UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A DATA-DRIVEN APPROACH FOR MONITORING AND PREDICTIVE DIAGNOSIS OF SUCKER ROD PUMP SYSTEM

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

NGOC LAM TRAN
Norman, Oklahoma
2022

A DATA-DRIVEN APPROACH FOR MONITORING AND PREDICTIVE DIAGNOSIS OF SUCKER-ROD PUMP SYSTEM

A THESIS APPROVED FOR THE
MEWBOURNE SCHOOL OF PETROLEUM AND GEOLOGICAL ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Hamidreza Karami, Chair

Dr. Catalin Teodoriu, Co-Chair

Dr. Xingru Wu

Acknowledgments

I would like to thank Dr. Karami and Dr. Teodoriu for their continuous support and mentorship. Both professors have been teaching me since my undergraduate years. The knowledge and insights that they have taught motivate me to become a better engineer. I am grateful for the research opportunities provided by Dr. Karami and Dr. Teodoriu, especially in production and drilling areas that are very challenging for women engineers back in my home country. Dr. Karami has encouraged me to pursue and explore data science prowess, which has been instrumental to my thesis completion. I am honored to be among the first students to work under Dr. Karami's research group. The safety lesson from Dr. Teodoriu is something that I will never forget. I also would like to thank Dr. Wu for his feedback and guidance to better this work.

I very much appreciate Dr. Bello and Aditya for taking the time to assist me in running the experimental work and troubleshooting at the facility.

I would like to thank my family for being with me throughout this journey. I wouldn't become myself today without the countless sacrifices that my parents made to give their children a better future.

Finally, I also have a few words for my husband Ishank. He has stood beside me through many challenges. He also introduced me to machine learning research in the oil and gas industry. It would not be possible for me to achieve this far without his constant support.

Table of Contents

List of Tables

List of Figures

Abstract

Given its long operational history, a sucker-rod pump (SRP) has been widely utilized as a lifting solution to bring reservoir fluids to the surface with low cost and high efficiency. However, debugging the rod pump issues requires on-site activities that could cost time and money for operators. With the vast dataset collected from years of operation, numerous companies are looking to turn these engineering processes into automated systems in the oilfield network, despite the complexity of data and lack of knowledge. The integral approach is to develop real-time diagnostics for downhole conditions. The emerging artificial intelligence and big-data analytics have provided relatively precise downhole condition forecasting based on available data, enabling better decision-making. This thesis focuses on collecting representative data and utilizing machine learning techniques to predict operational anomalies of sucker-rod pumps.

An experimental design at the University of Oklahoma, referred to as Interactive Digital Sucker Rod Pumping Unit (IDSRP), was used to facilitate a data-driven solution and monitor SRP performance and diagnostics. The physical framework includes a 50-ft transparent casing and tubing with a downhole rod pump at the bottom. A linear actuator provides the rod string's reciprocal movement and simulates different surface units and operating scenarios. This facility uses proper instrumentation and a data acquisition system for signal sensor readings. A workflow is developed to translate surface dynamometer cards to downhole ones and train predictive models in time-driven pressure and rate data. Though primarily focusing on the normal pump operation, the test matrix varies in stroke length, pump speed, and rod movement shape.

The tests validate the model to classify and detect various operational conditions in sucker-rod pumps. The model dynamically categorizes the pumps into key states of ideal condition and

over-pumping with a regression fit of accuracy higher than 0.7 and overall classification accuracy of 92%. Moreover, the real-time model anticipates an event in which the pump experiences a slight pumping-off that could potentially deteriorate the rod. The results also help understand key features that drive sucker rod pump performance prediction and help detect anomalous pump behavior. The machine learning algorithms, developed by the physics-based inputs, generate predictive models, thus classifying operational conditions or failures of the pump. The diagnosis for the pump's anomalies is also predicted by a real time analysis. The visualization enables to recognize the patterns and abnormal phases early. The explainable machine learning (i.e. Shapley additive explanation) helps decoding the predictive models with feature importance, local and global sensitivities in categorizing SRP conditions.

The developed unit has the capability of working with different well conditions, combining with real-time training and applying models, to initiate early warnings. The developed processes and workflows have the potential of becoming a generic optimizing and monitoring model for rod pumps. The novelty of this setup consists not only in its mechatronic design but also in through monitoring of the pump operations.

Chapter 1: Introduction

Artificial Lift includes various categories of production enhancement techniques applied to adjust the well bottom-hole pressure and increase the production of reservoir fluids. Sucker rod pumps (SRP) or beam pumps are widely considered the most applicable lifting method in the industry. With the ever-increasing number of declined wells, the popularity of rod pumps is making them even more popular. **Figure 1.1** shows the data of SRP usage considering the number of wells and production in comparison with other lift methods and natural flow. SRP accounts for the highest numbers of wells with 21%, after naturally flowing wells and plunger-lifted gas wells. The typical well candidates for SRP have low to medium producing capacity, resulting in a smaller production share of only 7%. However, SRP can produce as high as 8000 barrels of fluid per day (Takacs 2015).

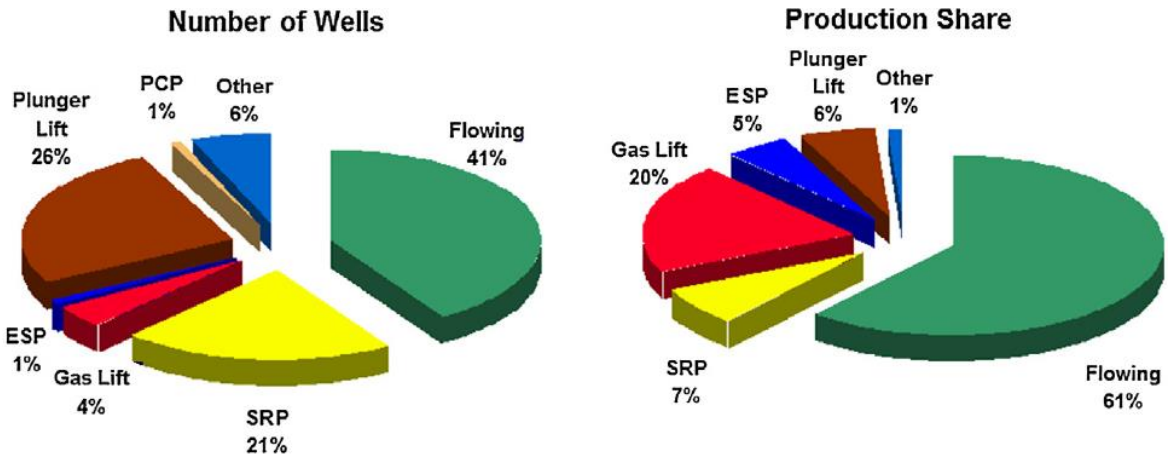


Figure 1.1 Artificial lift shares based on numbers of operating wells and their shares from overall oil production (Takacs 2015)

A crucial reason for most operators to select SRP lies in system efficiency, defined as the ratio of total input energy used to lift reservoir fluids to the surface. The efficiency may vary over a wide range of values because of several factors in the pumping unit, the energy consumed by the pumping activity, or power losses in the well and surface. **Figure 1.2** shows the efficiency ranges

for various artificial lift techniques. SRP is considered to be a relatively more effective artificial lift method than the others, with the exception of PCP. However, the maintenance issues with the PCP make it a more limited choice, leaving SRP as the most commonly used pumping technique.

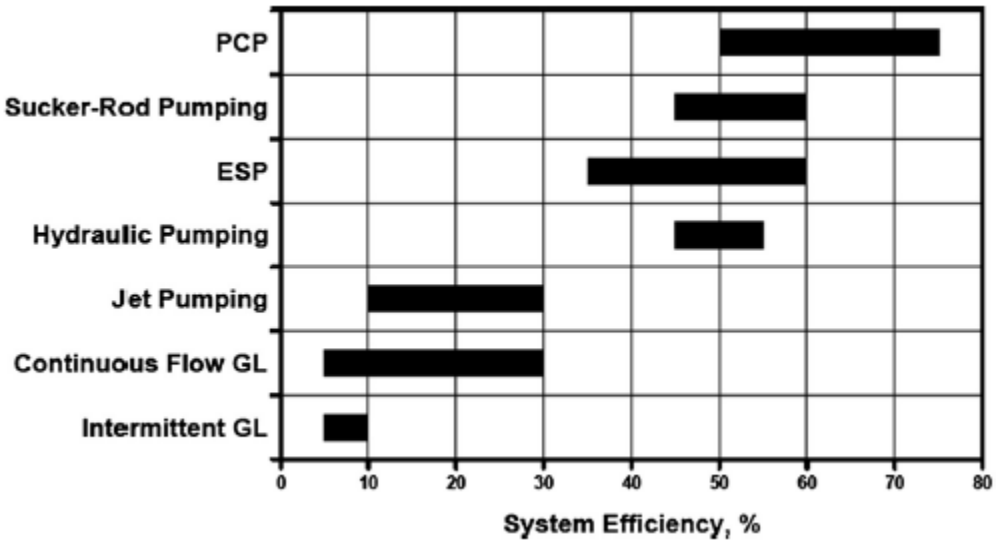


Figure 1.2. System efficiency ranges for various artificial lift methods (Takacs 2015)

With the long history of operation and dynamic growth of technology, various solutions have been developed for longer production and reliability. However, their online monitoring and optimization come with many challenges. One of the main drawbacks of SRP is the industry's tendency to move towards increased well depths with exponential production decline. The limitation of material properties and increasing well complexity hinder robust solutions suitable for various well candidates. Therefore, individually engineered artificial methods are imminent to provide the solutions for maintaining long-term production.

A large dataset is crucial for making the systems more automated and optimized, especially failure detection and performance monitoring. Given vast datasets collected from years of operation, many operators enact digital technology to generate automated artificial lift systems in the oilfield networks. The key to make sucker-rod pumps operate effectively lies in the downhole

condition diagnostics. The emerging artificial intelligence and big-data analytics aim to provide relatively precise downhole condition forecasting based on the available data, enabling better decision-making.

This study presents an experimental and analytical workflow for monitoring a sucker-rod pump and performing diagnostics. A lab-designed Interactive Digital Sucker-Rod Pumping Unit (IDSRP) is used to replicate field operations and associated data analysis. This unit consists of a vertical 50-ft long facility with a rod pump at the bottom and proper instrumentation, capable of simulating a rod-pumped wellbore. The surface dynamometer cards, and time-driven pressure and rate data are collected to process and train a cloud-based analytics software platform. Some of the tested scenarios are normal pump operations with varying rates. The applied online prototype is designed to provide a step towards digitized automation systems.

1.1 Objectives

The earlier work done by Pienknagura Dolberg (2019), and Bello et al. (2020) laid out the conceptual foundation of SRP facility design and device assembly. However, various objectives were not addressed thoroughly, especially the experimental development of an autonomous SRP simulation through an associated software environment. Therefore, these primary objectives continue in this study. The main objectives of this work are:

- Operate and further the implementation of digital sucker-rod pumping unit developed at University of Oklahoma. This includes developing a manual control testing matrix and validating the baseline model of operation. From there, datasets are generated for big data analytics training and testing.

- Model the elastic behavior of the rod string, digitize the dynocards collected for diagnostic mode through implementing signal processing algorithm, and transforming surface to downhole cards by solving Gibbs equation. The algorithm is validated based on the facility condition to ensure the reliability of the operation.
- Develop an analytical model with training and testing datasets obtained from the experiments. The predictive model is trained and validated to diagnose the pump condition, used for SRP health and performance monitoring.

1.2 Methodology

This study is an attempt to integrate an experimental Interactive Digital Sucker-rod Pump (IDSRP) facility with an analytical workflow to build the automated diagnostic systems for its operational framework. The workflow and processes developed in this study will serve as a prototype for artificial lift automation in the industry.

Figure 1.3 highlights the experimental and analytical approaches implemented. The experimental setup is completed to mimic the operation of a sucker-rod pump in the field scale. A physical baseline model is implemented to correlate the actuator's speed with the linear reciprocating movement of the rod string.

The data obtained from the experiments are treated for feature selection in the diagnostic model. The treatment stage includes resolving missingness and handling outliers in the raw dataset. The dataset is split in training set for model validation and test set for model evaluation. The model is retrained, if the resulting evaluation parameters (i.e., accuracy, percentage errors, etc.) are lower than the common goals. The end stage involves deployment of the workflow to detect the imminent performance of the rod unit.

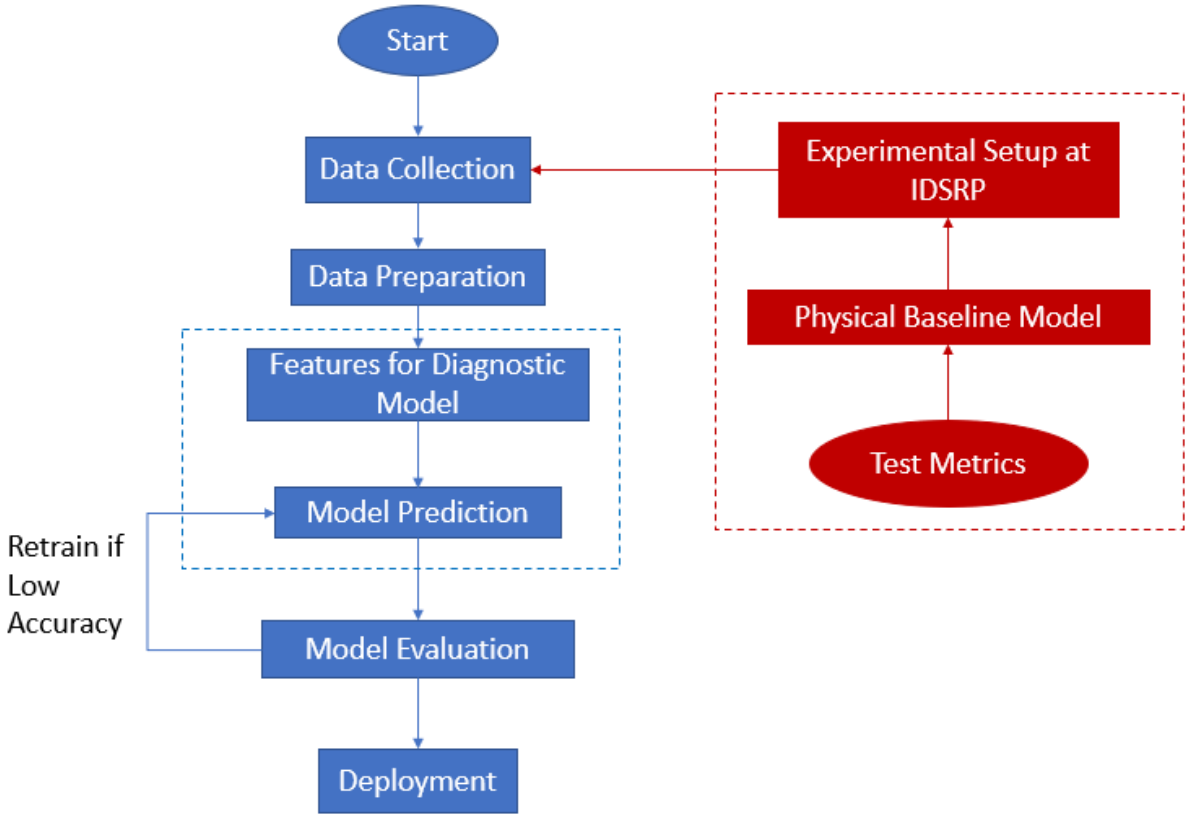


Figure 1.3 Workflow schematic combining experimental and data analytics approaches

1.3 Outline

This thesis is structured to present the objectives and outcomes of the study clearly and thoroughly. First, in **Chapter 2**, the fundamental concepts of sucker-rod pump units are addressed, and some previous work that deploy data-driven methods to diagnose and monitor SRP operations are reviewed. The IDSRP facility setup and experimental design are discussed in **Chapter 3**. A description of the three modes in a full-scale facility, data acquisition method and physical baseline calibration are explained in this chapter. An overview of experimental results and digitization of dyno card transformations utilizing Modified Everitt-Jennings are shown in **Chapter 4**. **Chapter 5** provides further analysis of parametric properties towards classifying pump operation motion and attempts to deploy predictive model to cluster pump operation condition based on regression

and classification methods. Finally, **Chapter 6** highlights the important outcomes of this study and notable additional work that could possibly advance the study and address its inherent challenges.

Chapter 2: Literature Review

This chapter provides a summary of the literature survey status with regards to the main topic of this study. First, a general review is conducted on sucker-rod pumps as the most common artificial lift technique in the industry. The advantages, limitations, and malfunction detection techniques of rod pumps are discussed. Then, the recent advances in data-driven models applied to rod pumps and current gaps in the literature are summarized.

2.1 Sucker Rod Pump Limitations

Artificial lift's main objective is to recover more reservoir fluids to surface. It serves the purpose of providing sufficient pressure to the fluids to push them up the wellhead. Sucker-rod pump or beam pump is one of the most generic artificial lift methods for delivering oil production practically and affordably. It is considered effective and simple in onshore fields with the ability to operate with very low reservoir pressures. This mechanical lifting method also suits varying fluid properties, even with high reservoir temperatures and viscosities. Moreover, the setup requires minimal cost when changing to other wells. However, the mechanical setup has some limitations, such as excessive friction in deviated wellbores, solids handling, gas interference, and over rod capacity.

One of the common problems when adding SRP in deviated borehole lies in the preventative measure to avoid severe dog legs (Ryan 1988, Gibbs 1991). If the rod fails to pass the build section in tubing, the tolerance capacity between barrel and plunger elements significantly decreases, leading to imminent damage. The continuous pumping at high speed also

intensifies the negative impact of wear action or corrosion in deviated well or doglegs (Ryan 1988), not to mention the presence of frictional loss when the rod is located at the horizontal section of the well. Several practices in the field suggest utilizing different coatings, spray materials in couplings to remediate the frictional wear (Caslin 1987 and Jackson et al. 2018).

The issue of solids handling in SRP occurs in sandy well operation when the pump suffers from sand abrasion that leads to plunger sticking in barrel, pump sticking in tubing, the damages of barrels, plunger, and valves due to sand corrosion, sand packing off, and sticking valves (Ghareeb and Beck 2012). The case study from water polymer flooding well also found that the presence of solids/sands is the leading case for plunger/barrel failure and causes a significant increase in corrosion rate (Al-Sidairi et al. 2018). The preventative field measures for solids handling include maintaining long stroke, using continuous and uniform rod diameter, and combining different materials and filtration/screening systems.

As for the gas issue in rod pump, the gas breakout in the pump prevents the movement of plunger into the barrel, leading to wear in rod guide and rod. The resulting compression also affects the well and causes rod buckling (Dove and Smith 2016). Based on the second plot from **Figure 2.1**, the sharp decrease in load as the plunger begins to move downward during the downstroke shows up as hyperbolic shape due to gas compression. During the upstroke, the gas presence at pump intake fills in the pump barrel, and the pump load is less than the max fluid load in the plunger as gas slowly carries out the load. The repeated pattern will occur as the rate of pressure build is directly proportional to the gas rate (Rowland and McCoy 2015).

When the pump runs at a very high rate or experiences low downhole pressure, the large differential pressures between the pump and the barrel during upstroke also trigger the gas dissolution in the rod pump and, as a result, not enough liquid to fill the barrel. Also, the exceeding

production threshold in reservoir from displacement also generate the incomplete pump fillage that can cause well being pump-off (Rowland and McCoy 2015). Continuing the first plot in **Figure 2.1** for fluid pound issue, the compressed force is introduced all the load due to insufficient liquid level in the annulus, leading to the sharp reduction in load during the closing of standing valve and the opening of traveling vale. This generates the sudden agitation, which could snap the rod and damage the pump.

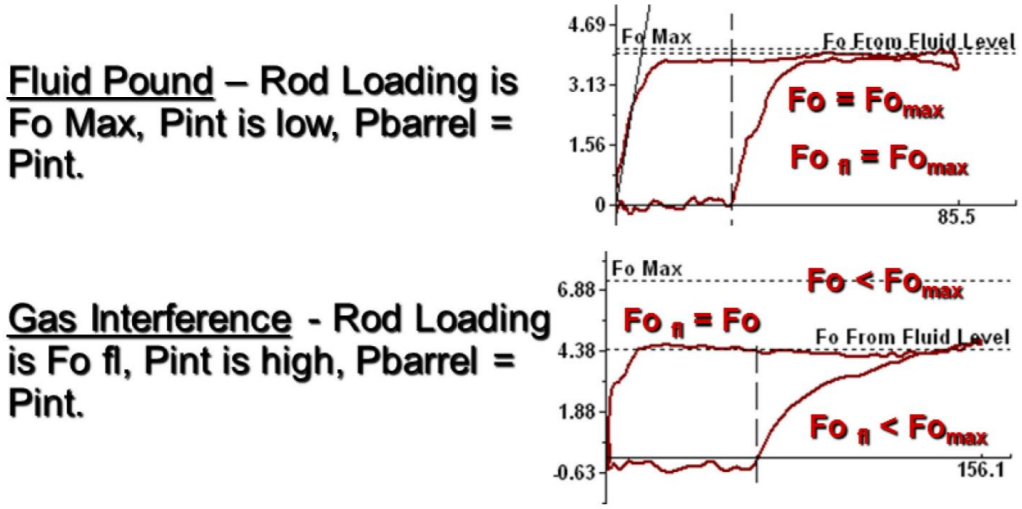


Figure 2.1 Dynamometer Cards from Pump experiencing Fluid Pound (1) and Gas Interference (2) (Rowland and McCoy 2015)

The recent more data-driven work by Zhao et al. (2019) studies the load data and incorporates the operational database by looking into the effects of stroke length, stroke frequency, diameter, or pump depth, to build sensitivity analysis for the eccentric wear prevention program. Note that the data from this approach only obtained from a particular oilfield operation, and the limitation of this approach also lies in the proprietary software used in the analysis, which is not generally available in the public domain. Therefore, the digital solutions, through data obtained from sensors and pattern visualization in software interface at IDSRP, allow to generalize the database and develop the predictive model to recognize pumping problems.

2.2 Understanding Pumping Units

The basic understanding of different pumping unit types is imperative to help understand the physical setup at IDSRP facility. The pumping unit geometries include conventional (type C), air-balanced unit (type A), Lufkin Mark II (type M) (**Figure 2.2**) and long stroke motion (**Figure 2.3**). The primary difference lies in the counterbalance setup. The counterbalance in Type-C unit is located on the crank arm or rear end of the beam, making it possible to run in both clockwise and counter clockwise directions. The counterbalance in Type-M unit is located on the opposite side of the crank at a specific angle. It also has an additional rotary counterweight, creating a more uniform net torque variation during the upstroke and downstroke cycles (Takacs 2015). Type-A unit utilizes the air-balanced mechanism of an air-compressed unit. The air pressure of the cylinder can be adjusted to maintain the balance. These differences lead to changes in mechanical characteristics, the crank angle at top and bottom of the polished rod, and walking beam length for example. This results in distinctive kinematics discussed in both Gray (1963) and Takacs (2015) books. Note that the selection of a specific pump geometry provides further mechanical and physical insight regarding the rod displacement behavior, which is discussed in the subsequent chapters.

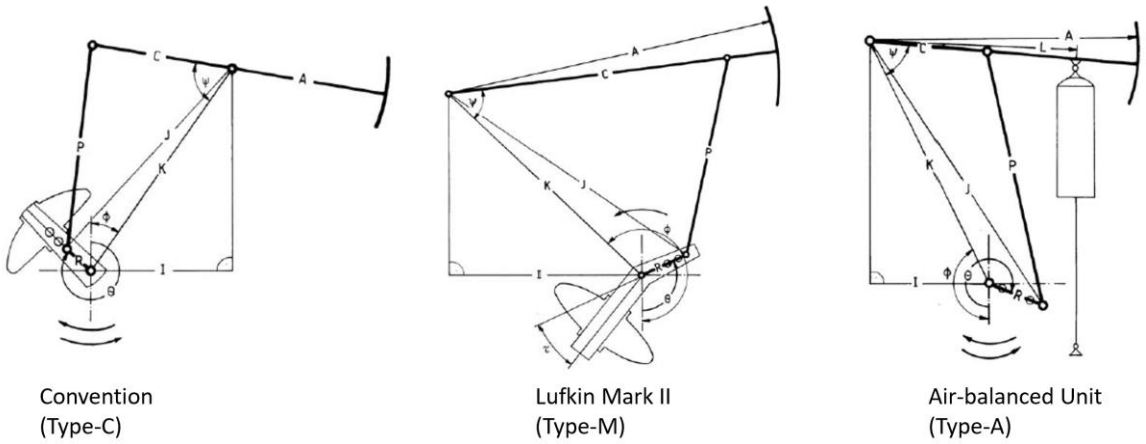


Figure 2.2 Convention, Lufkin Mark II and Air-balanced Pumping Units (Takacs 2015)

Long stroke units are developed to combine the advantages of the three pump units above. These units offer long-stroke capacity, constant rod string velocity, and fewer strokes per barrel, making them suitable candidates for high-volume, high-load, deviated, and unconventional horizontal wells. These characteristics increase the life span of pumping and allow more complete pump fillage and lower loading. They hence avoid common issues of gas interference and exceeding stress and strain thresholds from high peak rod load (Kennedy and Ghareeb 2017). Two principal types for the long stroke unit include mechanical and hydraulic pumping units (**Figure 2.3**). The mechanical long stroke unit converts the rotation in prime mover into the reciprocating movement that lifts formation fluids. The hydraulic long stroke units have a power unit that provides the hydraulic power to the piston. This power lifts the peak load during the upstroke and helps release hydraulic pressure in the downstroke with a controlled rate and extended cylinder (Ghareeb et al. 2017).

