Application of Time Series Analytics to Assess Performance of Artificial Lift Systems Deployed in Coal Seam Gas (CSG) Wells

By:

Fahd Saghir

B. Eng. (Hons)

A thesis submitted for the degree of Doctor of Philosophy (PhD.)

Supervisors:

Maria Gonzalez Perdomo

Peter Behrenbruch

School of Chemical Engineering. Discipline of Petroleum Engineering.

Faculty of Sciences, Engineering and Technology (SET)

The University of Adelaide, Australia

October 2023

Abstract

Artificial Lift Systems (ALS) play a crucial role in producing natural gas from Coal Seam Gas (CSG) wells in Australia. These systems are employed in over five thousand wells located in the Bowen and Surat Basins of Queensland. Operators face significant challenges in managing and maintaining ALS-supported production due to regular failures caused by factors like coal fines. Failure of ALS can have a detrimental impact on meeting both local and international gas export commitments; hence, effective management and maintenance of ALS-supported production are paramount.

The thesis highlights the importance of utilizing real-time data and time series analytics to evaluate ALS performance. Real-time data can help manage CSG wells with artificial lift proactively and with greater insight. Petroleum and well surveillance engineers' expertise is combined to enhance the analysis of time series data.

The research presents an innovative approach that involves transforming time series data into images through Symbolic Aggregate Approximation (SAX). SAX serves as a feature extraction technique that converts time series data into a symbolic representation, which is then translated into performance heatmap images. Petroleum and well surveillance engineers label these SAX-generated performance heatmap images with expert precision. By incorporating domain-specific insights and utilizing novel time series analytics techniques, operators can detect abnormal ALS behavior, proactively address performance issues, and improve overall production efficiency.

This research enabled the creation of a tailor-made ALS analytics application that helps monitor an extensive network of CSG wells, detect abnormal ALS behavior early, and provide insights for proactively managing performance issues, thereby imparting a significant economic impact on CSG operations in Australia.

Table of Contents

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signature:

Date:

30th November, 2023

Dedication

This thesis is dedicated to my family.

To my dear father, Saghir. Thank you for being a pillar of humility and constant inspiration throughout my journey.

To my dear mother, Azra. Thank you for your unwavering prayers and faith that have guided me on my journey. Your love has been a constant source of strength and inspiration in my life.

To my beloved wife, Ayesha. Thank you for being my rock and my anchor. Your amazing support has been my source of motivation and encouragement, and I cannot thank you enough for all that you do. Your love and commitment have constantly reminded me why I am so blessed to have you in my life.

And to my precious kids, Mariam and Omar. Thank you for your unconditional love and for filling my life with joy and laughter. Your presence in my life is a constant reminder of the beauty and wonder of the world. You are a true blessing, and I am so grateful for the love and light you bring into my life.

Acknowledgement

This research work would not have been possible without the guidance of my two supervisors **Mrs. Mary Gonzalez** and **Mr. Peter Behrenburch**.

I am grateful to Mary for improving the quality of my research. She provided me with valuable advice on various aspects, such as the approach to be taken, the right journal for publication, methods to complete my thesis, and revisions required for my manuscript. Her guidance has truly been instrumental in shaping the outcome of my research project.

Moreover, Peter's industry experience was instrumental in guiding me through the early stages of my research. His advice on software development and how to incorporate an industry-driven approach played an instrumental part in guiding my research methodology.

I want to express my gratitude to both supervisors for enabling this research work.

List of Publications

Published Journal Papers

- Saghir, F., Perdomo, M. G., & Behrenbruch, P. (2020). Application of machine learning methods to assess progressive cavity pumps (PCPs) performance in coal seam gas (CSG) wells. The APPEA Journal, 60(1), 197-214.
- Saghir, F., Gonzalez Perdomo, M. E., & Behrenbruch, P. (2023). Application of streaming analytics for Artificial Lift systems: a human-in-the-loop approach for analyzing clustered time-series data from progressive cavity pumps. Neural Computing and Applications, 35(2), 1247-1277.
- Saghir, F., Perdomo, M. G., & Behrenbruch, P. (2023). Performance analysis of artificial lift systems deployed in natural gas wells: A time-series analytics approach. Geoenergy Science and Engineering, 230, 212238.

Published Conference Papers

- Saghir, F., Gonzalez Perdomo, M. E., & Behrenbruch, P. (2019, October). Application of exploratory data analytics EDA in coal seam gas wells with progressive cavity pumps PCPs. In SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition. SPE.
- Saghir, F., Gonzalez Perdomo, M. E., & Behrenbruch, P. (2019, September). Converting time series data into images: an innovative approach to detect abnormal behavior of progressive cavity pumps deployed in coal seam gas wells. In SPE Annual Technical Conference and Exhibition (p. D021S020R006). SPE.
- Saghir, F., Gonzalez Perdomo, M. E., & Behrenbruch, P. (2019, November). Machine Learning for Progressive Cavity Pump Performance Analysis: A Coal Seam Gas Case Study. In SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology Conference (p. D021S013R003). URTEC.

List of Figures

List of Tables

Nomenclature

AE	Auto Encoders
ALS	Artificial Lift Systems
ALSAA	Artificial Lift Systems Analytics Application
CAE	Convolutional Auto Encoder
CSG	Coal Seam Gas
EDA	Exploratory Data Analytics
ESP	Electric Submersible Pump
ESPCP	Electric Submersible Progressive Cavity Pump
HDBSCAN	Hierarchical Density-Based Spatial Clustering
LSTM	Long Short-Term Memory
ML	Machine Learning
PCA	Principal Component Analysis
PCP	Progressive Cavity Pump
RNN	Recurrent Neural Network
RTU	Remote Telemetry Unit
SAX	Symbolic Aggregation Approximation
SCADA	Supervisory Control and Data Acquisition
t-SNE	t-Distributed Stochastic Neighbor Embedding

1. Contextual Statement

1.1. Research Rationale and Background

This thesis is motivated by the confluence of two critical challenges encountered in the domain of ALS within CSG wells. First, the sheer volume of ALS-equipped wells, totalling approximately eight thousand five-hundred (Figure 1 and Figure 2), presents a formidable challenge in terms of real-time monitoring and performance analysis. Second, there is a notable absence of research dedicated to effectively utilizing time series data to enhance ALS performance, further compounded by issues related to data labelling. These challenges collectively underscore the need for an innovative approach that can bridge these gaps and pave the way for improved ALS management in CSG wells.

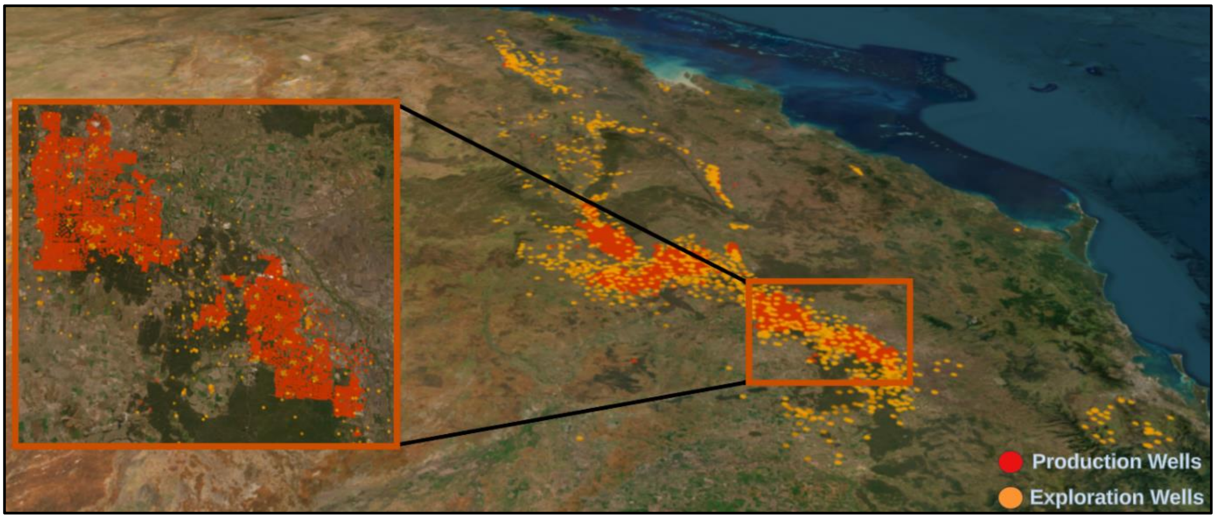


Figure 1 – An overview of Queensland's CSG wells (Production and Exploration). The map depicts the high density of wells across the Surat and Bowen basins.

To address the aforementioned challenges, this research adopts a novel approach. The primary focus is transforming time series data into images using Symbolic Aggregate Approximation (SAX) as a feature extraction technique. This innovative method streamlines the process of generating interpretable images, enabling easy labelling by experienced petroleum and well surveillance engineers. Their expertise in labelling these images unlocks valuable insights into the underlying patterns, anomalies, and trends inherent in the time series data. Moreover, the participation of petroleum and well surveillance engineers in labelling the images provides valuable insights and helps handle unlabelled data, making the analytical approach more robust. The analysis of the labelled images not only improves the understanding of ALS performance but also helps to optimize it. This optimization leads to better failure mitigation, which ultimately facilitates improved gas production efficiency.

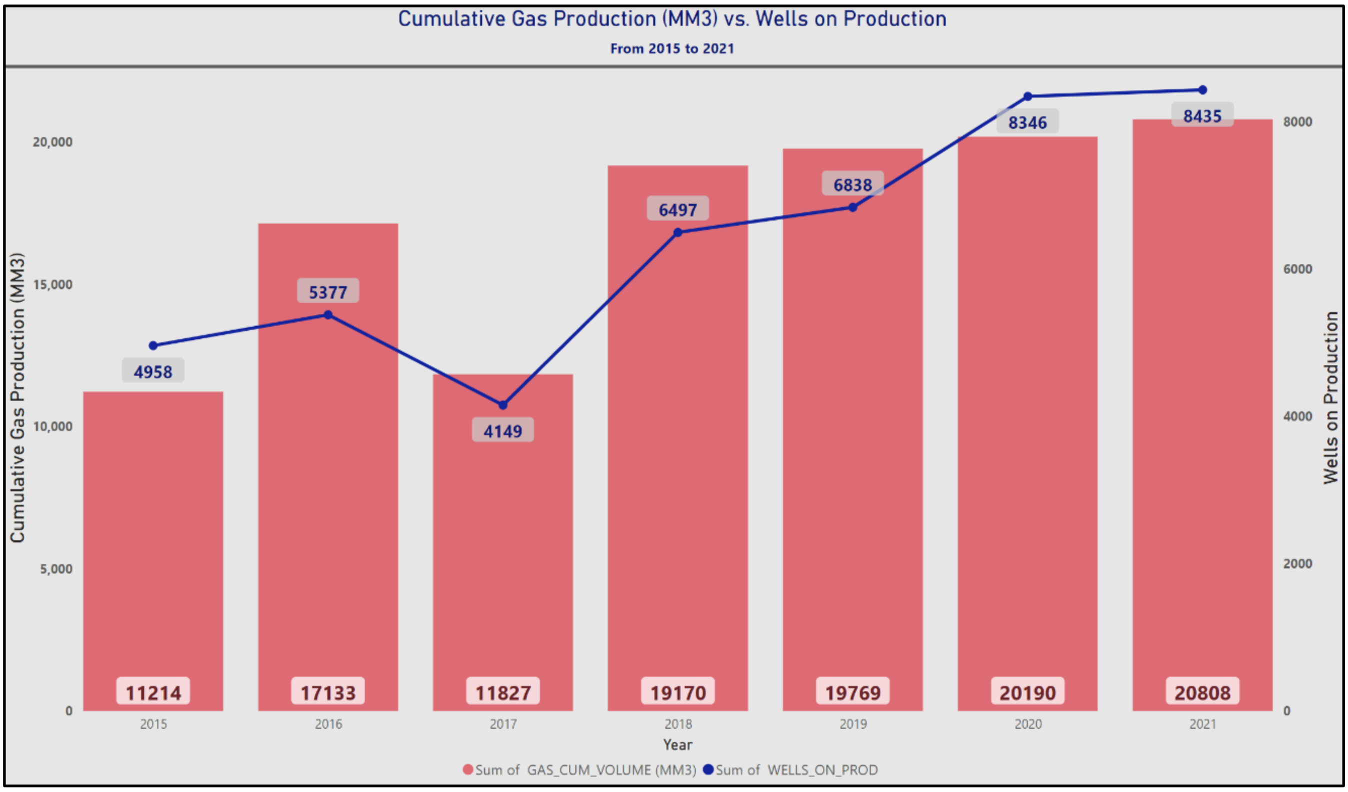


Figure 2 – Number of CSG wells and Cumulative Gas Production from 2015-2021 in Queensland.

This SAX based time-series image conversion represents a pivotal advancement, extracting actionable insights from time series data while harnessing the knowledge of domain experts. The ultimate objective is to elevate the overall performance and management of ALS in the unique context of CSG wells, addressing the challenges posed by a multitude of operational ALS wells.

1.2. Research Objectives

The research objective is to establish a comprehensive time series analytics methodology designed to assess the performance of ALS in near real-time and support well-informed decision-making in CSG production. The methodology detailed in this research adopts an innovative approach, simplifying the interpretation of complex multivariate time series data by converting it into SAX-derived images. This transformation into images streamlines the process for petroleum and well surveillance engineers, allowing them to easily label events of interest, which can serve as early indicators of actionable events, providing operators with opportunities to implement corrective measures that can effectively mitigate failures or enhance the performance of the Artificial Lift Systems.

The research objectives are outlined as follows:

1. Develop an efficient and near real-time methodology utilizing the SAX technique to convert raw time series data from ALS into easily interpretable images.
2. Thoroughly test and validate the proposed time series conversion method using historical ALS data to ensure its accuracy and reliability in capturing relevant features and patterns.
3. Create a user-friendly software tool, with an intuitive interface to empower production and well-surveillance engineers to annotate SAX-derived images, allowing them to label and offer valuable insights into the behavior and performance of ALS.
4. Validate the effectiveness of the annotated SAX-derived images through comprehensive testing on historical datasets, ensuring the accuracy and consistency of the labelling process.
5. Develop a comprehensive ALS analytics platform that integrates the converted time series data and annotated images, enabling near real-time monitoring of well operations. The platform should identify events of interest and provide actionable insights to optimize ALS performance and enhance operational efficiency.

The research objectives described above aim to showcase the significance of SAX derived heatmap images in enhancing time series analytics of coal seam gas (CSG) wells. The primary goal of this study is to illustrate how implementing a novel approach to time series analysis can be beneficial in real-world situations for monitoring and managing a large number of CSG wells. This research will provide a comprehensive understanding of the importance of SAX-derived heatmap images in analyzing CSG wells and how it can help in managing the wells more efficiently. Additionally, the study will demonstrate the practical application of the innovative approach to time series analysis and its potential to improve the performance of CSG wells in the long run.

1.3. Thesis Structure

This is a PhD thesis by publication.

The thesis consists of three (3) principal sections: literature review (Chapter 2), development of novel time series analytics method (Chapter 3, 4 & 5) and real-time analytics tool development (Chapter 6, 7 & 8). Chapters 3 through 8 comprise published papers that address the research gap.

The literature review section provides a comprehensive overview of how ALS is currently used in CSG wells and how real-time data is collected. Various machine learning methods currently used to detect anomalies in time series data are also delved into. The explanation includes how these methods work, their strengths and limitations. Additionally, significant gaps identified in these methods by recent studies are highlighted, and potential solutions are discussed to address these gaps. Finally, SAX, a mathematical technique to convert time series data into performance heatmap images, is delved into. The explanation includes how this technique works, its advantages over other machine learning methods, and how it helps to identify anomalies in real-time data more accurately and efficiently.

In the second section of the thesis, which comprises Chapters 3, 4, and 5, the methods employed to develop a novel approach to time series performance analytics are comprehensively discussed. The discussion begins by detailing the initial exploratory data analysis work, which involved examining the data to identify the key trends and patterns. The SAX technique used to extract useful features from the data, enabling the generation of time series performance heatmaps, is then described. This section also covers the work with PCPs and the leveraging of machine learning methods to cluster the time series heatmap images created through the novel approach. The discussion provides a detailed account of the clustering process used, including the algorithms employed, the parameters set, and the results obtained.

The third section of this thesis, spanning Chapters 6, 7, and 8, presents a comprehensive analysis of the development process for a time series analytics tool. This tool comprises various sub-components designed to cater to the specific needs of experts, particularly Petroleum and Well Surveillance Engineers. It explains how the tool effectively incorporates feedback from these professionals, utilizing their insights for event and sequence labelling.

These labelled events and sequences subsequently serve as the basis for generating real-time alerts, facilitating management-by-exception of multiple CSG wells.

The third section also delves into the discussion of two additional artificial lift methods. Firstly, the functioning of ESPCPs is expounded upon, a technology employed for fluid and gas extraction from wells, while underscoring their advantages compared to other artificial lift methods. Secondly, ESPs, another artificial lift method used in CSG production, are explored in detail. Furthermore, a comprehensive overview is provided of how additional multivariate parameters from these ALS are harnessed to create time series performance heatmap images. These images are pivotal in delivering precise results for real-time performance analysis.

1.4. Chapter Overview

The thesis is built upon six (6) papers published in highly ranked peer-reviewed journals and distinguished conferences, as indicated in Table 1. In its entirety, the thesis consists of nine (9) chapters, and the details of each are as follows:

Chapter 1: Introduces the research rationale, background, and objectives, discusses the structure of the dissertation, and outlines the relationship and contribution of the papers to the thesis.

Chapter 2: Presents a detailed literature review surrounding the works of this thesis, including discussions of the theoretical background, application and the state of anomaly detection methods in time series analytics.

Chapter 3: This section delves into the initial exploratory data analytics conducted on historical time series data collected from 42 wells.

Chapter 4 marks the initial introduction of the innovative approach, wherein time series data is transformed into performance heatmap images.

Chapter 5: Presents how clustering ALS time series data can help with labelling anomalous events to understand Progressive Cavity Pump performance in real-time.

Chapter 6: This chapter provides a comprehensive examination of how Machine Learning methods can be applied to performance heatmap images and how the outcomes from the ML models can be utilized for monitoring the performance of PCPs.

Chapter 7: Presents a method of clustering time series data based on performance heatmap images and showcases a data annotation tool to identify abnormal PCP performance.

Chapter 8: This section introduces the full Artificial Lift Analysis Tool and demonstrates its utilization by experts in the CSG industry to acquire valuable insights and mitigate ALS performance issues.

Chapter 9: This chapter summarizes the research conducted and offers recommendations for future work.

Table 1: Published Papers Status

Chapter	Paper Title	Status
Chapter 3	Application of Exploratory Data Analytics EDA in Coal Seam Gas Wells with Progressive Cavity Pumps PCPs	Published
Chapter 4	Converting Time Series Data into Images: An Innovative Approach to Detect Abnormal Behavior of Progressive Cavity Pumps Deployed in Coal Seam Gas Wells	Published
Chapter 5	Machine Learning for Progressive Cavity Pump Performance Analysis: A Coal Seam Gas Case Study	Published
Chapter 6	Application of machine learning methods to assess progressive cavity pumps (PCPs) performance in coal seam gas (CSG) wells	Published
Chapter 7	Application of streaming analytics for Artificial Lift systems: a human-in-the-loop approach for analyzing clustered time-series data from progressive cavity pumps	Published
Chapter 8	Performance analysis of artificial lift systems deployed in natural gas wells: A time-series analytics approach	Published

1.5. Addressing Research Gap through Published Papers

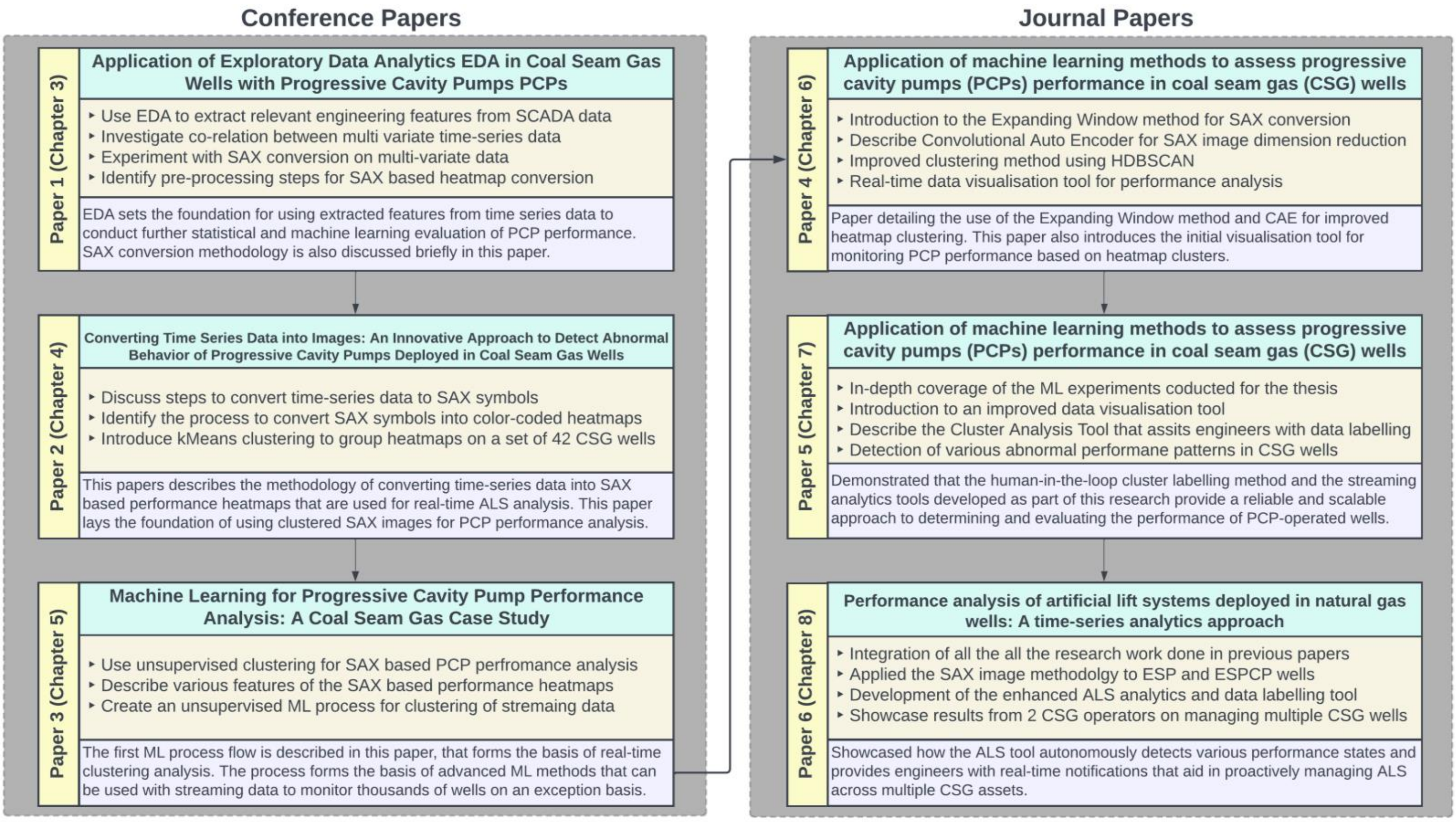
The published papers shown in Table 1 represent an extensive exploration of real-time performance assessment in the realm of ALS within the specific context of CSG operations. The combination of these six papers addresses the previously mentioned gap in this thesis, specifically the ability to monitor large numbers of wells through exception and the application of time-series analytics for ALS deployed in CSG wells. Chapter 3 lays the groundwork with an in-depth exploration of historical time series data from 42 wells, using exploratory data analytics. Chapter 4 introduces the novel approach that revolves around the innovative transformation of traditional time series data into performance heatmap images. These visual representations open new avenues for understanding and evaluating ALS operations. Progressing to Chapter 5, the focus shifts towards the crucial role of ML, specifically unsupervised clustering, in handling ALS time series data. The application of clustering techniques facilitates labelling anomalous events, contributing to a more comprehensive understanding of PCP performance dynamics.

