, LGB East, TGB North, TGB East, DT Pressure
LFdata = [off shaft raw(:,1) off shaft raw(:,2) off shaft raw(:,3)...
          off shaft raw(:,4) off shaft raw(:,5)];
% imported data has a 0 MW running condition that needs to be removed to
```
% match all the other data. Remove first 245,760 data points:
LFdata = LFdata(245761:end,:);
clear('offshelfraw');

% Global Inputs
points = 32;  % number of output points for runpts() function
runcond = 17;  % number of running conditions
srateoff = 1330000;  % samples/sec for off shaft data
srateon = 1000000;  % samples/sec for on shaft data
sratel = 10000;  % " " " for low f data
acclow = 1000;  % acc band pass cut off low in Hz
acchigh = 40000;  % acc band pass cut off high in Hz
aelow = 1000;  % AE band pass cut off low in Hz
aehigh = 400000;  % AE band pass cut off high in Hz

% Step 1: Generate Cavitation Sensitivity Parameters

% AE sensor on shaft first:
% Apply overall bandpass filter, then apply two narrower filters:
filtaeon1 = buttbandfilter(AEdataon,srateon,aelow,aehigh);
filtaeon2 = buttbandfilter(AEdataon,srateon,50000,aehigh);
filtaeon3 = buttbandfilter(AEdataon,srateon,1000,50000);

% create rms features from each of the three different twf above, then
% create peak-peak, crest factor and kurtosis features:
Faeon1 = runpts([filtaeon1 filtaeon2 filtaeon3],points,runcond,@rms);
Faeon2 = runpts(filtaeon1,points,runcond,@max);
Faeon3 = runpts(filtaeon1,points,runcond,@peak2rms);
Faeon4 = runpts(filtaeon1,points,runcond,@kurtosis);
clear('AEdataon','filtaeon1','filtaeon2','filtaeon3');

% AE sensor off shaft is next (same steps as above):
filtaeoff1 = buttbandfilter(AEdata,srateoff,aelow,aehigh);
filtaeoff2 = buttbandfilter(AEdata,srateoff,50000,aehigh);
filtaeoff3 = buttbandfilter(AEdata,srateoff,1000,50000);

% create rms features from each of the three different twf above, then
% create peak-peak, crest factor and kurtosis features:
Faeoff1 = runpts([filtaeoff1 filtaeoff2 filtaeoff3],points,runcond,@rms);
Faeoff2 = runpts(filtaeoff1,points,runcond,@max);
Faeoff3 = runpts(filtaeoff1,points,runcond,@peak2rms);
Faeoff4 = runpts(filtaeoff1,points,runcond,@kurtosis);
clear('AEdata','filtaeoff1','filtaeoff2','filtaeoff3');

% Accelerometer on shaft features next:
filtacon1 = buttbandfilter(ACCdataon, srateon, acclow, 20000);
filtacon2 = buttbandfilter(ACCdataon, srateon, 20000, 30000);
filtacon3 = buttbandfilter(ACCdataon, srateon, 30000, 100000);

Faccon1 = runpts([filtacon1 filtacon2 filtacon3], points, runcond, @rms);
Faccon2 = runpts(filtacon1, points, runcond, @peak2peak);
Faccon3 = runpts(filtacon1, points, runcond, @max);
Faccon4 = runpts(filtacon1, points, runcond, @kurtosis);
clear('ACCdataon', 'filtacon1', 'filtacon2', 'filtacon3');

% Accelerometer off shaft features next:
filtaccof1 = buttbandfilter(ACCdata, srateon, acclow, 30000);
filtaccof2 = buttbandfilter(ACCdata, srateon, 35000, 65000);
filtaccof3 = buttbandfilter(ACCdata, srateon, acclow, 100000);

Faccoff1 = runpts([filtaccof1 filtaccof2 filtaccof3], points, runcond, @rms);
Faccoff2 = runpts(filtaccof1, points, runcond, @max);