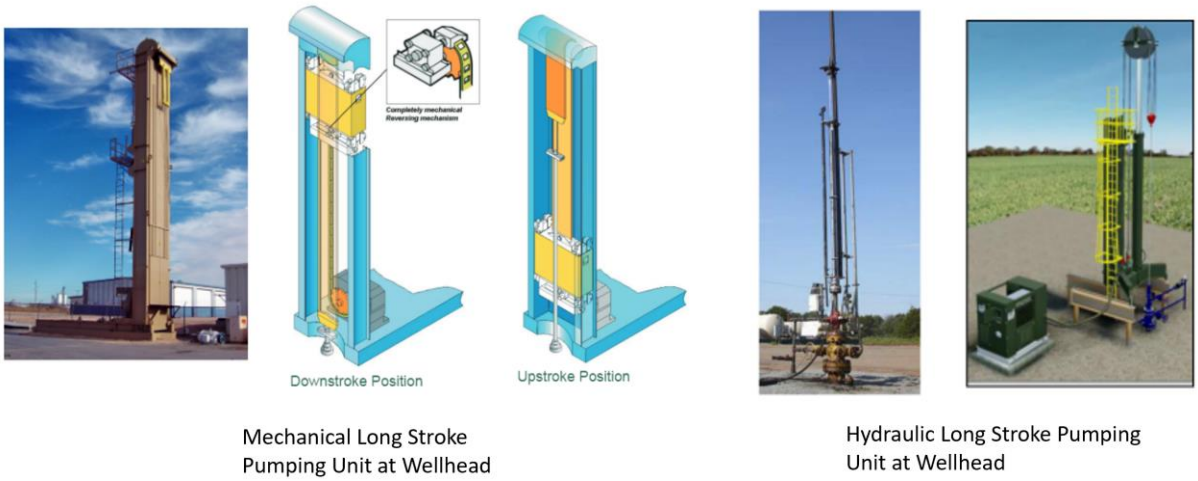


Figure 2.3 Long Stroke Pumping Units at Well Head: Mechanical and Hydraulic Pumping Types (Weatherford 2007, Kennedy and Ghareeb 2017, Ghareeb et al 2017)

An example of a dynamometer card for long stroke pumping units reveals the sloped lines due to the changes of acceleration and direction (**Figure 2.4**). The automatic control of strokes per minute and acceleration during the upstroke and downstroke transition and the linear movement simplifies the predictive diagnosis of the card (Hollenbeck 1980). With a smaller facility footprint, the IDSRP could follow the rotation-to-linear mechanism in both conventional and long stroke units and capture field-scale acceleration, speed, and displacement properties into the simulated environment. The automatic control mechanism from the long-stroke unit maps out the idea of creating a platform that tracks the operational input to maintain the longevity of the pumping life span.

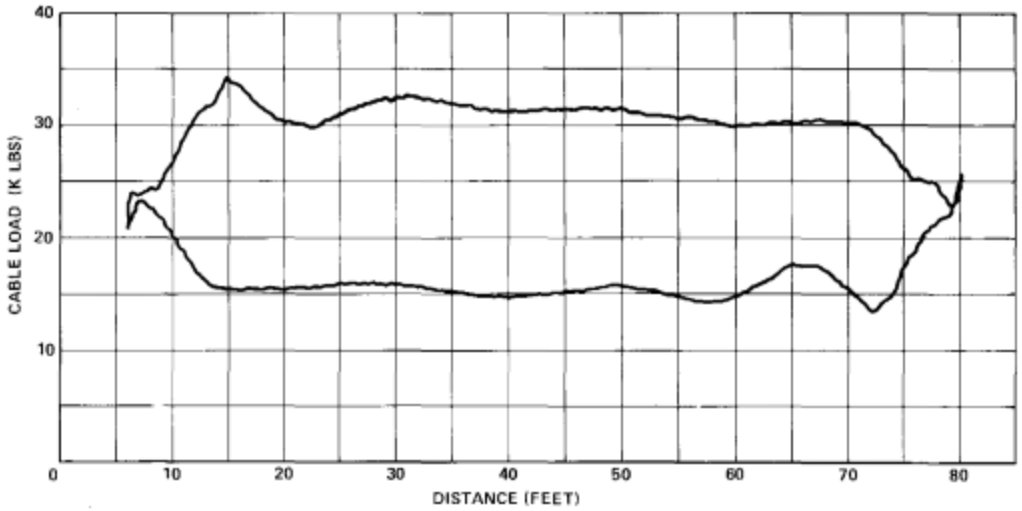


Figure 2.4 Example of dynamometer card obtained from a long stroke SRP unit (Hollenbeck 1980)

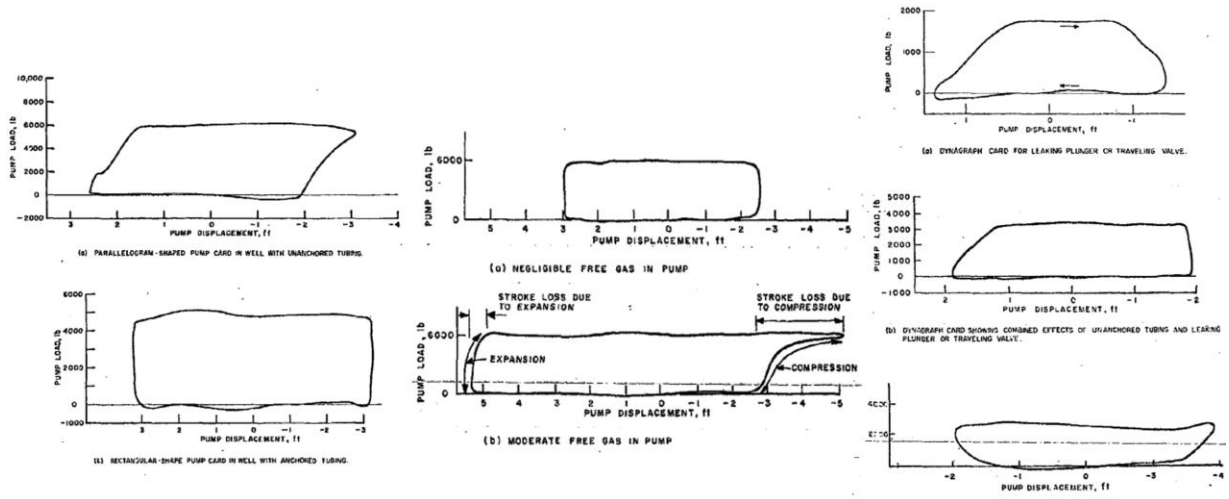
2.3 Data-Driven Models

Given its operational advantages and disadvantages, it is essential to monitor the conditions of sucker-rod pumps for sustaining operations. One of the common methods is to extract the surface dynamometer card, which records rod displacement and load or axial force in a full cycle. By looking at the dynocard, operators can monitor the performance and identify sucker-rod pump's possible faults (Tripp 1989).

To make an accurate identification of a rod pump's condition, Gibbs and Neely (1966) developed an algorithm to back-calculate the surface card to a downhole one by utilizing the damped-wave equation. The axial force on the rod string represents a stress wave, and thus, the rod string's movement physically resembles a 1-D wave propagation. The Gibbs wave equation for downhole card calculation was later improved by utilizing a finite-difference model. The modified Everitt-Jennings model allows the calculation of space and time discretization. Hence, stress can be computed at any finite difference node (Everitt and Jennings 1992). The modified algorithm also considers the iterations of damping, fluid load line and pump fillage calculations,

resulting in more accurate downhole data (Pons 2014). The pump card is used to predict the pump behavior, including the fluid dynamics and kinematics of the pumping unit.

The Gibbs method enables detection of various pump conditions through the downhole dynocard analysis (Gibbs and Neely 1966), such as normal condition, unanchored tubing detection, fluid pound (**Figure 2.1**), leaks in the pump and production string, and gas interference (**Figure 2.5**). The normal condition, i.e., full pump, indicates a rectangular shape, while the abnormal conditions display different forms of deformed shapes from the regular one. In the case of unanchored tubing, the movement causes the changes in displacement during the upstroke and downstroke transitions along the vertical line (**Figure 2.5 (A)**). The leaky pump takes longer to pick up the fluid at the upstroke and releases the load before the downstroke starts, showing an arch shape on top or bottom of stroke (**Figure 2.5 (C)**). Operators can detect the leak in tubing by observing the drop in production rate from the calculated fluid rate based on the downhole pump card conversion (Gibbs and Neely 1966). The gas compression effect also leads to a shape decrease in fluid load when the pump transitions from upstroke to downstroke, as mentioned in 2.1 and **Figure 2.5 (B)**.



(A) Downhole pump card of unanchored tubing (above) comparing to regular pump card of anchored tubing (below)

(B) Downhole pump card having gas interference (below) comparing to pump card of negligible gas (above)

(C) Downhole pump card different types of leaking pumps

Figure 2.5 Downhole card shapes for pump conditions: (A) unanchored tubing (above) compared to regular pump of anchored tubing (below), (B) Gas interference (below) compared to negligible gas (above) (C) Leak in plunger or traveling valve, unanchored tubing and leaking traveling valve, leaking standing valve (Gibbs and Neely 1966)

The detection, diagnosis, and prediction of events leading to system downtime and sub-optimal performance are leveraged by the analysis of subject matter experts (Pennel, Hsiung and Putcha 2018). However, this manual process becomes outdated due to high-level expertise requirements and time limitations. Given a long-term usage resulting in large data collection, the process of acquisition and transmission of surface and downhole dynocards has transformed from a manual to an automated process. The pattern recognition method was developed by computer-based systems, enabling the surveillance for hundreds or thousands of wells (Abdalla 2020 et al). This method includes 2 processes of feature extraction and classification. The Fourier-series method is developed for feature extraction, filtering noise in the pump displacement to recover the sinusoidal behavior in rod displacement. It also adds noise harmonic in pump load to recover rod load in the downhole (Chen et al. 2018).

Numerous pattern-recognitions workflows involve classification techniques, supporting more proactive decision-making. The common classification approach is to implement mathematical tools for similarity observation. These tools are supported by the descriptor calculation, summarized in **Table 2.1**.

Table 2.1. Summary of literature employing mathematical functions to categorize the similarity in rod pumps

Authors (year)	Methodology	Classifier	No. of faults detected	Samples
Dickinson & Jennings (1990)	Grid method Position-based, Curvature-based Fourier descriptors, String matching	Cost function	7	Dynamometer card limited in 7 wells
Lima et al. (2012)	Centroid Curvature descriptor K-curvature	Euclidean distance Pearson correlation	5	1500 cards
Yu et al. (2013) & Chen et al. (2018)	Fourier descriptors Geometric vector Level matrix	N/A	N/A	N/A

The machine learning and artificial intelligence algorithms are employed to train computer-based models. The objective of these algorithms is to partition feature space into class labeling. Two approaches for solving classification problems include supervised learning and unsupervised learning. The first rule-based expert system to diagnose rod pump unit operation was developed by Derek et al. (1998). Afterward, various classifier techniques were explored, including artificial neural network (ANN), support vector machine (SVM), AdaBoost, and Bayesian network.

As a few examples, a self-organizing map network proposed by Xu et al. (2007), a technique to classify labels based on multi-dimensional components, classifies five pumping operational categories with the load values as the features. A semi-supervised model, combining Decision Trees, Bayesian Network, and SVM, was trained to predict rod pump tubing failures (Liu

et al. 2010). A Bayesian Network model, coupled with AdaBNet boosting algorithm, is trained with different weights to define the stronger boosted ensemble model to predict rod pump failures (Liu et al. 2011). The sparse multi-graph regularized extreme learning machine (SMELM), which contains intra-class and inter-class graphs to train the data, enforcing different outputs associated with various faults (Gao et al. 2015). This method improves other classification methods by 90-95% accuracy in fault diagnosis.

The prowess of artificial neural network (ANN) was furthered in subsequent works from Martinez et al. (1999), Osman et al. (2005), and Mohammadpoor et al. (2010) to predict the bottom-hole pressure in oil wells or to analyze downhole dynamometer cards. Backpropagation neural network (BPNN) and SVM approaches were trained in 300 samples to diagnose nine pump operating conditions (Rogers et al. 1990, Nazi et al. 1994). The advantage of the ANN method lies in its backpropagation calculation capability that can handle the noise issues in data. The input produces an associated output, which can identify the pump faults. A variety of ANN algorithms is described in the work of Boguslawski et al. (2018), providing the automated pattern recognition for rod pump abnormal state recognition. Improved ANN models, which are determined by genetic algorithms, are furthered by the works from Abdalla et al. (2020) and Rashidi et al. (2010). The genetic algorithm is a search heuristic method that also optimizes other features, including optimal tubing size, injection rate, and artificial lift installation depth. Another heuristic search approach for pattern recognition and classification is implemented in Liu and Patel (2013), using the maximum load during upstroke as a feature to differentiate the failures. The study identifies the well failure with 82-86% of positive prediction and only 11-15% of false alarm. **Table 2.2** highlights the additional previous works discussed in this section.

Table 2.2. Previous studies applying machine learning algorithms

Authors (year)	Features	Classification Algorithm	No. of faults detected	Samples	Results
Liu et al. (2011)	Card area production, daily run time, stroke per minutes, etc	Semi-supervised learning with AdaBNet	4	100 cards	Accuracy 90 %
Xu et al. (2007)	Space feature points	Unsupervised learning, self-organizing map	5	6377 cards	Accuracy 93.4%
Liu et al. (2010)	Multivariate timeseries Card area, daily tun time, cycle, approved oil production rate	Decision Tree SVM Bayesian Network	6	42 train wells 32 test wells	Error rates 0.008 to 0.029
Bezerra et al. (2009)	Space features	Supervised learning BPNN	8	300	Classification error=1.31%
Abdalla et al. (2020)	Dynamometer cards	ANN with genetic algorithm	4	4467	Precision and recall values over all pump faults 93.2%

Apart from Liu et al (2010), these mentioned works would cause concerns in time, memory space, and cost due to utilizing super-computing power to train the deeper network model for pattern recognition. The huge drawback of deploying deep learning network lies in the requirement of complex data engineering program for image processing, giving those different operators utilize different proprietary software which could not be adapted in public or classroom settings. No existing literature in artificial lift mentions the possibility of generating models that learn from multiple operational parameters to predict the performance of pump. Therefore, the discussion of data-driven model prompts this study to explore less complex and more real-time machine learning model that could learn from the multivariate operational inputs to establish the predictive mode of generic and normal operation, which later is helpful for anomalies detection when the test datapoints fallout from the established trend.

Chapter 3: Experimental Design

This chapter discusses the experimental design for the Interactive Digital Sucker-rod Pump (IDSRP) facility, located at the University of Oklahoma, and used for the experiments of this study. The facility setup is completed with an objective to mimic the operation of a sucker-rod pump in the field scale. This chapter describes the design in more depth with the data acquisition method, testing procedure, and calibration for the actuator physical baseline model. This baseline model is implemented to correlate the actuator's speed to the pump's reciprocal movement. The test matrix is developed based on the calibration result.

3.1 Facility Setup

The rising application of the Internet of Things (IoT) in the upstream sector prompts the idea of integrating the physical setup and cloud-based architecture, providing the motivation for the IDSRP facility. The setup follows industrial standards from various wellsite instrumentation, including force and stroke transducers, electric motor controller, sensors, and their output parameters (Bello et al. 2020).

The setup mimics the common field operation of a sucker-rod pump in an existing laboratory structure at the University of Oklahoma. The IDSRP is a vertical 50-ft tall facility that houses a linear actuator, load cell and polished rod in the upper frame, transparent casing, tubing, return line and rod string in the middle section, and downhole rod pump and liquid tank in the bottom section (Teodoriu et al. 2020, Sharma et al. 2021).

Three modes are involved in the IDSRP experimental setup, including physical SRP mode, offline mode, and online mode. Together, these modes enable the implementation of the IoT principle in resembling field-scale operations and generating the AI model (Teodoriu et al. 2020). Overall, the IDSRP facility is designed to investigate the common field application using a variety of wellsite instrumentation such as force and stroke transducers, electric motor controller, etc.

The ***physical SRP mode*** focuses on the design and construction of the unique experimental SRP setup in the laboratory. The setup is designed to mimic and simulate a real wellbore's operation. It includes three components for an autonomous SRP system: Upper Frame, Middle Section, and Bottom Section. Each one of these sections occupies a level inside the lab.

Upper Frame section includes the linear actuator, load cell, the return line, polished rod, and the wellhead (**Figure 3.1**). The linear actuator provides the reciprocal movement of the polished rod. Digital pressure gauge 1 is used to record the pressure at the top gauge (**Pressure T**). Pressure at the top gauge represents the wellhead pressure. After the wellhead, the liquid is recirculated to a liquid tank, passing a valve for flow control.

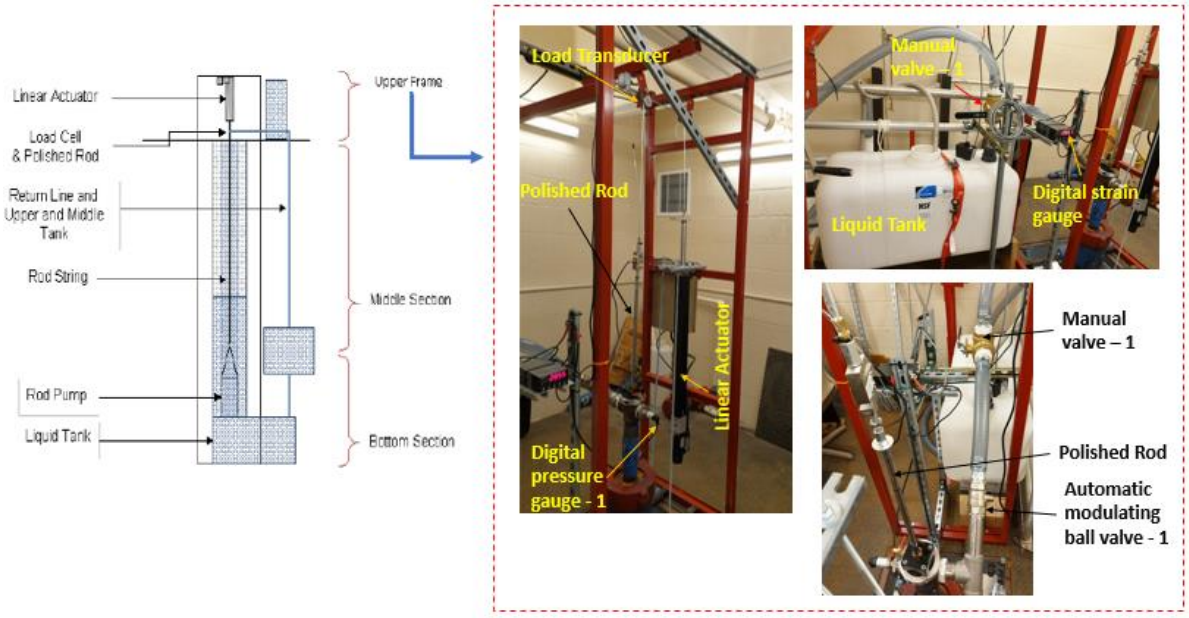


Figure 3.1 Upper Frame Facility of SRP facility setup

The linear actuator is mounted with a Pulse with Modulation (PWM) motor, monitored to allocate a particular point with the required angle and velocity. It can send the position and speed signal to the control driver for rotation and displacement. PWM is a technique to obtain the analog results with digital means based on modulating “on time” duration, or called pulse width, within signal spend-ons and signal spend-offs. According to Hirzel (2018), the process occurs in a very fast manner, making it difficult to notice the difference between on and off signals, thus no torque is lost virtually and more power efficiency as power source doesn’t need to accommodate for changing voltage (Pienknagura Dolberg 2019). Giving the constant voltages, meaning constant velocity in the machine, the changes of various speed, acceleration, and load intensities enable a clearer picture of the linear actuator motion.

Figure 3.2 illustrates the process of linear actuator motions during the accelerating and decelerating. The motion begins with accelerating downward motion, then accelerating upward motion, until both accelerating upward and downward motions become equal in the middle, and then decelerate before reaching the end of the stroke. With the movement of rod string following a Type C SRP model, the linear movement through beam with variable speed, and acceleration transforms into a circular motion. Therefore, velocities, acceleration and displacement follow a sine or cosine wave function (Pienknagura Dolberg 2019).

Changing the loads also impacts the upstroke and downstroke movements. During the upstroke movement, the opposing forces by the fluid weight in the rod string may cause a slower motion in the machine. On the other hand, as the traveling valve opens and the standing valve closes in the downstroke movement, the fluid load is transferred to the tubing, leaving only rod weight as the remaining downward force. Considering these force components in upstroke and

downstroke, we assume that no serious mechanical concern could prevent the rod from moving fast, which could negatively impact the integrity of obtaining analog results for PWM.

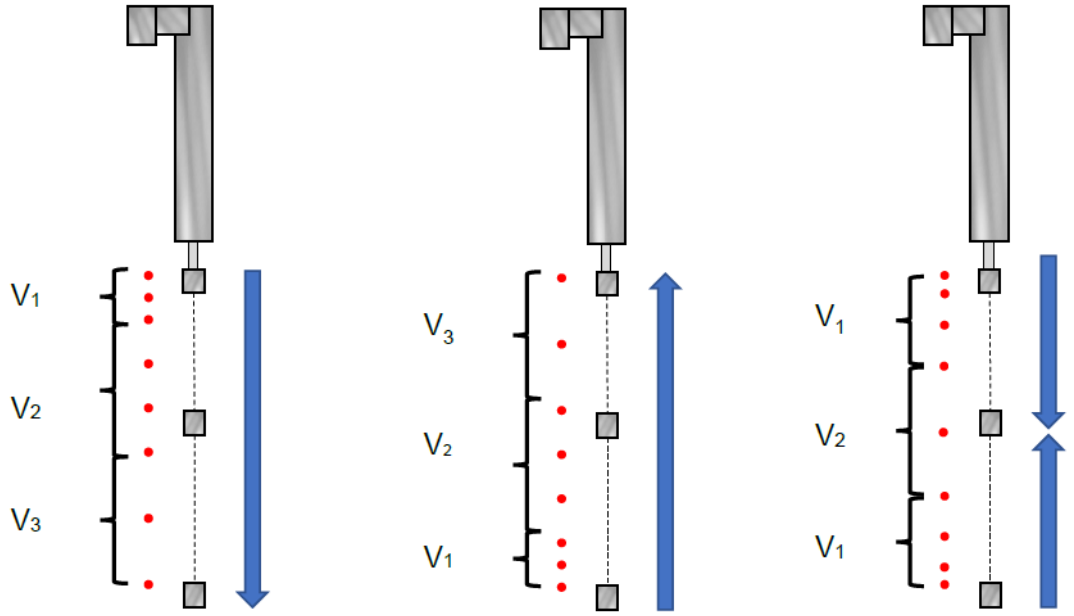


Figure 3.2 Linear Actuator Motions with Different Speeds and Accelerations (Pienknagura Dolberg 2019)

The Middle Section of the facility houses the continuation of rod string inside the tubing and return line. This section provides access to other areas of the facility through a ladder. The return line sends the liquid from the wellhead back into the tank to ensure a closed-loop system. The water level in the return line can provide hydrostatic pressure by adjusting a manual valve at the bottom of the line (Bello et al. 2020). Additional valves are installed in the return line to monitor and control the fluid flow and enable the facility operator to bypass the tank and send the liquid directly to the pump. During the operation, the valves are opened to maintain a closed flow loop within the pump and bottom tank through the return line. The bottom tank provides the static fluid column feeding the simulated wellbore. **Figure 3.3** displays the schematic for the Middle Section of the setup

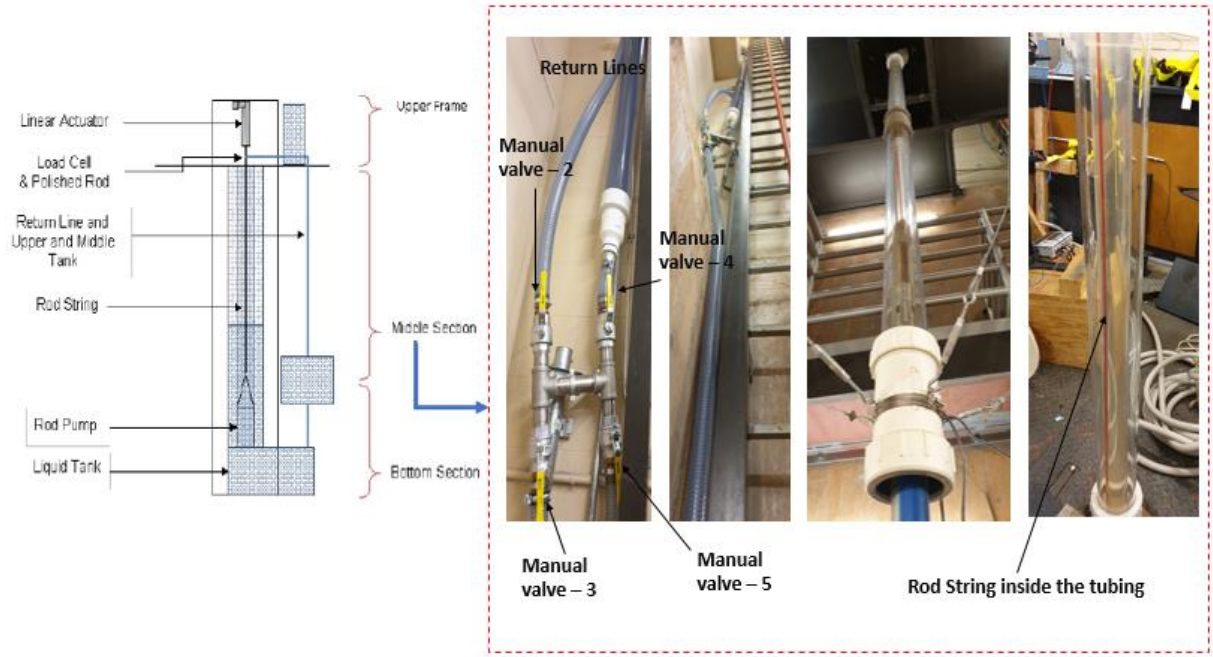


Figure 3.3 Middle Section of SRP facility setup

Two pressure gauges are installed in the Bottom Section of the facility to record the suction and discharge pressures, namely **Pressure B1** and **Pressure B2**. In terms of these gauges' locations, the pressure gauge **B1** is installed right at the intake of the pump, while the pressure gauge **B2** is set at the outlet of the pump, approximately 3.5 ft from the pump outlet. The suction pressure (**B1**) reveals the impact of changing the liquid level in the tubing inlet resulting in pressure buildup during the pump's movement cycle. The **pressure B2** shows the discharge pressure of the pump. The difference between top gauge pressure (**T**) and the 2nd bottom gauge pressure (**B2**) reflects the pressure drop across the facility's test section. As for the 3.4-ft long and 80 gallons liquid tank in the setup, it supplies the liquid to the test section. The tank in the bottom section serves as an active water source for the system under normal operation. A tube or line is added with bottom tank to connect the intake of the pump to the tank, combined with a filter pad to prevent solids from entering the pump. The initial purpose of the upper tank is to serve as a separator in case of simultaneous gas and water production (Pienknagura Dolberg 2019). However,

the operation discussed in this study only includes single-phase liquid flow. Hence, a static fluid column is maintained in the bottom tank, and closed loop within the bottom tank and the pump is ensured. The upper tank is kept empty. The schematic of the bottom section is illustrated in **Figure 3.4**.