Chapter 6 demonstrates the practical application of ML methods to utilize performance heatmap images for time series analytics using data from 359 wells. This chapter introduces three innovative concepts that significantly enhance the research methodology. First, it presents the expanding window technique for time series data, allowing for the comprehensive assessment of PCP performance from the inception of operations. This approach offers valuable insights into the entire lifecycle of PCPs. Second, autoencoders are employed to reduce the dimensionality of performance heatmap images effectively. This is a pivotal step aimed at alleviating the computational burden that stems from processing vast quantities of data gathered from 359 wells, without compromising the quality and accuracy of the analysis. Finally, Hierarchical Density-Based Spatial Clustering (HDBSCAN) is introduced, offering superior clustering of PCP performance. One advantage of HDBSCAN is that it eliminates the need to predefine the number of clusters, making it ideal for situations where the number of clusters is not known in advance. Also, in Chapter 6, the foundation for real-time PCP performance monitoring through visual analytics tools is established.

Chapter 7 addresses two essential tasks of this research. Firstly, it meticulously elaborates on the machine learning methods introduced in Chapter 6, providing a step-by-step guide on the clustering and labelling procedures. It demonstrates the practical implementation of these techniques and their effectiveness. Secondly, Chapter 7 introduces a performance analysis tool. This tool is utilized by petroleum and surveillance engineers to label the clustered performance heatmaps effectively. The chapter underscores how the grouping of performance heatmap clusters enables surveillance and production engineers to discern abnormal patterns in PCP performance. It provides a holistic view of the complete streaming analytics approach, showcasing how this methodology equips engineers with the necessary tools to monitor wells by exception and maintain optimal ALS performance.

To conclude the research work, Chapter 8 unveils the full Artificial Lift System Analytics Application (ALSAA) — a culmination of the previous chapters' work. The analytics platform is meticulously showcased in extensive detail within practical scenarios in the CSG industry. The research provides real-world results and insights obtained from two operators in Queensland. These practical demonstrations underline the platform's applicability and effectiveness in monitoring and optimizing ALS in the field. This chapter also showcases how the SAX performance heatmap images can be applied to ESPCPs and ESPs. It serves as a practical solution for experts and engineers to harness the insights gained from performance heatmap images and mitigate ALS performance issues in real-time. This paper serves as a culmination of the research presented in the preceding chapters and illustrates how the identified research gap is effectively addressed through real-world adaptation in the CSG industry.

Figure 3 provides a flow chart representation of the published papers and how the outcomes of each paper contribute to addressing research gaps.



Conference Papers

Paper 1 (Chapter 3)

Application of Exploratory Data Analytics EDA in Coal Seam Gas Wells with Progressive Cavity Pumps PCPs

- Use EDA to extract relevant engineering features from SCADA data
- Investigate co-relation between multi variate time-series data
- Experiment with SAX conversion on multi-variate data
- Identify pre-processing steps for SAX based heatmap conversion

EDA sets the foundation for using extracted features from time series data to conduct further statistical and machine learning evaluation of PCP performance. SAX conversion methodology is also discussed briefly in this paper.

Paper 2 (Chapter 4)

Converting Time Series Data into Images: An Innovative Approach to Detect Abnormal Behavior of Progressive Cavity Pumps Deployed in Coal Seam Gas Wells

- Discuss steps to convert time-series data to SAX symbols
- Identify the process to convert SAX symbols into color-coded heatmaps
- Introduce kMeans clustering to group heatmaps on a set of 42 CSG wells

This papers describes the methodology of converting time-series data into SAX based performance heatmaps that are used for real-time ALS analysis. This paper lays the foundation of using clustered SAX images for PCP performance analysis.

Paper 3 (Chapter 5)

Machine Learning for Progressive Cavity Pump Performance Analysis: A Coal Seam Gas Case Study

- Use unsupervised clustering for SAX based PCP perfromance analysis
- Describe various features of the SAX based performance heatmaps
- Create an unsupervised ML process for clustering of stremaing data

The first ML process flow is described in this paper, that forms the basis of real-time clustering analysis. The process forms the basis of advanced ML methods that can be used with streaming data to monitor thousands of wells on an exception basis.

Journal Papers

Paper 4 (Chapter 6)

Application of machine learning methods to assess progressive cavity pumps (PCPs) performance in coal seam gas (CSG) wells

- Introduction to the Expanding Window method for SAX conversion
- Describe Convolutional Auto Encoder for SAX image dimension reduction
- Improved clustering method using HDBSCAN
- Real-time data visualisation tool for performance analysis

Paper detailing the use of the Expanding Window method and CAE for improved heatmap clustering. This paper also introduces the initial visualisation tool for monitoring PCP performance based on heatmap clusters.

Paper 5 (Chapter 7)

Application of machine learning methods to assess progressive cavity pumps (PCPs) performance in coal seam gas (CSG) wells

- In-depth coverage of the ML experiments coducted for the thesis
- Introduction to an improved data visualisation tool
- Describe the Cluster Analysis Tool that assits engineers with data labelling
- Detection of various abnormal performane patterns in CSG wells

Demonstrated that the human-in-the-loop cluster labelling method and the streaming analytics tools developed as part of this research provide a reliable and scalable approach to determining and evaluating the performance of PCP-operated wells.

Paper 6 (Chapter 8)

Performance analysis of artificial lift systems deployed in natural gas wells: A time-series analytics approach

- Integration of all the all the research work done in previous papers
- Applied the SAX image methodolgy to ESP and ESPCP wells
- Development of the enhanced ALS analytics and data labelling tool
- Showcase results from 2 CSG operators on managing multiple CSG wells

Showcased how the ALS tool autonomously detects various performance states and provides engineers with real-time notifications that aid in proactively managing ALS across multiple CSG assets.

Figure 3 – Flow Chart showing the work done for each published paper and how the findings from each paper help in addressing the research gap.

2. Literature Review

2.1. Artificial Lift Systems Used in CSG Wells

During this research, a comprehensive analysis was conducted on the operational mechanisms of three different types of artificial lift systems in CSG wells. The main aim was to develop a deeper understanding of each lift type and explore the potential for operators to optimize pump performance using time series analytics.

a. Progressive Cavity Pumps

PCPs have gained prominence as a reliable artificial lift method in CSG operations [1]. The presence of solids (coal fines) in CSG makes PCPs ideal for dewatering natural gas wells. PCPs utilize a helical rotor-stator configuration, creating a continuous cavity that enables the movement of fluids containing solids. The gentle conveying action and low shear rate within the pump mechanism contribute to its impressive solid handling capability. The stator's elastomeric material and the rotor's precision design further enhance the pump's ability to manage abrasive materials without excessive wear. At present, the majority of ALS used in CSG wells is composed of PCPs. Figure 4 shows the various components of a PCP and how they are deployed in CSG wells.

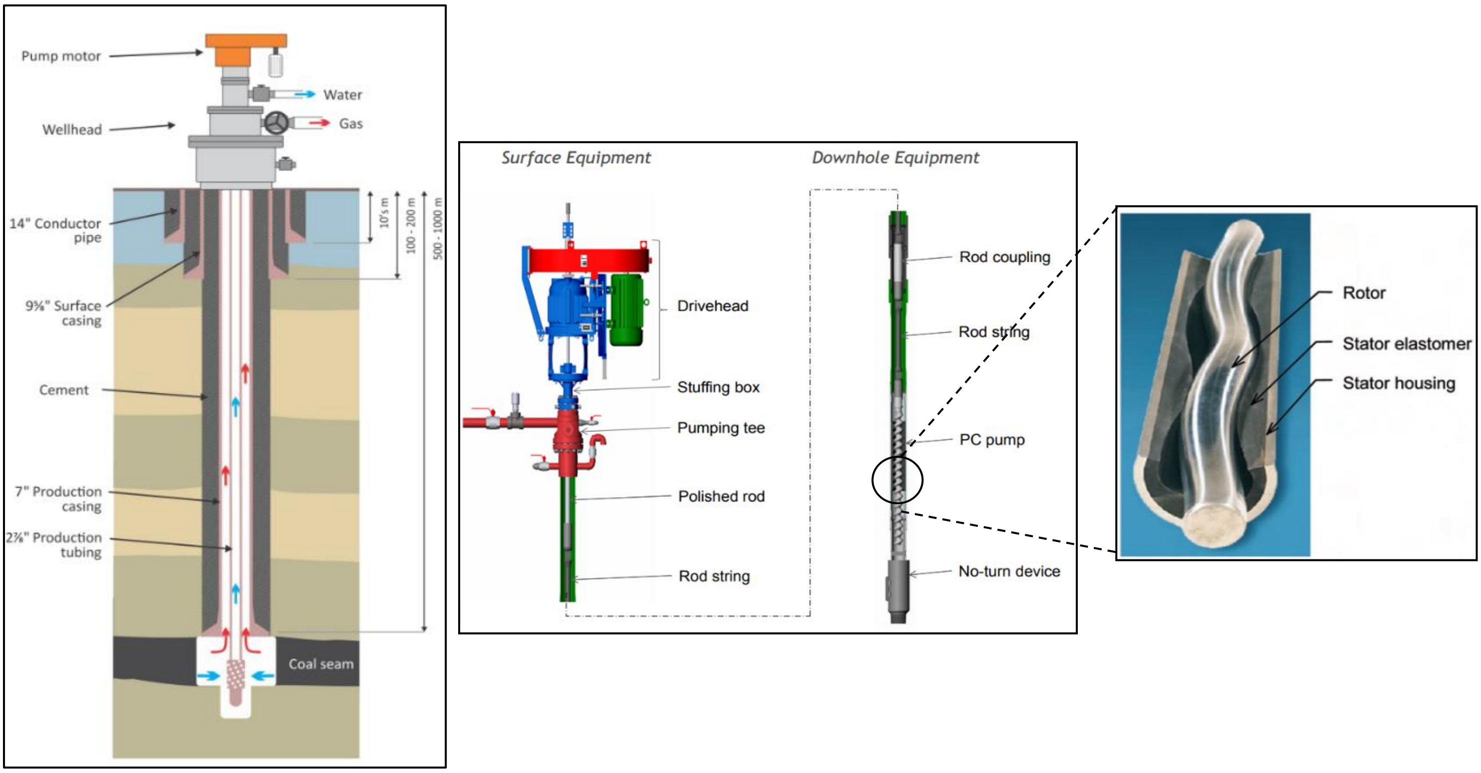


Figure 4 – (Left) Natural Gas Production from Coal Seam Gas (CSG) wells. (Centre) Main Components of a PCP system. (Right) Cut-out view of PCP Rotor and Stator.

b. Electric Submersible Progressive Cavity Pumps

Although PCPs are highly adept at managing solid production, the accumulation of solids from the interburden can become a significant concern. This problem can potentially lead to PCP failure, significantly impeding the operational reliability of the system. Moreover, the formidable torque exerted by the solids-laden fluid exacerbates the vulnerability of rods to mechanical stress, elevating the risk of rod failure. As a result, addressing the interplay between interburden-induced solid buildup, torque imposition, and rod integrity is crucial to sustaining the operational longevity of CSG wells.

Furthermore, natural gas producers operating many wells (typically more than five hundred) have recently looked at lateral well completions. Lateral wells minimize surface footprint by consolidating multiple wells in close proximity. This strategy minimizes land disturbance and complies with regulatory directives to preserve the environment. Moreover, the lateral wells enhance gas recovery by accessing a broader reservoir area, thereby offering improved production rates.

Recently, natural gas producers with a high number of CSG wells have begun utilizing ESPCPs to avoid rod failures and make use of lateral wells [2-5]. One major advantage of these pumps is that they do not require rods to transfer motor energy to the rotor. Figure 5 shows the various components of an ESPCP.

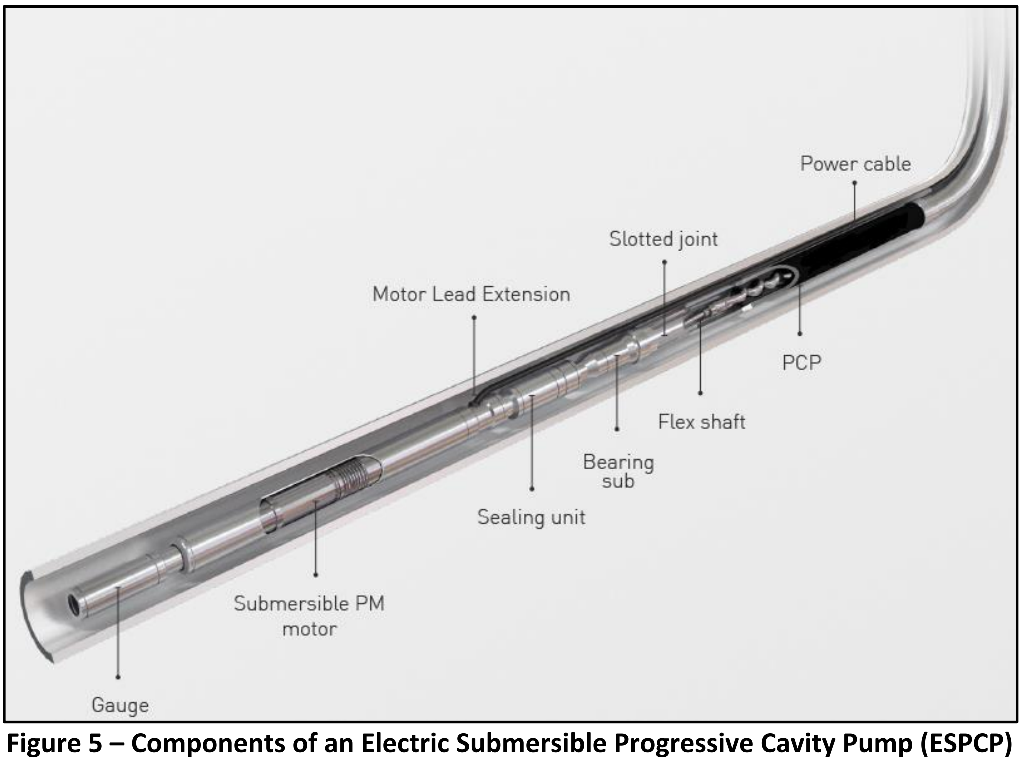


Figure 5 – Components of an Electric Submersible Progressive Cavity Pump (ESPCP)

c. Electric Submersible Pumps

Electric Submersible Pumps (ESPs) are centrifugal pumps used for artificial lift in CSG operations due to their ability to deliver high flow rates. They are chosen as an alternative to ESPCPs where natural gas operators require faster water drawdown rates. However, ESPs do have limitations versus ESPs where they are not designed to manage solid contents in produced fluids. Hence, ESPs are only used in lateral configurations, where the well-completion design allows for minimal solid encroachment into the produced water. The components of an ESP are shown in Figure 6.

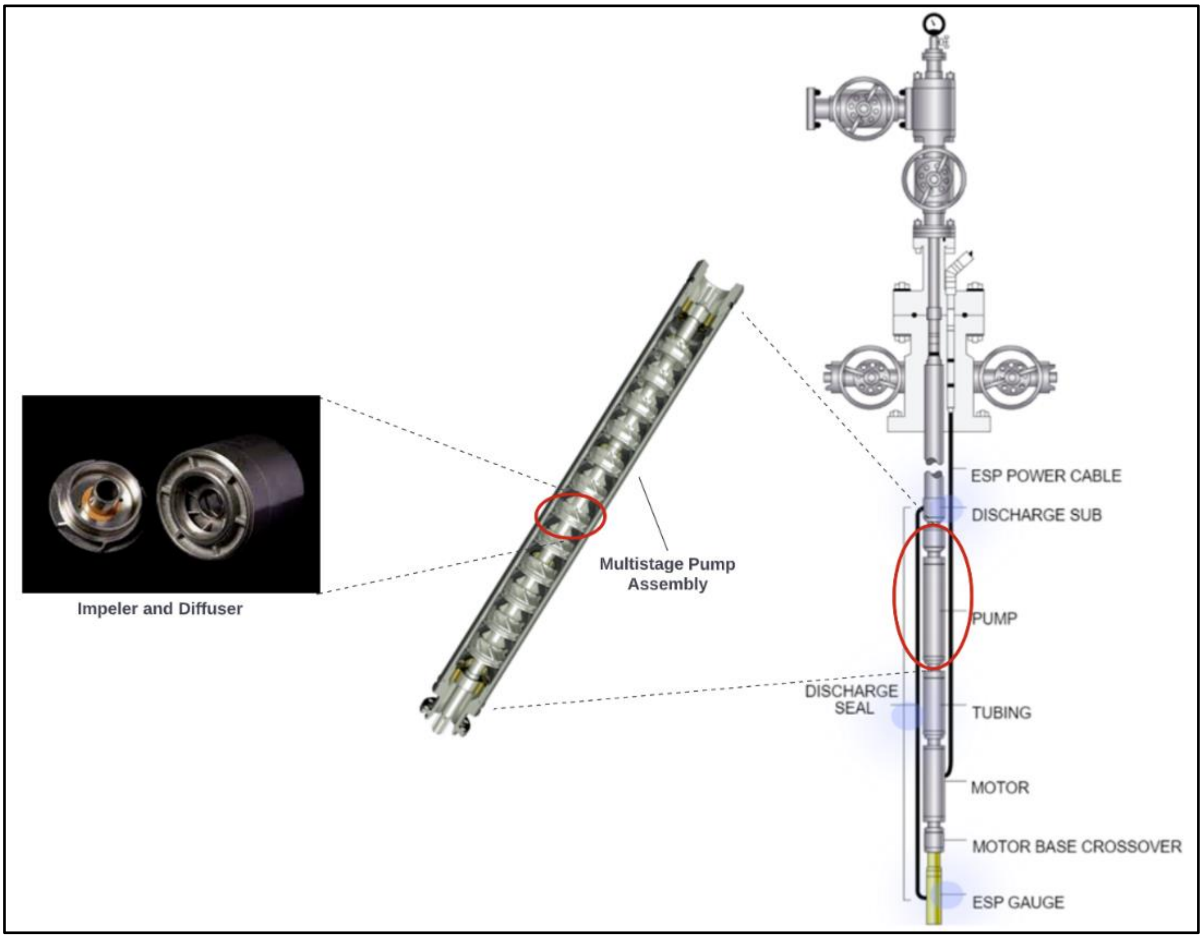


Figure 6 – Components of an Electric Submersible Pump (ESP)

2.2. Automation and Surveillance of Artificial Lift Systems

Supervisory Control and Data Acquisition (SCADA) systems are predominantly used to automate and monitor ALS deployed for oil and gas production [6]. Remote Telemetry Units (RTUs) are typically installed on wellheads, connecting to various sensors and electrical systems. The RTUs are critical to the SCADA system as they collect data from multiple sensors and transmit it to a central control through available communication media. Moreover, operators can also create and deploy logic on the RTUs to autonomously control ALS.