Faccoff3 = runpts(filtaccof1, points, runcond, @peak2rms);
Faccoff4 = runpts(filtaccof1, points, runcond, @kurtosis);
clear('ACCdata', 'filtaccof1', 'filtaccof2', 'filtaccof3');

flf = [runptsRB(LFdata,10000,40,1000,points,runcond,@rms)...
runptsRB(LFdata,10000,1,40,points,runcond,@rms)...
runptsRB(LFdata,10000,40,1000,points,runcond,@max)...
runptsRB(LFdata,10000,40,1000,points,runcond,@peak2rms)...
runptsRB(LFdata,10000,40,1000,points,runcond,@kurtosis)];
clear('LFdata');

%% Step 2: Build the Cavitation Feature Matrix
% Features will be arranged in the following order:
% Acc1 = Bearing, Acc2 = Stem, Acc3 = On Shaft
% AE1 = Bearing, AE2 = Stem, AE3 = On Shaft
% PP1 and PP2 = lower bearing proximity probes
% PP3 and PP4 = upper bearing proximity probes
% Pr1 = Pressure transducer
% CSP order = [rms1 rms2 rms3 p2p cf kurt]

Faeon = [Faeon1 Faeon2 Faeon3 Faeon4];
Faeoff = [Faeoff1(:,1) Faeoff1(:,3) Faeoff1(:,5) Faeoff2(:,1) Faeoff3(:,1)...
         Faeoff4(:,1) Faeoff1(:,2) Faeoff1(:,4) Faeoff1(:,6) Faeoff2(:,2)...
         Faeoff3(:,2) Faeoff4(:,2)];

Faccon = [Faccon1 Faccon2 Faccon3 Faccon4];
Faccoff = [ Faccoff1(:,1) Faccoff1(:,3) Faccoff1(:,5) Faccoff2(:,1) Faccoff3(:,1) ... Faccoff4(:,1) Faccoff1(:,2) Faccoff1(:,4) Faccoff1(:,6) Faccoff2(:,2) ... Faccoff3(:,2) Faccoff4(:,2) ];

flf = flf(:,[1 6 11 16 21 2 7 12 17 22 3 8 13 18 23 4 9 14 19 24 ... 5 10 15 20 25]);

F32 = [ Faccoff Faccon Faccoff Faeon flf ];

%% Create Feature Labels

flabels = {'Acc1_1' 'Acc1_2' 'Acc1_3' 'Acc1_4' 'Acc1_5' 'Acc1_6' ... 'Acc2_1' 'Acc2_2' 'Acc2_3' 'Acc2_4' 'Acc2_5' 'Acc2_6' ... 'Acc3_1' 'Acc3_2' 'Acc3_3' 'Acc3_4' 'Acc3_5' 'Acc3_6' ... 'AE1_1' 'AE1_2' 'AE1_3' 'AE1_4' 'AE1_5' 'AE1_6' ... 'AE2_1' 'AE2_2' 'AE2_3' 'AE2_4' 'AE2_5' 'AE2_6' ... 'AE3_1' 'AE3_2' 'AE3_3' 'AE3_4' 'AE3_5' 'AE3_6' ... 'PP1_1' 'PP1_2' 'PP1_3' 'PP1_4' 'PP1_5' ... 'PP2_1' 'PP2_2' 'PP2_3' 'PP2_4' 'PP2_5' ... 'PP3_1' 'PP3_2' 'PP3_3' 'PP3_4' 'PP3_5' ... 'PP4_1' 'PP4_2' 'PP4_3' 'PP4_4' 'PP4_5' ... 'Pr1_1' 'Pr1_2' 'Pr1_3' 'Pr1_4' 'Pr1_5'};

%% Step 3: Normalize features for PCA

%% Normalize data. PCA is used to find data dimensionality and not evaluate
%% scale. Values that are relatively large (like crest factor) will dominate
%% PCA unless everything is normalized. Z-score normalization is used which
%% centers each row to have 0 mean and a standard deviation of 1.

ZF32 = zscore(F32);

%% Step 4: Apply PCA to Features

%% for PCA analysis, the data of interest is as follows:
%% score = matrix of projection vectors onto each principal component
%% latent = Principal component variances
%% percent = percent of variance in each principal component

[coef32, score32, latent32, tsquared32, percent32] = ...
    pca(ZF32, 'Centered', false);

% Normalized the principal component scores:
Zscore32 = zscore(score32);
%% Step 5: Evaluate PCA scores

%% First Plot Eigenvalues (Scree Plot):
na32 = latent32(1:12);
x = 1:12;

figure();
gopty = gca;
plot(x,na32,'bo',x,na32,'b-');
hold on
xlabel('Principal Component Number'), ylabel('Eigenvalue');
gopty.FontSize = 11;
gopty.LineWidth = 1.5;
gopty.Box = 'off';
hold off