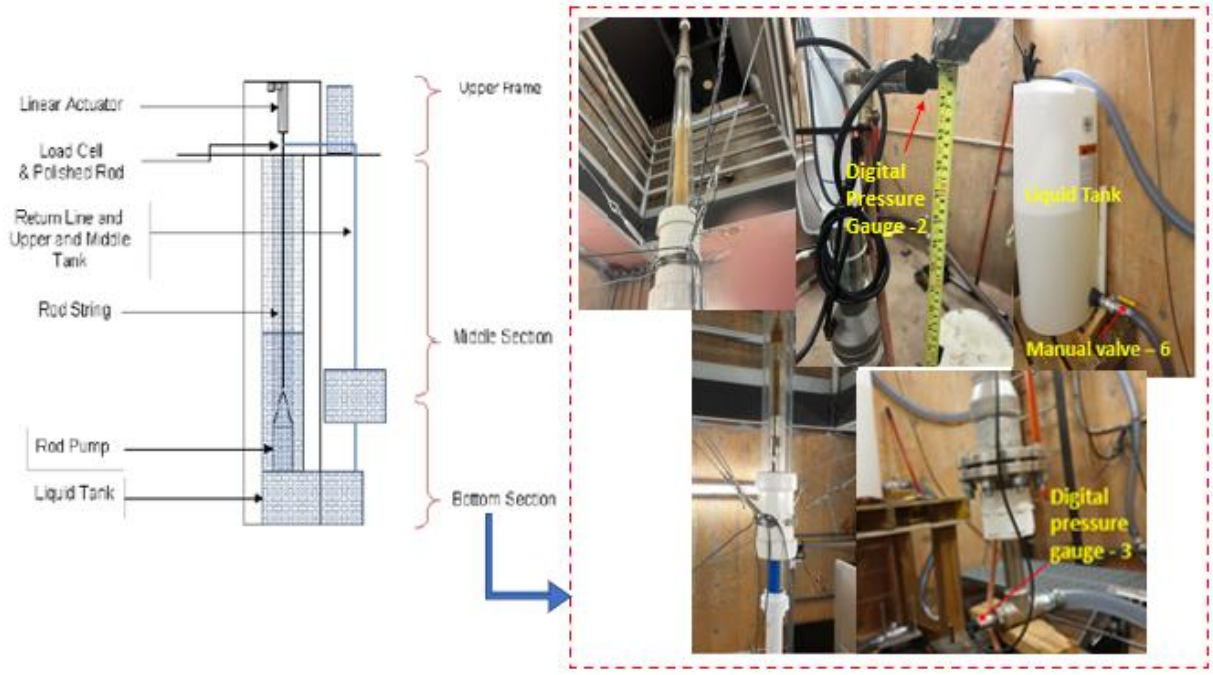


Figure 3.4 Bottom section of the SRP facility setup

Overall, the physical mode enables a state-of-the-art sensor system through gauges, control system via monitoring valves and data acquisition hardware connected between the physical mode and offline mode. Two of the principal elements in this setup include the linear actuator and motor. Transmitting linear forces like acceleration affects the dynamometer cards. Modeling the linear movement of the rod string can help simulate different pumping units. The setup has an SRP Simulator interface designed in the LabVIEW environment and connected to the physical SRP model. The facility uses water as the principal fluid, which helps achieve steady-state operation in a timely manner.

The ***offline mode*** carries out experimental verification and performance testing (**Figure 3.5**). This unit's operation is completed by integration of physical SRP setup and digital SRP simulator in LabVIEW environment, enabling data transmission and acquisition from sensors. The records from this setup are aggregated into a designed interactive digital solution, using the LabView environment. This interface records all the sensors' readings and visualizes the pressure variations and dynocards in real time. The integral components of the mode include the linear actuator and fluid tank. The sensor selection and installation procedures are described in Pienknagura Dolberg (2019) and Bello et al. (2020). The three representative groups of sensors installed in the setup include sensors for subsurface and surface criteria, representing reservoir properties, sensors attached to the electric prime drive, representing power supply, and sensors attached to the polished rod, wellhead parameters, and downhole pump, representing SRP primary operational body. The sensors measure vibrations, acoustics, pressure, stress and strain on the rod, flow rates, and loads. The load measurement builds up the surface dynamometer cards.

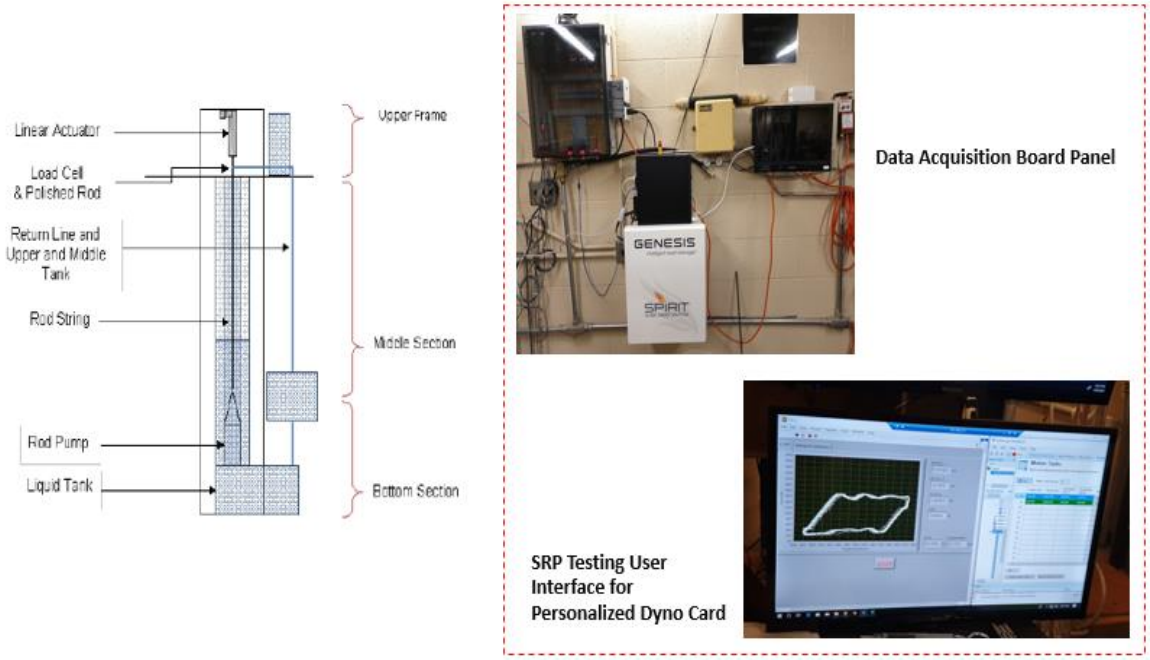


Figure 3.5 SRP setup's user interface for personalized dynocard (Teodoriu et al. 2020)

A real-time monitoring and diagnosis system is built in the ***online mode***. The graphical interface provides early warning and simultaneous monitoring for multiple wells. This mode enables data preprocessing and manipulation through programmable units. The recorded surface dynamometer cards and time-driven data are processed and trained in a cloud-based analytics platform. The algorithm plays a role as a diagnostic model to determine the operating condition of the rod unit. The cloud-based analytical solution allows production optimization and improves the decision-making process for abnormality detection and immediate maintenance activities. This mode contributes as a continuation and upgrades work from the proposed Model Predictive Controller (MPC) discussed in Bello et al. (2020). The predictive analysis using a machine learning algorithm is described in the next section.

3.2 Facility Design Discussion

The design for data collection and facility testing involves the development of a data acquisition (DAQ) system to obtain measurements and dynamometer card, a testing procedure to operate the facility, and a calibration baseline model to validate the measurements. The sensors and the interfaced SRP simulator designed in LabVIEW software help collect the pressure readings, load force, displacement, and liquid level.

3.2.1 Data Acquisition

The DAQ system is a continuation of offline mode to collect data through operating the facility and recording measurements from sensors. The specifications for sensor devices applied in the IDSRP are described in detail by Sharma et al. (2021) and in **Table 3.1** . The accuracy of these sensors ranges from 0.25 to 0.5%, which impacts the potential errors for theoretical calculation in later chapters. The noise reduction is considered through adding the low-pass, high-

pass, or bandpass filters in the input line and testing these filters for inference, acquisition frequency, and sampling rate. This process maintains the clear signal input lines. Each location in which a sensor is installed plays a role in acquiring principal measurements needed for the analysis. The load cell sensor receives a live signal through a digital output to turn on or off. The measurement of this sensor is decoded using the analog input port. Combining with motor sensor, these sensors form the load force and displacement components for a dynamometer card on the surface. The liquid level sensors detect the changes of liquid level in the top and bottom tanks in response to the loads of pump and rod string. The encoder of the sensor also measures the rotation of the shaft inside the linear actuator. Once the motor is operated, the associated drive will calculate the speed, acceleration, and position. The signal data recorded by these sensors are accessed and measured through the LabVIEW software.

Table 3.1 Sensors specifications utilized in the IDSRP Facility (Sharma et al. 2021)

Sensor name	WIKA C-10	PSD S1	TE Connectivity SP2-25
Range	0-100 psi	0-220 lbs.	0-25 inches
Excitation	12 Vdc	10-15 Vdc	30 Vdc
Output range	0-5 V	2 mV	94% ±4% of input voltage
Placement	Pressure Transducer, at Upper Frame and Bottom Sections	Load cell, at top of rod string	Displacement sensor, at rod string
Accuracy	% of span ≤ 1.0 (limit point calibration)	± 0.5%	±0.25% FS

LabVIEW is a system engineering application software for testing, measuring, and monitoring data. Given the flow of data fed into the system, this software utilizes a graphical programming approach to visualize all application features. The graphical block diagram visualizes the feedback of the systems. According to the facility schematic discussed in Pienknagura Dolberg (2019), the LabVIEW installation helps to control the motor and linear

actuator movements and supports the DAQ device in sensor data gathering. The setups for output voltage in the DAQ system and analog motor for linear actuator are also mentioned in the discussion. LabVIEW provides an interface for data gathering with the constant analog signal.

The Case Structure for data storage generates a 1-D array of data grouped by a displacement value and captures all data points in every iteration as fast as possible. The output file for LabVIEW is under a coma separated-value structure, which can be parsed and programed under the set of cloud software programming used in the study, including MATLAB, Excel, R, and Python. Continuing the proposed sensor data installation in the previous work of Pienknagura Dolberg (2019), the resulting visualization of force load and displacement data points in each iteration is displayed in **Figure 3.6**. In this view from **Figure 3.6**, we can observe and monitor the positions, velocity, and changes of force and displacement with respect to upstroke and downstroke movement in the physical unit of SRP. The top and bottom tank levels are also shown in the interface to monitor the fluid flow in the closed-loop and allow immediate action in case of leaks.

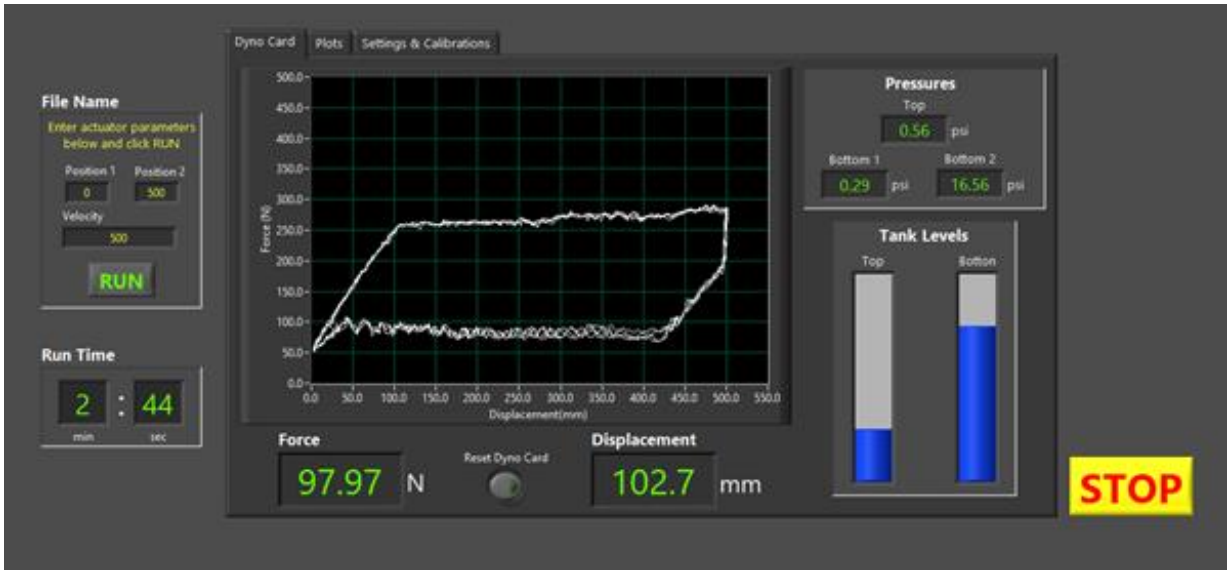


Figure 3.6. The visualization of SRP unit in LabVIEW during the operation

Following the conclusion and recommendation from Pienknagura Dolberg (2019), the testing procedure in the pumping unit follows by data input in the LabVIEW interface. The cyclic stack of load and displacement curves shown during the procedure resembles the field SRP cards, which is helpful for the analytical solution in online mode.

3.2.2 Testing Procedure

In the testing procedure, the applicable ranges of data input are addressed. The threshold of the existing material's mechanical properties may not be exceeded to avoid snapping the rod string movement. A model Kollmorgen enables the linear actuator to withstand a load capacity of 810 lb. The initial result from the Early Model on SRP of API RP11L (Takacs 2015) shows a peak polished load of 113 lbs., a minimum polished rod load of 45 lbs., and a peak net torque of 206 in.lbf. While the stroke length can reach a maximum of 24 in (609 mm), the material of rod string and the structural laboratory setup allows 17 in (450 mm) to 2 in (50 mm) as the limits of top and bottom positions for the rod string displacement. This means that the rod displacement ranges up to 15 in (400 mm) for stroke length. Apart from the generic stroke per minute (SPM) ranging from 3 to 8 SPM, SPM numbers in the facility operation can go up to more than 10 SPM due to the sudden change of rod upward and downward movements and minimal frictional factor loss due to homogenous fluid inside the flow loop.

Figure 3.7 provides the algorithmic flow of the experimental work design from Pienknagura Dolberg (2019). LabVIEW provides the interface for parameter inputs to perform the operation at IDSRP. The inputs include top and bottom positions of the linear actuator during the upstroke and downstroke movements, constant acceleration of 167 RPM/sec, and the actuator speed (RPM). The Data Collection is carried out when the IDSRP operates for a certain duration. The physical baseline model continues the progress of the calibration, suggested by Teodoriu and

Erik Pienknagura (2018), Pienknagura Dolberg (2019), Bello et al. (2020), and Sharma et al. (2021). The machine considers the limitations in system parameters for a reliable and safe operation, similar to what is applied in the oilfield. The calibration provides a linear correlation of the pump's strokes per minute and the actuator's RPM. The linear model is considered in calibration mode because the movement of the SRP follows a Type-C (Conventional) motion. Type-C pump is the most common SRP type applied in the oilfield. The calibration mode provides the standard values for operating parameters used to develop the test matrix for the experimental work.

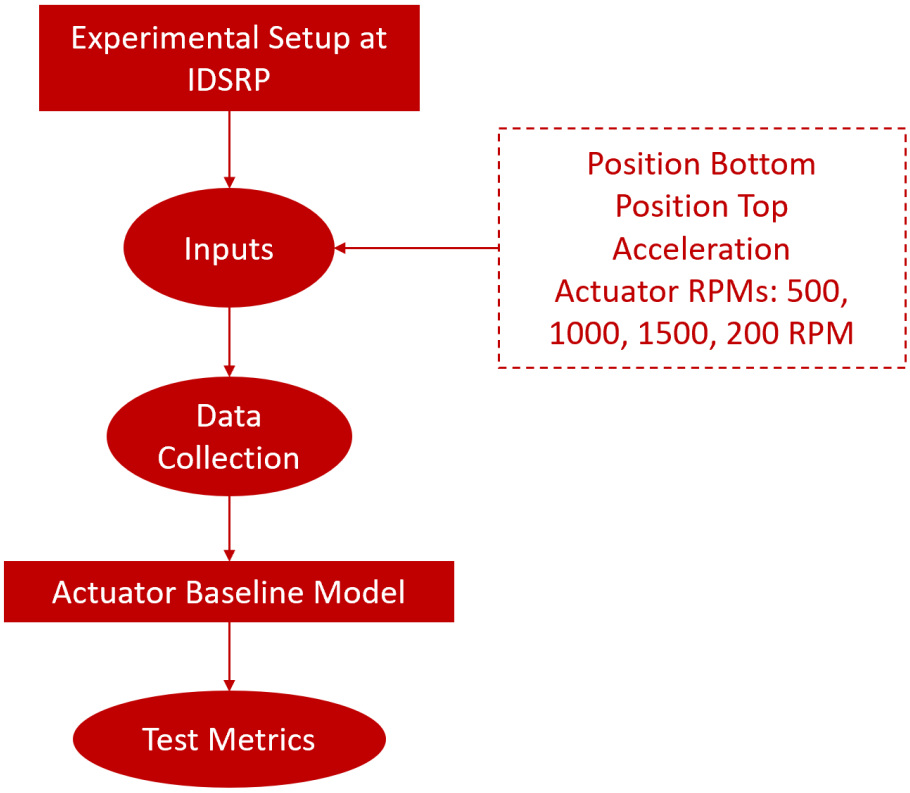


Figure 3.7. Algorithmic flow of the experimental work

A testing procedure fully details a step-by-step process based on **Figure 3.7**, from starting the software to shutting down the operation: The procedure follows the 3 of 4 IDSPR operation

modes, discussed in Bello et al. (2020) and Pienknagura Dolberg (2019): 1. Calibration, 2. Manual and 3. Graph Builder.

1. The first step is to start the machine and the associated network that enables the SRP operation, including LabVIEW and Kollgormen Workbench. The experimental check is also conducted to ensure a full and static fluid level in the bottom tank and ensure no leakage in the tubing and flowline.
2. The second stage is to adjust the inputs in Motion Tasks on Kollgormen Workbench so that the linear actuator moves within the range of top and bottom positions and the given RPM (**Figure 3.8)**. The array of inputs includes a range of speed from 500 to 2000 RPM and the rod displacement of 150, 200 or 250 mm. The acceleration and deceleration remain at a constant value of 166.67 RPS/S. Depending on the harmonic or sharp changing displacement type of the rod string, the profile column is set as "One-to-one" (sinusoidal) or "Trapezoidal" to initiate the pattern in Graph Builder mode. The "Next Task" column ensures the cyclic motion of the linear actuator and gears the operation towards the first 2 rows with values of 0 and 1 in each row of position, respectively. The program initiates the tests by transmitting the inputs to the facility by enabling "Update Motion Task" underneath the input table.
3. The third step is to conduct the IDSRP operation by entering position and speed inputs for file name storage and entering the "RUN" tab on the LabVIEW interface (**Figure 3.6**). The limit conditions of these inputs are addressed after the Calibration mode. This stage follows the "Manual" mode discussed earlier by Bello et al. (2020) and Pienknagura Dolberg (2019) to manually control machine and pre-program with defined motion and balanced geometry based on Step 2.

4. The fourth step is to operate the operation with different time constraints, which can be observed in “Run Time” in LabVIEW (**Figure 3.6**). This step enables enough amount of data collection for training and testing predictive modeling. The Graph Builder mode is also conducted in this stage to visual reference for a drawn dynamometer card.
5. After a defined duration, the operation is completed by clicking the STOP button at the right corner in LabVIEW and also the STOP in the upper left of the input table in Kollgormen Workbench (**Figure 3.6** and **Figure 3.8**). The establish the physical baseline model is also developed from this stage and established based on the relation between SPM and displacement. A test metrics, specified in **Table 3.2**, is also developed to gather more data points for further predictive analysis.

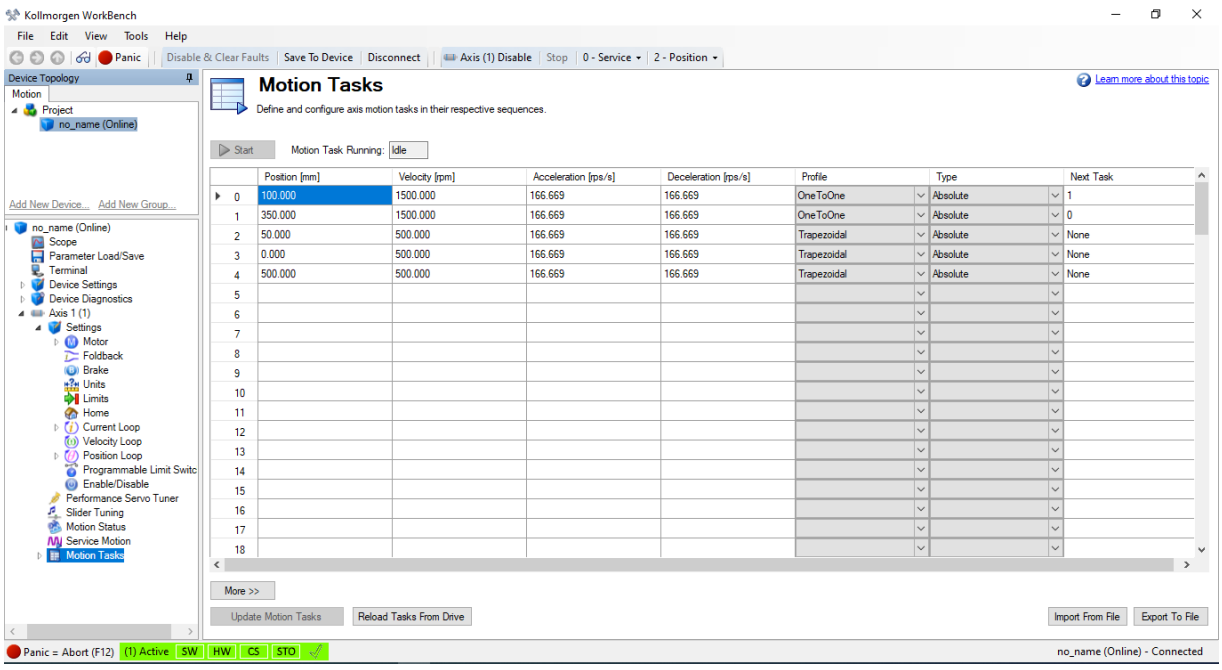


Figure 3.8 Kollmorgen interface for displacement, speed, and acceleration Inputs

3.2.3 Actuator Baseline Model Calibration

The experiments are carried out based on the relation between the number of pump strokes in a minute and the linear actuator's RPM. The movement of the actuator is simulated with trapezoidal and sinusoidal designs. The trapezoidal shape reflects the ideal experimental condition with the rapid change of rod displacement in upward and downward motions, resulting in larger SPM numbers (**Figure 3.9**), while the sinusoidal shape mimics the actual operational scenario with more harmonic displacement in each cycle, following polished rod position type curve for the Class I geometry unit shown in Gray (1963) (**Figure 3.10**). **Figure 3.11** and **Figure 3.12** illustrate the linear correlation between the number of strokes per minute and the actuator's RPM for different stroke lengths. The difference between the figures lies in the input movement shape. The testing for **Figure 3.11** is carried out with the trapezoidal shape, while the tests for **Figure 3.12** are operated under the sinusoidal shape. Due to the movement shape and lower peak accelerations, the sinusoidal design results in lower SPM values compared to the trapezoidal shape. Also, a higher stroke length generates a lower number of strokes or cycles over time for a fixed actuator RPM. The linear trends show a simplified correlation between these two factors that can be used to estimate the pump's speed.