In the late 20th century, SCADA systems were introduced into the oil and gas industry to monitor production systems in real time [7]. However, it was not until the early 21st century that they became widely used. This was largely due to the implementation of digital oilfield programs by both international and national oil companies [8]. These programs helped to monitor and improve production processes, leading to increased efficiency and profitability.

Furthermore, SCADA systems were predominantly utilized for the monitoring of ALS in maturing onshore oilfields, where operators used real-time data to optimize pump speeds for improved hydrocarbon production rates [9]. Figure 7 and Figure 8 show the CSG well layout and SCADA data flow, respectively. This research investigated the process of gathering data from SCADA systems, its storage in corporate historians, and its subsequent analysis for both business and engineering purposes.

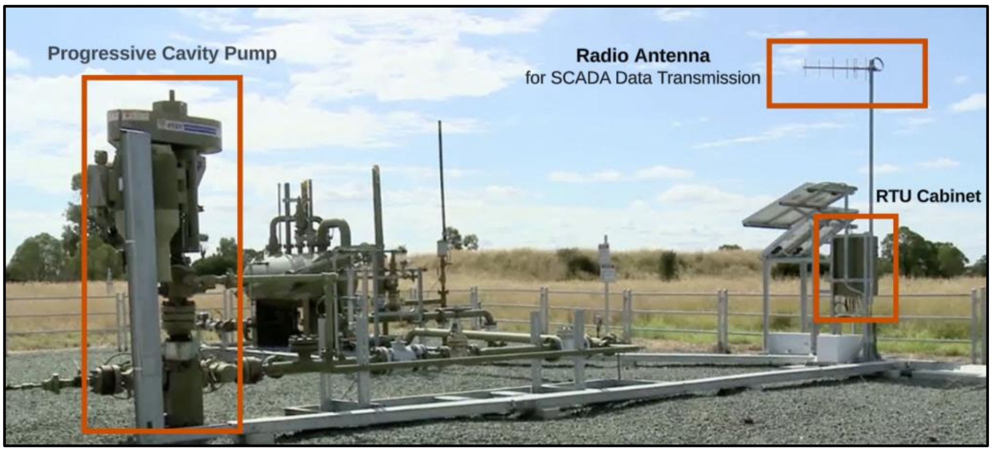


Figure 7 – CSG Well Layout depicting the PCP and RTU and Radio Antenna for data transmission.

A thorough analysis was conducted on how connectivity to SCADA (Supervisory Control and Data Acquisition) systems affects data frequency, a critical factor in time series analytics. This research explored the complexities of varying data frequencies and their impact on time-series analytics. It became evident that the effective management of these varying data frequencies was paramount to the methodology employed.

2.3. Exception-Based Surveillance in Coal Seam Gas Applications

Following a literature review on SCADA systems, an analysis was undertaken to examine how Coal Seam Gas operators, both domestically and internationally, leveraged real-time data to enhance ALS surveillance. Typically, SCADA data is utilized to optimize ALS control speed set points to increase production or prevent unnecessary shutdowns [10-13].

Furthermore, ALS surveillance systems are designed to monitor crucial parameters such as pressure, flow, torque and temperature. Operators are immediately alerted whenever any of these parameters exceed a certain threshold. This real-time information empowers operators to develop exception-based surveillance methods that efficiently monitor artificially lifted wells. By utilizing these methods, operators can streamline their procedures and ensure that any potential issues are identified and addressed in a timely manner[14].

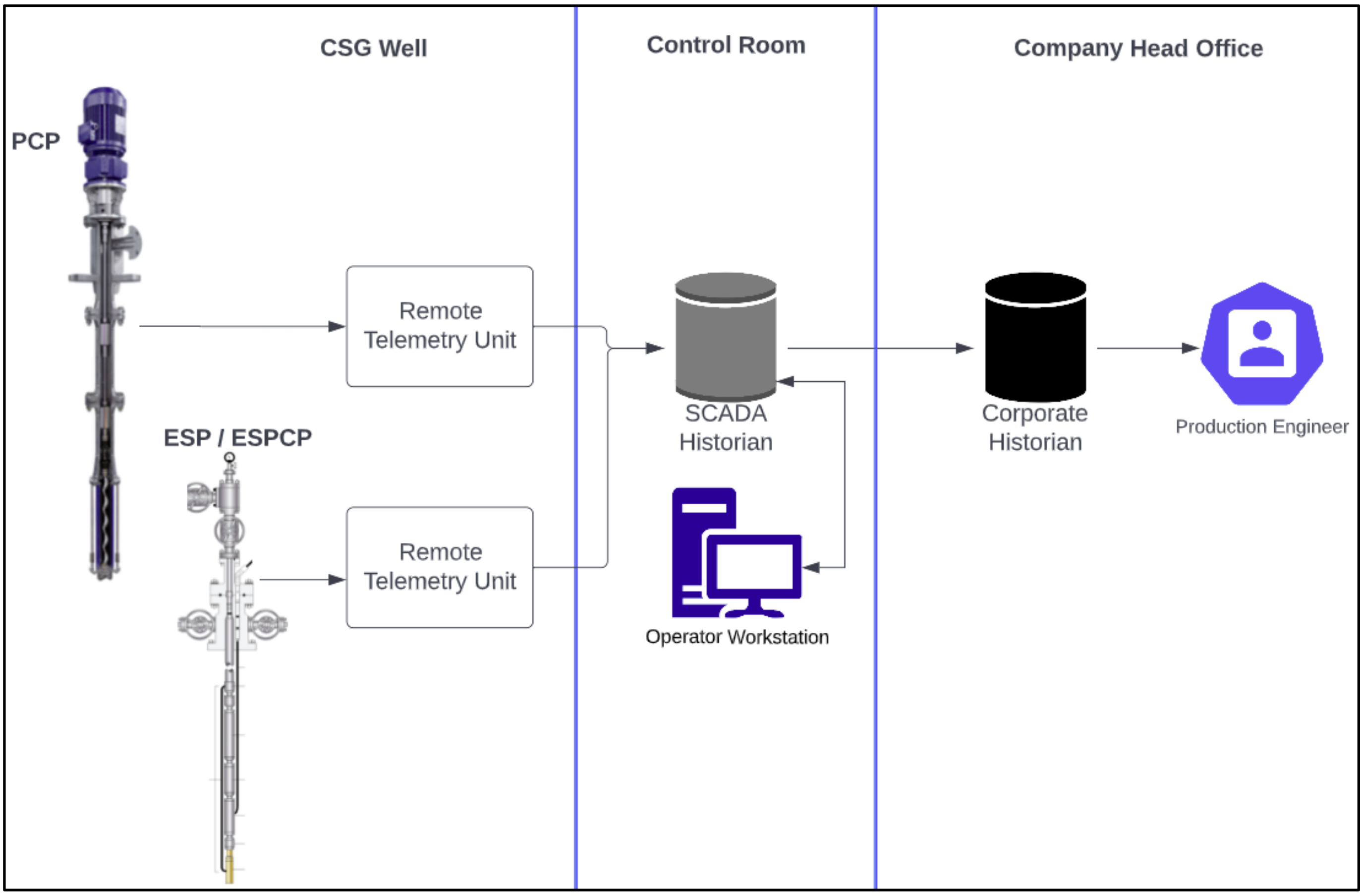


Figure 8 – Data Flow from CSG Wells to Company Head Office

Although exception-based surveillance methods are valuable for CSG operators, they do not provide early warning signs of abnormal ALS performance. Some of these methods also rely on downhole sensors, which are prone to failure during lifting operations [15]. Our analysis of exception-based surveillance revealed that providing context driven alerts for ALS performance would be more effective in managing ALS in coal seam gas wells.

2.4. Time Series Analytics

Time series analytics is a field of study that focuses on analyzing data that changes over time. It involves forecasting future trends and identifying any unusual patterns or outliers, which are known as anomalies. This field covers a broad range of topics, including machine learning algorithms, statistical modelling techniques, and data visualization methods. Extensive research is being conducted in this sought-after field by both academic and industrial sectors to uncover new insights and knowledge from time series data. Time series analysis aims to create models and methods that can accurately predict future trends and identify any anomalies that may arise, enabling businesses and organizations to make more informed decisions and stay ahead of the curve. Research in forecasting and anomaly detection has significantly increased with the rise of real-time data monitoring and the Internet of Things (IoT) in the past decade [16, 17]. In the case of ALS, this research focused on detecting changes in performance parameters and understanding trends in multivariate time series data; hence, the next step is to discuss notable methods used for anomaly detection.

a. Machine Learning Methods for Time Series Anomaly Detection

As SCADA data is often unlabeled, unsupervised machine learning methods are commonly used for analyzing time series data. The mutual theme among time series-based machine learning methods is to detect anomalies through clustering or classification. These methods help label events of interest that aid in improving asset performance monitoring. Following a comprehensive analysis of various machine learning techniques, two primary approaches were identified for time series clustering and anomaly detection: Convolutional Autoencoders and Long Short-Term Memory Autoencoders. These approaches are prominently featured in recent publications, and many other methods draw inspiration from these two approaches.

i. Convolutional Autoencoders (CAE)

Autoencoders, a class of neural networks, have gained prominence in time series anomaly detection due to their ability to capture complex patterns and representations within data. They have proven particularly effective in unsupervised settings where labelled anomaly examples are scarce or unavailable. One key advantage of autoencoders is their ability to learn a compact and informative representation of the input data, which can be exploited for anomaly detection purposes. Figure 9 shows a typical AE architecture.

Autoencoders can learn to compress data and then reconstruct it. This is done by mapping the input data into a lower-dimensional latent space through an encoder network and then attempting to reconstruct the original data from this compressed representation through a decoder network. In the context of time series data, an autoencoder works by taking a sequence of data points as input and then transforming it into a compressed representation using the encoder network. The decoder network then tries to reconstruct the original sequence from this compressed representation.

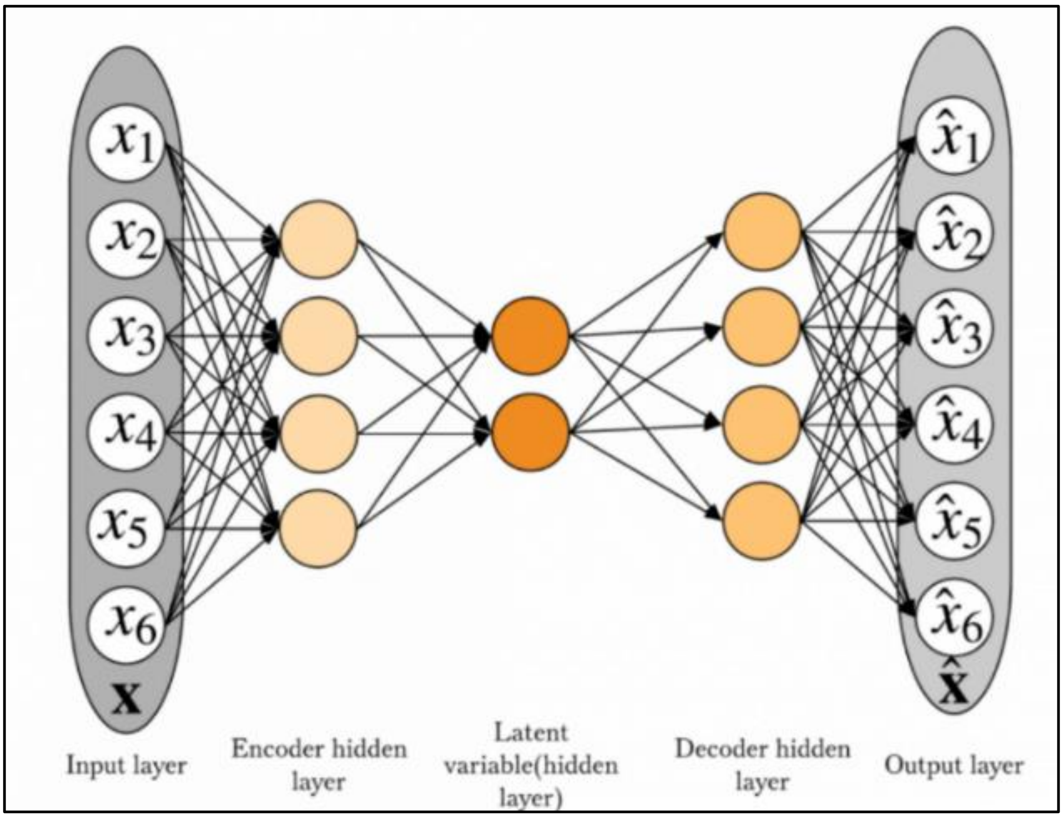


Figure 9 – A typical Autoencoder Neural Network Architecture

Throughout the training phase, the autoencoder acquires the ability to minimize the reconstruction error between the initial and reconstructed data. This process incentivizes the model to capture the most significant features and patterns inherent in the input data. By doing this, the autoencoder can identify the most important characteristics of the input data, which can be useful in a variety of applications such as anomaly detection, dimensionality reduction, and denoising. Figure 10 shows an overview of how time series data is encoded and decoded to produce a reconstructed time series signal.

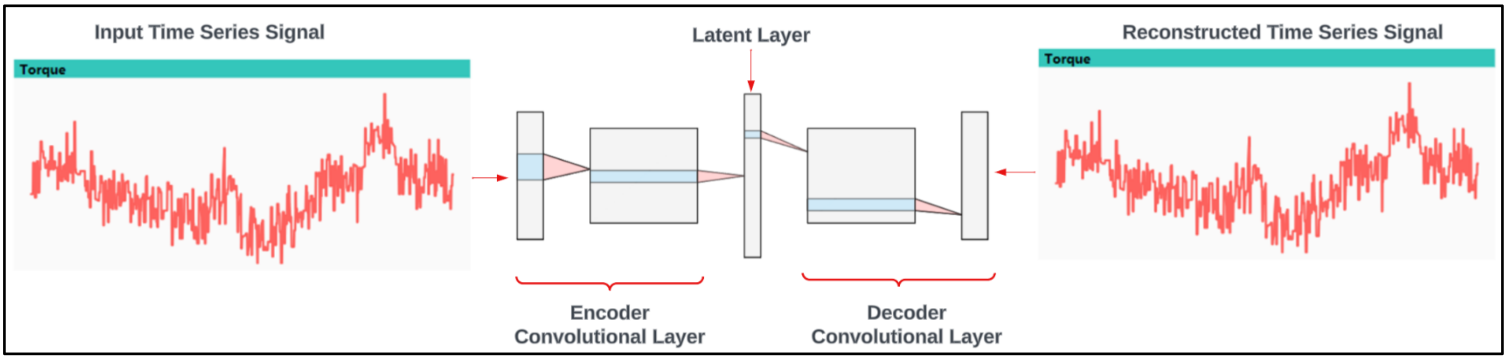

Input Time Series Signal
Torque
Latent Layer
Reconstructed Time Series Signal
Torque
Encoder Convolutional Layer
Decoder Convolutional Layer

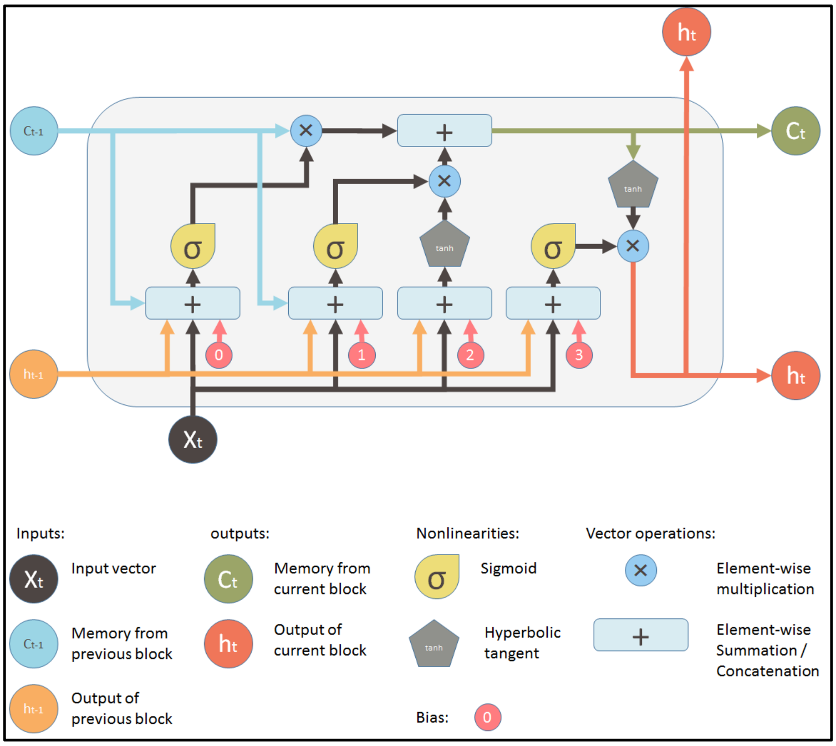
Figure 10 – Conversion of Time Series Signal to a Latent layer via an Encoder, and conversion to reconstructed Time Series signal via a Decoder. The difference between the reconstructed and input signal is used to determine the anomalies in the data.

Autoencoders are crucial in time series anomaly detection as they efficiently learn a representation of temporal data and identify deviations from learned patterns [18]. The autoencoder is trained on a dataset of "normal" time series sequences, capturing the underlying temporal dependencies and patterns. The encoder compresses the time series data into a lower-dimensional latent representation during training, while the decoder attempts to reconstruct the original data from this compressed representation. The primary objective of the autoencoder is to minimize the reconstruction error, ensuring that it learns to accurately encode and decode the normal data sequences.

Once the autoencoder is trained and verified, it can be employed for anomaly detection on streaming time series data. When presented with a new or unseen time series sequence, the model attempts to reconstruct it. If the sequence follows the learned patterns and is considered "normal," the reconstruction error is typically low. However, the reconstruction error tends to be significantly higher when the input sequence contains anomalies or deviations from the learned patterns. By setting an appropriate threshold for the reconstruction error, anomalies can be effectively identified. This mechanism allows autoencoders to excel in detecting various time series anomalies, including point and contextual anomalies, making them valuable tools in monitoring systems for deviations from expected temporal behavior [19]. Autoencoders' ability to capture intricate temporal dependencies, combined with their unsupervised nature, makes them particularly useful in scenarios where labelled anomaly data is scarce or when the nature of anomalies is not well-defined in advance.

ii. Long Short-Term Memory Neural Networks (LSTM) based Auto Encoders

LSTM is a type of recurrent neural network (RNN) specifically designed to capture and model data sequences. This makes it a great tool for analyzing time series data. Unlike traditional statistical methods that rely on predefined patterns and assumptions, LSTM networks can learn complex temporal dependencies from data [20]. This makes them highly adaptable to diverse and dynamic time series. One of the most significant features of LSTMs is the presence of memory cells within LSTM units. These memory cells allow LSTMs to capture and store information over extended time periods, as shown in Figure 11 [21].



h_t
C_{t-1}
C_t
tanh
σ
σ
tanh
σ
0
1
2
3
h_{t-1}
h_t
X_t
Inputs:
outputs:
Nonlinearities:
Vector operations:
X_t Input vector
C_t Memory from current block
σ Sigmoid
× Element-wise multiplication
C_{t-1} Memory from previous block
h_t Output of current block
tanh Hyperbolic tangent
+ Element-wise Summation / Concatenation
h_{t-1} Output of previous block
Bias: 0

Figure 11 – A typical LSTM Cell with various operation functions.

As a result, LSTMs can learn and represent long-range dependencies in sequential data. This feature makes them well-suited for tasks that involve complex temporal patterns. In addition, LSTMs operate sequentially, processing data point by point. This allows them to capture the intricate temporal dynamics of sequences, making them highly effective in recognizing patterns and trends. Furthermore, LSTMs are robust in handling irregularly sampled time series data, missing values, and noisy observations. They can effectively adapt to different time intervals between data points and imputing missing values. The adaptability of LSTMs, combined with their ability to automatically extract relevant features from data, makes them tools for time series forecasting and anomaly detection [22]. Figure 12 shows an overview of how time series data is encoded and decoded using LSTM cells to produce a reconstructed time series signal.

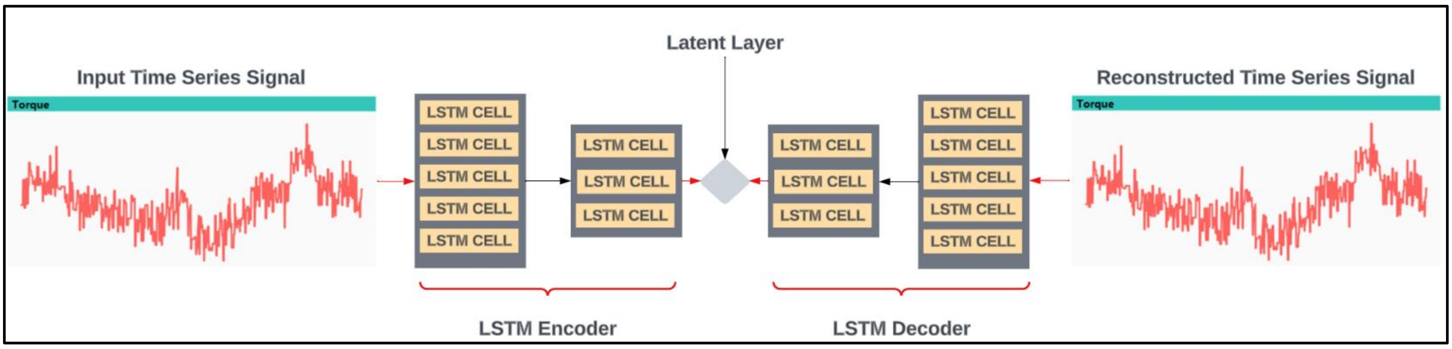


Figure 12 – Conversion of Time Series Signal to a Latent layer via an LSTM Encoder and conversion to reconstructed Time Series signal via an LSTM Decoder. The difference between the reconstructed and input signal determines the anomalies.