%% Based on Scree Plot, Analyze the first 4 PC's
%% pcplot() formats the plots into the running condition domain for
%% publication:

pcplot(Zscore32(:,1),32);
pcplot(Zscore32(:,2),32);
pcplot(Zscore32(:,3),32);
pcplot(Zscore32(:,4),32);

%% Step 6: Calculate Correlation Coefficients

Cordata32 = []; for i = 1:4;

%% Correlate first three principal components to features:
cor = corrcoef([Zscore32(:,i), ZF32]);
Cordata32(:,i) = cor(2:end,1);
end
clear('cor');

%% Plot Absolute Value of Correlation Coefficients
%% only look at correlation plots for PC scores 1 and 4
corbarplot(abs(Cordata32(:,1)))
corbarplot(abs(Cordata32(:,4)))

%% Eliminate Features with low correlation:

thrpc1 = 0.9;  %threshold correlation value for PC 1
thrpc2 = 0.4;  %threshold correlation value for PC 2

% 1st PC correlation with 32 point data:
ZF32pc1 = ZF32;
ZF32pc1(:, Cordata32(:, 1)< thrpc1) = NaN;
ZF32pc1_labels = labels(1, ~isnan(ZF32pc1(1, :)));
ZF32pc1 = ZF32pc1(:, ~isnan(ZF32pc1(1, :)));

% 4nd PC correlation
ZF32pc4 = ZF32;
ZF32pc4(:, Cordata32(:, 2)< thrpc2) = NaN;
ZF32pc2_labels = labels(1, ~isnan(ZF32pc4(1, :)));
ZF32pc4 = ZF32pc4(:, ~isnan(ZF32pc4(1, :)));

%% Step 8a: Calculate Disspersion
% dispersion is calculated by the standard deviation at the minimum and
% maximum values as determined by the PC score plot:
minval = (3*points+1):5*points;
maxval = (10*points+1):12*points;
minstd = std(ZF32pc1(minval, :));
maxstd = std(ZF32pc1(maxval, :));

%% Step 8b: Examine Minimum and Maximum Dispersion
% Evaluate standard deviation at minimum signal value:
stddevbarplot(minstd', ZF32pc1_labels, 1);
stddevbarplot(maxstd', ZF32pc1_labels, 1);

%% Step 8c: Apply initial thresholds based on Step 8b
% min threshold = 0.05, max threshold = 0.25
selections = ZF32pc1;
bothstd = [minstd; maxstd];
% remove any features that have a minimum standard deviation above the
% threshold or a maximum standard deviation above the threshold and
% update
% remove the associated labels and standard deviation values as well:
selections(:, minstd(1,:)>=0.05 | maxstd(1,:)>= 0.25) = NaN;
selections_labels = ZF32pc1_labels(1, ~isnan(selections(1,:)));
bothstd = bothstd(:, ~isnan(selections(1,:)));
selections = selections(:, ~isnan(selections(1,:)));

%% Step 8c: Rank remaining features and graph rankings:
% sort the remaining features, labels and standard deviations based on the
% sort the combined standard deviation values in ascending order:
[~, I] = sort(sum(bothstd));

% 'I' now is the new index that is used to sort everything else:
bothstd = bothstd(:, I);
selections = selections(:, I);
selections_labels = selections_labels(:,I);

stddevbarplot('bothstd', selections_labels,2);

%%% Heat Map of Cluster Thresholds
%%% replace maximum value of feature into cluster 3 to display max value on
%%% heat map:
thpc1 = 0.1; %based on visual inspection

[THselections,aa,bb,threshindex] = threshcluster(selections,thpc1,32,0);

threshmap(THselections,32,selections_labels);
APPENDIX B - MATLAB CODE FOR CAVITATION DETECTION

MATLAB scripts used for analysis in Chapter 4. (Listing B.2)

Listing B.2: Code used for cavitation detection