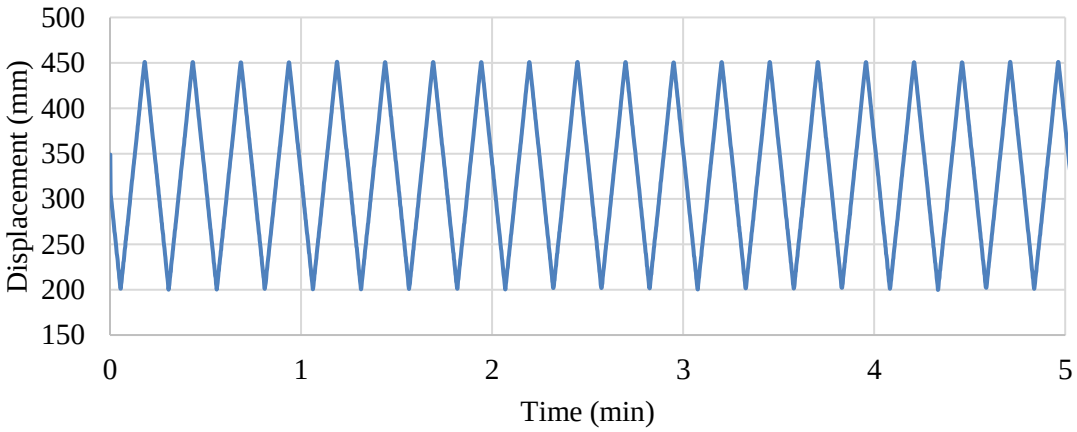


Figure 3.9 A sample of trapezoid-shape displacement plot against the running time of 5 minutes at stroke length of 250 mm, actuator RPM of 500 and SPM of 4

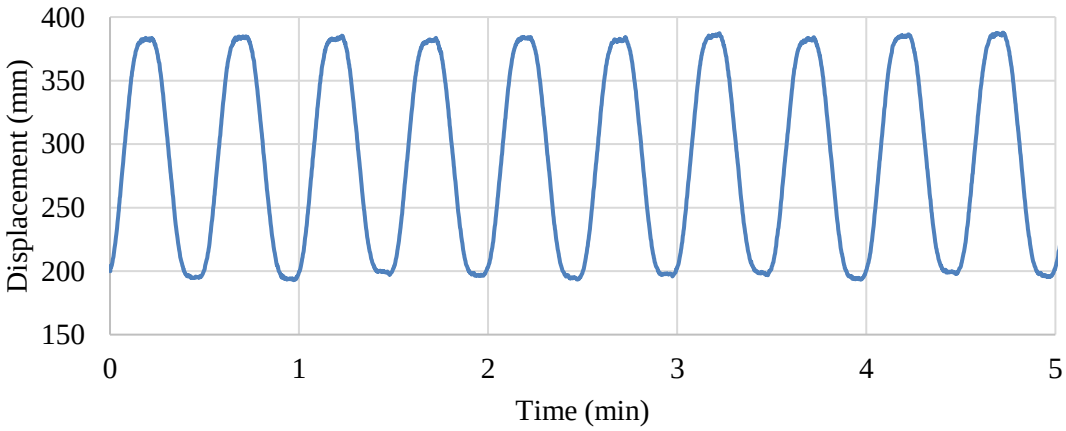


Figure 3.10 A sample of trapezoid-shape displacement plot against the running time of 5 minutes at stroke length of 250 mm, actuator RPM of 500 and SPM of 4

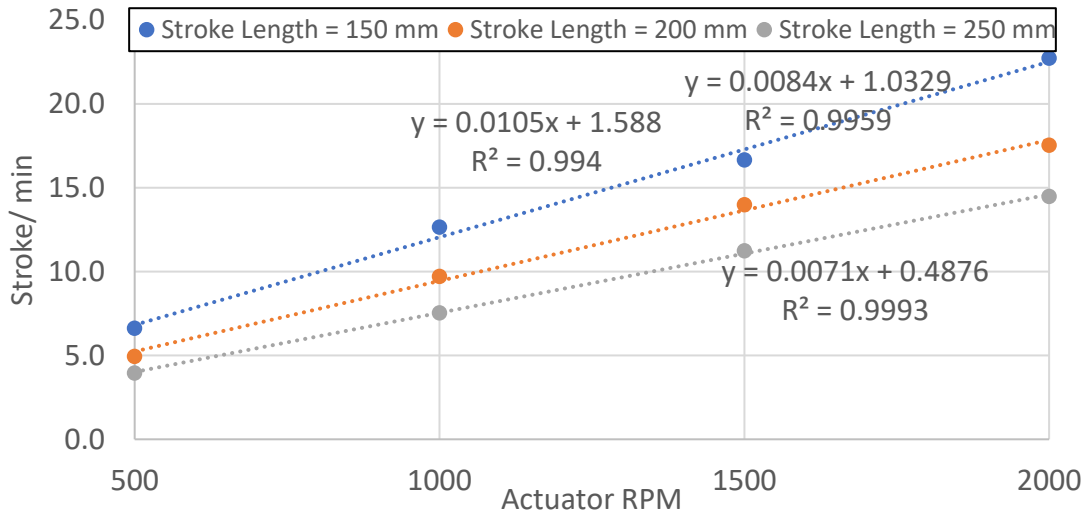


Figure 3.11. Linear correlation between the pump SPM and actuator RPM for varying stroke lengths and trapezoidal shape

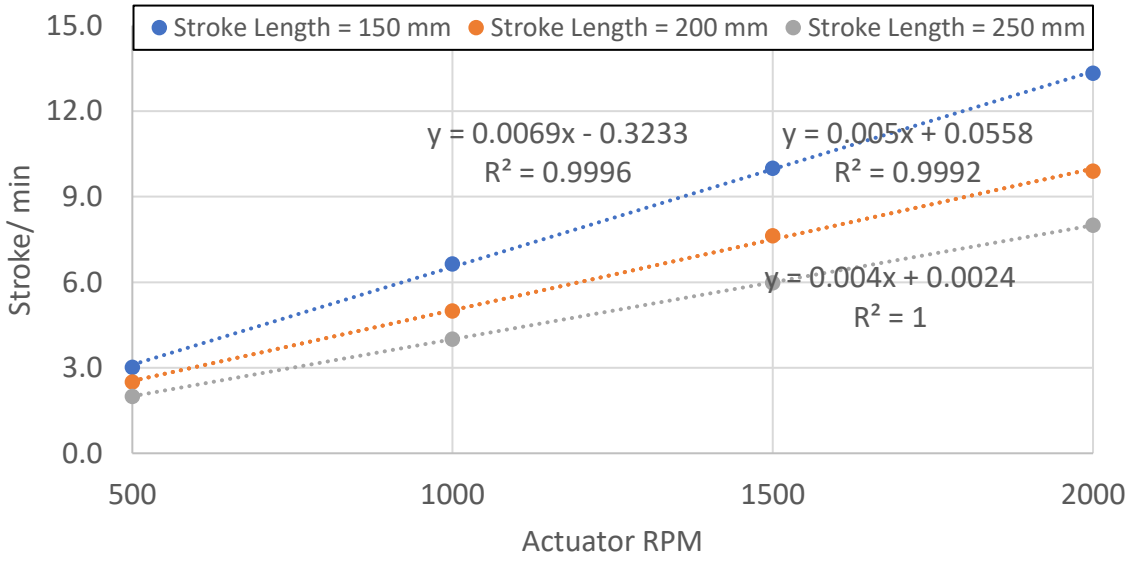


Figure 3.12. Linear correlation between the pump SPM and actuator RPM for varying stroke lengths and sinusoidal shape

3.3 Test Matrix

Based on the result of physical baseline model calibration, the designed test matrix is described in **Table 3.2**. The operation of the SRP is simulated in short-term and long-term durations (5-20 minutes), different levels of actuator speed (RPM), and magnitudes of the rod

movement (stroke length). The short-term 5 minutes is carried out to calibrate the machine, following the plots in 3.2.3. Long term 20-min duration tests help generate more data points for testing and validation of the analytical model in online mode. The linear baseline model is used to correlate the actuator's RPM and the rod pump's number of strokes per minute. The acceleration of the actuator is maintained at 167 rpm/sec. The tests generate the surface dynamometer cards of load force vs. displacement and pressure readings in the top and bottom of the rod-pump unit.

Table 3.2. Current study's experimental test matrix.

Stoke Length (mm)	Duration (min)	Actuator RPM	Shape	Acceleration (rpm/sec)
150, 200, 250	5 to 20	500, 1000, 1500, 2000	Trapezoidal and Sinusoidal	167

Chapter 4: Experimental Results

Following the experimental design and operational test matrix**, Chapter 4** will dive deeper into the data obtained from the experimental work. This includes an exploratory overview, and more importantly, transformation calculations from surface cards to the downhole ones. The visualization from the statistical analysis provides underlying physical feedback from the system operations. As the measurements for rod displacement and load force are in surface conditions due to the sensor installation, a downhole dynocard transformation algorithm is introduced and carried out. The applied algorithm follows the modified Everitt-Jennings procedure from literature to solve the Gibbs wave equation. The resulting visualization is helpful for pattern recognition of the SRP performance in a given system.

4.1 Results Overview

The stroke per minute (SPM) parameter is an important metric to determine the speed of the sucker-rod pump's movement. This is a result of the actuator's speed input into the system, in RPM, and therefore proportional to it. It can be visually determined by counting the number of displacement peaks over a period of time (n), or as the reciprocal of the average time difference between two peaks. The calculation of SPM is based on **Equation 4.1**.

$$SPM = \frac{1}{\sum_{n}^{i=0} \frac{t_{i+1} - t_i}{n}} \tag{4.1}$$

where t_i is a time at the displacement peak reaches, t_{i+1} is time at the next displacement peak reaches, n is numbers of peak observed/ selected.

The resulting SPM for both shapes are presented in **Table 4.1, Table 4.2,** and also summarized in **Figure 3.11** and **Figure 3.12.** SPM metric contributes as the driven vector for exploratory analysis for other parameters of force load or pressure. The SPM values are consistent in any duration of the testing. The impact of SPM on the rod operation, as well as operation classification, is furthered in subsequent sections.

Table 4.1: Resulting SPM based on rod displacement data collected for trapezoidal geometry

Stroke Length (mm)	SPM for trapezoidal geometry (SPM)			
250	4	8	11	14
200	5	10	14	18
150	7	13	18	23

Table 4.2: Resulting SPM based on rod displacement data collected for sinusoidal geometry

Stroke Length (mm)	Equivalent SPM for sinusoidal geometry (SPM)			
250	2	4	6	8
200	2	5	7	10
150	3	7	10	13

4.2 Statistical Analysis For Parameters

The experimental pressure measurements recorded from three digital gauges placed at the top and bottom of the test section are shown in **Figure 4.1**, **Figure 4.2**, and **Figure 4.3.** The top pressure gauge is placed at the upper frame section (**Figure 4.1**), while the two bottom pressure gauges are placed at the suction (**Figure 4.2**) and discharge (**Figure 4.3**) of the pump. The recordings capture the changes of suction and discharge pressures during the SRP movement between the upstroke and downstroke. Based on these plots, different RPM tests generate distinct pressure measurements. The distributions of these pressure measurements are relatively normal. The B1 (Pump suction) pressure indicates the cyclic nature of the movement, with higher pump speeds generating lower absolute values of pressure during the upstroke. Negative pressure values are observed in pump suction for these cases, implying that the casing is empty of liquid. When

the surplus fluid is removed from the casing annulus, the fluid column level becomes very low. Therefore, the possibility of a pump-off situation is imminent. On the other hand, the tests with lower pump speeds illustrate larger fluctuations in top gauge pressure measurements due to longer time durations taken for each cycle.

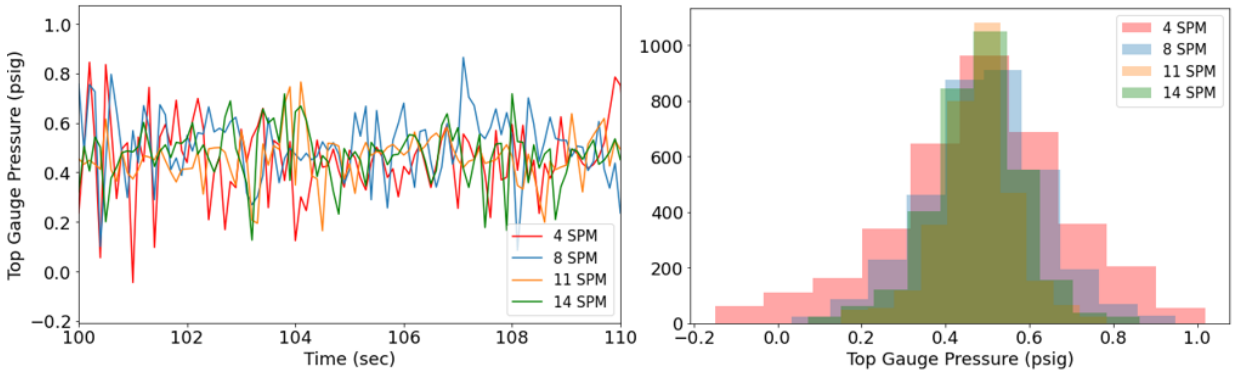


Figure 4.1 Pressure recorded at the top gauge with time and the histogram of its distribution, obtained at stroke length of 250 mm, trapezoidal shape

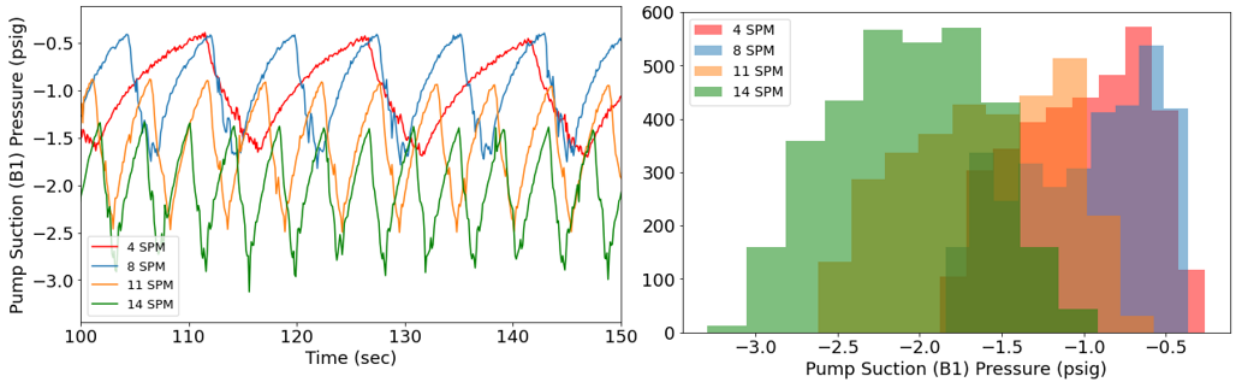


Figure 4.2. Pressure recorded at the first bottom gauge with time and the histogram of its distribution, obtained at stroke length of 250 mm, trapezoidal shape

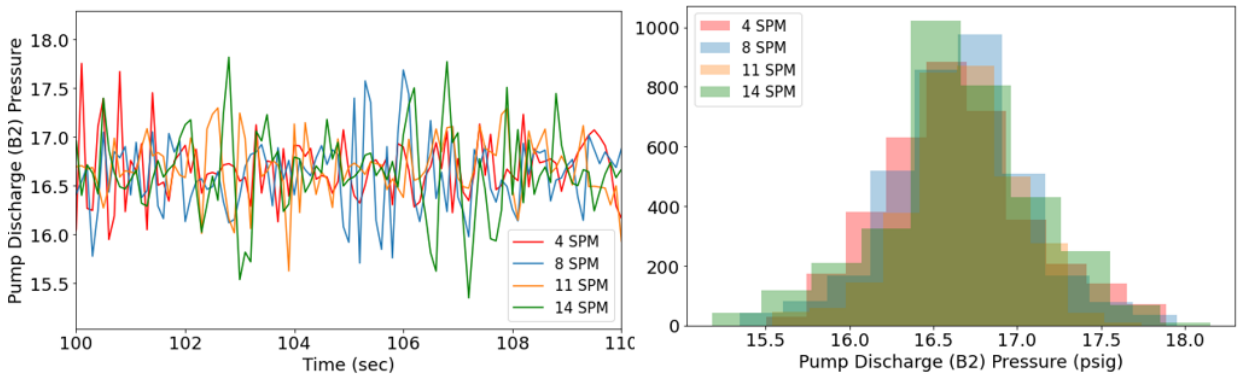


Figure 4.3. Pressure recorded at the second bottom gauge with time and the histogram of its distribution, obtained at stroke length of 250 mm, trapezoidal shape

The analysis for whether the mean differences between various tests are significant is carried out through a statistical test called a t-test. Given the null hypothesis that the two populations' means are equal, if the resulting p-value (the probability of observing as or more extreme values assuming the null hypothesis) returns smaller than a threshold of 5%, this null hypothesis of equal population means can be rejected. The descriptive analysis of pressure records at the first bottom gauge pressure, i.e., p values for the t-test, are summarized in **Table 4.3**. The resulting table quantifies the distinctive differences in the distributions of pressure records at the pump's suction pressure for tests with varying speeds. Therefore, the speed (SPM) could be the driven parameter for modeling input to predict the pump performance because it also impacts in the level of water feed into the pump per cycle. The incomplete pump fillage due to the rod moving too fast causes the pump-off issue that would deteriorate the rod.

Table 4.3: p-values from t-tests among pressure records at the first bottom gauge with varying equivalent SPMs

Pressure Records at Top Gauges at	p-value	Reject Null Hypothesis
4 SPM vs 8 SPM	0.0	Reject
4 SPM vs 11 SPM	0.0	Reject
4 SPM vs 14 SPM	0.016	Reject
8 SPM vs 11 SPM	0.0	Reject
8 SPM vs 14 SPM	0.0	Reject
11 SPM vs 14 SPM	$8.746\ E^{-310}$	Reject

4.3 Downhole Dynocard Transformation

Figure 4.4 illustrates the analytical approach applied to build a real-time monitoring and diagnostic solution in the online module. The objective of this project is to employ data-driven techniques for pattern recognition and classification of a SRP's performance. The surface dynamometer card and time-driven pressure and force load data are processed and trained in a cloud-based analytics software system. They are then used for pattern recognition purposes.

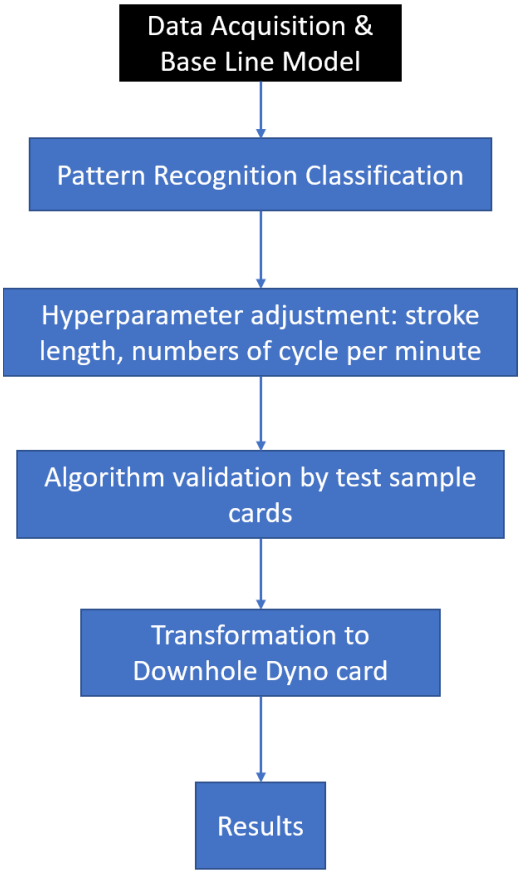


Figure 4.4. Predictive analysis using data-driven approaches for online IDSRP solution

The initial step is to resolve the outliers observed in the raw dynamometer card by using a z-score that indicates the vicinity of the data to the mean. The feature extraction and normalization are implemented as the data processing steps. The Gibbs wave equation is deployed to transform the surface card to a downhole one. The result is used for pattern recognition of the dynocard and determining the operational condition of the SRP. The implementation of the Gibbs wave equation in the downhole calculation is discussed in Everitt-Jennings (1992). It is important to adjust the inputs for the wave equation, such as the number of nodes, material properties (Young Modulus, density, length of rod string), etc., based on the laboratory conditions.

4.3.1 Modified Everitt-Jennings Algorithm

The principal method to diagnose the operating condition of pumping systems is to simulate the rod string's behavior and convert the surface dynocard to a downhole one. The transformation approach considers the rod string's elasticity, which dampens the load and is impacted by surface and pump intakes. The wave equation is modeled based on the change of tension force with depth along the rod string, enabling transmission of the impulses from the surface. The breakdown of forces includes damping force (F_d), the buoyant weight of rod (W), tension/ pulling from above the rod element (F_x), and tension from below the rod element ($F_{x+\Delta x}$) (**Equation 4.2**). By substituting the rod load and stresses in sections x and $x + \Delta x$, determining damping force, and incorporating all elements, Gibbs reached the one-dimensional damped wave equation (**Equation 4.3**). The equation must be solved to convert the surface card to a downhole one.

$$m\frac{\delta^2 u}{\delta t^2} = F_x - F_{x+\Delta x} + W - F_d \qquad 4.2$$

$$v^2\frac{\delta^2 u}{\delta x^2} = \frac{\delta^2 u}{\delta t^2} + c\frac{\delta u}{\delta t} \qquad 4.3$$

where $u = u(x,t)$ is the displacement of the rod string as a function of position (x) and time (t). Acoustic velocity of rod material is calculated by $v = \sqrt{\frac{144Eg}{\rho}}$, g is the gravitational acceleration, c is a constant damping factor, and ∂u/∂x is the rod strain or the change of rod displacement over its length.

Gibbs's method focuses on solving the damping force, which is a complex sum of forces opposed to the movement of rod string. The 1-D equation, shown in **Equation 4.4**, is extended to a linear hyperbolic differential equation by considering rod string material properties: cross section area (A), density (ρ), friction coefficient (k), and elasticity modulus (E).

$$EA\frac{\delta^2 u}{\delta x^2}(x,t) = \frac{\rho A}{144g}\frac{\delta^2 u}{\delta t^2}(x,t) + c\frac{\rho A}{144g}\frac{\delta u}{\delta t}(x,t) \quad 4.4$$

Using a finite-difference model, the modified Everett-Jennings method defines the solving approach for **Equation 4.4** (Everitt & Jennings 1992). The boundary conditions for solving this equation are obtained from the relations between surface position and load with time. Some factors needed for this finite-difference model are the number of recorded surface data points (N), total number of finite-difference nodes along the rod string (M), vector of finite-difference nodes along the rod string $\{i\}_1^M$, vector of sample points obtained at the surface $\{j\}_1^N$, discrete functions of rod position vs. time $\{g_{PR}\}_1^N$ and rod load versus time $\{f_{PR}\}_1^N$.

The finite difference approach for wave equation involves space and time discretion, i.e. mesh formation. Since the mesh progresses down the rod string, the direction is positive downward. The time increments between the surface data readings follows $u_{1,j} = \frac{f_{PR,j}.\Delta x}{EA} + u_{0,j}$ direction, which splits into smaller sections. Therefore, the following algorithm is developed based on **Equation 4.5** and the finite difference approach (Everitt & Jennings 1992). The algorithm involves the steps of initiation and back-calculation. Following this algorithm, the load and stress can be computed at any finite difference node by utilizing space and time discretization. The initialization equation for *j* values of *1-N* is:

$$u_{0,j} = g_{PR,j} \quad 4.5$$

The Hooke's law gives for $j = 1, \ldots, N$:

$$u_{1,j} = \frac{f_{PR,j}.\Delta x}{EA} + u_{0,j} \quad 4.6$$

Applying the finite difference approach on **Equation 4.4** for $i = 2, \dots, M$ results in **Equation 4.7**.

$$u_{i+1,j} = \frac{1}{\left(\frac{EA}{\Delta x}\right)^{+}}\{[\alpha(1 + c\Delta x)].u_{i,j+1} - \left[\alpha(2 + c\Delta t) - \left(\frac{EA}{\Delta x}\right)^{+} - \left(\frac{EA}{\Delta x}\right)^{-}\right].u_{i,j} + \alpha.u_{i,j-1} - \left(\frac{EA}{\Delta x}\right)^{-}.u_{i-1,j}\} \quad 4.7$$

where:

$$\alpha = \frac{\overline{\Delta x}}{\Delta t^2}\left[\frac{\left(\frac{\rho A}{144g}\right)^{+} + \left(\frac{\rho A}{144g}\right)^{-}}{2}\right] \quad 4.8$$

$$\overline{\Delta x} = \frac{1}{2}(\Delta x^{+} + \Delta x^{-}) \quad 4.9$$

The resulting force ($F_{pump,j}$) and displacement ($u_{pump,j}$) at the pump, or at node M, are displayed on **Equations 4.10** and **4.11**. The arrays of force and displacement form the resulting downhole dynocard.

Displacement:

$$u_{pump,j} = (1 + c\Delta t).u_{M-1,j+1} - c\Delta t.u_{M-1,j} + u_{M-1,j-1} - u_{M-1,j} \quad 4.10$$

Force:

$$F_{pump,j} = \frac{EA}{2\Delta x}\left(3u_{M,j} - 4u_{M-1,j} + u_{M-2,j}\right) \quad 4.11$$

The stability condition associated with the finite difference diagnostic model is $\frac{\Delta x}{v\Delta t} \leq 1$. In this equation, Δx (ft) is the space between two finite-difference nodes of a particular taper, Δt (s) is the time spacing between consecutive surface sampling points, and v (ft/s) is the velocity of sound in the rod string.

4.3.2 Algorithm Validation

The dynocard conversion algorithm is applied for a few sample cards extracted from the literature. The purpose of this comparison is to verify the conducted calculations and the developed algorithm. The sample cards are recorded in real field scales, as shown in **Figure 4.5** and **Figure 4.6**. For both cases, the left plots show the surface and downhole cards picked up from the literature. The right plots show the comparison of the published digitized downhole card and the calculated one using this study's program. The percentage errors in the area under the curve between the real downhole cards and the calculated ones are analyzed to validate the model (**Table 4.4**).

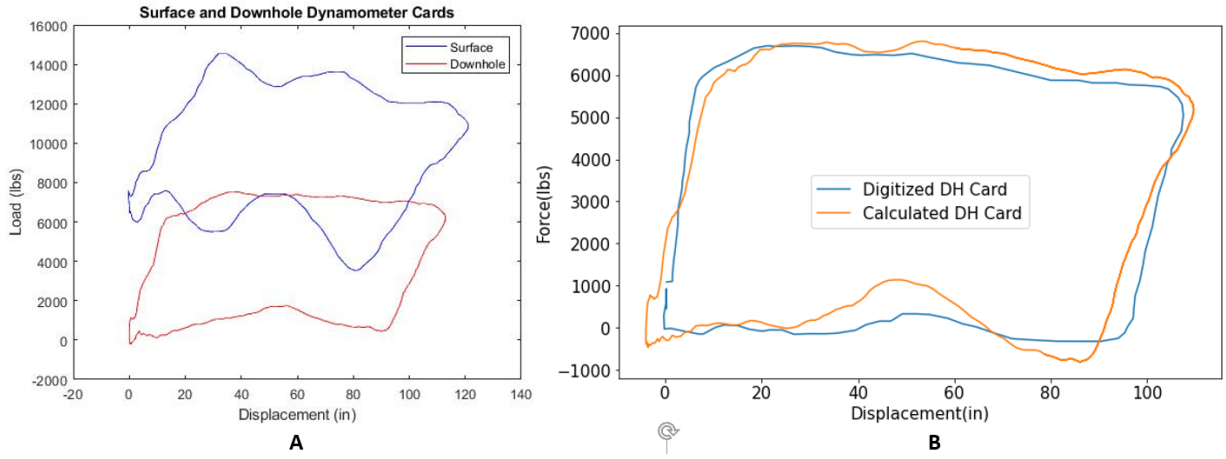


Figure 4.5 (A) Extracted surface and downhole cards, sample 1 (Everitt, Jennings 1992, Ercolino 2011) (B) Actual (digitized) and calculated downhole cards comparison

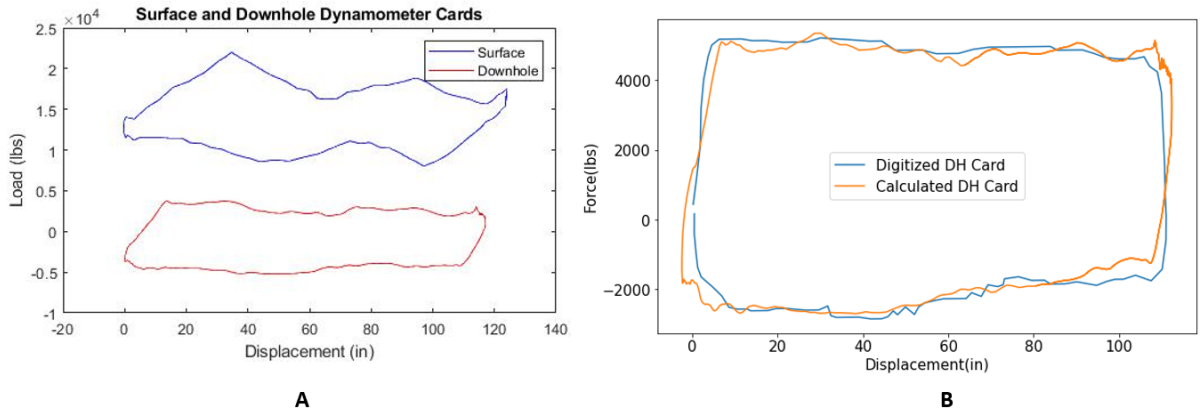


Figure 4.6 (A) Extracted surface and downhole cards, sample 2 (Dawsey 2020) (B) Actual (digitized) and calculated downhole cards comparison

Table 4.4 reveals the percentages errors between actual and calculated downhole cards. The error is calculated by looking at the discrepancy of the area under these cards. The errors are small and less than 1.5%, which justifies the confidence in the data and conversion algorithm.