LSTM autoencoders are a powerful solution for a wide range of data analysis tasks, including sequence data compression, feature extraction, and anomaly detection [19]. They combine the strengths of autoencoders and LSTM neural networks, resulting in an efficient data representation. These models contain an encoder that compresses the input data into a lower-dimensional latent representation and a decoder that reconstructs the original data from this representation. However, what sets LSTM autoencoders apart is the inclusion of LSTM units in the encoder and decoder components. This addition equips the model with the ability to capture and model intricate temporal dependencies and sequential patterns present in the data, allowing it to maintain the sequential context while learning an efficient representation. LSTM autoencoders are particularly effective when dealing with time series and sequential data, preserving the temporal aspects of the data that standard autoencoders may overlook. They can capture complex temporal dependencies, making them highly adept at preserving the sequential information in data, which is essential for tasks like sequence reconstruction, forecasting, and anomaly detection. They are also capable of feature extraction and dimensionality reduction while maintaining the temporal context, offering a more comprehensive representation of the data.

b. Limitations in Machine Learning Methods for Time Series Data

Despite the abundance of machine learning methods available for analyzing time series data, studies conducted in the past few years have revealed that ML-based methods for multivariate time series data have noteworthy limitations [17, 23-28]. These limitations compromise the accuracy and practical value of insights derived from such methodologies. When applying machine learning techniques to the analysis of time series data, it is crucial to carefully acknowledge and factor in the limitations. This holds true even for well-published methods that use AE and LSTM autoencoder approaches. Several fundamental reasons account for the limitations in these methods, and through an extensive literature review, the most prominent gaps have been identified, as outlined below. There are several underlying reasons for the limitations of these methods, and based on our extensive literature review, we were able to identify the most notable gaps, which are presented below.

i. Handling Missing Data

The presence of missing data can significantly impact machine learning-based time series analysis, introducing challenges that must be carefully addressed. In time series data, missing values often occur due to various reasons, such as sensor malfunctions, data transmission issues, or irregular sampling intervals. This is true for SCADA systems that collect data from artificial lift pumps, as these gaps can disrupt the continuity of the time series, potentially leading to inaccurate anomaly detection and unreliable results.

Handling missing data in time series analysis is essential because it can affect the model's ability to capture temporal dependencies, make accurate predictions, or detect anomalies. Various techniques, including imputation methods, handling strategies, and deep learning models, are employed to mitigate the effects of missing data [29]. However, these methods make anomaly detection from ML methods even more unreliable [30].

One of the primary drawbacks of data imputation is the potential introduction of bias and distortion into the dataset. Imputation methods, whether they involve simple techniques like mean imputation or more advanced methods such as interpolation, introduce values that may not accurately represent the underlying reality of the time series. This can lead to misleading interpretations and incorrect conclusions, particularly when the missing data points are not missing at random, and their absence carries meaningful information or patterns. Moreover, imputing missing data can artificially reduce the variability in the time series, which can affect statistical analyses and lead to inaccurate forecasts or anomaly detection results. Additionally, imputation assumes that the relationships between variables remain constant over time, which may not hold in dynamic and evolving systems, further compromising the integrity of the analysis. Therefore, while imputation is a practical approach to handle missing data, its disadvantages necessitate careful consideration and validation of the imputed values' impact on the overall analysis and decision-making processes in time series analytics.

ii. Managing Multivariate Data

Multivariate data, characterized by multiple variables or features measured over time, can have a profound impact on the performance and applicability of AEs and LSTMs. When applied to multivariate time series data, these models need to grapple with the increased complexity and dimensionality, which can present multiple challenges with neural network architecture design.

When dealing with AEs, managing multivariate data requires encoding and decoding multiple variables simultaneously. This increases computational demands, especially for high-dimensional data. Additionally, processing multiple variables in parallel can limit the capture of interdependencies and correlations among variables. Therefore, it is essential to carefully engineer and preprocess features to ensure that the autoencoder can effectively learn relevant patterns within the multivariate time series. Furthermore, the choice of loss function and evaluation metrics should align with the multivariate nature of the data.

LSTM networks, when applied to multivariate time series, can simultaneously model the temporal dependencies across multiple variables. This ability makes them well-suited for capturing complex interactions and patterns within the data. However, increased dimensionality can lead to model training and interpretation challenges. Proper architecture design and hyperparameter tuning become critical to ensure that the LSTM network effectively captures the relevant temporal dependencies. Additionally, handling missing data, ensuring proper normalization, and dealing with varying scales among different variables are essential preprocessing steps.

iii. Capturing Domain Context

AEs and LSTM autoencoders can struggle to capture domain-specific context, which is often crucial for accurate modelling in specific applications. LSTMs, although skilled at capturing temporal dependencies in sequential data, may not inherently understand the semantics or domain-specific meaning of the data they analyze. They rely solely on patterns and relationships learned from the data, potentially missing out on domain-specific nuances that a human expert might recognize. This can lead to suboptimal performance when dealing with data where contextual understanding is essential, such as medical diagnoses or natural language understanding.

Similarly, AEs, while proficient at feature extraction and data reconstruction, do not inherently possess domain knowledge. They learn to compress and reconstruct data based on statistical patterns without understanding the underlying meaning or context. Therefore, integrating domain-specific knowledge and context into these models often requires additional techniques and human expertise to ensure that the insights derived from these models are meaningful and relevant in the specific domain of interest.

iv. Sampling Window Size

Selecting an appropriate window size is a crucial yet challenging aspect of time series analysis when using machine learning models like AEs and LSTMs. The window size determines the temporal context that the model can consider when making predictions or detecting patterns in the data. However, choosing the right window size is far from a one-size-fits-all task and involves several challenges.

One of the primary challenges is balancing the trade-off between capturing local and global temporal patterns. A smaller window size allows the model to focus on fine-grained, short-term patterns but may overlook longer-term trends or seasonality. Conversely, a larger window size can capture broader trends but may blur or dilute the impact of shorter-term fluctuations. Deciding on the appropriate window size often requires a deep understanding of the specific domain and the underlying temporal dynamics. Furthermore, the choice of window size can impact the model's computational requirements and memory consumption, as larger windows lead to more extensive feature vectors and potentially longer training times. Therefore, users must carefully consider the intended use case and objectives of the analysis to strike the right balance and select an optimal window size for their machine learning models in time series analysis.

Another challenge in choosing the right window size is dealing with irregular or missing data. Time series data often exhibit irregular sampling intervals or missing values, which can complicate the selection of a suitable window size. Irregular data may lead to misalignment between windows and data points, requiring interpolation or data preprocessing to address gaps. Additionally, selecting an inappropriate window size in the presence of missing data can result in either information loss or excessive noise in the analysis. Thus, practitioners must carefully handle data irregularities and consider how the window size interacts with the data's temporal characteristics to ensure meaningful and accurate results in time series analysis using machine learning models.

v. Reconstruction Error Threshold

The reconstruction error threshold presents a challenge in time series analysis for AEs and LSTM autoencoders. This threshold determines when the model flags an observation as an anomaly based on the difference between the original data and its reconstruction. However, setting an appropriate threshold can be highly challenging due to several factors.

The choice of threshold in anomaly detection can be subjective and context-dependent. There is a trade-off between sensitivity (the ability to detect true anomalies) and specificity (the ability to avoid false alarms). If the threshold is set too low, it may result in many false positives, while some anomalies may be missed if it is too high. The optimal threshold can vary for different datasets and use cases, or even over time as the data distribution changes. As a result, practitioners must carefully consider these trade-offs and domain-specific requirements when establishing the reconstruction error threshold.

It is worth noting that the distribution of reconstruction errors can be quite complex and multi-modal. As a result, not all anomalies will necessarily stand out as clear outliers in the reconstruction error distribution. Some anomalies may have subtle deviations that are difficult to distinguish from normal variations. This complexity can make it challenging to define a single, fixed threshold that effectively captures all types of anomalies. It is often necessary to use advanced techniques, such as adaptive or percentile-based thresholds, to handle these complexities. Additionally, the presence of noise or outliers in the training data can affect the reconstruction error distribution, making the threshold selection process even more complicated. Thus, the challenge lies in developing thresholding strategies that can account for the diverse nature of anomalies and their corresponding reconstruction errors in time series data analysis using autoencoders and LSTM autoencoders.

vi. Flawed Anomaly Detection Benchmarks

Over the past decade, there has been a significant increase in work related to time series analysis. Many studies [17, 28, 31-35] have been conducted to evaluate the effectiveness of machine learning-based time series anomaly detection methods.

In a paper covering a comprehensive evaluation of time series anomaly detection, Schmidl et al. [35] observed that, despite the increased computational resources required during training, deep learning methodologies are currently not competitive in the field of time series anomaly detection; this includes AEs and LSTM AEs. The study also confirms the principle that simpler techniques can yield performance results almost on par with more complex approaches. Furthermore, no single machine learning algorithm comprehensively outperforms the others. Various algorithms exhibit specific strengths, but the overall findings call for further exploration in three critical domains.

Schmidl et al. highlight three (3) areas for further research. Firstly, the importance of flexibility, as no single algorithm or algorithmic family universally dominates all anomaly detection scenarios, urging the pursuit of hybrid systems to enhance anomaly detection in diverse time series settings. Second, it highlights the necessity for more research on the reliability and scalability of these algorithms, given that only a few were able to process time series data error-free within common resource constraints. Lastly, the research underscores the challenge of parameter sensitivity in many anomaly detection algorithms and advocates for the development of auto-configuring and self-tuning algorithms to simplify parameter selection, which is particularly vital in practical use cases lacking training data for parameter optimization.

Another notable paper by Wu and Keogh [32] highlights four significant flaws in publicly available time series datasets that are utilized to train anomaly detection models based on machine learning. These flaws include triviality, unrealistic anomaly density, mislabeled ground truth, and run-to-failure bias. By identifying these issues, the paper emphasizes the need for reliable and accurate time series datasets to train robust machine learning models for anomaly detection.

c. Matrix Profile for Time Series Anomaly Detection – A non-ML based approach

During the course of this research, Matrix Profile [36] based anomaly detection garnered prevalent adoption in the time series analytics domain. The Matrix Profile is a way to represent the similarity between subsequences in time series data. It is very efficient in capturing complex patterns and irregularities, which makes it an excellent tool for detecting anomalies in various applications. It can be used for monitoring industrial processes or identifying unusual behaviors in sensor data.

Matrix profile-based anomaly detection has a significant advantage in providing a detailed understanding of time series data. This technique enables the identification of both recurring patterns (motifs) and distinct patterns (discords), which helps to improve the interpretability of anomalies in different datasets. This interpretability is crucial in practical applications where specific patterns can reveal valuable insights, ensuring more accurate and informed decision-making.

However, matrix profile-based methods are not without limitations. Like the shortcoming of ML based anomaly detection methods, matrix profile-based methods are also influenced by various parameters that collectively impact their performance. The window size, which determines the length of subsequences used to compute the matrix profile, is a critical parameter. A larger window size can capture broader patterns but may overlook localized anomalies, while a smaller window size could be more sensitive to noise. The matrix profile length, or the granularity of pattern detection, is another important parameter, with longer profiles offering more detailed insights but requiring increased computational resources. The choice of distance metric to quantify similarity between subsequences is a pivotal decision, as different measures may be more suitable for specific data types. Setting thresholds for anomaly detection is crucial, with careful consideration needed to avoid false positives or negatives. Additionally, the normalization method applied to the time series data plays a role in ensuring consistent performance across datasets with varying scales and magnitudes.

For multivariate time series data, a different set of considerations emerges, including the definition of distance measures and the handling of multiple dimensions [37]. Adapting matrix profile-based methods to multivariate scenarios requires careful parameter selection to account for the complexities introduced by the additional dimensions. Optimal parameter choices are often dataset-specific, and fine-tuning based on the characteristics of the specific dataset is essential for achieving accurate and meaningful results in matrix profile-based time series analysis. Experimentation and thorough parameter tuning are critical steps to maximize the effectiveness of these methods in capturing and interpreting patterns in diverse datasets.

2.5. Symbolic Aggregation Approximation

In response to the limitations encountered with ML-based time series analysis methods, an exploration was undertaken to discover more effective alternatives for extracting contextual information from multivariate time-series data. The primary goal was to identify innovative approaches that could provide more meaningful insights and facilitate better decision-making based on the data. In the course of this research, Symbolic Aggregate Approximation (SAX) emerged as a pivotal data transformation technique that plays a critical role in the analysis of time series data. [38]. It operates by converting continuous time series data into a symbolic representation, enabling the application of various data mining and pattern recognition techniques. SAX offers several advantages over machine learning (ML) methods, making it a valuable tool in specific time series analysis scenarios.

One of the primary advantages of SAX is its ability to reduce the dimensionality of time series data while preserving essential information. By representing the data symbolically, SAX significantly reduces the data's dimensionality, making it more manageable for subsequent analysis [39-42]. This reduction simplifies the computational demands, especially when working with large-scale or high-dimensional time series datasets. Furthermore, the symbolic representation makes visualizing and interpreting the data easier, aiding in pattern discovery and anomaly detection tasks.

SAX also excels in handling noisy or uncertain time series data. ML methods often require clean, pre-processed data, which can be challenging to obtain in real-world scenarios. On the other hand, SAX is robust to noise and variations in the data, as it discretizes the time series into a predefined set of symbols. This robustness allows SAX to work effectively with data from domains like sensor networks, financial markets, and healthcare, where noise and irregularities are common. Additionally, SAX provides a compact representation of time series data, reducing the impact of outliers and anomalies on subsequent analysis.

Another advantage of SAX is its interpretability. The symbolic representation is intuitive and understandable (as shown in Figure 13), making it easier for domain experts to interpret and extract insights from the data. This interpretability is particularly valuable in fields where domain knowledge is critical. By simplifying the data representation and focusing on patterns within symbols, SAX enables domain experts to gain meaningful insights and make informed decisions based on the transformed data.

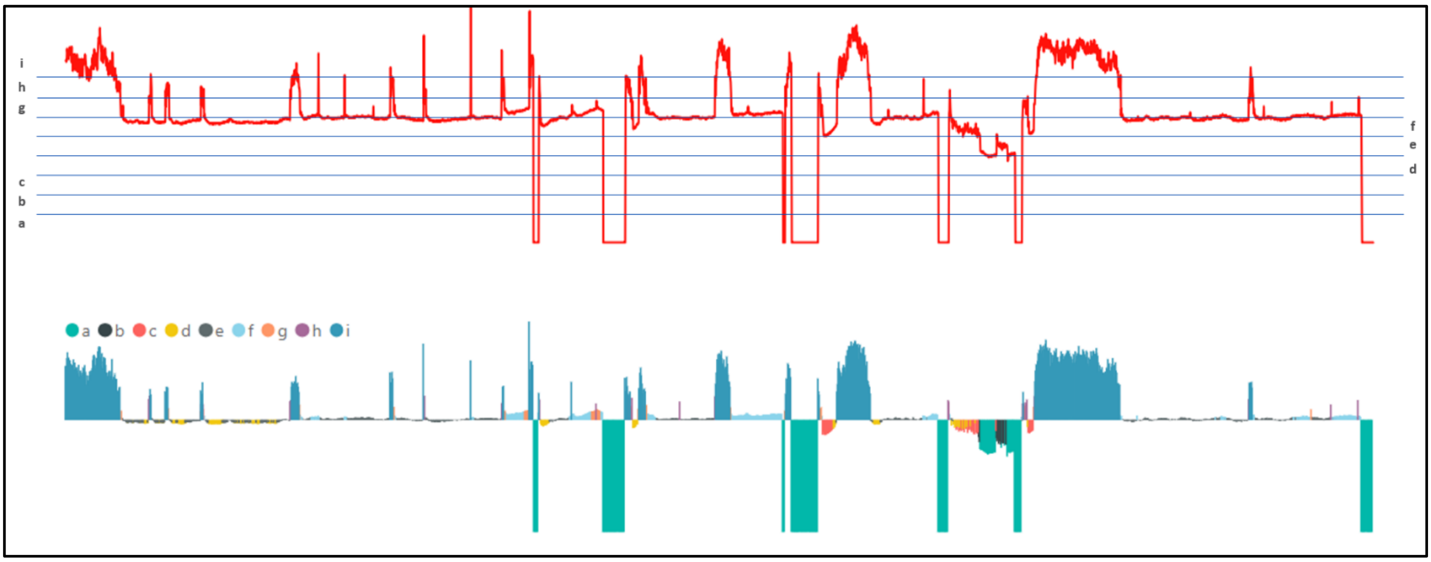


Figure 13 – A time series feature (above) is discretized based on SAX, and a plot is shown (below) with the relevant SAX labels.

SAX's ability to simplify and enhance the analysis of time series data makes it a valuable tool in various domains, where it complements machine learning approaches and aids in uncovering hidden patterns and insights in time-dependent data.

2.6. Summary

Monitoring real-time SCADA data from thousands of CSG wells is an intricate task, and as a result, there is a pressing need for a more simplified exception-based approach to discern which wells require heightened attention. Although ML-based solutions have been proposed and may seem promising, they come with a host of limitations. These limitations include the complexity of ML model training and interpretation, sensitivity to parameter settings, and the significant requirement for labelled data, which can be particularly challenging to obtain in industrial settings. Consequently, harnessing SAX-based time series analysis emerges as the most compelling approach for effective real-time well performance monitoring. SAX's ability to reduce data dimensionality while preserving critical information renders it highly robust against noisy data, all while maintaining interpretability. Given the paramount importance of timely and accurate insights in real-time SCADA systems for optimizing well operations, the plan is to utilize SAX as the foundation of the research. The research aims to develop a novel approach that transforms time series data into performance images, offering a simplified and accessible solution for identifying anomalies and optimizing well performance in real time. This approach will help CSG operators manage vast numbers of wells by exception, avoiding limitations of existing ML based anomaly detection methods.

2.7. References

1. Kalinin, D., et al. *Alleviating the Solids Issue in Surat Basin CSG Wells*. in *SPE Asia Pacific Oil and Gas Conference and Exhibition*. 2018.
2. Ming, L., et al. *CoalBed Methane Pad Wells Completion and Artificial Lift Optimizations: Case Study From Australia Surat Basin DS Gas Field*. in *International Petroleum Technology Conference*. 2021.
3. Rajora, A., et al. *Deviated Pad Wells in Surat: Journey So Far*. in *SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology Conference*. 2019.
4. Lin, X., et al. *Increasing Coal Seam Gas Field Productivity with Horizontal Well Technology: A Case Study*. in *SPE Asia Pacific Oil and Gas Conference and Exhibition*. 2018.
5. Krawiec, M.B., et al. *Dewatering Coalbed Methane Wells Using ESPCPs*. in *Canadian International Petroleum Conference*. 2008.
6. Dunham, C.L., *Production Automation in the 21st Century: Opportunities for Production Optimization and Remote Unattended Operations.* Journal of Petroleum Technology, 2003. **55**(07): p. 68-73.
7. Bohannon, J.M., *Automation in Oilfield Production Operations.* Journal of Petroleum Technology, 1984. **36**(08): p. 1239-1242.
8. Crompton, J. *The Digital Oil Field Hype Curve: A Current Assessment the Oil and Gas Industry's Digital Oil Field Program*. in *SPE Digital Energy Conference and Exhibition*. 2015.
9. Ormerod, L., et al., *Real-Time Field Surveillance and Well Services Management in a Large Mature Onshore Field: Case Study.* SPE Production & Operations, 2007. **22**(04): p. 392-402.
10. Bybee, K., *Case Study: A Coalbed- Methane Automation System.* Journal of Petroleum Technology, 2001. **53**(04): p. 72-73.
11. Robertson, S.K., et al. *Case Study: Coalbed Methane Automation System for River Gas Corporation*. in *SPE Asia Pacific Oil and Gas Conference and Exhibition*. 2000.
12. Robertson, S.K., et al. *Case Study: Coalbed Methane Automation System for River Gas Corporation*. in *SPE Eastern Regional Meeting*. 1999.
13. Robertson, S.K., et al. *Automation System Case Study of Coalbed-Methane Development*. in *SPE/CERI Gas Technology Symposium*. 2000.
14. Knafl, M., et al. *Diagnosing PCP Failure Characteristics using Exception Based Surveillance in CSG*. in *SPE Progressing Cavity Pumps Conference*. 2013.
15. Rathnayake, S., A. Rajora, and M. Firouzi, *A machine learning-based predictive model for real-time monitoring of flowing bottom-hole pressure of gas wells.* Fuel, 2022. **317**: p. 123524.
16. Ao, S.-I. and H. Fayek, *Continual Deep Learning for Time Series Modeling.* Sensors, 2023. **23**(16): p. 7167.
17. Belay, M.A., et al., *Unsupervised Anomaly Detection for IoT-Based Multivariate Time Series: Existing Solutions, Performance Analysis and Future Directions.* Sensors, 2023. **23**(5): p. 2844.
18. Chen, Z., et al. *Autoencoder-based network anomaly detection*. in *2018 Wireless Telecommunications Symposium (WTS)*. 2018.
19. Yin, C., et al., *Anomaly Detection Based on Convolutional Recurrent Autoencoder for IoT Time Series.* IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022. **52**(1): p. 112-122.
20. Hochreiter, S. and J. Schmidhuber, *Long Short-term Memory.* Neural computation, 1997. **9**: p. 1735-80.
21. Yan, S., *Understanding LSTM and its diagrams*. 2016, Medium.
22. Widiputra, H., A. Mailangkay, and E. Gautama, *Multivariate CNN-LSTM Model for Multiple Parallel Financial Time-Series Prediction.* Complexity, 2021. **2021**: p. 9903518.
23. Lafabregue, B., et al., *End-to-end deep representation learning for time series clustering: a comparative study.* Data Mining and Knowledge Discovery, 2022. **36**(1): p. 29-81.