```matlab
%% Script Outline:
% 1) Cavitation Feature Selection:
%    Step 1: Collect Ramp−down Data
%    Step 2: Calculate the Variance of Each Frequency
%    Step 3: Select CSP Frequency Ranges
% 2) Create Training Sets:
%    Step 1: Calculate CSPs
%    Step 2: Select Baseline Data for MD calculation
%    Step 3: Calculate the standardized MD
%    Step 4a: Plot and select thresholds
%    Step 4b: Apply K−Means to find thresholds
% 3) Cavitation Feature Classification:
%    Create Test Data
%    Train and Test SVM versus Naive Classifiers

%% 1) Feature Selection − Ramp−Down Data:

% Step 1: Collect ramp−down data (imported as a variable)
load('rampdown1.mat'); %load ramp−down data

%User Inputs:
chan = 4; %data channels
srate = 10000; %data sample rate in S/s
ds = 1; %length of sample divisions (sec)

% calculate the number of data segments
sdiv = (length(rampdown1)/srate)/ds;
x = 1:sdiv; % sample numbers for plotting

% calculate channel data segment ranges
srange = [1:sdiv; sdiv+1:2*sdiv; 2*sdiv+1:3*sdiv; 3*sdiv+1:4*sdiv];

%% 1) Feature Selection − Calculate Variance of each Frequency:
% Step 2: Divide ramp down data into blocks, remove DC Trend and center
% data around 0, Calculate FFTs, Calculate Variance
```
% Break up data into columns with the correct number of samples so each column has 'ds' time length of data. Also, remove the dc variation using 'detrend()'.

rampbits = reshape(rampdown1(:,1:4), [srate*ds sdiv*chan]);
rampbits = detrend(rampbits);

% peel out and average the MW for later use:
mw = reshape(rampdown1(:,5), [srate*ds sdiv*1]);
mw = mean(mw);
mw = reshape(mw, [sdiv 1]);

% Calculate FFT:
% calculate the normalized FFT. frange and res are used for plotting.

[rampfft,frange,res] = normfft3(rampbits,srate,0,0,.02);

% Find the Frequency Variance of the Ramp-Down Data % Separate FFT results by Sensor (PP1 – PP4)
% Calculate variance of each frequency across all ramp-down ffts

pp1fft = rampfft(:,srange(1,:));
vpp1fft = var(pp1fft');

pp2fft = rampfft(:,srange(2,:));
vpp2fft = var(pp2fft');

pp3fft = rampfft(:,srange(3,:));
vpp3fft = var(pp3fft');

pp4fft = rampfft(:,srange(4,:));
vpp4fft = var(pp4fft');

% 1) Feature Selection – Select Frequencies for CSP

% spectrum plot of first 200Hz for proximity probe 1
figure()
plot(frange(1:(200/res)),vpp1fft(1:200/res),'b')
hold on
plot(frange(1:(200/res)),vpp2fft(1:200/res),'r')
plot(frange(1:(200/res)),vpp3fft(1:200/res),'g')
plot(frange(1:(200/res)),vpp4fft(1:200/res),'c')
% gopt = gca;
% y1 = gopt.YLim;
% hold on
% plot([50 50],y1,'r',[90 90],y1,'r',[3 3],y1,'r',[30 30],y1,'r');
legend('Prox_Probe_1','Prox_Probe_2',...
'Prox_Probe_3','Prox_Probe_4')
xlabel('Frequency (Hz)');
ylabel('Variance');
hold off

% 2) Create Training Sets – Create CSPs

% design butterworth filters. 4 pole arranged by band based on variance
% spectrum results above.
f1 = 2;  % low pass cut off frequency range in Hz
f2a = 50;  % band pass frequency range in Hz
f2b = 90;
f3 = 30;
f4 = 3;

[z1 p1] = butter(4, f1/(srate/2), 'low');  % Frequency Range 1
[z2 p2] = butter(4, [f2a f2b]/(srate/2), 'bandpass');  % Frequency Range 2