Table 4.4 Percentage errors between actual and calculated downhole cards

Sample	Error (%)
1	0.74 %
2	1.34 %

4.4 Results

The surface to downhole dynocard conversion algorithm was applied to some sample cards obtained from the experiments in normal operating conditions. The results of the surface and downhole dynamometer cards are shown in **Figure 4.7**, **Figure 4.8**, **Figure 4.9** and **Figure 4.10**. These figures are obtained from tests with varying stroke lengths and SPM values. The surface dynocards from **Figure 4.7** and **Figure 4.8** are obtained with 250 mm stroke length and with 2 SPM and 4 SPM, respectively. Meanwhile, the surface dynocards from **Figure 4.9** and **Figure 4.10** are recorded from tests with 200 mm stroke length and 2 SPM and 5 SPM, respectively.

The surface cards are selected from one cycle of the pump's operation. The converted downhole cards from these figures are shown in red. The noise and discrepancy during the

transition from upstroke to downstroke is not commonly observed in a field-scale operation. This phenomenon can be due to the lower modulus of elasticity and buckling in the rod string of the experimental setup. The low modulus of elasticity affects the static elongation, where the natural vibration opposes the motion of the polished rod at the peak of dynamometer card. Kendrick and Corneluis (1937) also described this phenomenon by observing the polished rod returning to the top of the stroke before completing the downstroke. The fluid pressure surrounding the rod string increases the magnitude of the compressive force that the rod string must support. A slim rod string can withstand a certain amount of compressive load before the buckling threshold, which results in contact between rod string and tubing (Lukasiewicz and Knight 2006).

Overall, the transformation algorithm is supposed to help convert the surface cards into the downhole ones. The resulting downhole pattern enables operational malfunction recognition for each test. However, the abnormal rod string stretching or buckling may alter the shape of the dynocard compared to the field scale cases. As a result, the transformation may not be readily translated from the lab to field conditions.

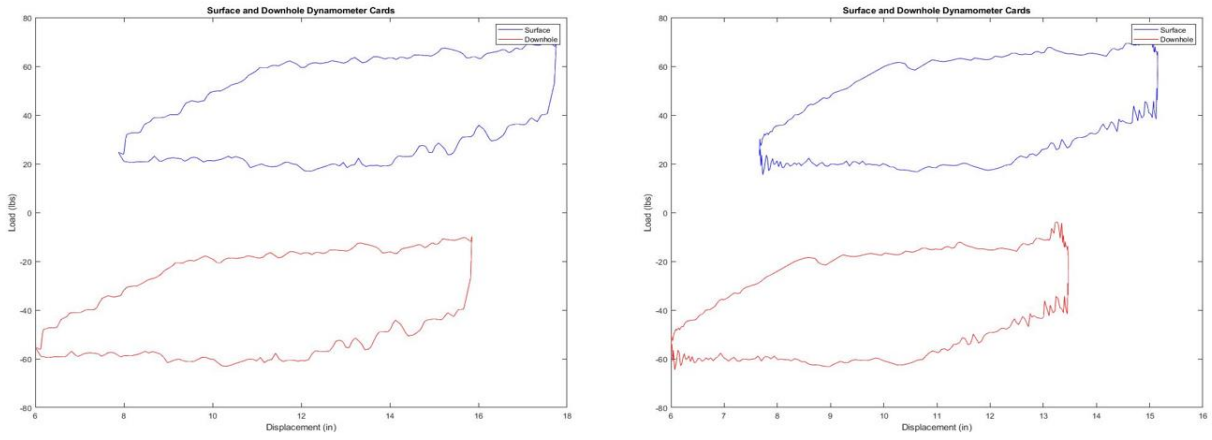


Figure 4.7. Conversion of surface to downhole dynamometer cards with input stroke length of 250 mm, machine SPM of 4

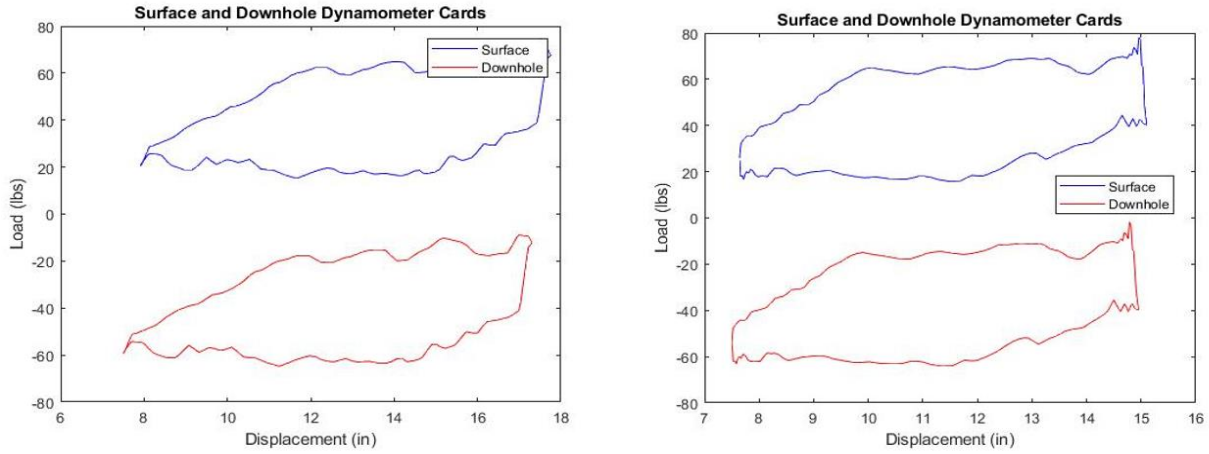


Figure 4.8. Conversion of surface to downhole dynamometer cards with input stroke length of 250 mm, machine SPM of 8

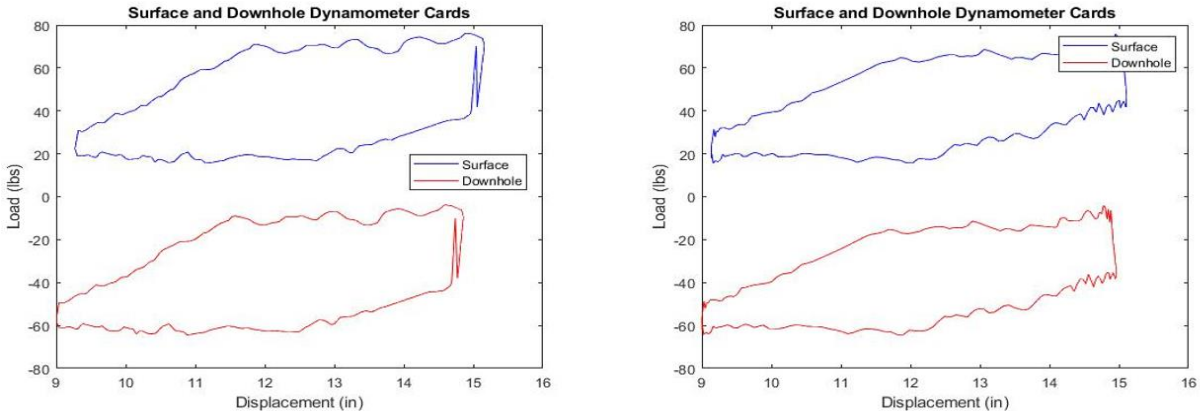


Figure 4.9. Conversion of surface to downhole dynamometer cards with input stroke length of 200 mm, machine SPM of 4

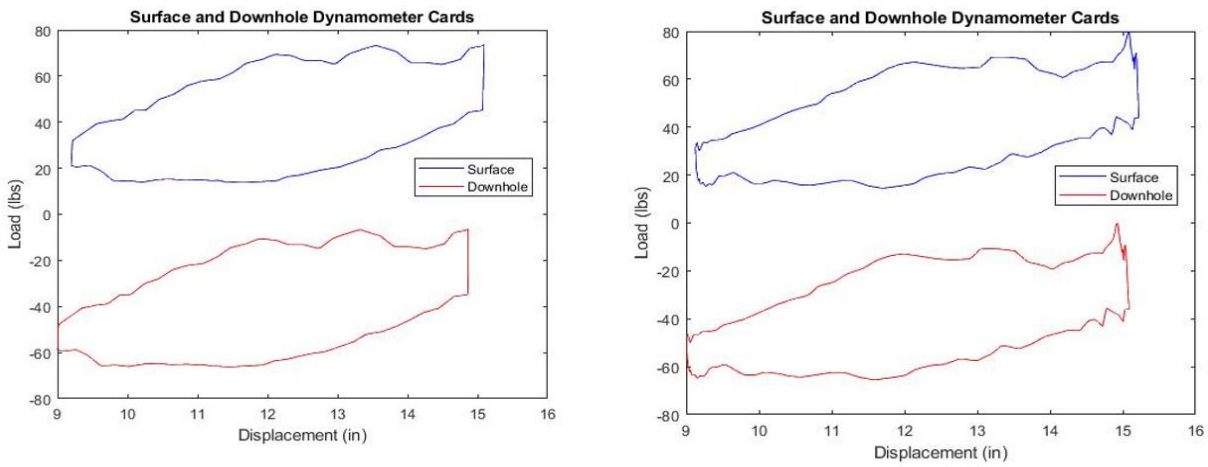


Figure 4.10. Conversion of surface to downhole dynamometer cards with input stroke length of 200 mm, machine SPM of 5

Chapter 5: Data Analytics

Chapter 5 discusses implementing analytical approaches to diagnose the operational conditions through developing predictive multiple input and multiple output regression models. This chapter furthers feature calculations based on the results obtained in the experimental design. The physics-based analysis helps understand the impacts of experimental inputs on the operation of SRP, categorizes the performance, and helps select important features for models. The results and model evaluation are presented in the chapter.

5.1 Preliminary Analysis: Feature Generation

The conclusion from **Chapter 4** suggests a more time-driven analysis based on the parametric signal data processing. The pressure and movement data over time provide further insights about the pump performance, including force range, liquid flow rate delivered by the pump plunger per cycle, differential and frictional pressure losses. These calculations help generate additional features for the model input. When putting any of these features in the time dimension, the changes of the feature can be used to evaluate the SRP performance.

This section focuses on the calculation of additional feature generation. Each resulting data point represents a pump card operation. The pump delivery, or the liquid flow rate provided by the plunger, is primarily driven by the upward and downward motions of the rod pump with time (i.e., number of strokes per minute) and the effective plunger stroke length, as shown in **Equation 5.1**

. The plunger's mechanical properties and fluid properties are taken from the early model developed by Pienknagura Dolberg 2019 (**Equation 5.1** and **5.2**). The formation fluid factor value

is 1, and the volumetric efficiency of the plunger is 80% because the system utilizes water as the testing fluid and slippage factor.

$$q = 0.1484 \frac{A_p N S_p E_v}{B_o} \ (STB/D) \tag{5.1}$$

Where A_p is 3.167 in^2, N is the number of strokes per minute (SPM), $B_0$is equal to 1 with water utilized as the testing fluid in the facility.

$$S_p = S - (e_r + e_t) + e_o \tag{5.2}$$

Where $e_r = 0.0035$, $e_t = 1 \times 10^{-6}$, e_o=0.000015. (Pienknagura Dolberg 2019)

The resulting liquid flow rates for varying inputs of linear actuator's RPM, stroke length and shapes are displayed in **Figure 5.1**. The facility can achieve flow rates of 6 to 54.1 STB/D. The increasing trends of flow rate are observed with the increase in input RPM and stroke length. As the rod accelerates more rapidly with trapezoidal shape, a higher amount of fluid is delivered by plunger over time than the sinusoidal shape. The impacts of flow rate on pressure parameters are also considered in the following discussion.

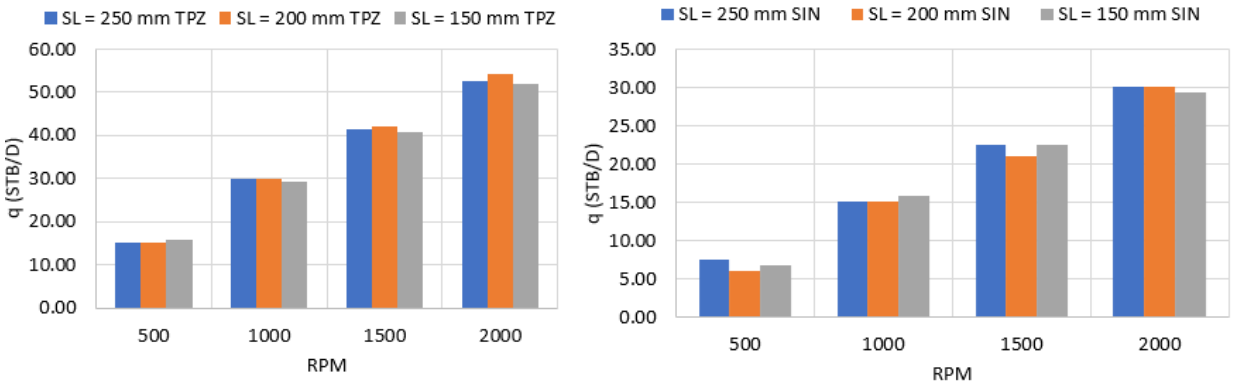


Figure 5.1 Liquid flow rates with varying inputs of RPM, stroke length and card shapes

The pressure difference between the top and pump discharge sensors shows the flow's pressure gradient, implying the frictional pressure loss after subtracting the hydrostatic column

(**Equations 5.3** and **5.4)**. The resulting frictional losses are shown in **Figure 5.2** with respect to flow rate, combining with systematic and random uncertainties. The lower flow rates lead to lower values of frictional losses, which are no more than 2 psi. As the flow rate is too low with the shorter displacement range (**Figure 5.1**), a possibility of flow back could not occur due to minimal or too low frictional factor.

$$\Delta P = P_T - P_{B_2}(psi) \tag{5.3}$$

Where P_T is the pressure record at the top sensor (psi) and P_{B_2} is the pressure recorded at second bottom or pump's discharge sensor (psi).

$$\Delta \mathrm{P}_{\mathrm{friction}} = \Delta P - P_{hydrostatic}(psi) \tag{5.4}$$

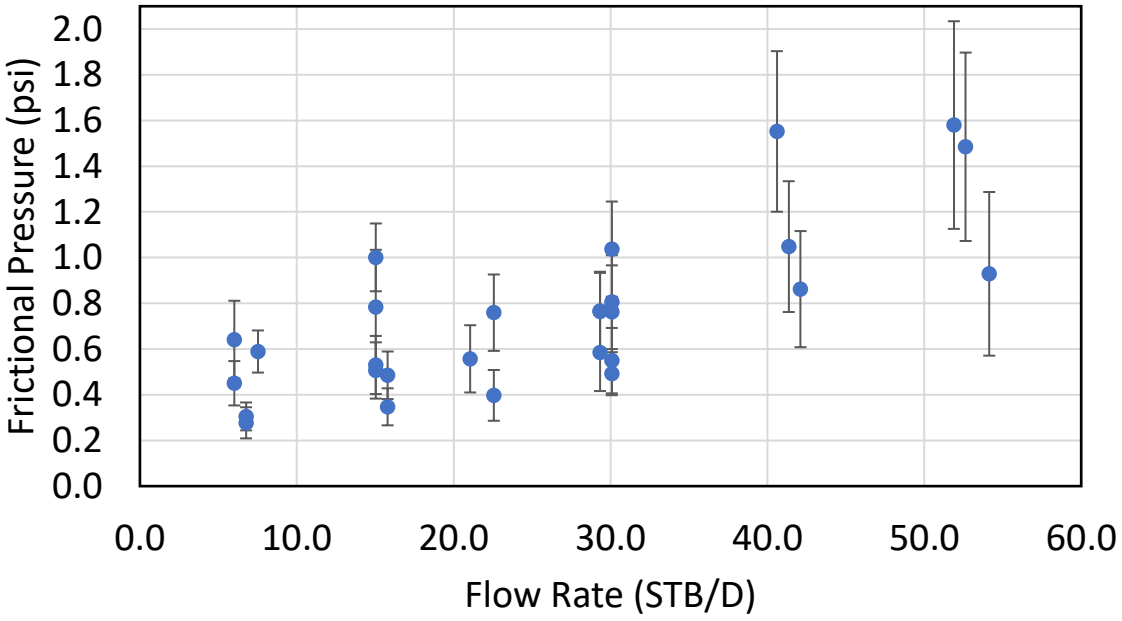


Figure 5.2 Frictional pressure results with liquid flow rate delivered by plunger

The cyclic trend in the suction pressure recorded at the first bottom sensor (**Figure 4.2**) provides the information about the amount of water column pumped at the intake. The changes in

the hydrostatic column of water can be inferred from the difference between maximum and minimum pressure values recorded at upstroke and downstroke (**Equations 5.5** and **5.6**). The frictional pressure results are significantly smaller than the hydrostatic pressure column, as seen in **Figure 5.2**, and can be neglected for this estimation. **Figure 5.3** reveals the values of B1 pressure range or intake pressure (ΔP_{B1}) and resulting water column per cycle at the pump intake with respect to liquid flow rate. The datapoint are scattered in the plot because flow rate is a function of both stroke length and SPM, while water column primarily depends on stroke length and pressure B1. The generalized linear regression (GLM) is added to improve the fitting. The larger range of pressure B1 means a higher column of water fed into the system during the upstroke, thus a higher flow rate.

$$\Delta P_{B1} = P_{B_1 max} - P_{B_1 min}(psi) \tag{5.5}$$

$$Water\ Column\ at\ Pump\ Intake\ per\ Cycle = \frac{\Delta P_{B1}}{0.433}\ (ft) \tag{5.6}$$

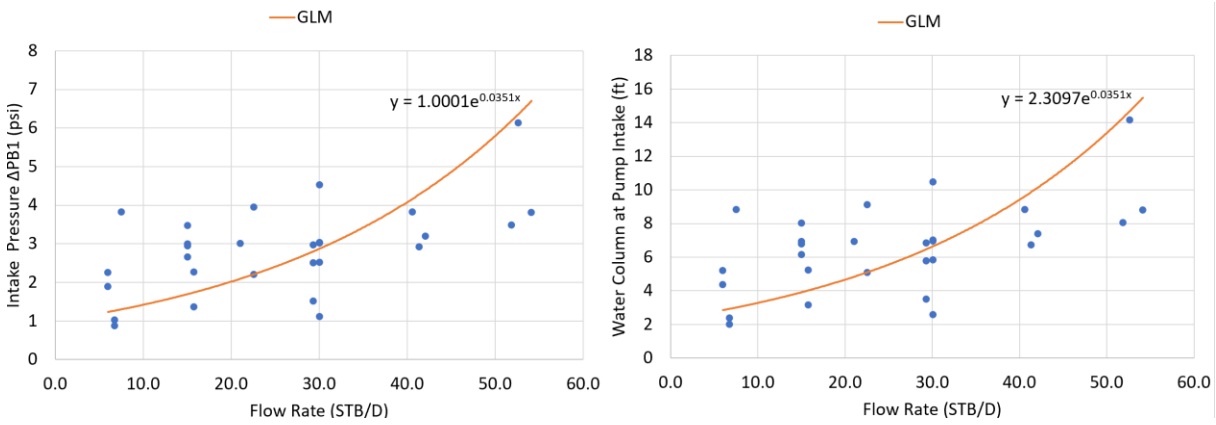


Figure 5.3 Pressure B1 range and resulting water column pumped per cycle at the intake with liquid flow rate

As the rod load indicates a recurring pattern with time (**Figure 5.4)**, the range of rod load is calculated as the difference between the peak load achieved on the upstroke and the minimum load recorded on the downstroke (**Equation 5.7**). This feature is an important parameter for the

predictive modeling, because the load range should be consistent with time when the pump operation reaches the steady state under ideal conditions.

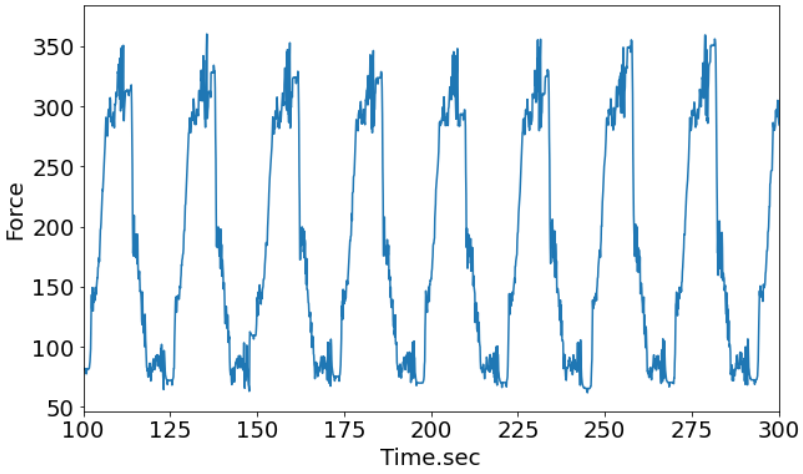


Figure 5.4 Rod load raw data with time, with 200 mm stroke length and 2.5 SPM

$$\Delta F = F_{max} - F_{min}(N) \quad 5.7$$

5.2 Data Preprocessing: Feature Selection and Pump Card Categorization

This section focuses on preparing the data and selecting target features for the predictive modeling by analyzing the changes in parameters with time. As for data cleaning, the data points from the first-time cycle are omitted because the operation has not reached the steady state. The analysis in this section is carried out by selecting pivotal data points based on each time cycle. When the SRP operates in a continuous normal mode, it maintains steady-state conditions with constant pressure, continuous inflow into wellbore and fluid lifted to the surface (Rowlan and McCoy 2018). Due to the instability of SRP operation, the negative outcome alters the consistent trend of parametric features with time. Following the statistical analysis concluded that the rod speed (SPM) could be the driving factor to determine pump operations, impacts of fast or slow rod motion are analyzed to categorize normal operations and check the possibility of pumping off.

The plots of differential pressures and resulting water column filled at the pump intake with respect to time are shown in **Figure 5.5** and **Figure 5.6**. With the combined changes of stroke length and SRP speed, the pressure difference between wellhead and bottom is maintained consistent with slight fluctuations of less than 1 psi. The trends in differential pressure are not significant. On the other hand, the increase in rod speed (SPM) alters the suction pressure trends and the resulting water column filled at the pump intake. As time goes by, the continuing high rod speed removes an excessive amount of liquid in the casing annulus, delaying the liquid fallback to get into the suction point, thus reducing the B1 (suction) pressure ranges and the liquid column. As a result, the suction pressure and liquid column fall off from the constant trend with time. As suction pressure range and resulting liquid column at pump intake strongly correlate with time, these features are selected for output candidates when the predictor variables are the experimental inputs. **Figure 5.6** prompts a potential real-time predictive mode for the expected SRP performance. When the actual data behaves differently from the predicted model, anomaly occurrence is detected.

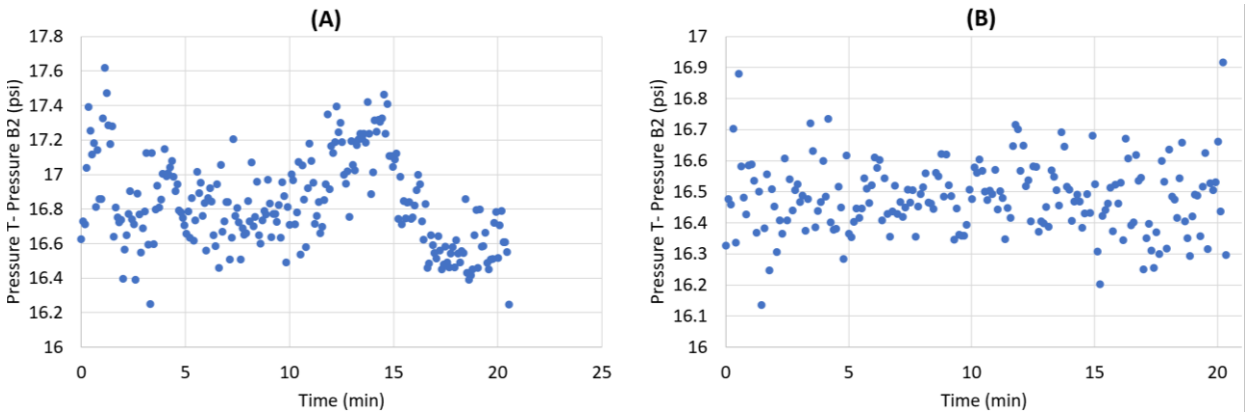


Figure 5.5 Differential Pressures obtained from tests with (A) 250 mm stroke length with 9.6 SPM, and (B) 200 mm stroke length and 7.8 SPM with time

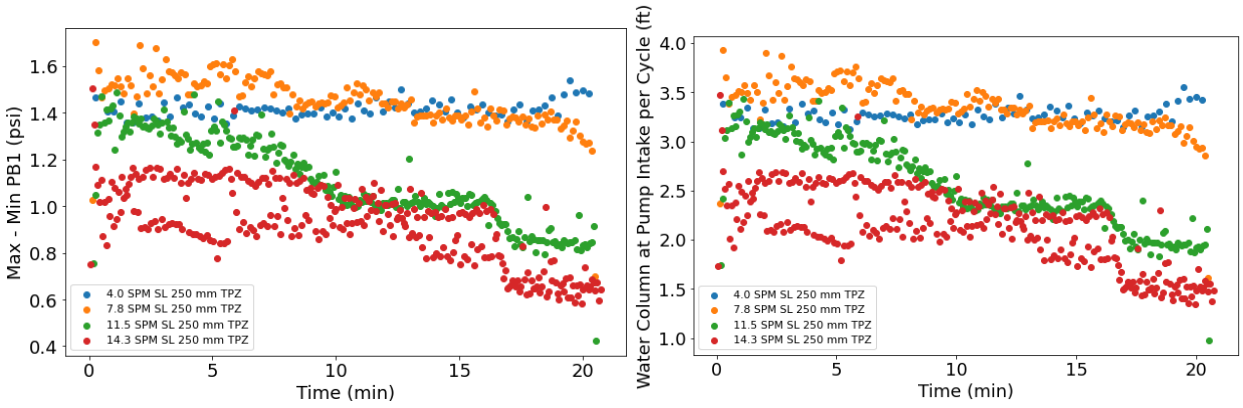


Figure 5.6 The difference of maximum and minimum intake pressure (B1) and resulting water column at pump intake per cycle, with 250 mm stroke length in trapezoidal shape

The tests with IDSRP were planned for normal operation with continuous fluid input to the pump. However, for the tests with high SPM values with different displacement shapes, fluid recirculation back to the pump could not always keep up. As a result, the potential for incomplete pump fluid fillage was raised. Before training the predictive mode, the pump cards for various tests were categorized by looking at the intake pressure trends with time throughout the test.