24. Xu, C., H. Huang, and S. Yoo. *A Deep Neural Network for Multivariate Time Series Clustering with Result Interpretation*. in *2021 International Joint Conference on Neural Networks (IJCNN)*. 2021.
25. Alqahtani, A., et al., *Deep Time-Series Clustering: A Review.* Electronics, 2021. **10**(23): p. 3001.
26. Tadayon, M. and Y. Iwashita, *A clustering approach to time series forecasting using neural networks: A comparative study on distance-based vs. feature-based clustering methods.* arXiv preprint arXiv:2001.09547, 2020.
27. Javed, A., B.S. Lee, and D.M. Rizzo, *A benchmark study on time series clustering.* Machine Learning with Applications, 2020. **1**: p. 100001.
28. Rewicki, F., J. Denzler, and J. Niebling, *Is It Worth It? Comparing Six Deep and Classical Methods for Unsupervised Anomaly Detection in Time Series.* Applied Sciences, 2023. **13**(3): p. 1778.
29. Lim, S., et al., *A deep learning-based time series model with missing value handling techniques to predict various types of liquid cargo traffic.* Expert Systems with Applications, 2021. **184**: p. 115532.
30. Hasan, M.K., et al., *Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021).* Informatics in Medicine Unlocked, 2021. **27**: p. 100799.
31. Choi, K., et al., *Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines.* IEEE Access, 2021. **9**: p. 120043-120065.
32. Wu, R. and E. Keogh, *Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress.* IEEE Transactions on Knowledge and Data Engineering, 2021: p. 1-1.
33. Darban, Z.Z., et al., *Deep learning for time series anomaly detection: A survey.* arXiv preprint arXiv:2211.05244, 2022.
34. Goswami, M., et al., *Unsupervised model selection for time-series anomaly detection.* arXiv preprint arXiv:2210.01078, 2022.
35. Schmidl, S., P. Wenig, and T. Papenbrock, *Anomaly detection in time series: a comprehensive evaluation.* Proc. VLDB Endow., 2022. **15**(9): p. 1779–1797.
36. Yeh, C.C.M., et al. *Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets*. in *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 2016.
37. Yeh, C.C.M., N. Kavantzas, and E. Keogh. *Matrix Profile VI: Meaningful Multidimensional Motif Discovery*. in *2017 IEEE International Conference on Data Mining (ICDM)*. 2017.
38. Lin, J., et al., *A Symbolic Representation of Time Series, with Implications for Streaming Algorithms*. 2003. 2-11.
39. Keogh, E., J. Lin, and A. Fu. *HOT SAX: efficiently finding the most unusual time series subsequence*. in *Fifth IEEE International Conference on Data Mining (ICDM'05)*. 2005.
40. Lin, J., et al., *Experiencing SAX: a novel symbolic representation of time series.* Data Mining and Knowledge Discovery, 2007. **15**(2): p. 107-144.
41. Keogh, E., et al., *Finding the Unusual Medical Time Series: Algorithms and Applications.* IEEE Transactions on Information Technology in Biomedicine - TITB, 2005.
42. Kumar, N., et al., *Time-series Bitmaps: a Practical Visualization Tool for Working with Large Time Series Databases*. 2005.

3. Paper 1: Application of Exploratory Data Analytics EDA in Coal Seam Gas Wells with Progressive Cavity Pumps PCPs

This paper explores the application of EDA within the realm of CSG wells equipped with PCPs. The primary aim is to enhance the understanding and optimize PCP performance, addressing unique challenges posed by multivariate time series data within the oil and gas industry.

The application of EDA methodologies on a three-year time series dataset gathered from 42 CSG wells is demonstrated, utilizing the Python programming language and its supporting libraries.

The critical role of EDA as a preliminary step before embarking on real-time analytics cannot be overstated. EDA is used to address data discontinuity, which is prevalent in SCADA data, and generalize multivariate SCADA data with time series analytics. This helps to comprehend the behavior of time series data and extract pertinent features, especially in the context of multivariate datasets.

The methodology encompasses several key steps, including the sorting and normalization of raw time series data, interpolation to handle missing values, data filtering to remove non-representative data, and data decomposition to reveal underlying trends. Correlation analysis is explored as a means to assess PCP performance over different time periods, employing sliding window techniques. Additionally, a novel approach to data visualization is introduced, leveraging SAX to create HEATMAP images for real-time monitoring of PCP performance.

The results discussed in the paper demonstrate how managing SCADA data and converting time-series data to SAX HEATMAP are necessary for real-time PCP performance analysis.

Statement of Authorship

Title of Paper	Application of Exploratory Data Analytics EDA in Coal Seam Gas Wells with Progressive Cavity Pumps PCPs
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Saghir F, Gonzalez Perdomo ME, Behrenbruch P (2019) Application of exploratory data analytics EDA in coal seam gas wells with progressive cavity pumps PCPs. In: SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition. 2019, Society of Petroleum Engineers: Bali, Indonesia, p. 10.

Principal Author

Name of Principal Author (Candidate)	Fahd Saghir		
Contribution to the Paper	Conduct data analysis write and record experiments, create test reports, tabulate results and write paper.		
Overall percentage (%)	75%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	19/09/2023

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Mary Gonzalez Perdomo		
Contribution to the Paper	Assisted with paper structure, writing and paper review (20%)		
Signature		Date	18/09/2023

Name of Co-Author	Peter Behrenbruch		
Contribution to the Paper	Assisted with paper structure, writing and paper review (5%)		
Signature		Date	10-09-2023

Please cut and paste additional co-author panels here as required.

SPE-196528-MS

Application of Exploratory Data Analytics EDA in Coal Seam Gas Wells with Progressive Cavity Pumps PCPs

Fahd Saghir, M. E. Gonzalez Perdomo, and Peter Behrenbruch, University of Adelaide

This paper was prepared for presentation at the SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition held in Bali, Indonesia, 29-31 October 2019.

Abstract

Artificial lift methods typically drive Coal Seam Gas (CSG) wells, and Progressive Cavity Pump (PCP) is the preferred method of lift with Australian CSG operators. CSG wells in Australia are typically equipped with necessary instrumentation and automation systems to provide real-time data gathering for monitoring and control purposes. Real-time data gathered from CSG wells presents an opportunity to better understand PCP performance by identifying anomalous pump behavior.

However, before undertaking any real-time analytics exercise, it is pertinent to carry out Exploratory Data Analytics (EDA) to understand time series data behavior and extract relevant features; and this exercise is particularly important with multi-variate data sets. Obtaining significant data features from multivariate time series data can help define which analytics and machine learning methods could be exploited to analyze PCP performance in near real time.

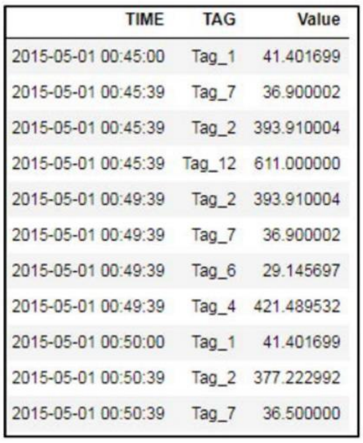
This paper will discuss EDA methodologies that can help streamline time-series data normalization and feature extraction techniques. A three (3) year time-series dataset, gathered from forty-two (42) CSG wells, will be used to showcase EDA methodologies utilized as part of this research. All EDA activities covered in this paper are based on the Python programming language and its supporting libraries.

Introduction

Time Series Data in Production Systems - Challenges and Opportunities

Not all time series datasets are created equal. Although, standardization of Supervisory Control and Data Acquisition (SCADA) systems in upstream oil and gas has paved the way for real-time data gathering; the data collection and storage methodology may still be unique to each SCADA technology provider. Heterogeneous datasets, produced by varying data collection and storage methodologies, pose a challenge for analytics and machine learning platforms in terms of data ingestion[1].

Common issues encountered with heterogeneous datasets include, but are not limited to, unsynchronized data measurements, missing values due to network communication failures, and values captured when sensors are faulty[2]. Another drawback with time series repositories is that they store multivariate data sequentially, and this too is not suitable for analytics and machine learning applications. Figure 1 shows a sample sequentially stored time series dataset, which was used as part of this project.



TIME	TAG	Value
2015-05-01 00:45:00	Tag_1	41.401699
2015-05-01 00:45:39	Tag_7	36.900002
2015-05-01 00:45:39	Tag_2	393.910004
2015-05-01 00:45:39	Tag_12	611.000000
2015-05-01 00:49:39	Tag_2	393.910004
2015-05-01 00:49:39	Tag_7	36.900002
2015-05-01 00:49:39	Tag_6	29.145697
2015-05-01 00:49:39	Tag_4	421.489532
2015-05-01 00:50:00	Tag_1	41.401699
2015-05-01 00:50:39	Tag_2	377.222992
2015-05-01 00:50:39	Tag_7	36.500000

Figure 1—Sequentially Stored Time Series Dataset

It is important that sequentially arranged and heterogeneously collected multivariate data sets are sorted and normalized, so they may be used for further analysis. EDA presents an opportunity not only to discover a streamlined methodology to cleanse, filter and normalize multivariate time series data but also to experiment with a variety of feature extraction methodologies that enable a better understanding of PCP performance in CSG wells.

Methodology

Step 1: Sort Raw Time Series Data

As with any EDA exercise, it is pertinent first to sort data in a format that is acceptable within the software platform; in this case, Python programming language. Statistical and Machine Learning libraries associated with Python accept data input in the shape of *n_samples, n_features* (Figure 2), where *n_samples* represents the increasing timestamp rows, and *n_features* represents data column unique to each sensor value.

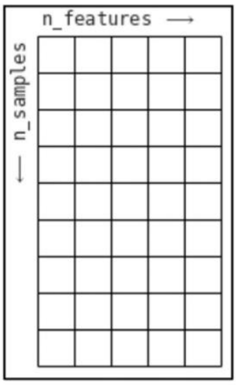
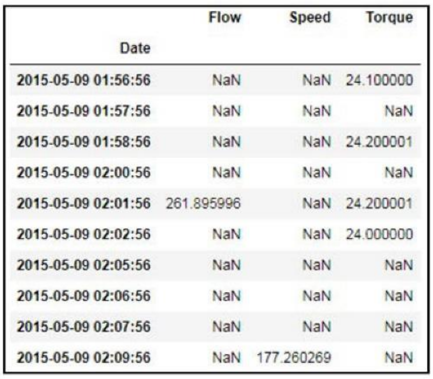


Figure 2—Python Data Format for Statistical and Machine Learning Libraries[3]

Raw time series data shown in Figure 1, can be re-arranged and sorted as per timestamp (Date) values, with each sensor value segregated into columns. This transformation is shown in Figure 3. However, there are drawbacks when raw time series data is converted into *n_samples* and *n_features* format. Sensor values which are not recorded concurrently produce *NaN* entries in the time series dataset.



	Flow	Speed	Torque
Date			
2015-05-09 01:56:56	NaN	NaN	24.100000
2015-05-09 01:57:56	NaN	NaN	NaN
2015-05-09 01:58:56	NaN	NaN	24.200001
2015-05-09 02:00:56	NaN	NaN	NaN
2015-05-09 02:01:56	261.895996	NaN	24.200001
2015-05-09 02:02:56	NaN	NaN	24.000000
2015-05-09 02:05:56	NaN	NaN	NaN
2015-05-09 02:06:56	NaN	NaN	NaN
2015-05-09 02:07:56	NaN	NaN	NaN
2015-05-09 02:09:56	NaN	177.260269	NaN

Figure 3—Raw Time Series Data sorted into *n_samples, n_features* format

Figure 4 shows a plot diagram of the dataset, where gaps in time series data are visible in the Flow and Speed trends. The Torque values are recorded with a higher frequency. Hence fewer gaps are visible in this trend.

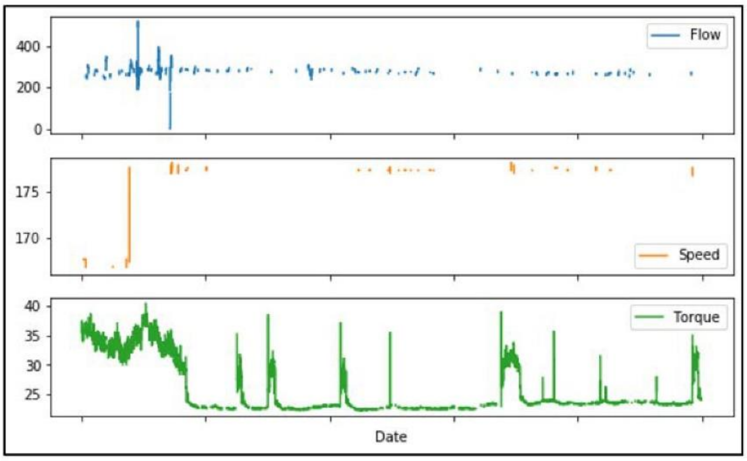


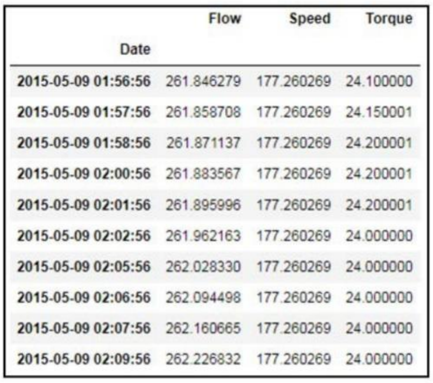
Figure 4—Plotted Raw Time Series Data with NaN entries

Step 2: Interpolate Time Series Data

To ensure all *NaN* values are replaced with useful data, it is best to interpolate the sensor values. Interpolation enables estimation of missing data points by using either a linear or a polynomial method of calculation. Both methods of interpolation are available through the *SciPy*[4] library in Python. Before deciding which interpolation techniques can be applied to time series data, it is best to understand the characteristics of the measured values. From a PCP production perspective, Flow and Torque values are measured by an instrument. Speed, however, is a control setpoint, which is altered based on pump control methodology.

As Flow and Torque demonstrate a more dynamic behavior, it is best to interpolate these values using the *cubic* interpolation method. For Speed, *linear* interpolation works best as there is no dynamic value change between two (2) recorded values.

Figure 5 shows the dataset where *NaN* entries were replaced with interpolated values. Figure 6 is the plotted time series data with interpolated values. An interpolated dataset provides a better visual representation of the multivariate time series data.



Date	Flow	Speed	Torque
2015-05-09 01:56:56	261.846279	177.260269	24.100000
2015-05-09 01:57:56	261.858708	177.260269	24.150001
2015-05-09 01:58:56	261.871137	177.260269	24.200001
2015-05-09 02:00:56	261.883567	177.260269	24.200001
2015-05-09 02:01:56	261.895996	177.260269	24.200001
2015-05-09 02:02:56	261.962163	177.260269	24.000000
2015-05-09 02:05:56	262.028330	177.260269	24.000000
2015-05-09 02:06:56	262.094498	177.260269	24.000000
2015-05-09 02:07:56	262.160665	177.260269	24.000000
2015-05-09 02:09:56	262.226832	177.260269	24.000000

Figure 5—Interpolated Time Series Data

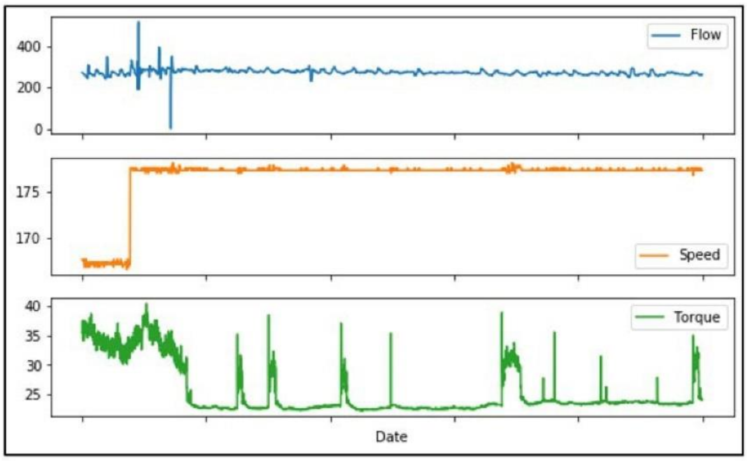


Figure 6—Plotted Interpolated Time Series Data

Step 3: Filter and Normalize Time Series Data

Next step in the EDA process is to filter out any data that is not representative of the PCP performance. Figure 7 shows unfiltered time series data, and it is obvious that the high peaks in the Flow trend are not representative of the actual production performance. These non-characteristics values can easily be filtered out by confirming peak PCP flow design, which is done in the initial production design stage. In this example, the PCP theoretical flow rate cannot exceed 2000bbl/day of water. Hence, any values greater than this number can be filtered out. Likewise, any non-representative sensor values can be filtered out based on technical design limit of the artificial lift system.

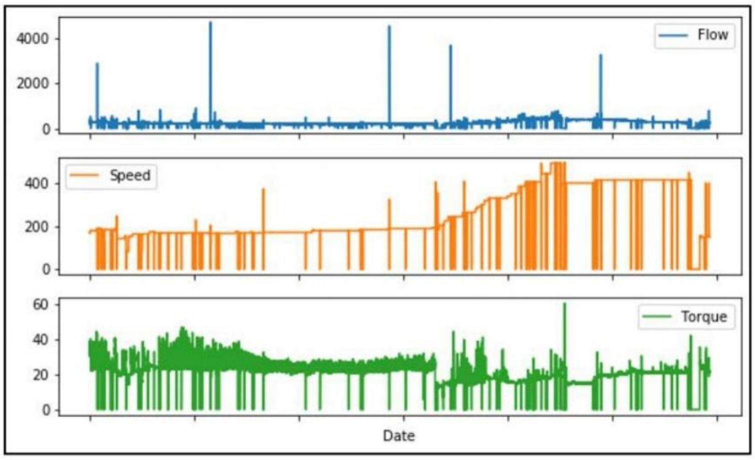


Figure 7—Unfiltered Time Series Data

Once the data is filtered, it should be normalized to remove sensitivities based on the measurement scale (y-axis). In the context of machine learning, data normalization removes dependencies on measurement scales which may otherwise produce skewed results. *Scikit-learn*[5] library in Python provides various data normalization techniques that can be chosen based on data characteristics.

In the case of CSG operated PCPs, normalizing data by removing the mean and scaling to unit variance will not suffice, as certain pump operating conditions must be factored into data normalization. During pump startup, torque is usually high either due to PCP polymer swell or solids settling over the pump. Another operating method applied to PCP wells is sudden speed bursts which are required to clear any solid contents in the pump stator. These events create measurement readings which highly deviate from the mean and hence require specialized normalization. The *RobustScaler* method from *Scikit-learn* is best suited for PCP data normalization where unique events cannot be ignored during the data normalization process.

Figure 8 shows plotted PCP dataset that has been filtered and normalized based on the procedure described above.

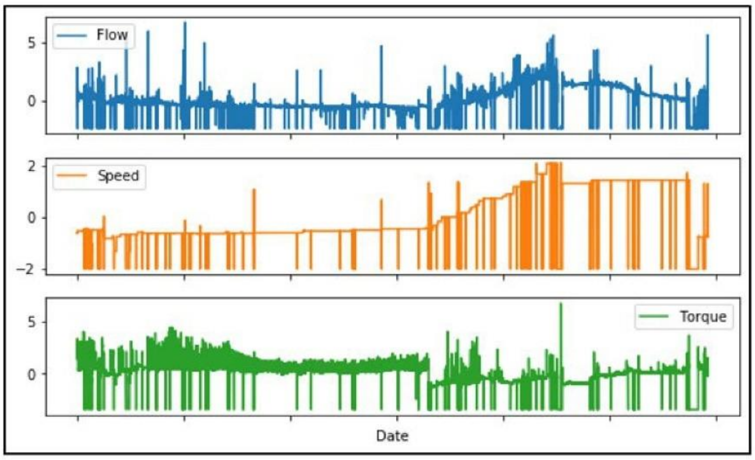


Figure 8—Filtered and Normalized Data

Step 4: Data Decomposition

Decomposition breaks down time series data into *Trends, Seasonality*, and *Residue*. As PCP data is void of any seasonal characteristics, we will work with the Trend and Residue decomposition components of time series data.

The extracted *Trend* plots reveal the underlying characteristics of the measured values. As seen in Figure 9, the *Trend* plot for Flow, Torque, and Speed depicts how these measurements change over the production period.