[z3 p3] = butter(4, f3/(srate/2), 'low');  % Frequency Range 3
[z4 p4] = butter(4, f4/(srate/2), 'high');  % Frequency Range 3

% Calcu late RMS
Framp1 = rms(filter(z1, p1, rampbits));
Framp2 = rms(filter(z2, p2, rampbits));
Framp3 = rms(filter(z4, p4, filter(z3, p3, rampbits)));

Framp1 = reshape(Framp1,[100 4]);
Framp2 = reshape(Framp2,[100 4]);
Framp3 = reshape(Framp3,[100 4]);

% Standardize and plot to select baseline data for use in Ma distance
% calculation
Zramp1 = zscore(reshape(Framp1,[100 4]));
Zramp2 = zscore(reshape(Framp2,[100 4]));
Zramp3 = zscore(reshape(Framp3,[100 4]));

% 2) Create Training Sets – Select Baseline Data
figure();
plot(x,Zramp1(:,1),'o--b',x,Zramp3(:,1),'x--r',x,...
     Zramp2(:,1),'*--g')

legend('Frequency Range 1', 'Frequency Range 2', 'Frequency Range 3')
xlabel('Sample Number');
ylabel('Standardized RMS Amplitude');

% Based on inspection, baseline data is from sample 55:100
% 2) Create Training Sets – Calculate Standardized MD
bline = x(55:end);
Mramp1p1 = zscore(sqrt(mahal(Framp1(:,1),Framp1(bline,1))));
Mramp2p1 = zscore(sqrt(mahal(Framp2(:,1),Framp2(bline,1))));
Mramp3p1 = zscore(sqrt(mahal(Framp3(:,1),Framp3(bline,1))));
Mramp1p2 = zscore(sqrt(mahal(Framp1(:,2), Framp1(bline,2))))
Mramp2p2 = zscore(sqrt(mahal(Framp2(:,2), Framp2(bline,2))))
Mramp3p2 = zscore(sqrt(mahal(Framp3(:,2), Framp3(bline,2))))

Mramp1p3 = zscore(sqrt(mahal(Framp1(:,3), Framp1(bline,3))))
Mramp2p3 = zscore(sqrt(mahal(Framp2(:,3), Framp2(bline,3))))
Mramp3p3 = zscore(sqrt(mahal(Framp3(:,3), Framp3(bline,3))))

Mramp1p4 = zscore(sqrt(mahal(Framp1(:,4), Framp1(bline,4))))
Mramp2p4 = zscore(sqrt(mahal(Framp2(:,4), Framp2(bline,4))))
Mramp3p4 = zscore(sqrt(mahal(Framp3(:,4), Framp3(bline,4))))

% consolidate data and create variable labels
Mramp = [Mramp1p1, Mramp1p2, Mramp1p3, Mramp1p4, Mramp2p1,...
         Mramp2p2, Mramp2p3, Mramp2p4, Mramp3p1, Mramp3p2, Mramp3p3, Mramp3p4];

varnames2 = {'PP1_1', 'PP2_1', 'PP3_1', 'PP4_1', 'PP1_3', 'PP2_3', 'PP3_3',
             'PP4_3', 'PP1_2', 'PP2_2', 'PP3_2', 'PP4_2'};

% create table for using matlab trainer app if needed
Trainer2 = array2table(Mramp, 'VariableNames', varnames2);

%% 2) Create Training Sets – Plot and Manually Select Threshold
figure();
plot(x, Mramp2p1, 'o-b', x, Mramp2p2, 'x-r', x,...
     Mramp2p3, '*-g', x, Mramp2p4, '+:c')
legend('Prox_Probe_1', 'Prox_Probe_2', 'Prox_Probe_3',
       'Prox_Probe_4');
xlabel('Sample_Number');
ylabel('Standardized_Mahalanobis_Distance');
go = gca;
hold on

% Create training labels/classes based on plot inspection
classes1 = zeros(length(mw),1);
classes1(12:38) = 1;   %Select sample range for cavitation
classes1(1:11) = -1;   %Select sample range for no cavitation
classes1(39:end) = -1;

Mind = find(classes1==1,1,'last');

plot(get(gca,'xlim'),[Mthrall(2) Mthrall(2)],'r--');
hold off

%% 2) Create Training Sets – Select Threshold with K-means
% 3) Cavitation Feature Classification – Create Test Data
load 'lowfrawU1.mat';
% BE CAREFULL: low f variable name is still called 'offshaftrain'
% Variable Order: LGB North, LGB East, TGB North, TGB East, DT Pressure