If pump fillage reduces with time, the intake pressure gradually reduces, and the pumping-off issue occurs. It means that the slope of intake pressure with time becomes more negative than the tests with steady-state normal conditions. It is initially assumed that the rod SPM is the primary driving factor for this condition. The plot of **Figure 5.7** provides the SPM values corresponding to the linear actuator's RPM and stroke length, from which 5 levels of SPM values are established. The details of the 5 groups are summarized in **Table 5.1**, including highest, upper high, middle, lower, and lowest SPMs. The groups of upper high SPMs are primarily results of abrupt trapezoidal changes in rod displacement and faster linear actuator's RPM. The largest SPM value that sinusoidal shape with 250 mm stroke length can achieve is 8 SPM.

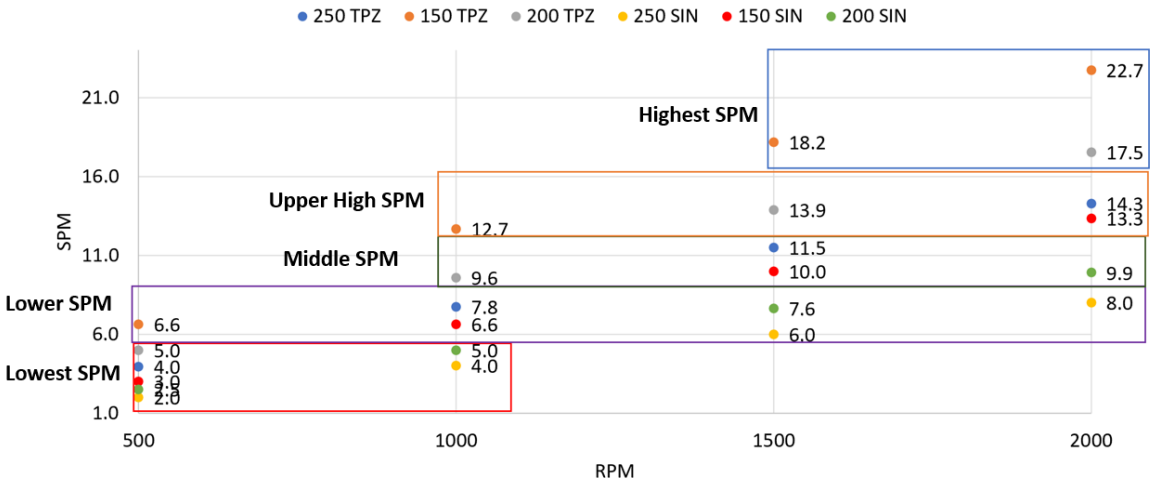


Figure 5.7 Five levels of SPM values based on SPM vs. RPM plot

Table 5.1 Groups of stroke per minute test levels

Group Type	SPM values
Group 1: Highest SPM	17.5, 18.2, 22.7
Group 2: Upper High SPM	12.7, 13.3, 13,9, 14.3
Group 3: Middle SPM	9.6, 9.9, 10, 11.5
Group 4: Lower SPM	6.0, 6.6, 7.6, 7.8, 8.0
Group 5: Lowest SPM	2.0, 3.0, 3.0, 5.0

In order to analyze the intake pressure trends with time, a linear regression line is fitted into the B1 pressure range (intake pressure) and the resulting water column filled per cycle at pump intake. These regression lines are shown in **Figure 5.8** and **Figure 5.9** for the tests in group 3 of SPM values as an example. The resulting slope is obtained and shown in **Table 5.2**. Although these pump cards have relatively close SPM values, the pump data with 250 mm stroke length indicate a more negative slope, meaning that the filled liquid column and intake pressure decline faster. Suppose this test is conducted for a longer time. In that case, the pump fillage could reduce even more, amplifying the severity of pumping-off and the possibility of the deteriorated pump with excessive rod displacement and speed. Therefore, within the tests in this group, the test with

250 mm stroke length, 1500 RPM, and trapezoidal shape is considered to indicate the early signs of pumping-off.

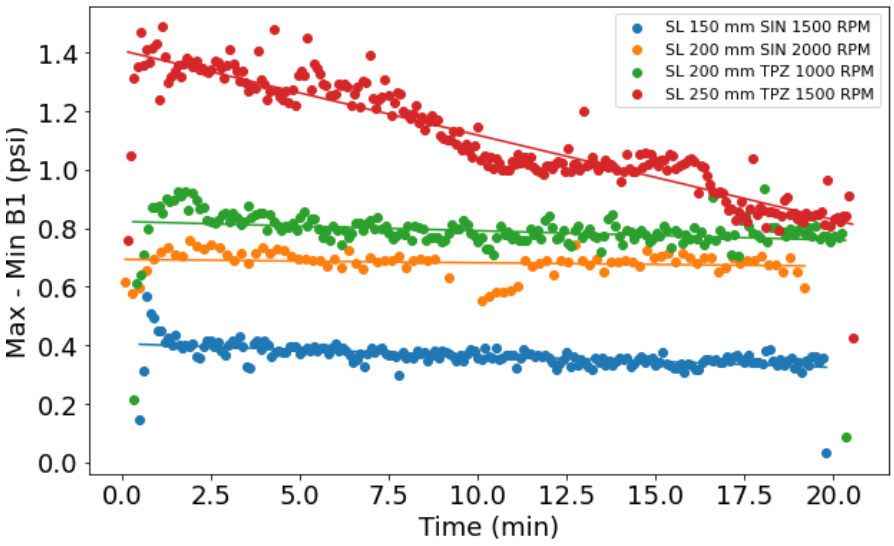


Figure 5.8 B1 pressure trends per cycle with time for tests in group 3 middle SPM values

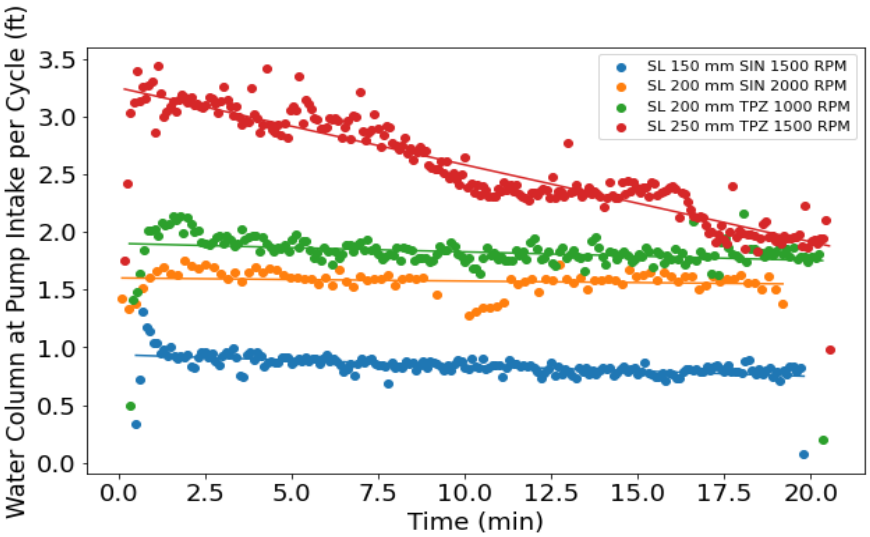


Figure 5.9 Water column at pump intake per cycle with time for group 3 middle SPM tests

Table 5.2 Slope results obtained from linear regression fit for group 3 middle SPM tests

SPM	Stroke Length	RPM	Shape	Slope ΔP_{B1}	Slope of water filled at pump intake
9.6	200	1000	TPZ	-0.00317	-0.00732
9.9	200	2000	SIN	-0.00115	-0.00265
10.0	150	1500	SIN	-0.00405	-0.00936
11.5	250	1500	TPZ	-0.0289	-0.0667

A similar process was carried out to find slopes through linear regression fitting in the remaining SPM groups. A more negative slope is observed if the pump card indicates early pumping-off signs. As the water column depends on the intake pressure through **Equation 5.6**, only the representative regression fit results of the water column filled at the pump intake are shown for the remaining groups in **Figure 5.10**. There are slight instabilities observed in a few pump cards at upper high and lower SPM groups. Even though the production slightly declines with time, the amount of water filled at pump intake is maintained relatively constant for SRP systems operating at lower speed groups.

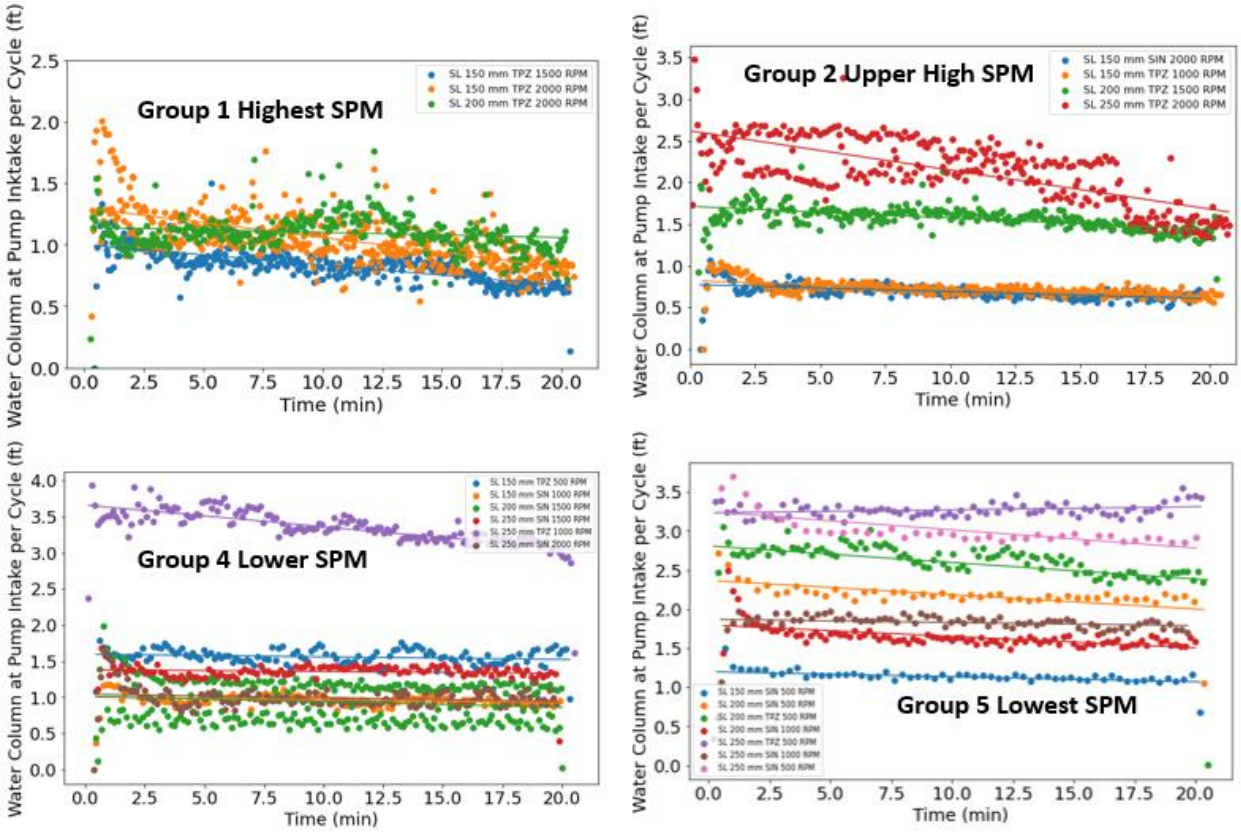


Figure 5.10 Water column at pump intake per cycle with time for the highest, upper high, lower and lowest SPM groups

The specific input that determines the SRP cards of normal and pump-off operations remains inconclusive. Grouping the tests based on only SPM values does not fully encompass other factors, such as displacement shape, stroke length, or linear actuator's RPM. Therefore, the product of speed (SPM) and the stroke length is used as the proportional length of the liquid column filled per minute. Using this parameter establishes a threshold to determine the possibility of pumping off. If the total length of the water column getting pumped at a given time (mm/min) exceeds the feed water amount, the water level at the pump intake decreases. This occurrence reduces the suction pressure over time, and pumping-off may occur. Therefore, a negative slope for the intake water column plot with time can be interpreted as pumping off.

Figure 5.11 demonstrates the changes in the length of water column pumped (mm/min) with respect to the slopes of intake pressure and resulting water column with time. When the water column length being pumped exceeds 2000 mm/min, more anomalous cases are observed with more negative slopes. Thresholds of -0.0104 for the slope of the intake pressure range and -0.024 for the slope of the water column at pump intake are defined as the onsets of pumping off. With this observation, four of the experiments result in slightly pump-off cases.

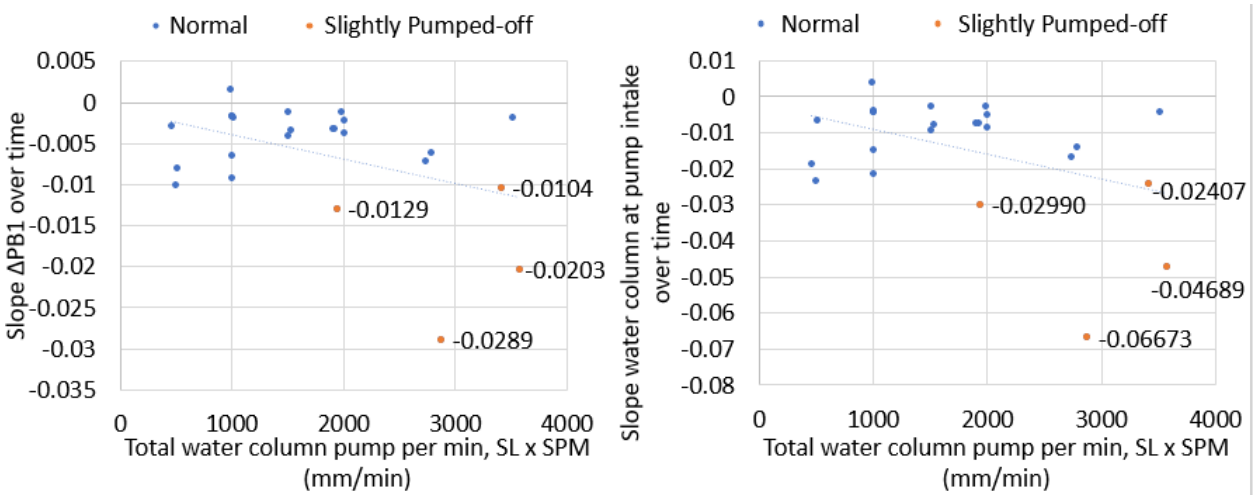


Figure 5.11 Slopes of pressure B1 range (ΔP_{B1}) and water column at the pump intake in response to total flow rate (mm/min)

Even though the physical threshold is established, we would like to validate input combinations that categorize normal operation or slightly pump-off by the data analytics approach. The objective for building a machine learning model is to create the predictive regression-based and classification models to predict normal operation performance and detect the abnormal case if there is any deviation, based on multiple experimental inputs. For validating the pump classification from **Figure 5.11**, the model is trained to fit the operation performance while detecting abnormality by deploying in slightly pump-off cases. As shown in **Figure 5.12**, all the pump cards are split with 70% used for model training and 30% for model testing. As the regression model will differentiate between normal and anomalous cases, the sets of input from

slightly pumped-off cases will predict the trained model to determine whether the actual issue is deviated from the expected normal operation. The data from slightly pumped off cases will be utilized as diagnosis cases to detect the abnormality.

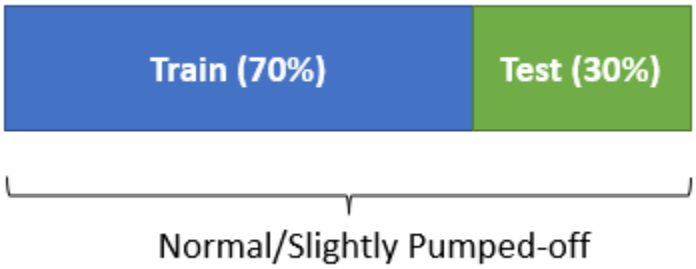


Figure 5.12 Data preparation for modeling

5.3 Multiple Input Multiple Output Model Discussion

Multiple-output regression is a method to predict two or more numerical values with multiple input examples. This technique has been widely implemented in the industry for production forecasting, drilling optimization, and formation evaluation. The studies from Guo et al. (2009) and Razali and Izni Mustafar (2011) provide the implications of utilizing multiple linear regression to predict oil production volume. The multiple output linear regression approach was also extended by Torlov et al. (2020), incorporating downhole ultrasonic formation tester data to determine the in-situ gas-oil ratio. The applications of multiple linear and non-linear regression models are also discussed in Yazitova and Yagis (2020) when the models are trained and evaluated in Measure While Drilling (MWD) log and rock properties data to estimate the drilling rate. With arrays of experimental test matrix and output collection, this study is an attempt to build a multiple input multiple output regression model towards real-time prediction of expected SRP performance and potential abnormal occurrences.

5.3.1 Multiple Output Linear Regression (LR)

According to Watt et al. (2020), the basic elements of multiple-output regression model include vector-valued input and output. At p[th] point (x_P, y_P), the input x_P has N dimensional inputs, i.e. $N \times 1$ columns, and the output y_p has associated C dimensions, i.e. $1 \times C$ row vectors (**Equation 5.8**).

$$(x_1, y_1), (x_2, y_2), \dots, (x_P, y_P) \tag{5.8}$$

The dimensionality of (x_P, y_P) is elaborated in **Equation 5.9**. The index of y_p starts from 0 instead of 1 in this case, because each input point x_p has a 1 stacked on top. The newly adjusted form of each point x_P is denoted as $\dot{x_p}$.

$$x_P = \begin{bmatrix} x_1, P \\ x_2, P \\ \vdots \\ x_N, P \end{bmatrix} \quad y_p = \begin{bmatrix} y_{0,p} \; y_{1,P} \; \dots y_{C-1,P} \end{bmatrix} \quad \dot{x}_p = \begin{bmatrix} 1 \\ x_1, p \\ x_2, p \\ \vdots \\ x_N, p \end{bmatrix} \text{ where } p = 1, \dots, P \tag{5.9}$$

With weight matrix W containing compressed weight vector of $[w_0 \; w_1 \; \cdots w_{C-1}]$, the linear relationship between $\dot{x_p}$ and the entire set of C elements of y_p can be written as:

$$\dot{x_p}^T W = \begin{bmatrix} \dot{x_p}^T w_0 \; \dot{x_p}^T w_1 \; \dots \; \dot{x_p}^T w_{C-1} \end{bmatrix} \text{ or } \dot{x_p}^T W \approx y_p \quad \text{where } p = 1, \dots, P \tag{5.10}$$

The goal of multi-output regression is to achieve the optimal likelihood of approximation in **Equation 5.10**. Therefore, the objective is to minimize the regression cost function, calculated by Least Squares or Absolute Deviations, displayed in **Equations 5.11** and **5.12**, respectively. The results of these cost functions are considered as the evaluation metrics to validate the models. Assuming that $\widehat{y_i}$ is a predicted value of i[th] samples and y_i is corresponding true value within $n_{samples}$, the Absolute Deviations use the mean absolute error (MAE) as the regression metric in

this study (**Equation 5.13**). Least Squares are also inferred to as Mean Square Error (MSE), corresponding to the square error of loss (**Equation 5.14**).

$$g(W) = \frac{1}{P}\sum_{p=1}^{P}\left\|\dot{x_p}^T W - y_p\right\|^2 = \frac{1}{P}\sum_{p=1}^{P}\sum_{c=0}^{C-1}\left(\dot{x}_p^T w_c - y_c, p\right)^2 \quad 5.11$$

where $p = 1, \dots, P$

$$g(W) = \frac{1}{P}\sum_{p=1}^{P}\left\|\dot{x_p}^T W - y_p\right\| = \frac{1}{P}\sum_{p=1}^{P}\sum_{c=0}^{C-1}\left|\dot{x_p}^T w_c - y_{c,p}\right| \quad 5.12$$

where $p = 1, \dots, P$

$$MAE(y, \hat{y}) = \frac{1}{n_{samples}}\sum_{i=0}^{n_{samples}-1}|y_i - \hat{y}| \quad 5.13$$

$$MSE(y, \hat{y}) = 1 - \frac{1}{n_{samples}}\sum_{i=0}^{n_{samples}-1}(y_i - \hat{y})^2 \quad 5.14$$

Other important metrics to evaluate the multi-input multi-output regression models include R-squared (R-2 or R^2) and explained variance score (Pedregosa et al. 2011). R-2 or R^2, also called the coefficient of determination, is a statistical measure for how close the data are to the regression model, i.e., goodness of fit. The metric is calculated by the proportion of explained variance as shown in **Equation 5.15**. The best possible score is 1.0. A score of 0 indicates a constant model, while negative values indicate arbitrary worse model performance. The explained variance score suggests the dispersion of errors or the proportion of model that accounts for the variation of the dataset (**Equation 5.16**).

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad 5.15$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$$\text{explained variance } (y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}} \quad 5.16$$

where *Var* is the square of standard deviation.

Besides Linear Regression (LR), which is commonly deployed for multiple input multiple output models, this study also explores the possibility of utilizing k-Nearest Neighbor (KNN) and Decision Tree (DT) to optimize the predictive models. The reason is that the input parameters include multiple categorical variables, such as stroke length, shape, SPM, which can generate either set of neighbors based on the shortest distance or the set of decision rules to split these input nodes to predict the target leaf.

5.3.2 k-Nearest Neighbor (KNN)

KNN is a distance-based machine learning method. In this method, each training data point is mapped to a multidimensional space, where each axis corresponds to a distinctive variable. In a regression context, the predicted value $\hat{y}$ can be mean or weighted mean of y values for the neighbors. These values are represented as distance $d(q, x_c)$ and k neighbors (**Equation 5.17**) (Naomi et al. 1992, Cunningham et al. 2020). The mechanism of this technique is to predefine the number of nearest neighbors to determine the $\hat{y}$ value of new observation. The validation of the model is performed using the test data.

$$\hat{y} = \frac{1}{k}\sum_{c=1}^{k}\frac{1}{d(q, x_c)^P}y_c \qquad 5.17$$

where $d(q, x_c)$ is a distance between unknown example q and training sample x_c.

5.3.3 Decision Tree (DT) Regression

Decision Tree (DT) is a data mining technique that establishes different decision rules to classify a particular target (Morgan and Sonquist 1963). The instance for DT regression is to split two or more groups by internal node and a set of decision rules. It follows the divide and conquer greedy algorithm that splits data into smaller parts in a recursive manner (Pekel 2020). Apart from the classification tree, the regression tree takes ordered values at each split point and establishes

the regression model that fits each node to generate a predicted output y (Loh 2011). The DT regression model deployed in this study follows Classification And Regression Trees (CART) algorithm that involves greedy splitting and pruning as the stopping criterion (Breiman et al. 2017). The pseudocode is established as below:

1. Start at root node.
2. Find each feature's best plot and node's best split. For each X, find the set S that minimizes the sum of node impurities in the 2 child nodes and chooses to split $\{X^* \in S^*\}$ that gives the minimum overall X and S. In CART algorithm, the impurity function uses the generalization of binomial variance, i.e., Gini index (**Equation 5.18**). As for regression purpose, the stopping criterion is the mean squared error (MSE) (**Equation 5.14**).

$$i(t) = \sum_{i,j} C(i|j)p(i|t)p(j|t) \qquad 5.18$$

   where $C(i|j)$ is a cost of misclassification of a class j case as a class i, and $p(i|t)$ and $p(j|t)$ are probabilities of cases in class i (or j) falling into node t.
3. If stopping criterion, or pruning, is reached, exit. Otherwise, step 2 is applied recursively to each child node in turn.

Figure 5.13 provides a sample of how Decision Tree Regressor generates a set of rules in this dataset. In each node, variable features try all possibilities within the value range and keep splitting in different subsets in which the MSE is as small as possible. The tree continues splitting into many child nodes until reaching the leaf node. The stopping criterion is the mean squared error (MSE) (**Equation 5.14**).

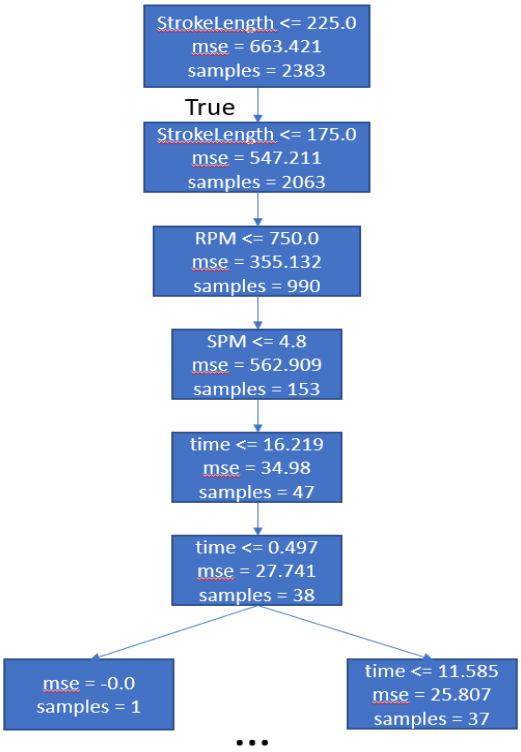


Figure 5.13 A sample of decision tree regression

5.4 A Discussion on Classification Models

Classification is a machine learning technique to identify the different label outputs from predictor features. The two classification types include supervised and unsupervised learnings. Supervised learning establishes the relationship between the known label and the entities, while unsupervised learning builds a clustering based on the input variables. The targets of the classification problems could be binary or multiple classes. The evaluation primarily focuses on the metrics derived from the confusion matrix, including the actual and predicted values in response to each label. Some performance metrics include accuracy, precision, the areas under the curves of true and false positive rates, and the trade-off between precision and recall.