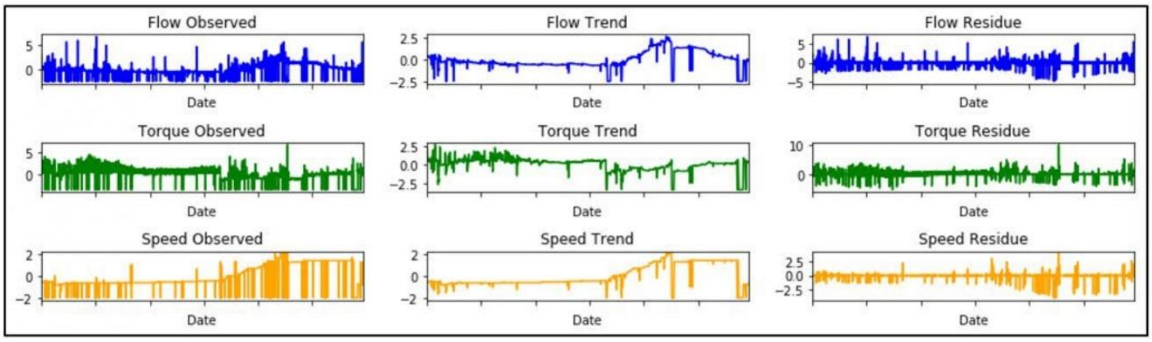


Figure 9—Decomposing Time Series Data to Extract Trend and Residue from Observed Values

Decomposition is achieved by using the *Statsmodels* library, where the *seasonal_decomposition* function is used to split the observed data in *Trend* and *Residue*.

Step 5: Analyze Correlations in Multi-Variate Data

Correlation analysis is conducted as part of the EDA exercise to establish an association between multivariate sensor readings. Trend decomposition values are used from Step 4 to create *Speed vs. Flow*, and *Speed vs. Torque* correlation plots. These plots help analyze PCP performance over various time periods. Figure 10 shows a sample *Speed vs. Flow* correlation plot, which is created using the *jointplot* function from the *Seaborn* library in Python. Pearson coefficient, along with a regression plot, is also calculated to show the linear relationship between the measured time series variables.

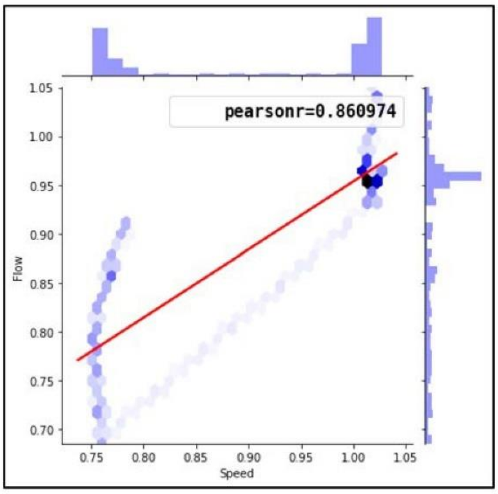


Figure 10—Correlation Plot: Speed vs. Flow

Step 6: Data Approximation

Approximating time series data into bins is a dimensionality reduction technique where measured sensor values are converted into alphabet characters. This character-based conversion is derived from the Symbolic Aggregation Approximation (SAX) methodology [6], where time series data is divided into bins based on Gaussian Distribution.

Figure 11 shows a time series trend converted to nine (9) SAX characters. Approximation based dimensionality reduction aids with improved time series data visualization, where character-labeled trends can distinctly illustrate the change in PCP performance. SAX-based trends will be shown in the Results section of this paper.

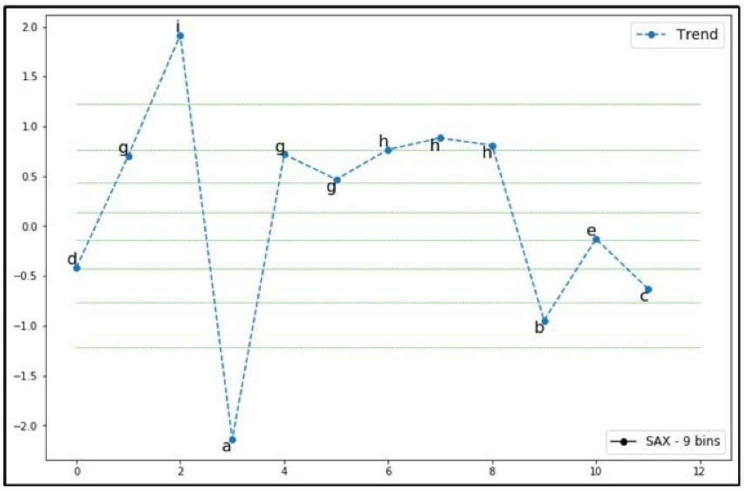


Figure 11—SAX Based Breakdown of Time Series data into 9 Bins (Alphabet Characters)

Results

Correlation-Based PCP Performance Analysis

The correlation analysis technique mentioned in Step 5 can be used to gauge PCP performance based on a sliding window technique, which is shown in Figure 12. A correlation plot is created for a one (1) day time window, with a half-day stride. By doing so, a correlation plot is created every twelve (12) hours to gauge overall PCP performance. This allows CSG operators to observe the regression and Pearson coefficient trend across progressive correlation plots.

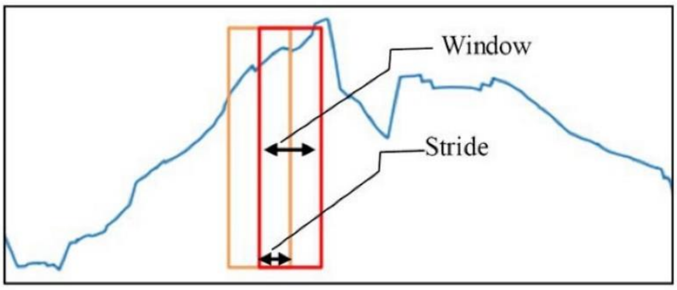


Figure 12—Sliding Window Mechanism for Creating Correlation Plots

Figure 13 shows the progression of correlation plots over a life of CSG operated PCP well. In the figure below, the left column shows *Speed vs. Flow* and *Speed vs. Torque* during an early stage of pump life where both plots have a positive linear regression plot. The Pearson coefficient is also positive and closer to one (1).

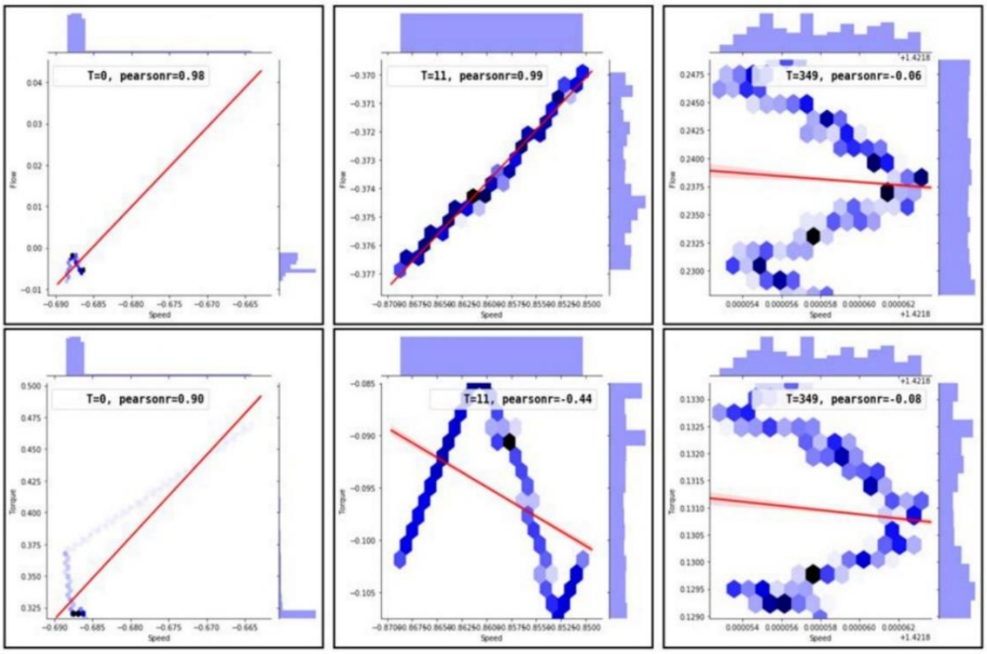


Figure 13—Sliding Window Based Correlation Plots

The center column depicts PCP performance characteristic that requires further investigation as the *Speed vs. Torque* and *Speed vs. Flow* have the opposite correlation. Ideally, both correlation plots should have the same regression trend (either increasing or decreasing), and the opposite trend is an indication of uncharacteristic PCP performance. Cases where *Speed vs. Torque* correlation is negative, and the *Speed vs. Flow* correlation is positive indicates a solid intake problem with PCP. In a reverse scenario, where *Speed vs. Torque* correlation is positive, and *Speed vs. Flow* correlation is negative, this is an indication of either plugged pump intake/discharge or a pump-off condition.

The last column in Figure 13 shows *Speed vs. Flow* and *Speed vs. Torque* towards the end of PCP run life. Both correlations now have a negative linear regression and a slightly negative Pearson coefficient. This indicates that PCP efficiency has decreased due to elastomer degradation.

Correlation plots can be adjusted to characterize PCP performance over a shorter time window. Reducing the time window adds granularity to the multivariate time series analysis, which allows operators to observe changes in correlation over shorter time periods.

SAX Based Data Visualization

SAX-based data approximation technique aids with improved visualization of multivariate time series trends. Both the time series trends and SAX-based plots can be generated in real-time to capture changes in PCP performance profile.

In Figure 14, the SAX-based trend for Flow shows regions where the water measurements start increasing above the mean, then plateau off, followed by measurements which start falling below the mean. Such visualization help track PCP performance in real-time, where observing a transition in SAX characters can easily identify overall PCP performance. Abnormal events can also be gauged with this method where a non-linear transition of SAX characters can indicate events which are uncharacteristic of normal PCP performance, for example, characters changing from *d* to *g* without transitioning through *e* and *f*.

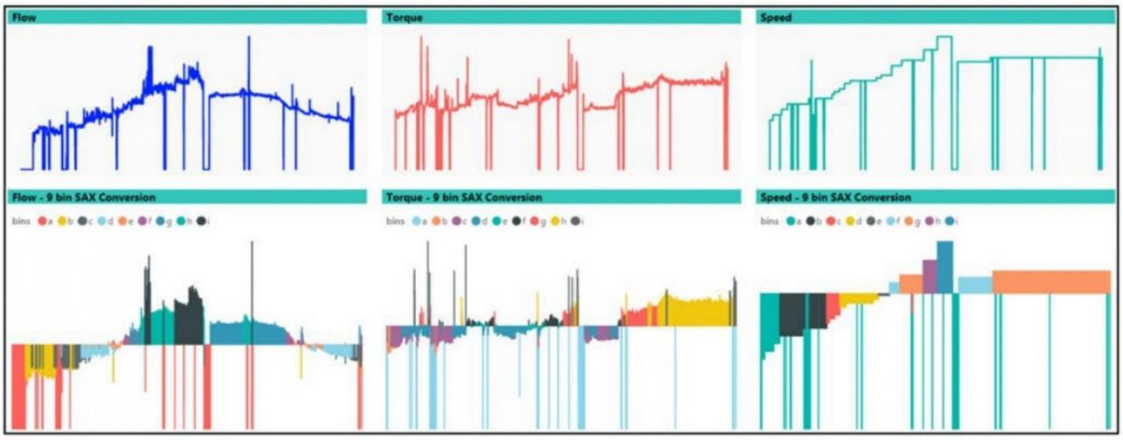


Figure 14—SAX Based Visualization of Time Series Data

SAX Based HEATMAP Conversion

Another method to visualize PCP performance is by converting the SAX-based characters into HEATMAP images[7]. Converted images enable the use of supervised and unsupervised machine learning methods to tag PCP performance in near real-time autonomously. This methodology also assists with the detection of anomalous events based on HEATMAP color.

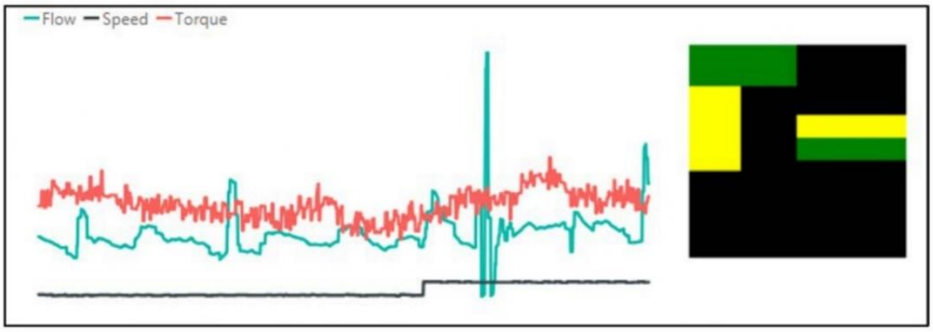


Figure 15—Multivariate Time Series Data Converted to SAX-based HEATMAP

Conclusion

In this paper, we have shown that Exploratory Data Analytics help extract significant information from time series data that can aid with a better understanding of PCP performance. EDA steps described in the Methodology section set the foundation for using extracted features from time series data to conduct further statistical and machine learning evaluation of PCP performance. These methods, although not exhaustive, are reusable with any sequential time series data set. Furthermore, these reusable methods can be standardized as part of a broader analytics and machine learning endeavor to obtain diagnostic and predictive insights on PCP performance in near real time.

References

1. Yan, J., et al, *Industrial Big Data in an Industry 4.0 Environment: Challenges, Schemes, and Applications for Predictive Maintenance*. IEEE Access, 2017. **5**: p. 23484–23491.
2. Wang, L., *Heterogeneous Data and Big Data Analytics. Automatic Control and Information Sciences*, 2017. **3**(1): p. 8–15.
3. VanderPlas, J., *Python Data Science Handbook: Essential Tools for Working with Data*. 2016: O'Reilly Media, Inc. 548.
4. Jones, E., et al *SciPy: Open Source Scientific Tools for Python*. 2001; Available from: http://www.scipy.org/.
5. Pedregosa, F., et al, Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011. **12**: p. 2825–2830.
6. Lin, J., et al, Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 2007. **15**(2): p. 107–144.
7. Saghir, F., M.E.G. Perdomo, and P. Behrenbruch, Converting Time Series Data into Images: An Innovative Approach to Detect Abnormal Behavior of Progressive Cavity Pumps Deployed in Coal Seam Gas Wells, in Annual Technical Conference and Exhibition. 2109, Society of Petroleum Engineers: Calgary, Canada.

4. Paper 2: Converting Time Series Data into Images: An Innovative Approach to Detect Abnormal Behavior of Progressive Cavity Pumps Deployed in Coal Seam Gas Wells

This paper introduces a novel approach to detect abnormal PCP behavior in CSG wells by converting time series data into heatmap images and utilizing machine learning techniques. The method involves converting multivariate time series data from PCPs into images using the SAX methodology, which helps to detect abnormal behavior autonomously. SAX is a technique for normalizing time series data and converting it into heatmap images. The methodology includes selecting relevant variables, creating a Gaussian distribution, and adding breakpoints to convert data into SAX symbols.

The paper highlights the significance of PCPs as a primary artificial lift method in CSG wells, and the challenges they face due to coal fines. Frequent pump failures caused by coal fines make it challenging to manage CSG wells. Hence, the proposed method has the potential to benefit the management of CSG wells in Australia significantly.

In the results section, the paper demonstrates the effectiveness of using heatmap images for detecting abnormal performance events in PCPs. The paper proposes an automated approach that involves K-Means clustering to segregate and label the clusters. Each heatmap cluster represents a specific type of PCP abnormal behavior, which is used to identify performance issues in real-time.

The findings outlined in this paper demonstrate the significance of using SAX-based heatmap images to enhance the accuracy of time-series clustering. By clustering time-series data, it is possible to automatically identify abnormal PCP behavior and label datasets to improve PCP performance analysis.

Statement of Authorship

Title of Paper	Converting Time Series Data into Images: An Innovative Approach to Detect Abnormal Behavior of Progressive Cavity Pumps Deployed in Coal SeamGas Wells
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Saghir F, Gonzalez Perdomo ME, Behrenbruch P (2019) Converting time series data into images: An innovative approach to detect abnormal behavior of progressive cavity pumps deployed in coal seam gas wells. In: SPE Annual Technical Conference and Exhibition. 2019, Society of Petroleum Engineers: Calgary, Alberta, Canada, p. 14.

Principal Author

Name of Principal Author (Candidate)	Fahd Saghir		
Contribution to the Paper	Conduct data analysis, write and record experiments, create test reports, tabulate results and write paper.		
Overall percentage (%)	75%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	19/09/2023

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Mary Gonzalez Perdomo		
Contribution to the Paper	Assisted with paper structure, writing and paper review (20%)		
Signature		Date	19/09/2023

Name of Co-Author	Peter Behrenbruch		
Contribution to the Paper	Assisted with paper structure, writing and paper review (5%)		
Signature		Date	10-09-2023

Please cut and paste additional co-author panels here as required.

SPE-195905-MS

Converting Time Series Data into Images: An Innovative Approach to Detect Abnormal Behavior of Progressive Cavity Pumps Deployed in Coal Seam Gas Wells

Fahd Saghir, M. E. Gonzalez Perdomo, and Peter Behrenbruch, University of Adelaide

This paper was prepared for presentation at the SPE Annual Technical Conference and Exhibition held in Calgary, Alberta, Canada, 30 Sep - 2 October 2019.

Abstract

Progressive Cavity Pumps (PCPs) are the predominant form of artificial lift method deployed by Australian operators in Coal Seam Gas (CSG) wells. With over five thousand CSG wells [1] operating in Queensland's Bowen and Surat Basins, managing and maintaining PCP supported production becomes a significant challenge for operators. Especially when these pumps face regular failures due to the production of coal fines.

It is possible to gauge the holistic production performance of PCPs with the aid of real-time data, as this allows for pro-active and informed management of artificially lifted CSG wells. Based on data obtained from two (2) CSG operators, this paper will discuss in detail how features extracted from time series data can be converted to images, which can then aid in autonomously detecting abnormal PCP behavior.

Introduction

Overview of PCPs in CSG Wells

In a CSG reservoir, methane is adsorbed to the surface of the coal, and gas is extracted by dewatering a CSG well [2]. As depicted in Figure 2 [3], dewatering is required at an early onset to produce gas from the reservoir. Progressive Cavity Pumps (PCPs) are utilized in the well to facilitate the dewatering process throughout the production lifecycle. Figure 3 shows the main components of a PCP system.