LFdata = [offshaftrain(:,1) offshaftrain(:,2) offshaftrain(:,3)...
          offshaftrain(:,4)];
clear ('offshaftrain');
LFdata = LFdata(245761:245760*18,:); %account for zero MW condition.

% Shape data to match ramp–down data:
LFdata = reshape(LFdata,[245760 17*chan]);
LFdata = LFdata(1:240000,:);
LFdata = reshape(LFdata,[240000/24 24*17*chan]);
LFdata = detrend(LFdata);

% Calculate Features:
Frun1 = rms(filter(z1,p1, LFdata));
Frun2 = rms(filter(z2,p2, LFdata));
Frun3 = rms(filter(z4,p4, filter(z3,p3, LFdata)));
Frun1 = reshape(Frun1,[size(LFdata,2)/4 4]);
Frun2 = reshape(Frun2,[size(LFdata,2)/4 4]);
Frun3 = reshape(Frun3,[size(LFdata,2)/4 4]);

% 3) Cavitation Feature Classification – Create Test Data

% MD Features from normal running data:
Mrun1p1 = zscore(sqrt(mahal(Frun1(:,1),Framp1(bline,1))));
Mrun1p2 = zscore(sqrt(mahal(Frun1(:,2),Framp1(bline,1))));
Mrun1p3 = zscore(sqrt(mahal(Frun1(:,3),Framp1(bline,1))));
Mrun1p4 = zscore(sqrt(mahal(Frun1(:,4),Framp1(bline,1))));
Mrun2p1 = zscore(sqrt(mahal(Frun2(:,2),Framp2(bline,1))));
Mrun2p2 = zscore(sqrt(mahal(Frun2(:,2),Framp2(bline,2))));
Mrun2p3 = zscore(sqrt(mahal(Frun2(:,3),Framp2(bline,3))));
Mrun2p4 = zscore(sqrt(mahal(Frun2(:,4),Framp2(bline,4))));
Mrun3p1 = zscore(sqrt(mahal(Frun3(:,1),Framp3(bline,1))));
Mrun3p2 = zscore(sqrt(mahal(Frun3(:,2),Framp3(bline,2))));
Mrun3p3 = zscore(sqrt(mahal(Frun3(:,3),Framp3(bline,3))));
Mrun3p4 = zscore(sqrt(mahal(Frun3(:,4),Framp3(bline,4))));
Mtest1 = [Mrun1p1, Mrun1p2, Mrun1p3, Mrun1p4, Mrun2p1,Mrun2p2,...
          Mrun2p3, Mrun2p4, Mrun3p1, Mrun3p2, Mrun3p3, Mrun3p4];
for intensity measurement
Mrin = sqrt(mahal(Frun2(:,2),Framp2(bline,2)));
Mrin2 = sqrt(mahal(Frun2(:,4),Framp2(bline,4)));

% 3) Cavitation Feature Classification – Create Test Class Labels
% test data class labels are based on previous analysis where power
% outputs
% between 40 and 65 MW are thought to be experiencing cavitation
runclasses = ones(length(Frun2),1)*-1;
runclasses(169:312) = 1;

% 3) Cavitation Feature Classification – SVM Manual
% A linear Support Vector Machine is used for classification. Training
data
% is labeled using the manually selected threshold

svmfit = [];
svmp = [];

% train and test an SVM model for every unique combination of proximity
% probe, and CSP to test the top performers:
for i = 1:k
    testindex{i} = dindex(bindex(i,:)==1);
    [svmfit(:,i) ptest(i,:)] = svmtraintest(Mramp(:,testindex{i}),...
        classes1, varnames2(testindex{i}), Mtest1, ...
        runclasses, varnames2);
end
tic

% Look at variable combinations with accuracy larger than 95%
topsvm = find(pertest >= 0.950, 50);
topper = pertest(topsvm);
vtovsvm{1} = varnames2(testindex{topsvm(2)});
vtovsvm{2} = varnames2(testindex{topsvm(3)});
vtovsvm{3} = varnames2(testindex{topsvm(1)});

intensity{1} = sum(Mrinn2(svmfit(:,topsvm(2)) == 1));
intensity{2} = sum(Mrinn2(svmfit(:,topsvm(3)) == 1));
intensity{3} = sum(Mrinn2(svmfit(:,topsvm(1)) == 1));

% 3) Cavitation Feature Classification – SVM K-means
Training data is labeled using the k−means clustering selected threshold.

Based on inspection, the threshold from the 2nd proximity probe was used