The common approaches to solve classification problems include logistic regression, decision tree and its ensembles. Apart from linear regression, logistic regression utilizes the

logistic function to determine the probability or predict the binary outcome (Wright 1995, Edgar et al. 2017). In the classification problem, the decision tree technique predicts the target variable by learning the decision rule. The pseudocode for classification follows **5.3.3** discussion, except using the Gini index as a stopping criterion (**Equation 5.18**). **Figure 5.14** provides an example of decision tree classification applied on this dataset. Each parametric feature keeps splitting in each node until the stopping criterion is reached. Random forest and gradient boosting are the improved forms of the decision tree due to multiple decision trees ensemble. Random forest randomly selects subsets of input attributes (experimental parameters) for splitting the tree at any node, creating multiple trees (Breiman 2001). The trees giving minimal errors are aggregated, or bagged, to provide the final model. The random nature of the model improves its robustness and reduces variance. Gradient boosting generates the succession of trees that learn from the mistakes of previous trees by minimizing the residual error (Breiman 1997). This boosted decision tree reduces both variance and bias of the model due to subsequent learning. In this thesis, the classification is conducted with comprehensive datasets from both 5 and 20-minute tests. It serves as a mean to justify the pump condition categorization based on the threshold criteria described in **5.2**.

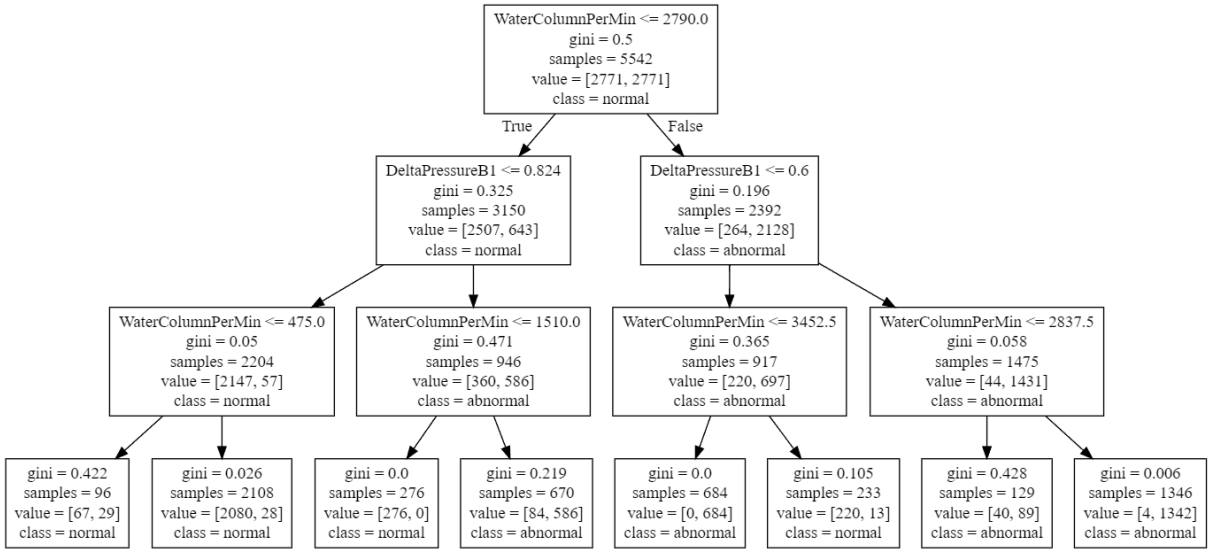


Figure 5.14 A sample of decision tree classification

5.5 Results and Discussion

5.5.1 Multiple Input Multiple Output (MIMO) Regression Evaluation

Following the algorithm discussion, the models are trained with the experimental inputs, including shape, RPM, SPM, stroke length, and test duration. **Table 5.3** and **Table 5.4** summarize the evaluation metrics on validation data for the three MIMO regression techniques. These metrics denote the capability of the models to fit and predict accurately. Due to the ability to utilize distance-based calculation to navigate the neighborhood, KNN performs the best. It shows a better evaluation score in predicting output values compared to the other two techniques, with the 0.87 goodness of fit and relatively low errors.

Table 5.3 Mean Absolute Error (MAE) and Mean Squared Error (MSE) metrics of validation data of multiple outputs for 3 MIMO regression techniques

Output	Water column filled at pump intake per cycle	Intake pressure (ΔP_{B1})	Rod load range (ΔF)
MAE LR	0.27	0.116	37.77
MSE LR	0.12	0.021	2483.6
RMSE LR	0.35	0.14	49.84
MAE KNN	0.10	0.044	14.12
MSE KNN	0.032	0.006	86.92
RMSE KNN	0.179	0.077	9.32
MAE DT	0.13	0.06	18.17
MSE DT	0.059	0.01	1464.3
RMSE DT	0.24	0.1	38.27

Table 5.4 R-2 evaluation score and explained variance of validation data for 3 MIMO regression techniques

Technique	R-2 Evaluation Score	Explained Variance
Linear Regression	0.5441	0.54445
K-nearest Neighbor	0.87439	0.87454
Decision Tree	0.78152	0.78164

The predictions of the three models are compared to the SRP features over time for a test with normal conditions, as shown in **Figure 5.15**. The predicted results from KNN and DT are

relatively close to the actual normal operation data. The visual observations imply that Linear Regression does not perform well due to the linear relationship assumption, which fails to adopt the cyclic nature of raw data.

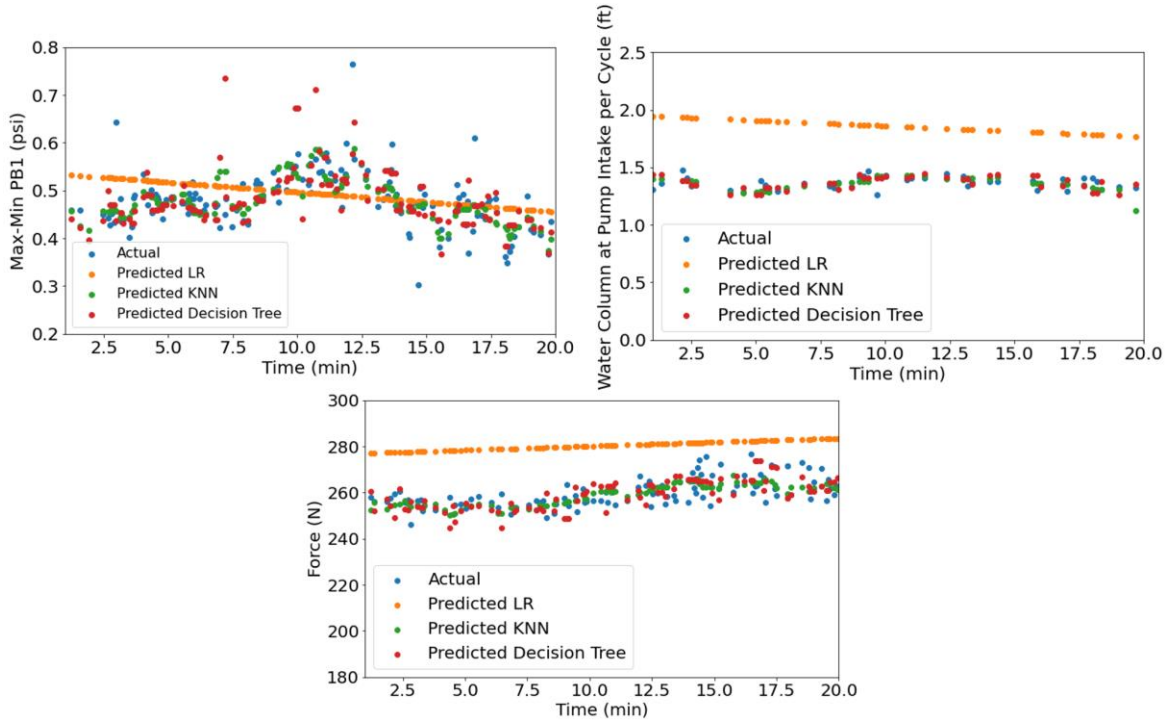


Figure 5.15 Visual comparison between the true (actual) normal operation data, predicted multiple outputs in test data

5.5.2 Model Deployment

The trained model is deployed in the slightly pump-off data. In this case, the prediction differentiates actual normal and slightly pump-off operations based on the same sets of experimental inputs. In the testing stage, KNN and DT models are selected based on the evaluation results from **5.5.1**. **Figure 5.16** exhibits the distinctive trend between normal and abnormal operations. The evaluation metrics include the negative R-squared (R-2/ R^2) of -3.56 for KNN and -4.78 for Decision Tree, and significant errors shown in **Table 5.5**. The quantitative measurement implies that the pump-off cases do not follow the predicted trend of normal operation in the steady condition. The predicted constant suction/ intake pressure and constant water level at pump intake

confirm the underlying physical basis in these MIMO regression models. In this case, the abnormal operation is detected when parametric performances of slightly pump-off cards deviate from the diagnostic plots of forecasted models. Overall, these predictive models can be applied to any experimental input set to diagnose the pump operation and monitor the early signs of abnormal occurrences (i.e., pumping-off issues) in IDSRP.

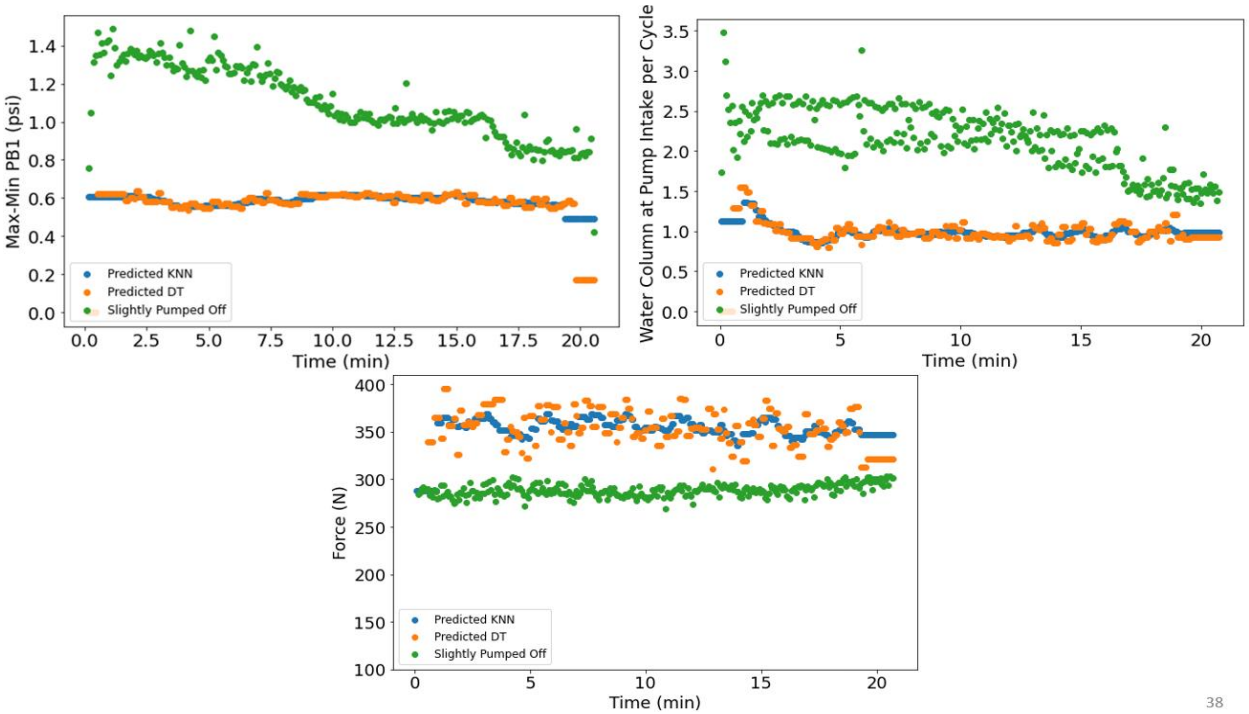


Figure 5.16 Visual comparison between the slightly pump-off case and the normal operation predicted from KNN and Decision Tree

Table 5.5 Mean Absolute Error (MAE) and Mean Square Error (MSE) metrics on testing data of multiple outputs for 2 MIMO regression techniques

Output	Water column filled at pump intake per cycle	Intake pressure (ΔP_{B1})	Rod load range (ΔF)
MAE KNN	0.892	0.386	53.894
MSE KNN	1.094	0.205	4000.48
RMSE KNN	1.046	0.453	63.249
MAE DT	1.093	0.473	53.089
MSE DT	1.805	0.338	4512.31
RMSE KNN	1.344	0.581	67.17

5.5.3 Supervised Learning Classification

MIMO regression approaches generate models to predict the normal performance of a SRP and detect anomalies. There is more confidence in quantifying the threshold parameters to identify the operational type of a pump card. The threshold parameters may be the stroke length and number of strokes per minute or the minimum slope of the intake pressure and the resulting water column at pump intake over time. The last step is to carry out a comprehensive supervised learning model to classify the SRP's operation based on the experimental parameters. Apart from the given test matrix, four additional tests with random inputs are included in the dataset to ensure the volume and variety qualities for big-data applications. Therefore, the supervised classification approach incorporates data from 52 pump cards with 5- and 20-minutes testing durations.

The labels used for the analysis are shown in **Figure 5.17** with 12 anomalous cards or 1741 associated data points in Label 1, and 40 normal cards or 4568 associated data points in Label 0. The crucial assumption for the classifier is that there are no intercorrelations (i.e., multicollinearity) among predictors. Instead of selecting both intake pressure (ΔP_{B1}) and resulting water column filled at the pump intake, only the intake pressure is considered as an element of predictor sets. Because the product of stroke length and numbers of stroke per minute (i.e., expected total water being pumped per minute) affects the operational outcome, the predictor set extends to this feature. Other parametric attributes include rod displacement, rod load range, time, and the pressure difference between the top and pump discharge sensors. The dataset is split into training and testing groups with a ratio of 60-40. The imbalanced data concern due to uneven distribution is addressed by up-sampling the minority class (i.e., Label 1) in the training set. The up-sampling technique prevents the model from bias prediction inclining towards the majority class (i.e., Label 0).

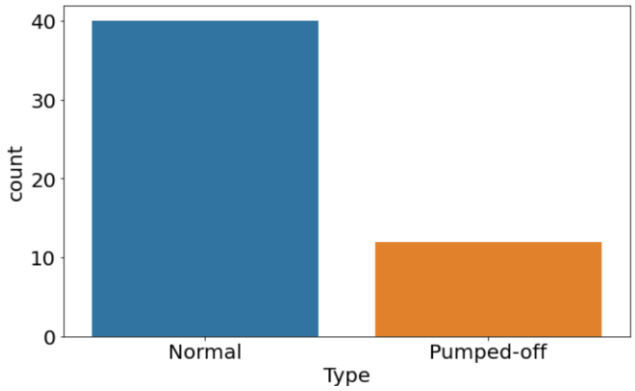


Figure 5.17 Resulting labels for 52 pump cards obtained from IDSRP testing based on the slope of intake pressure with time

As the output is binary, the four machine learning models include Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. Once these models are trained, the prediction is evaluated using the test data. **Table 5.6** summarizes the evaluation for the four classification techniques. The accuracy metric represents the fraction of test data predicted correctly. Precision indicates the confidence level of predicting Label 1 (i.e., abnormal case) correctly. F-1 score is a harmonic mean of precision and recall (the proportion of all Label 1's or all Label 0's predicted correctly). The marginal overall accuracy metric implies the robustness of these classification techniques in determining normal and slightly pump-off cards.

Table 5.6 Different evaluation metrics on test dataset for supervised classification methods

Methods	**Overall Accuracy**	**Precision (confidence predicting Label 1)**	**F-1 score Label 1**	**F-1 score Label 0**	**AUC**
Logistic Regression	0.86	0.69	0.78	0.89	0.94
Decision Tree	0.83	0.63	0.77	0.86	0.94
Gradient Boosting	0.91	0.79	0.86	0.94	0.97
Random Forest	0.92	0.79	0.87	0.94	0.98

The Receiver Operating Characteristic (ROC) curve analysis provides further insight into the performance of binary classification at various discrimination thresholds (Gupta et al. 2020).

The ROC curve in **Figure 5.18** shows the relation between true positive rate (probability of correctly predicting that SRP indicates issues) and false positive rate (probability of falsely predicting issues while the SRP operates normally) cases. The resulting area under the curve (AUC) demonstrates the degree of the predictive model separating or classifying the two labels. In this case, the ROC curve could be smoother if more testing or more pump cards are generated for the training set. Considering that the data is imbalanced by nature, precision-recall is another helpful metric to determine the success of prediction. The higher area under the precision-recall curve (AP) indicates both high recall and high precision (Pedregosa 2011).

The AUC and AP results prove Random Forest to be the best classifier among the four techniques. Following this best predictive model, the resulting confusion matrix displays the numerical values of actual, predicted, and misclassified labels in the testing data (**Figure 5.19**). 206 data points are misclassified in this case, while 1606 data points for normal cards and 712 data points for slightly pump-off cards are accurately predicted. It is worth noting that the data are primarily collected from the laboratory design described in **3.3**. The classification models serve as a prototype predictive mode and could be improved by simulating more cases in the facility, and especially by using the other lab or field-scale data.

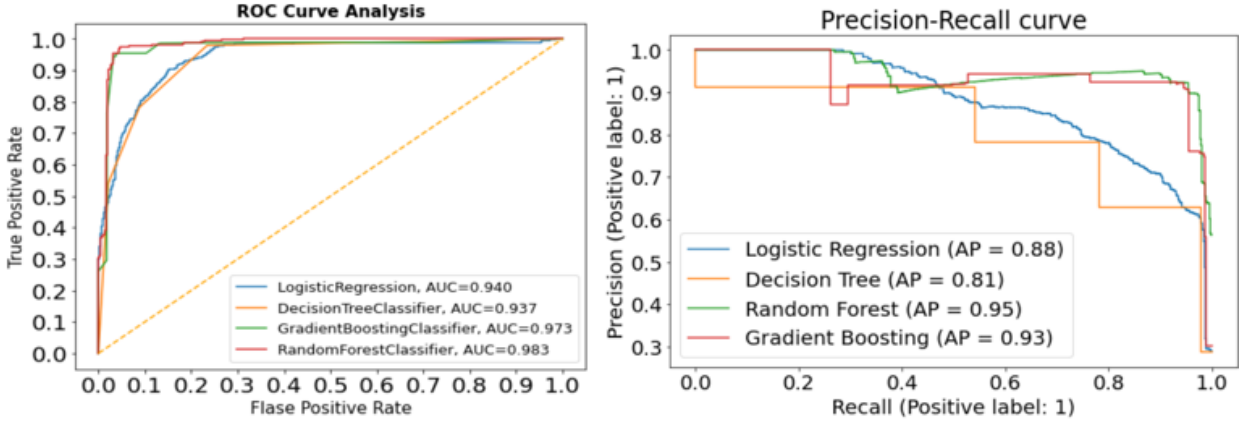


Figure 5.18 ROC curve analysis and Precision-Recall curve document AUC and AP values for the four classification techniques

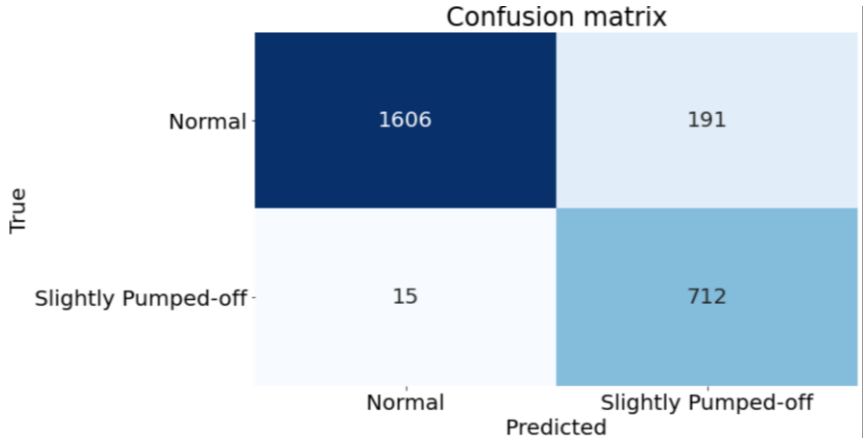


Figure 5.19 Confusion Matrix showing True and Predicted Labels of Random Forest

The final stage of supervised learning is to analyze and quantify the importance of experimental parameters. The feature importance decodes the 'black-box machine learning' and explains how these features generate confidence in the model. The Shapley values, calculated by using Shapley Additive Explanation, indicate the importance of each predictor in predicting class/ label of 0 or 1. In the statistical context, the Shapley values represent the average of marginal contribution across all combinations of features (Lundberg and Lee 2016, Tran et al. 2020, Lam Tran et al. 2020). This method helps operators to determine key physical parameters for predictive modes in production systems.

Figure 5.20 reveals that the total water column being pumped per minute, intake pressure, and pressure difference between the top and discharge sensors are the top three predictors in determining labels 0 and 1. This verifies the selection of both multiplication of SPM and Stroke Length and the changing rate of intake pressure with time as determining factors in labeling SRP condition. As the pump operates under the stabilized conditions, the performance is independent of time, making the time feature not as important as the other entities. The rod load (ΔF) indicator is anticipated to be the primary factor in an actual field scenario. However, the dynocard

transformation results from **4.4** imply the presence of buckling in the rod string and possibly significant uncertainties in the pump card data. Hence, for the case of this study, the rod load factor is not a significant contributor to classify normal and slightly pumped-off classes.

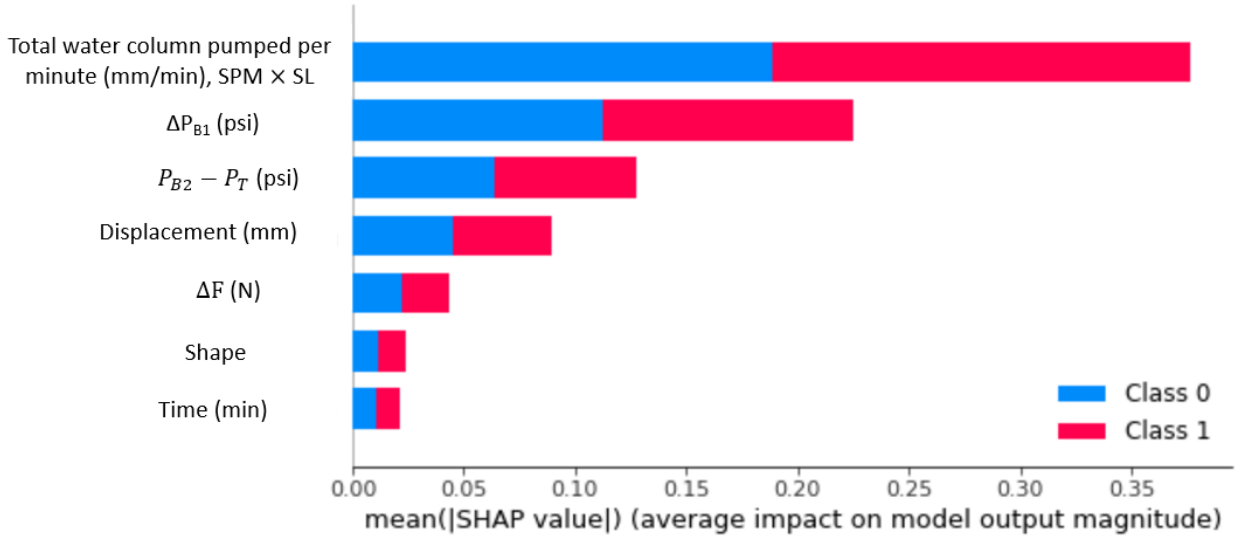


Figure 5.20 Variable importance plot based on Shapley values for normal operation (Class 0) and slightly pumped-off anomaly (Class 1)

Analyzing Shapley values for each label helps in understanding the impacts of operational parameters on SRP conditions. The tornado chart in **Figure 5.21** displays the Shapley values or the significance of all predictors and abnormal card labels in descending order. The colors also represent the intensity of the parametric features. In the x-axis for each predictor, the higher Shapley value means a higher tendency for a pumping-off situation. Therefore, the cases with a higher pumping-off possibility mostly include higher total water column being pumped per minute ($SPM \times Stroke\ Length$), higher intake pressure changing slope, and higher differential pressure cases. A high total water column being pumped per minute is a key attribute for abnormal cards. The observation from **Figure 5.11** also confirms more pump-off cards for the tests with total water columns higher than 2000 mm/min. The results confirm the underlying physical basis of predictive models in categorizing operational type for IDSRP.

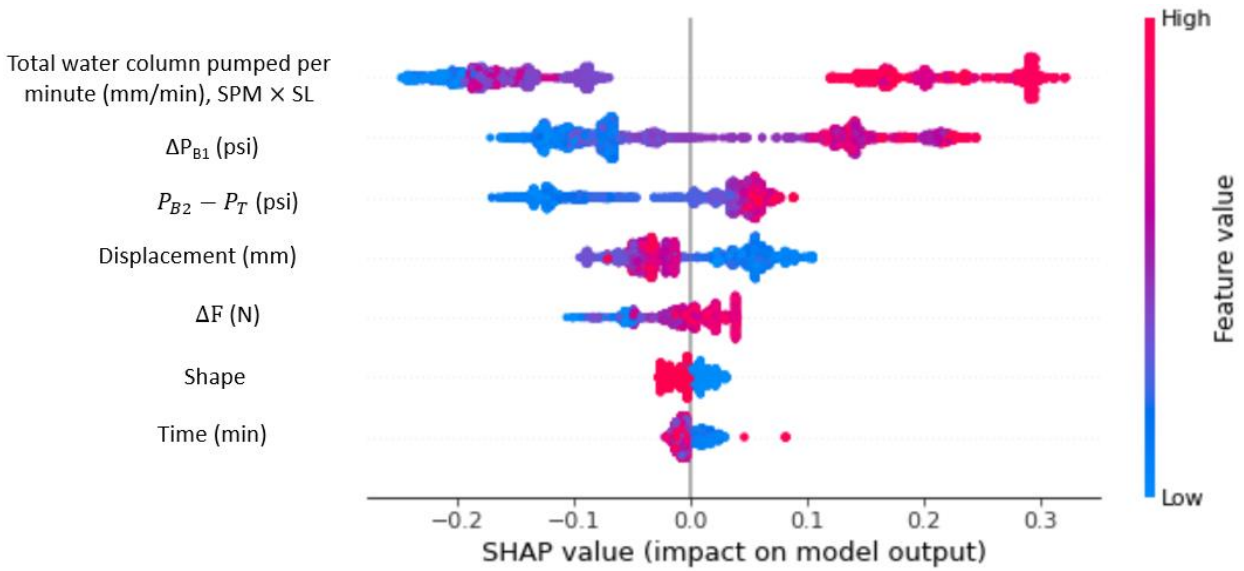


Figure 5.21 Shapley values for label 1 or slightly pumped-off data, varying with different predictors

Chapter 6: Discussions and Conclusions

This chapter provides a brief overview of the main takeaways from this study. First, the challenges of the study are listed and described. Then, after a short discussion on the work's methodology, the main conclusions of the work are listed.