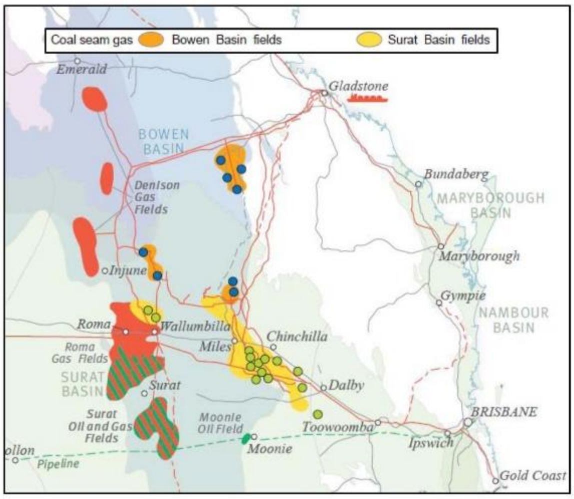


Figure 1—Map of CSG fields in Queensland, Australia [1]

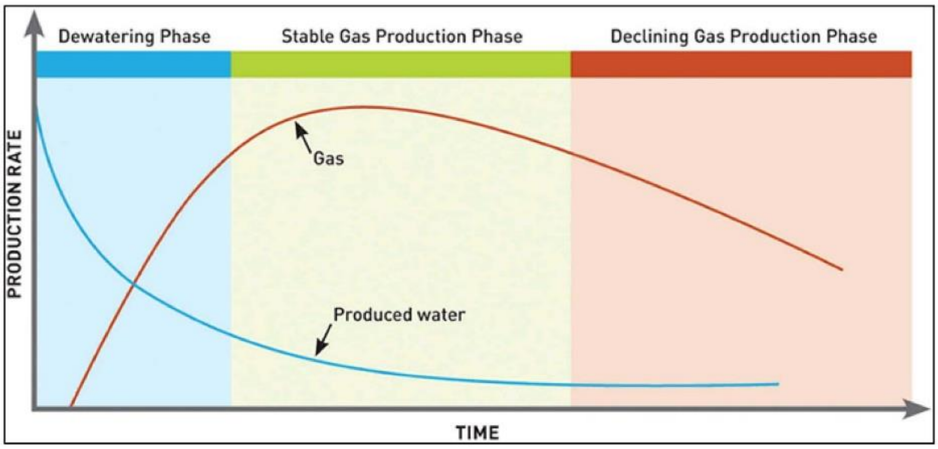


Figure 2—CSG Well Production Life Cycle [3]

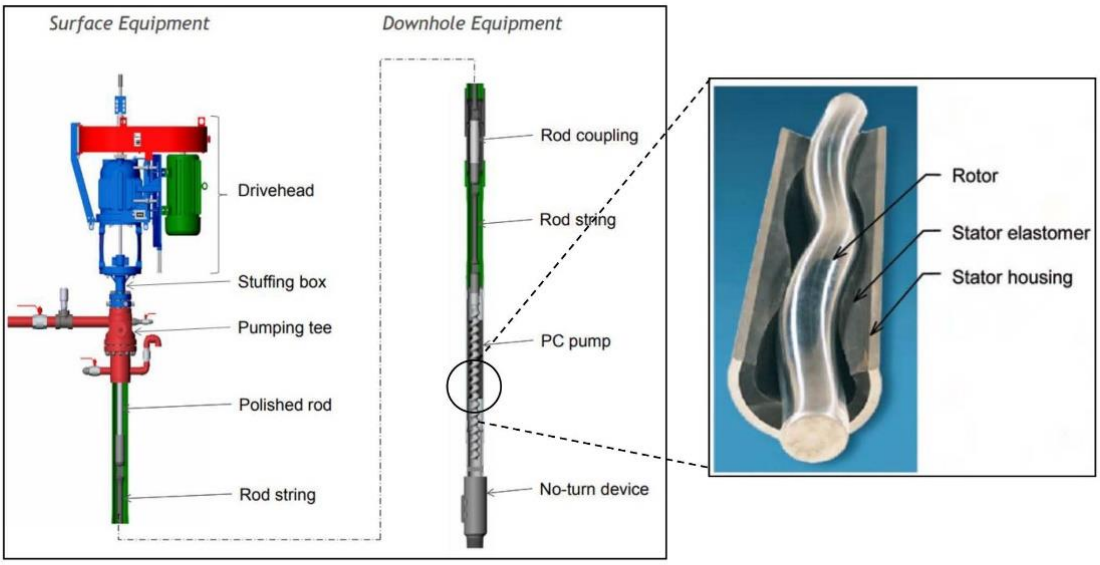


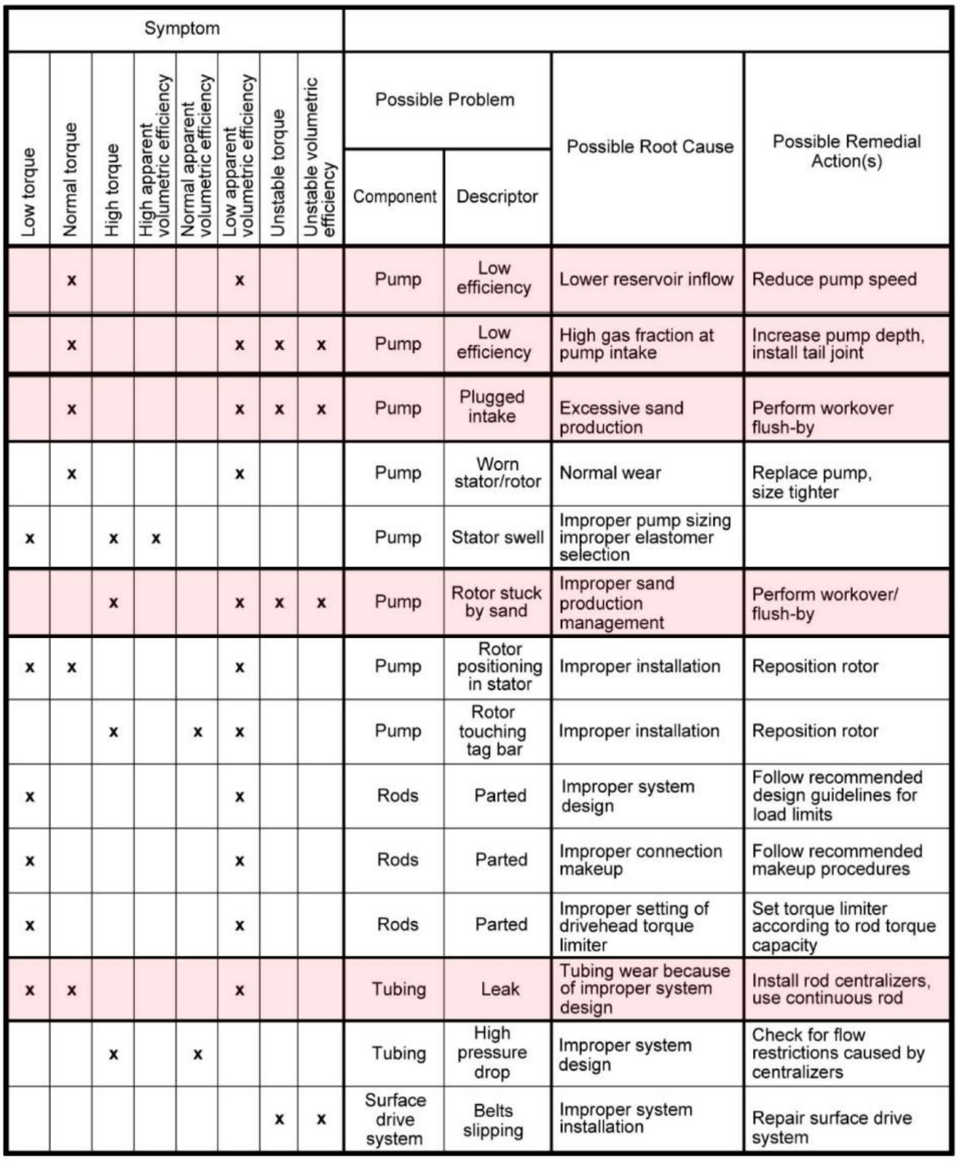
Figure 3—(Left) Main Components of a PCP system. (Right) Cut out view of PCP Rotor and Stator. [4]

Progressive Cavity Pumps (PCPs) have been deployed in CSG wells since the mid-1980s [2]. PCP is a positive displacement pump, where elastomer with cavities acts as a stator, and a helix shape metallic rotor acts as a stator. The unique design makes this Artificial Lift method resilient to solids, and hence can be deployed in wells with high solid contents.

Common PCP Failures in CSG Wells

Table 1 [4] captures the most common PCP failures, their symptoms, and possible root causes. Although their mechanical design makes PCPs resilient to solid production, the presence of coal fines and interburden incursion in CSG wells presents significant pump performance challenges. Failures highlighted in Table 1, capture problems that are common in CSG operations. Multiple CSG operators have documented PCP failures and have provided best practices and remedial actions to improve pump life [5–7]. Except for gas interference at pump intake (caused either by setting the pump above perforation or reduction in the liquid level below pump intake), all operators agree that coal fines and interburden intrusion are the root cause for the majority of PCP failures [5–7]. These failures are inclusive of, but not limited to, plugged intake, plugged discharge, parted rods and hole in tubing.

Table 1—List of PCP Failures and their Root Cause [4]



Symptom											
								Possible Problem		Possible Root Cause	Possible Remedial Action(s)
Low torque	Normal torque	High torque	High apparent volumetric efficiency	Normal apparent volumetric efficiency	Low apparent volumetric efficiency	Unstable torque	Unstable volumetric efficiency	Component	Descriptor		
	x				x			Pump	Low efficiency	Lower reservoir inflow	Reduce pump speed
	x				x	x	x	Pump	Low efficiency	High gas fraction at pump intake	Increase pump depth, install tail joint
	x				x	x	x	Pump	Plugged intake	Excessive sand production	Perform workover flush-by
	x				x			Pump	Worn stator/rotor	Normal wear	Replace pump, size tighter
x		x	x					Pump	Stator swell	Improper pump sizing improper elastomer selection	
		x			x	x	x	Pump	Rotor stuck by sand	Improper sand production management	Perform workover/ flush-by
x	x				x			Pump	Rotor positioning in stator	Improper installation	Reposition rotor
		x		x	x			Pump	Rotor touching tag bar	Improper installation	Reposition rotor
x					x			Rods	Parted	Improper system design	Follow recommended design guidelines for load limits
x					x			Rods	Parted	Improper connection makeup	Follow recommended makeup procedures
x					x			Rods	Parted	Improper setting of drivehead torque limiter	Set torque limiter according to rod torque capacity
x	x				x			Tubing	Leak	Tubing wear because of improper system design	Install rod centralizers, use continuous rod
		x		x				Tubing	High pressure drop	Improper system design	Check for flow restrictions caused by centralizers
						x	x	Surface drive system	Belts slipping	Improper system installation	Repair surface drive system

Automation Systems in CSG Operations

Majority of CSG wells operated in Australia are equipped with automation and control systems, which allows operators to gather real-time data for operational and process control purposes. These systems are also responsible for safety and control of water/gas production process. A typical CSG well, as shown in Figure 4, comprises of three main components:

- Water and Gas Separator
- Progressive Cavity Pump
- A combined Variable Speed Drive (VSD) and Remote Telemetry Unit (RTU) Panel

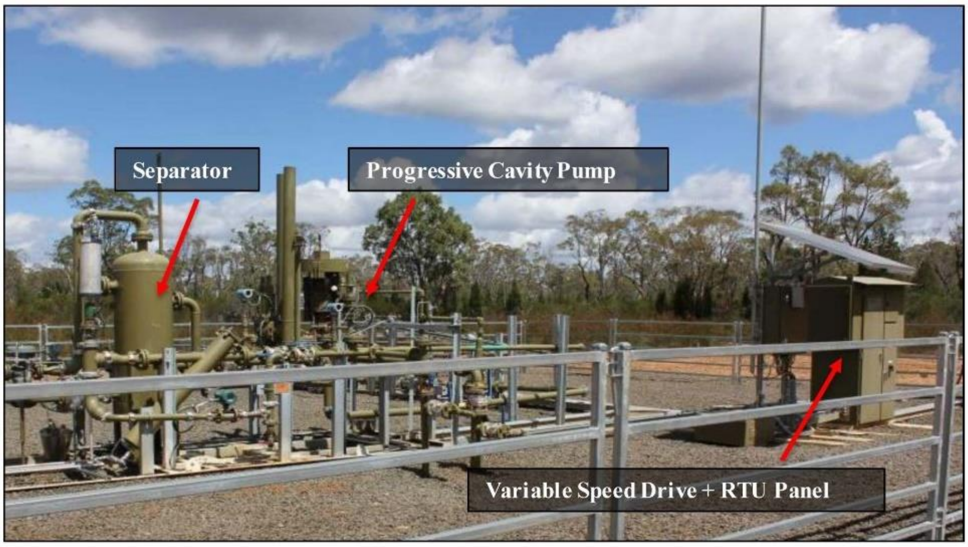


Figure 4—A typical CSG Well Facility Layout [8]

Primarily, real-time data enables operators to monitor and control production operations at well sites. With the aid of RTUs, the automation system controls the well in a desired operating envelope and invokes a shutdown or pump speed change when an unwanted situation is detected; such as pump-off [9], high torque, high pressure or any process related stoppage. Although such systems aid operators in ensuring that production is maintained at an optimal level; however, managing hundreds of PCPs can prove challenging, primarily if optimization must be achieved across multiple wells in a streamlined fashion.

Challenge with Monitoring CSG Wells in Near Real Time

Although exception-based surveillance has been used in the past to characterize PCP failure [10], there remains a significant challenge associated with detecting anomalous or abnormal PCP behavior in near real-time. Due to the sheer volume of data collected from wells, an approach is required to classify pump performance in near real-time and streamline identification of abnormal pump behavior. This near real time pump performance analysis can be achieved by implementing Symbolic Aggregation Approximation analysis on time-series data.

Symbolic Aggregation Approximation (SAX) is a normalization technique that enables improved performance analysis of time series data collected by automation systems. This technique segregates time series data into symbols, which are then converted into Heatmap images. These images provide an improved indication of PCP performance as they aid with the identification of abnormal pump behavior. This behavior is easily identified by a change in Heatmap image shape and color, hence enabling the detection of abnormal events.

Methodology

Research Data

Real-time PCP production data provided by two (2) CSG operators was analyzed as part of this research. Failure data provided by operator 1, covers a four (4) years period and shows 52% failure (Figure 5) in two-hundred (200) wells related to pump-off (pumped dry) and coal fine conditions (coal fines, torqued up, plugged intake and hole in tubing). A three (3) year data set provided by operator 2 shows almost 80%

failures (Figure 6) in forty-two (42) wells, where the symptoms are caused by coal fines (hole in tubing, plugged intake and plugged discharge).

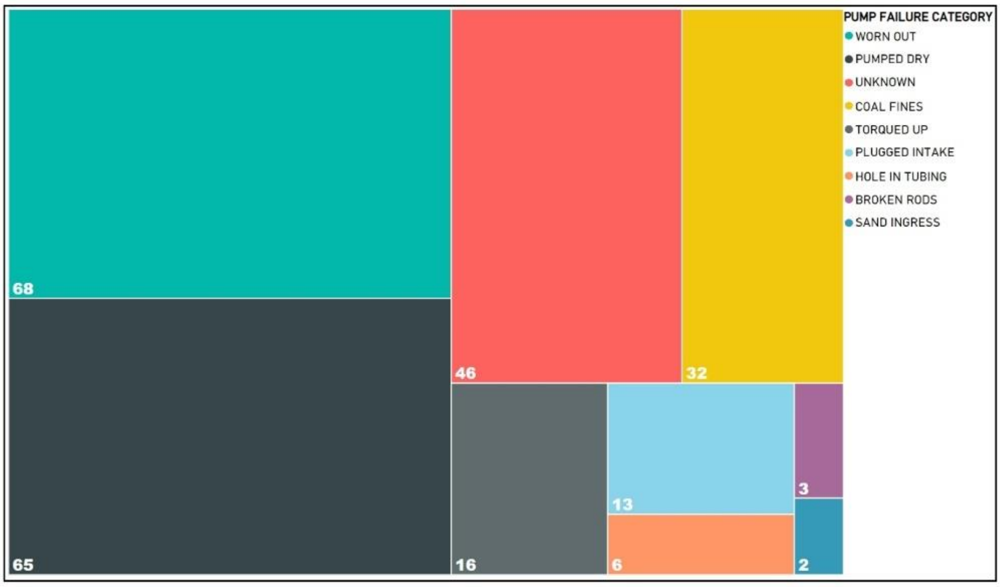


Figure 5—Pump Failures Data from Operator 1

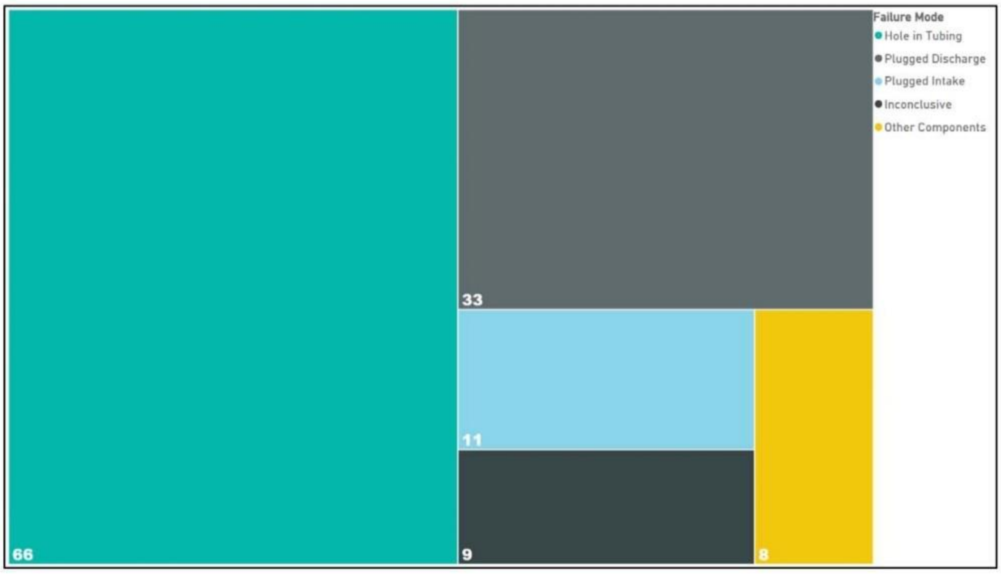


Figure 6—Pump Failures Data from Operator 2

Converting Time Series Data to SAX Symbols

Time-series data gathered by automation systems is usually stored in data historians. This data is sequentially organized and requires pre-processing before any analytics can be carried out. This section will describe a step-by-step process of converting time series data to heatmap images. It is assumed that the time series data has been quality checked and normalized as part of the exploratory data analytics exercise.

Selecting the right variables from a multivariate data-set. To achieve tangible outcomes from time series analysis, it is vital to choose a subset of variables from the data-set that account for maximum influence on the pump performance. Based on Table 1, it is observed that the behavior of torque and efficiency are primary indicators of PCP failure symptoms. As pump efficiency is not directly measured, we can replace it with the measured flow since efficiency is the ratio of actual versus theoretical flow. Furthermore, pump design calculations confirm a correlation between speed, torque, and flow parameters.

Correlation between flow and speed is shown in equation 1 [4].

$$q_{th} = s\,\omega \tag{1}$$

Where,
q_{th} *= theoretical flow,* s *= pump displacement,* ω *= rotational speed*

Correlation between torque and speed is shown in equation 2 [4].

$$T_{pr} = \frac{P_{pmo}E_{pt}}{C\omega} \tag{2}$$

Where,
T_{pr} *= polished rod torque,* P_{pmo} *= prime mover power,* E_{pt} *= power transmission efficiency,* C *= constant,* ω *= rotational speed*

Based on information from Table 1 and the correlation between speed, torque, and flow, we will use these three (3) measured variables to convert time series data into heatmap images.

Gaussian Distribution. This method is commonly used towards providing standard deviation and mean distribution in sequential datasets. Equation 3 [11] provides an overview of a bell curve based on the Gaussian Distribution function.

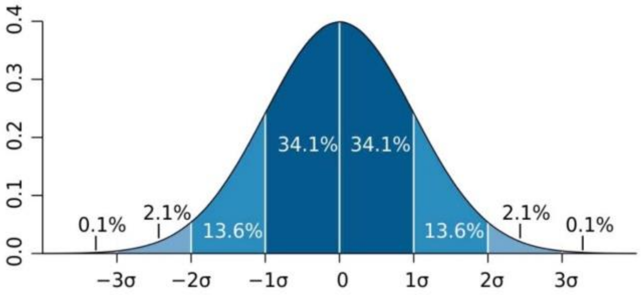


Figure 7—Gaussian Distribution (bell curve) based on Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\,e^{\frac{-(x-a)^2}{2\sigma^2}} \tag{3}$$

Where,
σ *= Standard Deviation,* a *= Mean*

Adding Breakpoints to Gaussian Distribution – Conversion to SAX Symbols. Once data is normalized, the Gaussian Distribution is divided into breakpoints. The number of breakpoints is determined based on the data spread, which is then divided into equidistant regions. These regions are based on the equation below [12],

$$\beta_{i+1} = \frac{1}{a} \tag{4}$$

Where,
a *= number of breakpoints or bins,* β *= Breakpoint (cut-off)*

Exploratory data analytics must be carried out on time series data before determining an optimal number of bins or SAX symbols. Mean, standard deviation, and outliers should be considered when deciding the number of bins. Figure 8 shows an example of time series data converted into regions of three (3) SAX characters.

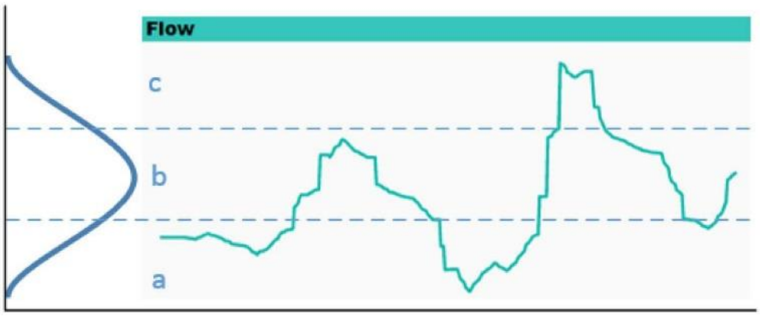
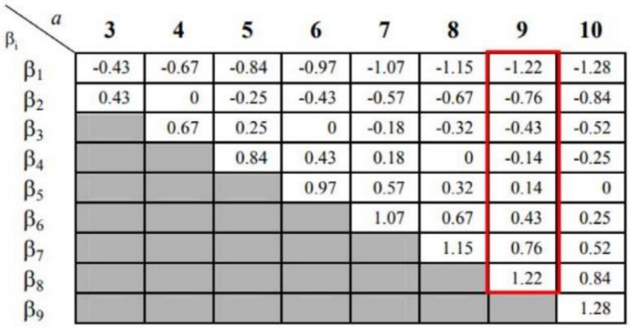


Figure 8—Time Series Data divided into 3-character bins

Table 2 shows the breakpoint regions for up to ten (10) bins. Nine (9) character bins are chosen for the dataset used in this research.

Table 2—Breakpoint regions based on Gaussian Distribution [12]



a β_i	3	4	5	6	7	8	9	10
β_1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15	-1.22	-1.28
β_2	0.43	0	-0.25	-0.43	-0.57	-0.67	-0.76	-0.84
β_3		0.67	0.25	0	-0.18	-0.32	-0.43	-0.52
β_4			0.84	0.43	0.18	0	-0.14	-0.25
β_5				0.97	0.57	0.32	0.14	0
β_6					1.07	0.67	0.43	0.25
β_7						1.15	0.76	0.52
β_8							1.22	0.84
β_9								1.28

Once the number of bins is decided, we can then convert our time series data into SAX character representation. Figure 9 shows an example where torque data is converted into nine (9) SAX characters. Each symbol is denoted by a unique color, which provides an informative view of torque performance during PCP run time.