```
svmfit2 = [];
pertest2 = [];
```

```
n = 12; %number of prox probe and CSP pairs
k = 2^n−1; %total number of unique combinations
dindex = 1:n;
bindex = decimalToBinaryVector(1:k); %binary conversion makes it easy
tic
for i = 1:k
    testindex{i} = dindex(bindex(i,:) ==1);
    [svmfit2(:,i) pertest2(i,:)] = svmtrain(Mramp(:,testindex{i})
        ,..., kclasses(:,3), varnames2(testindex{i}), Mtest1, ...
        runclasses, varnames2);
end
toc
```

```
% Look at variable combinations with accuracy larger than 93.7%
topsvm2 = find(pertest2 > 0.985, 50);
topper2 = pertest2(topsvm2);
```

```
vtopsvm2{1} = varnames2(testindex{topsvm2(2)});
vtopsvm2{2} = varnames2(testindex{topsvm2(3)});
vtopsvm2{3} = varnames2(testindex{topsvm2(4)});
vtopsvm2{4} = varnames2(testindex{topsvm2(1)});
```

```
intensity2(1) = sum(MrunI2(svmfit2(:,topsvm2(2)) == 1));
intensity2(2) = sum(MrunI2(svmfit2(:,topsvm2(3)) == 1));
intensity2(3) = sum(MrunI2(svmfit2(:,topsvm2(4)) == 1));
intensity2(4) = sum(MrunI2(svmfit2(:,topsvm2(1)) == 1));
```

```
% Manually selected threshold:
```
```
```
thrttest = ones(length(Frun2),4)*−1;
```
```
```
thrttest(Mrun2p1 > Mthrall(2),1) = 1;
```
```
```
thrttest(Mrun2p2 > Mthrall(2),2) = 1;
```
```
```
thrttest(Mrun2p3 > Mthrall(2),3) = 1;
```
```
```
thrttest(Mrun2p4 > Mthrall(2),4) = 1;
```
```
```
correct(1) = (length(runclasses(thrttest(:,1) ==...
        runclasses)))/length(runclasses);
correct(2) = (length(runclasses(thrttest(:,2) ==...
        runclasses)))/length(runclasses);
correct(3) = (length(runclasses(thrttest(:,3) ==...
```

119
runclasses() / length(runclasses);
correct(4) = (length(runclasses(thrtest(:,4) ==...
runclasses())) / length(runclasses);

intensity3(1) = sum(MrunI2(thrtest(:,1) == 1));
intensity3(2) = sum(MrunI2(thrtest(:,2) == 1));
intensity3(3) = sum(MrunI2(thrtest(:,3) == 1));
intensity3(4) = sum(MrunI2(thrtest(:,4) == 1));

% 3) Cavitation Feature Classification – Testing Naive K-means
% Manually selected threshold:
thrtest2 = ones(length(Frun2),4)*-1;
thrtest2(Mrun2p1 > kthr(2),1) = 1;
thrtest2(Mrun2p2 > kthr(2),2) = 1;
thrtest2(Mrun2p3 > kthr(2),3) = 1;
thrtest2(Mrun2p4 > kthr(2),4) = 1;

correct2(1) = (length(runclasses(thrtest2(:,1) ==...
runclasses())) / length(runclasses);
correct2(2) = (length(runclasses(thrtest2(:,2) ==...
runclasses())) / length(runclasses);
correct2(3) = (length(runclasses(thrtest2(:,3) ==...
runclasses())) / length(runclasses);
correct2(4) = (length(runclasses(thrtest2(:,4) ==...
runclasses())) / length(runclasses);

intensity4(1) = sum(MrunI2(thrtest2(:,1) == 1));
intensity4(2) = sum(MrunI2(thrtest2(:,2) == 1));
intensity4(3) = sum(MrunI2(thrtest2(:,3) == 1));
intensity4(4) = sum(MrunI2(thrtest2(:,4) == 1));

% Classification Plots:
plotbiclass(Mrun2p4, svmfit(:, topsvm(5))
plotbiclass(Mrun2p4, svmfit2(:, topsvm2(2))
plotbiclass(Mrun2p4, thrtest(:,2))
plotbiclass(Mrun2p4, thrtest2(:,2))
plotbiclass(Mrun2p4, runclasses)