6.1 Challenges

The applicability of any data analytics study depends on whether or not the dataset and model would meet the 4V's of the Big Data, namely volume, variety, velocity, and veracity. Some challenges, primarily related to data uncertainty and the data range limitations, could hinder the proposed data-driven models from being applied to the field operations right away:

- The limited lab scale raises the concern over data quality since some physical factors (e.g. frictional pressure loss, the presence of bubbles) are omitted. Though noise reduction actions are taken care of, the signal processing still requires further data refinement for noise handling.
- The limitation of lab scale extends to the rod buckling during the testing, which increases the uncertainty in load measurement data, and the transformation of the surface to the downhole card. Changing the rod string's material to steel can potentially resolve this issue. Since steel has a much lower elasticity, the errors in downhole conversion will be significantly reduced. With the improvement in translation to downhole cards in the experiments, the possibility of further pattern recognition works could be extended.

- The test duration is not long enough to entirely reflect the severity of the pumping-off issue. Changes in the intake pressure with time are used in this study as an indicator for pumping-off. However, longer-duration tests can help us better understand the trends.
- With the proposed test matrix and limited lab-scale environment, arbitrary threshold values have been used to separate normal and pump-off cases. These values need to be validated by added datasets and data analytics applications. It is anticipated that the assumption to use the slope of the intake pressure remains valid for longer tests. However, the thresholds may need to be adjusted.
- As the load and displacement sensors are linked to the surface actuator, the limitation of surface dynocard transformation can be improved by vibration monitoring. This can avoid rod string's abnormal stretch and buckling. Also, implementing downhole sensors enables applying the subsurface parameters in a real-time manner while conducting the tests.
- A major challenge of this study is the unavailability of field data to compare with the data collected from IDSRP. The comparison of field and lab data validates the labeling of SRP conditions in the facility. Several regression models were carried out to confirm the normal operational trends and detect anomalies. The additional field data help shorten and validate the labeling process and the supervised classification. If the field data are added to the experimental data, the 'variety' quality of this study would be improved.

6.2 Discussions

The data-driven approaches implemented in this study demonstrate the real-time capability of IDSRP facility in mimicking field-scale operations to develop big data applications. The observations from pump categorization in **Chapter 5** confirm the pumping-off occurrence, also

described in **Chapter 4**. This is because the rate of the water column being pumped is too high to maintain the complete pump fillage, resulting in negative intake pressures due to the low liquid level in the casing. The concern of negative suction pressure values could be addressed by lowering the linear actuator's RPM and slower rod displacement. The test matrix from **Chapter 3** could be extended by reducing the linear actuator's speed to less than 500 RPM.

The model developed in **Chapter 5** provides a good prototype for anomaly detection. The linear regression and supervised classification learning tools recognize the slightly pumped-off cases before being validated by a machine learning model. However, only the cases with the normal condition and slightly pumped-off issue were observed in this study. The testing for the severity of pumped-off cases could be further analyzed if the running time is extended to more than 20 minutes. For this experimental design, the recommended maximum threshold of water rate being pumped is 2000 mm/min. After analyzing the trends, minimum rates of change in intake pressure and water column filled at pump intake with time are -0.0104 and -0.024, respectively, before we start facing pumping off issues in the tests. These threshold criteria are still arbitrary. With longer test running, we anticipate more robust indications of severe pump-off. This will increase the confidence of labeling the pump condition based on the rate of change in suction pressure range over time.

Besides the pumped-off case, more anomalies in rod pumping operations can be studied, including traveling/ stand valve leakage, gas influence, solids handling, and excessive friction. In this study, large amounts of data were generated for the normal condition as the base case before making any altercation. As the liquid utilized in the facility is primarily water, the possibility of introducing a gas line is recommended as an extension for analyzing the impacts of gas-liquid ratio on pump performance.

6.3 Conclusions

This study is an attempt to implement a generic digital design to monitor and diagnose SRP operations. A prototype is trained on the data collected from the experimental design, opening a new window towards remote control and smart automation. The base inputs (SPM, stroke length, displacement shape) could be altered to avoid the imminent failure instead of on-site reactive measures. Some of the keynotes for this study include:

- By validating a physical baseline model, the digital laboratory setup has achieved a practical goal to record large datasets of surface dynocards, recording more than 400000 data points out of 52 experiments. The applied algorithm to transform from surface to downhole dynamometer cards is replicated using the Gibbs equation and verified with errors ranging within 0.7%-1.34%.
- The study underlines the discrepancy in the experimental material and buckling during the rod movement, resulting in card noises during the transition from upstroke to the downstroke. Based on the calculated downhole cards, the digital sucker-rod pump unit simulates a realistic field operation of a rod pump with carrying levels of pump fillage.
- With the utilization of the Internet of Things (IoT) in the mechanical rod pump setup, real-time data are collected through sensors for data analytics solutions. Given the time-driven nature of the pump suction pressure profile and its changes with pump speed, the machine learning algorithm is trained based on multiple experimental inputs. The model predicts pressure, load, and associated features in SRP, monitoring the stabilized operation and detecting anomalies. These models achieve evaluation scores from 0.54 to 0.87, with KNN being the best model. The low error between predicted and true

outputs demonstrates the strong correlation between the machine learning model and actual operation. The testing proves the ability of analytical solutions to detect the slightly pumped-off cards. The comprehensive supervised learning from 4 classification models shows that the overall accuracy achieves up to 92% for the complete dataset of 5-minute and 20-minute testing, with random forest being the optimal classifier. The results from classification approaches improve more confidence in quantifying experimental parameters which determine operational performance in IDSRP. The key findings are that experimental inputs can be used as predictors to monitor the SRP performance and detect anomalies by forecasting pressure and load trends with time.

- The data sets generated by the setup and the analytical solutions developed in this study show the benefits of utilizing the physical SRP operation and IoT, cloud data storage, and processing. In this case, this will reduce the complexity of machine learning algorithms that could be impacted by field data noise. This study mimics the field operations and produces digital solutions that could be broadly applied without costly software processors.

Nomenclature

IDSRP	Interactive digital sucker-rod pump
SRP	Sucker-rod pump
u	Displacement of rod string as a function of position and time
$\frac{\delta u}{\delta x}$	Rod strain, change of rod displacement over its length
c	Damping factor, s^{-1}
g	Gravitational acceleration, ft/s^2
A	Rod string cross section area, in^2
E	Elastic modulus
ρ	Rod density, lbm/ft^3
k	Rod frictional coefficient
M	Number of recorded surface data points
N	Number of finite difference nodes along the rod string
$\{i\}_1^M$	Vector of finite difference nodes along the rod string
$\{j\}_1^N$	Vector of sample points obtained at the surface
$\{g_{PR}\}_1^N$	Discrete function of rod position vs. time
$\{f_{PR}\}_1^N$	Discrete function of rod load vs. time
Δx	Space between 2 finite difference nodes, ft
Δt	Time spacing between consecutive surface sampling points, s
v	Velocity of sound in rod string, ft/s
q	Liquid flow rate delivered by plunger pump. STB/D
B_o	Formation volume factor, rb/STB
E_v	Volumetric efficiency of the plunger
S_p	Effective plunger stroke length, in
A_p	Gross plunder cross-sectional area, in^2
e_r	Rod stretch, in
e_t	Tubing stretch, in
e_o	Plunger over-travel, in
S	Stroke length, in

P_{B_1}	First bottom gauge (B1) pressure, psi
P_{B_2}	Second bottom gauge (B1) pressure, psi
P_T	Top gauge (T) pressure, psi
ΔP_{B1}	Difference between maximum and minimum pressure B1 in each cycle, psi
ΔF	Difference between maximum and minimum rod load in each cycle, N
ΔP	Differential pressure between bottom and top gauge, psi
$\Delta P_{\text{friction}}$	Frictional pressure, psi
(x_p, y_p)	Vector-valued input and output at pth point
W	Weight matrix for multiple-input multiple-output linear regression
$g(W)$	Least Squares or Absolute Deviation
$\hat{y}$	Predicted value of output y
$n_{samples}$	Numbers of sample (x, y)
Var	Variance, square of standard deviation
$d(q, x_c)$	Distance between unknown example q and training sample x_c
k	Numbers of nearest neighborhoods
$i(t)$	Gini index
$C(i\|j)$	Cost of misclassification of a class j case as a class i
$p(i\|t)$	Probability of cases in class i failing into nodes
$p(j\|t)$	Probability of cases in class j failing into nodes
R^2 / R-2	Coefficient of determination
LR	Linear Regression
KNN	K-nearest neighbor
DT	Decision Tree
RMSE	Root mean-square error
MSE	Mean-square error
MAE	Mean absolute error
AUC	Area under Receiver Operating Characteristic curve
AP	Area under precision-recall curve

References

Al-Sidairi, Nada, Mahmoud Osman, Nasir Oritola, Farsi Ahmed, Khalid Zuhaimi, Mohsin Amri, Ghalib Hinai, and Majid Al-Mahrooqi. "Overcoming Artificial Lift Failures in Polymer Flooding Oil Field in the South of Oman." In *SPE EOR Conference at Oil and Gas West Asia*. OnePetro, 2018.

Abdalla, R., El Ela, M.A. and El-Banbi, A., 2020. Identification of downhole conditions in sucker rod pumped wells using deep neural networks and genetic algorithms. *SPE Production & Operations*, *35*(02), pp.435-447.

Altman, Naomi S. "An introduction to kernel and nearest-neighbor nonparametric regression." *The American Statistician* 46, no. 3 (1992): 175-185.

Bello, O., E. P. Dolberg, C. Teodoriu, H. Karami, and D. Devegowdva. "Transformation of academic teaching and research: Development of a highly automated experimental sucker rod pumping unit." Journal of Petroleum Science and Engineering 190 (2020): 107087.

Bezerra, Marco AD, Leizer Schnitman, Manuel de A. Barreto Filho, A. M. Jose, and Felippe de Souza. "Pattern Recognition for Downhole Dynamometer Card in Oil Rod Pump System Using Artificial Neural Networks." In *ICEIS (2)*, pp. 351-355. 2009.

Boguslawski, Bartosz, Matthieu Boujonnier, Loryne Bissuel-Beauvais, Fahd Saghir, and Rajesh D. Sharma. "IIoT edge analytics: Deploying machine learning at the wellhead to identify rod pump failure." In *SPE Middle East Artificial Lift Conference and Exhibition*. OnePetro, 2018.

Breiman, L. "Arcing the Edge Technical Report 486; Statistics Department." University of California: Berkeley, CA, USA (1997).

Breiman, Leo. "Random forests." Machine learning 45, no. 1 (2001): 5-32.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and regression trees*. Routledge, 2017.

Chen, Zengshi, Luther W. White, and Huimin Zhang. "Predicting sucker-rod pumping systems with Fourier series." *SPE Production & Operations* 33, no. 04 (2018): 928-940.

Cunningham, Padraig, and Sarah Jane Delany. "k-Nearest neighbour classifiers: (with Python examples)." *arXiv preprint arXiv:2004.04523* (2020).

Dawsey, S. Downhole Diagnostics. 2020. https://www.downholediagnostic.com/biography

Derek, H. J., J. W. Jennings, and S. M. Morgan. "Sucker rod pumping unit diagnostics using an expert system." In *Permian Basin Oil and Gas Recovery Conference*. OnePetro, 1988.

Dickinson, Roderick R., and James W. Jennings. "Use of pattern-recognition techniques in analyzing downhole dynamometer cards." *SPE Production Engineering 5*, no. 02 (1990): 187-192.

Dove, J., and Z. D. Smith. "Using Sucker Rod Pump Repair Data To Optimize Rod Lift Design." In *SPE North America Artificial Lift Conference and Exhibition*. OnePetro, 2016.

Edgar, Thomas W. "Chapter 3-Starting Your Research, Editor (s): Thomas W. Edgar, David O. Manz, Research Methods for Cyber Security, Syngress, 2017." (2017): 63-92.

Everitt, Thomas Aaron, and J. W. Jennings. "An improved finite-difference calculation of downhole dynamometer cards for sucker-rod pumps." SPE production engineering 7, no. 01 (1992): 121-127.

Ercolino, J (2011) Calculation of Downhole Dynamometer Cards for Sucker-Rod Pumps Finite Difference Wave Equation Diagnostic Analysis for Sucker-Rod Pumps. (Version 1.0.0.0) https://www.mathworks.com/matlabcentral/fileexchange/32342-calculation-of-downhole-dynamometer-cards-for-sucker-rod-pumps

Gao, Qian, Shaobo Sun, and Jianchao Liu. "Working Condition Detection of Suck Rod Pumping System via Extreme Learning Machine." In *2nd International Conference on Civil, Materials and Environmental Sciences*, pp. 434-437. Atlantis Press, 2015.

Gibbs, S.G. and Neely, A.B.: "Computer Diagnosis of Down-Hole Conditions In Sucker Rod Pumping Wells," JPT (Jan. 1966) 91-98; Trans., AIME, 237.

Gibbs, S. G. "Design and diagnosis of deviated rod-pumped wells." In *SPE Annual Technical Conference and Exhibition*. OnePetro, 1991.

Ghareeb, Mohamed, and Anthony Tony Beck. "Design of Sucker Rod Pumping Systems for Effectively Handling Solids and Sand." In *SPE International Production and Operations Conference & Exhibition*. OnePetro, 2012.

Ghareeb, Mohamed, Dave Kennedy, and Alvin Woo. "Intelligent Longest Stroke Reciprocating Pumping System for Production and Cost Optimization." In *SPE Annual Technical Conference and Exhibition*. OnePetro, 2017.

Gray, H. E. "Kinematics of oil-well pumping units." In *Drilling and Production Practice*.OnePetro, 1963.

Guo, Liang, and Xianghui Deng. "Application of improved multiple linear regression method in oilfield output forecasting." In 2009 International conference on information management, innovation management and industrial engineering, vol. 1, pp. 133-136. IEEE, 2009.

Gupta, Ishank, Ngoc Tran, Deepak Devegowda, Vikram Jayaram, Chandra Rai, Carl Sondergeld, and Hamidreza Karami. "Looking Ahead of the Bit Using Surface Drilling and Petrophysical Data: Machine-Learning-Based Real-Time Geosteering in Volve Field." *SPE Journal* 25, no. 02 (2020): 990-1006.

Hirzel, T. Arduino. PWM, https://www.arduino.cc/en/Tutorial/PWM (accessed 12 December 2021).

Hollenbeck, Alfred L. "An alternate approach to high volume, a long stroke pumper." In *SPE Annual Technical Conference and Exhibition*. OnePetro, 1980.

Jackson, M. S., D. A. Howell, J. R. Bailey, S. Rajagopalan, A. Ozekcin, G. Inglish, and C. Allen. "Wear Resistant, Friction Reducing Coatings Reduce Tubing Wear in Sucker Rod Couplings Application: Lab Testing and Field Trial Results." In *SPE Annual Technical Conference and Exhibition*. OnePetro, 2018.

Kollmorgen. Kollmorgen. EC2 Series, https://www.kollmorgen.com/en-us/products/linear actuators/electric-cylinders/ec2-series/ (accessed 12 December 2021)

Kennedy, Dave M., and M. Ghareeb. "Ultra-Long Stroke and Intelligent Rod Pumping System for Producing Difficult Wells and/or Fluids." In *SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition*. OnePetro, 2017.

Kendrick, John F., and Paul D. Cornelius. "The Sucker-rod Pump as a Problem in Elasticity." Transactions of the AIME 123, no. 01 (1937): 15-31.

Lam Tran, Ngoc, Ishank Gupta, Deepak Devegowda, Hamidreza Karami, Chandra Rai, Vikram Jayaram, and Carl Sondergeld. "Machine Learning Workflow to Identify Brittle, Fracable and Producible Rock in Horizontal Wells Using Surface Drilling Data." In *SPE Annual Technical Conference and Exhibition*. OnePetro, 2020.

Lima, F. S., L. A. Guedes, and D. Silva. "Comparison of Border Descriptors and Pattern Recognition Techniques Applied to Detection and Diagnosis of Faults on Sucker-Rod Pumping System." *Digital Image Processing. Rijeka, Croatia: InTechOpen* (2012).

Liu, Feilong, and Anilkumar Patel. "Well failure detection for rod pump artificial lift system through pattern recognition." In *International Petroleum Technology Conference*. OnePetro, 2013.

Liu, Shuping, Cauligi S. Raghavendra, Yintao Liu, K. Yao, Oluwafemi Balogun, Lanre Olabinjo, Ram Soma et al. "Automatic early fault detection for rod pump systems." In *SPE annual technical conference and exhibition*. OnePetro, 2011.

Liu, Yintao, K. Yao, Shuping Liu, Cauligi Srinivasa Raghavendra, Tracy Lynn Lenz, Lanre Olabinjo, B. Seren, Sanaz Seddighrad, and C. G. Dinesh Babu. "Failure prediction for rod pump artificial lift systems." In *SPE Western Regional Meeting*. OnePetro, 2010.

Loh, Wei-Yin. "Classification and regression trees." *Wiley interdisciplinary reviews: data mining and knowledge discovery* 1, no. 1 (2011): 14-23.

Lukasiewicz, S. A., and C. Knight. "On lateral and helical buckling of a rod in a tubing." *Journal of Canadian Petroleum Technology* 45, no. 03 (2006).

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions."

Advances in neural information processing systems 30 (2017).

Martinez, E. R., W. J. Moreno, V. J. Castillo, and J. A. Moreno. "Rod pumping expert system." In *Petroleum Computer Conference*. OnePetro, 1993.

McCaslin, K. P. "A Study of the Methods for Preventing Rodwear Tubing Leaks in Sucker Rod Pumping Wells." In *SPE Production Operations Symposium*. OnePetro, 1987.

Mohammadpoor, Mehdi, Kh Shahbazi, Farshid Torabi, and A. Qazvini. "A new methodology for prediction of bottomhole flowing pressure in vertical multiphase flow in Iranian oil fields using artificial neural networks (ANNs)." In *SPE Latin American and caribbean petroleum engineering conference*. OnePetro, 2010.

Morgan, James N., and John A. Sonquist. "Problems in the analysis of survey data, and a proposal." *Journal of the American statistical association* 58, no. 302 (1963): 415-434.

Mustafar, Izni. "A study on prediction of output in oilfield using multiple linear regression." (2011).

Nazi, G. M., Kaveh Ashenayi, J. F. Lea, and Frank Kemp. "Application of artificial neural network to pump card diagnosis." *SPE Computer Applications* 6, no. 06 (1994): 9-14.

Osman, El-Sayed A., Mohammed Abdalla Ayoub, and Mohamed Ahmed Aggour. "An artificial neural network model for predicting bottomhole flowing pressure in vertical multiphase flow." In *SPE middle east oil and gas show and conference*. OnePetro, 2005.

Palmer, Tyler, and Mark Turland. "Proactive rod pump optimization: Leveraging big data to accelerate and improve operations." In *SPE North America Artificial Lift Conference and Exhibition*. OnePetro, 2016.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." *The Journal of machine Learning research* 12 (2011): 2825-2830.

Pekel, Engin. "Estimation of soil moisture using decision tree regression." *Theoretical and Applied Climatology* 139, no. 3 (2020): 1111-1119.

Pennel, Mike, Jeffrey Hsiung, and V. B. Putcha. "Detecting failures and optimizing performance in artificial lift using machine learning models." In *SPE Western Regional Meeting*. OnePetro, 2018.

Pienknagura Dolberg, Erik. "Implementation of a Digital Sucker Rod Pumping Unit for Research and Educational Purposes." (2019).

Pons, Victoria. "Optimal stress calculations for sucker rod pumping systems." In *SPE Artificial Lift Conference & Exhibition-North America*. OnePetro, 2014.

Rashidi, Fariborz, Ehsan Khamehchi, and Hanieh Rasouli. "Oil field optimization based on gas lift

optimization." *ESCAPE20* 6 (2010).

Razali, Radzuan, and Izni Mustafar. "A study on prediction of output in oilfield using multiple linear regression." *International Journal of Applied Science and Technology (2011)*.

Rogers, J. D., C. G. Guffey, and W. J. B. Oldham. "Artificial neural networks for identification of beam pump dynamometer load cards." In *SPE Annual Technical Conference and Exhibition*. OnePetro, 1990.

Rowlan, O. L., and J. N. McCoy. "Pump Card Reference Load Lines Used for Analysis and Troubleshooting." In *SPE Production and Operations Symposium*. OnePetro, 2015.

Ryan, R. A. "Controlling rod body Corrosion failures with spray-metal plastic-coated sucker rods." In *Permian Basin Oil and Gas Recovery Conference*. OnePetro, 1988.

Sharma, Aditya, Opeyemi Bello, Catalin Teodoriu, and Hamidreza Karami. "Design and Implementation of a Laboratory Sucker Rod Pumping Unit Using Industry 4.0 Concepts." *Journal of Energy and Power Technology* 3, no. 2 (2021): 1-1.

Takacs, Gabor. Sucker-rod pumping handbook: production engineering fundamentals and long stroke rod pumping. Gulf Professional Publishing, 2015.

Teodoriu, C., Bello, O., Karami, H. and Devegowda, D. Automated Data-driven Systems for Predictive Diagnostic in Sucker Rod Pumps. University of Oklahoma, 2020.

Teodoriu, Catalin, and Erik Pienknagura. "Bringing the sucker rod pumping unit into the classroom with the use of the internet of things." In SPE Annual Technical Conference and Exhibition. OnePetro, 2018.

Torlov, Vladislav S., Asif Amin, and Mohammed El Sharaky. "Methodology to Derive Gas Oil Ratio and Other Formation Evaluation Applications of Downhole Ultrasonic Sound Speed Formation Tester Measurements." In Abu Dhabi International Petroleum Exhibition & Conference. OnePetro, 2020.

Tran, Ngoc Lam, Ishank Gupta, Deepak Devegowda, Vikram Jayaram, Hamidreza Karami, Chandra Rai, and Carl H. Sondergeld. "Application of Interpretable Machine-Learning Workflows To Identify Brittle, Fracturable, and Producible Rock in Horizontal Wells Using Surface Drilling Data." *SPE Reservoir Evaluation & Engineering* 23, no. 04 (2020): 1328-1342.

Tripp, H. A. "A review: analyzing beam-pumped wells." *Journal of petroleum technology* 41, no. 05 (1989): 457-458.

Watt, Jeremy, Reza Borhani, and Aggelos K. Katsaggelos. *Machine learning refined: Foundations, algorithms, and applications*. Cambridge University Press, 2020.

Weatherford, 2007. Rotaflex long stroke pumping units. Weatherford Artificial Lift Systems Brochure

Wright, Raymond E. "Logistic regression." (1995).

Yu, Yunhua, Haitao Shi, and Lifei Mi. "Research on feature extraction of indicator card data for sucker-rod pump working condition diagnosis." *Journal of Control Science and Engineering* 2013 (2013).

Yazitova, A., and S. Yagiz. "Estimation of drilling rate of diamond driller using rock properties and driller specifications." In *ISRM International Symposium-EUROCK 2020*. OnePetro, 2020.

Xu, Peng, Shijin Xu, and Hongwei Yin. "Application of self-organizing competitive neural network in fault diagnosis of suck rod pumping system." *Journal of Petroleum Science and Engineering* 58, no. 1-2 (2007): 43-48.

Zhao, Ruidong, Jinya Li, Zhen Tao, Meng Liu, Junfeng Shi, Chunming Xiong, Hongxing Huang, Chengyan Sun, Yufeng Zhang, and Xiaowen Zhang. "Research and Application of Rod/Tubing Wearing Prediction and Anti-Wear Method in Sucker Rod Pumping Wells." In *SPE Middle East Oil and Gas Show and Conference*. OnePetro, 2019.

Appendix

Table A.1 Slope values of ΔP_{B1} and ΔF over time for 20-minute testing

	SPM	RPM	shape	total amount of water pumped per min (mm/min)	Slope ΔP_{B1} over time	Slope ΔF over time
250 TPZ	4.0	500	trapezoid	988	0.001701	0.003928
	7.8	1000	trapezoid	1938	-0.01294	-0.0299
	11.5	1500	trapezoid	2874	-0.0289	-0.06673
	14.3	2000	trapezoid	3571	-0.0203	-0.04689
150 TPZ	6.6	500	trapezoid	993	-0.00156	-0.0036
	12.7	1000	trapezoid	1899	-0.00317	-0.00732
	18.2	1500	trapezoid	2727	-0.00712	-0.01644
	22.7	2000	trapezoid	3409	-0.01042	-0.02407
200 TPZ	5.0	500	trapezoid	995	-0.00917	-0.02118
	9.6	1000	trapezoid	1918	-0.00317	-0.00732
	13.9	1500	trapezoid	2778	-0.00601	-0.01387
	17.5	2000	trapezoid	3509	-0.00183	-0.00421
250 SIN	2.0	500	sinusoidal	499	-0.01007	-0.02326
	4.0	1000	sinusoidal	1003	-0.00174	-0.00403
	6.0	1500	sinusoidal	1497	-0.00109	-0.00252
	8.0	2000	sinusoidal	2000	-0.00212	-0.0049
150 SIN	3.0	500	sinusoidal	453	-0.00281	-0.01837
	6.6	1000	sinusoidal	997	-0.0017	-0.00392
	10.0	1500	sinusoidal	1500	-0.00405	-0.00936
	13.3	2000	sinusoidal	2000	-0.0037	-0.00854
200 SIN	2.5	500	sinusoidal	501	-0.00796	-0.00649
	5.0	1000	sinusoidal	1000	-0.00636	-0.0147
	7.6	1500	sinusoidal	1527	-0.00328	-0.00758
	9.9	2000	sinusoidal	1980	-0.00115	-0.00265