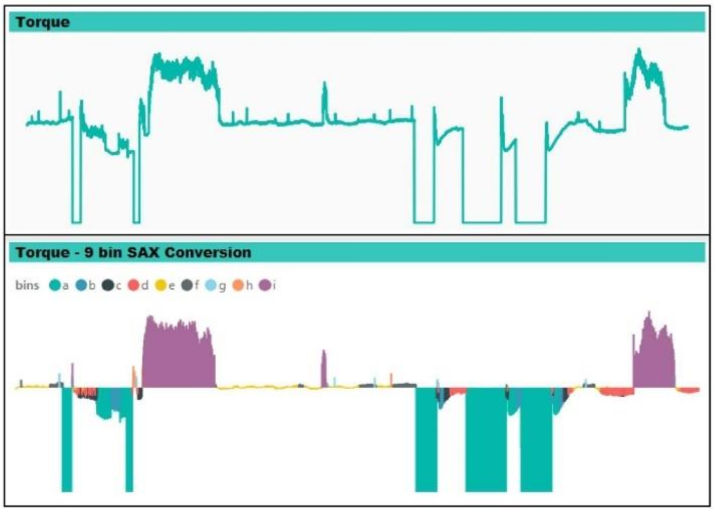


Figure 9—(Top) Time Series Torque Data, (Bottom) Time Series data converted to 9-character bins

Transforming SAX Derived Characters to Performance Heatmaps

SAX representations, obtained from time series data, are converted into a matrix representation. These matrix representations are then transformed into Heatmap images. Figure 10 (Left) shows the position of 'HIGH', 'NORMAL', and 'LOW' matrix cells, where each cell indicates the performance range based on SAX bins distribution. The top-tiered matrix cell signifies 'HIGH' performance range bin (*character i*), mid-tiered matrix cell signifies 'NORMAL' performance range bins (*characters h, g, f, e, d*), and bottom-tiered matrix cell signifies 'LOW' performance range bins (*character c, b, a*). Figure 10 (Center) shows the color-coding map where individual colors (red, yellow, green, and black)are applied to each matrix cell based on the time-based count of each symbol. A converted Heatmap is shown in Figure 10 (Right).

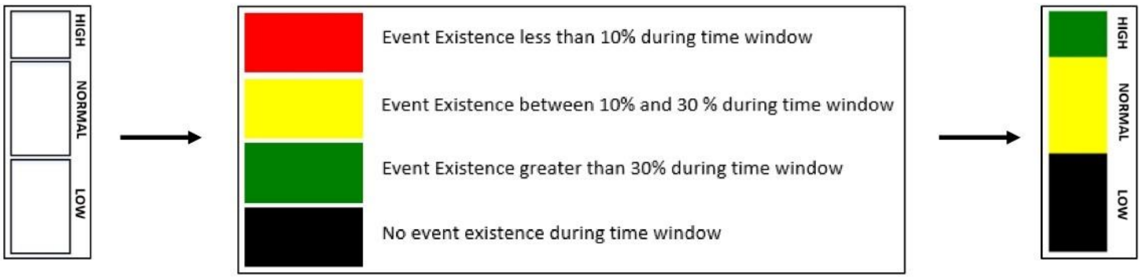

HIGH
NORMAL
LOW
Event Existence less than 10% during time window
Event Existence between 10% and 30 % during time window
Event Existence greater than 30% during time window
No event existence during time window
HIGH
NORMAL
LOW

Figure 10—(Left) Single Variable Heatmap showing HIGH, NORMAL and LOW-performance range, (Center) Time Distribution based Color Code for each Matrix Cell, (Right) Example of Converted Heatmap for Uni-variate Time Series Data

A sample conversion of flow time series data is shown in Figure 11, where the converted heatmap indicates performance varying between 'HIGH' and 'NORMAL' range. Character 'i' had the most counts (greater than 30%)in a single time-window and hence the 'HIGH' performance matrix cell is colored green. Characters 'e', 'f' and 'h' have a cumulative count in the range of 10% and 30%. Hence the 'NORMAL' performance matrix cell is colored yellow. A similar process for torque and speed conversion is shown in Figures 12 and 13, respectively.

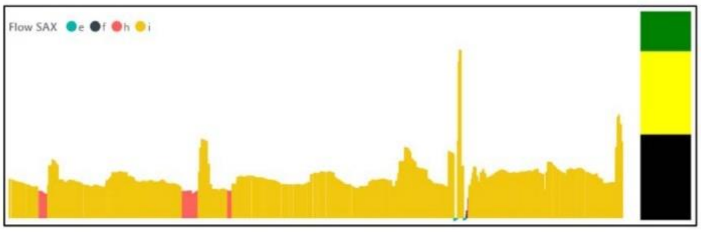


Figure 11—Flow Time Series Data conversion to HEATMAP Image

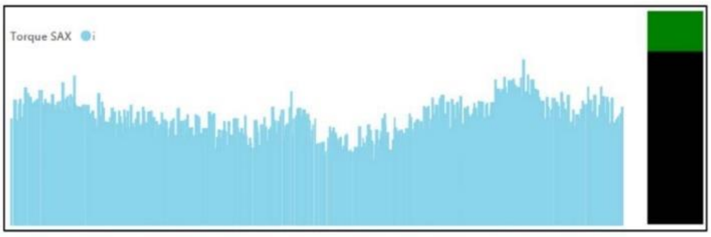


Figure 12—Torque Time Series Data conversion to HEATMAP Image

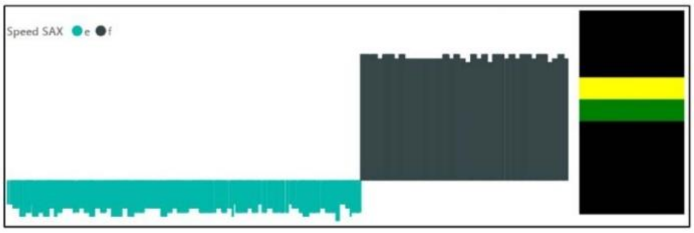


Figure 13—Speed Time Series Data conversion to HEATMAP Image

Individual color-coded Heatmaps for each variable (flow, torque, and speed) are then merged to form a multivariate Heatmap. Figure 14 shows a multivariate Heatmap for one (1) day time series data inclusive of flow, torque, and speed. Two events are observed during this twenty-four (24) hour period; the first event is a speed change, and the second event is unstable flow. Both events are captured by the multivariate Heatmap as per the color code shown in Figure 10. However, this Heatmap does not recognize if flow fluctuation occurred before or after the speed change. This drawback can be avoided if a smaller time window is used for SAX conversion and Heatmap generation.

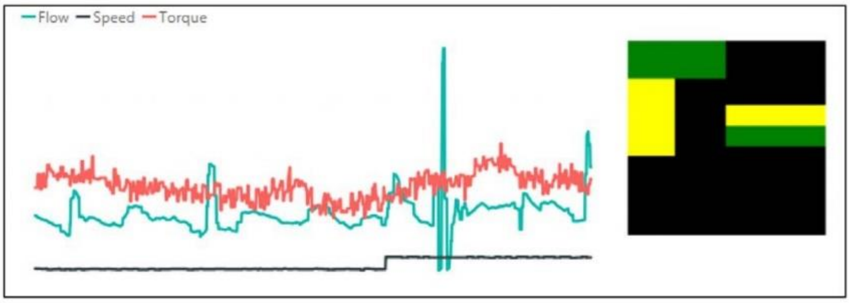


Figure 14—Multivariate Heatmap for a One (1) Day Time Window

Results

Detecting Abnormal PCP Performance Events with Heatmap Images

Once time series data is converted into a representation of multivariate Heatmap images, we can then use color-coded matrix cells to identify abnormal PCP behavior. Figure 15 shows the same one (1) day time series data, as in Figure 14, converted to twenty-four (24) Heatmap images.

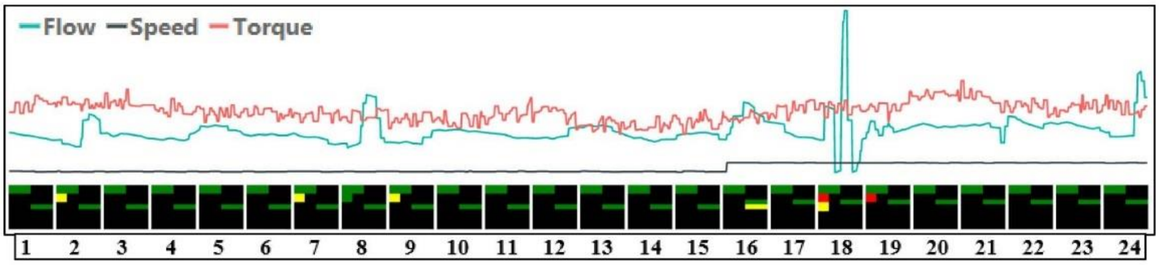


Figure 15—Multivariate Heatmap for a One (1) Hour Time Window over a One (1) Day Period

From the above figure, it is evident that Heatmaps generated for smaller time series windows better illustrate PCP performance. This added granularity provides improved visual analysis of the PCP performance, where minor changes in flow profile are easily recognized during a twenty-four (24) hour period. In hours 2, 7, and 9, the Heatmap images pick up variations in flow profile, which is characterized by the presence of yellow cells. Variation in flow is also detected during hours 18 and 19.A noticeable fluctuation in flow occurred during hour 18, which is highlighted by the presence of both yellow and red cells. These variations in flow performance occur without any change in speed or torque; hence, the behavior of PCP during hours 2, 7, 9, 18, and 19 can be categorized as abnormal.

Heatmap images can be tailored to capture events that have occurred for a fraction of the time during a one (1) hour period, i.e., events highlighted by red cells. This can be done by masking all other colors in the Heatmap except red and black. Figure 16 shows time series data where both original and masked Heatmaps are depicted with the multi-variate trend. Based on the masked Heatmaps, it is seen that torque events occur during hours 2, 11, 12, 16, and 24; whereas, a flow event occurs during hour 23 of this period. These events occurred without any change in pump speed; hence, these periods within the time series data can be marked as abnormal PCP performance and can be investigated further. Furthermore, torque event captured during hour 16 is not visible on the time series trend; however, SAX character conversion can capture minute variations in data that may be missed through visual inspection of data.

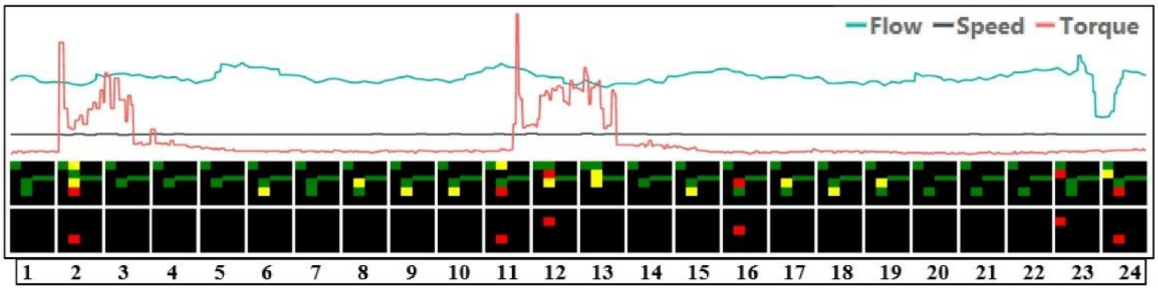


Figure 16—Original and Masked Heatmaps for Multi-variate Time Series Data

Heatmap masking technique can also be adapted to observe changes in majority events, i.e., events highlighted by green cells. This would allow operators to monitor change in PCP performance over the life of the pump by filtering out abnormal activity depicted by yellow and red cells.

Using K-Means Clustering to Automate Abnormal Event Detection

Heatmap images generated with SAX character conversion can be grouped based on the K-Means clustering method. This unsupervised machine learning method organizes unlabeled data-set (in this case Heatmap images) into clusters based on their similarity. This is done by computing centroids for each cluster, and images are then grouped in clusters based on the nearest centroid [13]. Machine Learning libraries in Python such as Sci-kit Learn provide access to the K-Means clustering algorithm which can be utilized to automatically group existing Heatmap images and match new Heatmap images to a cluster based on their distance from pre-computed centroids [14].

Figure 17 shows ten (10) Heatmap image clusters identified for forty-two (42) PCP wells. In this case, masked Heatmap images (red and black cells only) have been used to minimize the number of clusters. As the goal is to identify abnormal PCP behavior, newly created Heatmap images from streaming time series data can automatically be tagged as anomalous if they fall in clusters 2 through 10. Likewise, Heatmap images in cluster 1 can automatically be tagged as normal.

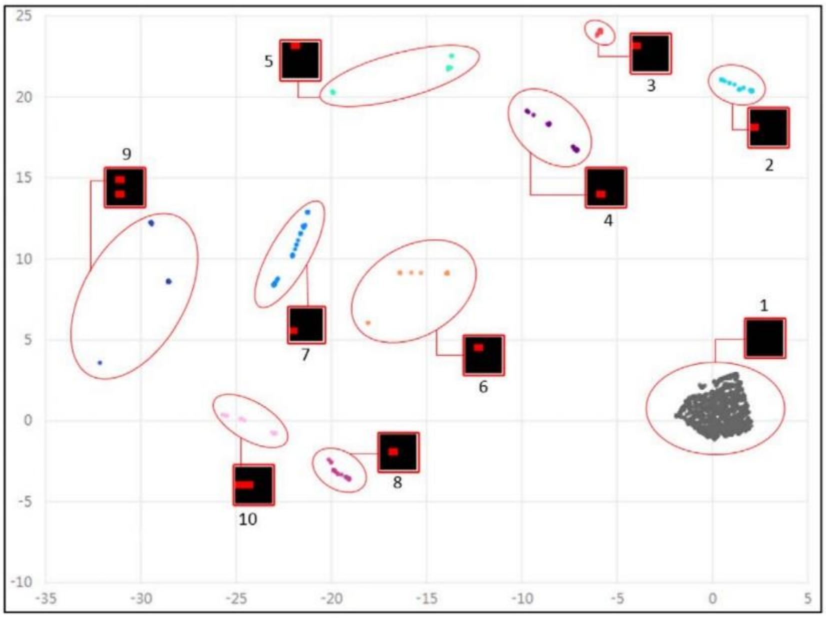


Figure 17—Heatmap Image Clusters Identified by K-Means Clustering

To improve abnormal event inference, clusters can be tagged based on the anomaly they represent. Clusters 4, 5, 6, 8 and 9 can be labelled "T" as they represent anomalous events due to torque. Similarly, clusters 2, 3, and 7 can be labelled "F" as they represent anomalous events due to flow. Cluster 10 can be labelled "T|F" as it represents an abnormal event due to flow and torque. Figure 18 shows an automatically labelled time-series trend based on the tag representation.

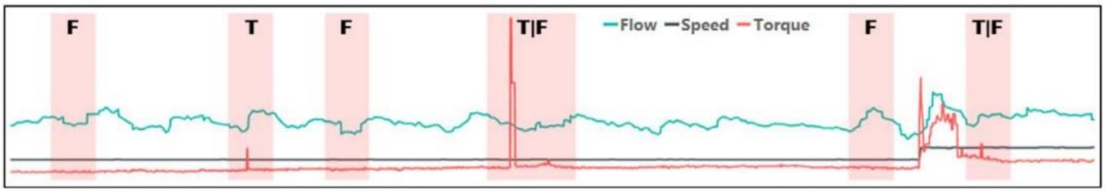


Figure 18—Automatically Labelled Time Series Data based on K-Means Clustering

Conclusion

In this paper, we have shown that converting time series data into SAX based Heatmap images not only provides improved visualization of real-time trends, but also enables identification of PCP performance events in near real-time. This unique methodology further allows the use of unsupervised machine learning techniques to cluster and label Heatmap images to automate abnormal event detection. By doing so, thousands of CSG wells can be analyzed autonomously in a streamlined manner, which can allow operators to focus on wells requiring attention and help capture performance trends that can in-turn assist with improving pump life and reducing failures.

Further Work

Research work is underway to examine how machine learning methods and image analysis techniques can further be improved to provide a more holistic PCP performance analysis. Supervised learning methods are being investigated where clustered images are labelled by an expert, and these labelled images can then be used to automate pump performance classification. Further image analytics techniques are also being investigated where algorithms can pick the difference in colored cells between two consecutive Heatmap images to more definitively categorize PCP performance factors such as drop or increase in flow rate or fluctuations in torque. Results from these investigations will be published in future conference papers and journals.

References

1. *Queensland's petroleum and coal seam gas*. 2017, Department of Natural Resources and Mines: www.dnrm.qld.gov.au.
2. Lea, J.F., Gas Well Deliquification. 2nd ed. *Gulf Drilling Guides*, ed. H.V. Nickens and M. Wells. 2011, Burlington: Elsevier Science.
3. Unsworth, N.J. and C. Sharratt, LNG industry faces challenge of monetizing huge CSG resources with Australia projects. *LNG Journal*, 2011.
4. Matthews, C.M., et al., Petroleum Engineering Handbook, in *Production Operations Engineering*. 2007, Society of Petroleum Engineers.
5. Gaurav, K., et al., Performance Analysis in Coal Seam Gas, in *SPETT 2012 Energy Conference and Exhibition*. 2012, Society of Petroleum Engineers: Port-of-Spain, Trinidad. p. 15.
6. Yaning, L., et al., A Case Study on Application of Progressive Cavity Pump in Coalbed Methane Wells, in SPE Asia Pacific Oil & Gas Conference. 2016, Society of Petroleum Engineers: Perth, Australia.
7. Vora, J., R. Singh, and S. Saxena, Novel Idea for Optimization of a Progressive Cavity Pump PCP System at Different Stages of Coal Bed Methane CBM Well Life, in SPE Oil and Gas India Conference and Exhibition. 2017, Society of Petroleum Engineers: Mumbai, India.
8. *Santos CSG well at the Narrabri Gas Project*. 2015: https://www.abc.net.au/news/2015-03-03/santos-csg-well/6277336.

9. Woolsey, K.A., Improving Progressing-Cavity-Pump Performance Through Automation and Surveillance. *Journal of Canadian Petroleum Technology*, 2012. **51**(01): p. 74–81.
10. Hoday, J.P., et al., Diagnosing PCP Failure Characteristics using Exception Based Surveillance in CSG, in SPE Progressing Cavity Pumps Conference. 2013, Society of Petroleum Engineers: Calgary, Alberta, Canada. p. 13.
11. Marx, M.L. and R. Larsen, *An Introduction to Mathematical Statistics and Its Applications*. 5th ed. 2012.
12. Lin, J., et al., Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 2007. **15**(2): p. 107–144.
13. Tan, P.N., M. Steinbach, and V. Kumar, *Cluster Analysis: Basic Concepts and Algorithms*. 2005. 487–568.
14. Pedregosa, F., et al., Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011. **12**: p. 2825–2830.

5. Paper 3: Machine Learning for Progressive Cavity Pump Performance Analysis: A Coal Seam Gas Case Study

The paper emphasizes the application of machine learning techniques in conjunction with innovative visualization methods for enhancing the analysis of performance in CSG wells. The methodology introduced in the paper centers around the conversion of time-series data into SAX Heatmap images. The Heatmaps serve as a powerful tool for visually depicting changes in PCP performance, allowing for a more intuitive understanding of the data.

The paper's standout feature is its application of machine learning techniques to the dimensionality reduction process. Principal Component Analysis (PCA) is employed for this purpose, reducing the complexity of the data. The resulting PCA components are visualized using t-Distributed Stochastic Neighbor Embedding (t-SNE), a machine learning technique that simplifies the representation of high-dimensional data in two dimensions.

The machine learning component is further advanced by the application of k-Means clustering, where the ideal number of clusters is derived using the *elbow method*. This clustering method categorizes the SAX Heatmap images based on their characteristics. It helps identify patterns and similarities within the data, ultimately enabling the detection of PCP performance and anomalies. These insights are gained in near real-time, offering significant advantages for operational management.

The visual representation of Heatmaps alongside time-series data plays a crucial role in understanding PCP performance changes. The clusters generated through k-Means are labelled based on the characteristics of the Heatmap images, such as torque or flow anomalies. This process allows for the automated tagging of new Heatmap images generated from streaming time-series data, making it a powerful tool for anomalous PCP performance detection.

The automated tagging of SAX images, based on the clustering method discussed, enhances the detection of anomalous PCP performance in streaming time-series data, highlighting the paper's contribution to effective anomaly detection and hence paving the way for manage-by-exception for a large fleet of CSG wells.

Statement of Authorship

Title of Paper	Machine Learning for Progressive Cavity Pump Performance Analysis: A Coal Seam Gas Case Study		
Publication Status	☑ Published □ Submitted for Publication	□ Accepted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style	
Publication Details	Saghir, F., Gonzalez Perdomo, M. E., and Behrenbruch, P. (2019c). Machine Learning for Progressive Cavity Pump Performance Analysis: A Coal Seam Gas Case Study. In: SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology Conference. 18–19 November,. Brisbane, Australia.		

Principal Author

Name of Principal Author (Candidate)	Fahd Saghir		
Contribution to the Paper	Conduct data analysis, write and record experiments, create test reports, tabulate results and write paper.		
Overall percentage (%)	75%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	19/09/2023

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Mary Gonzalez Perdomo		
Contribution to the Paper	Assisted with paper structure, writing and paper review (20%)		
Signature		Date	18/09/2023

Name of Co-Author	Peter Behrenbruch		
Contribution to the Paper	Assisted with paper structure, writing and paper review (5%)		
Signature		Date	10-09-2023

Please cut and paste additional co-author panels here as required.

URTEC-198281-MS

Machine Learning for Progressive Cavity Pump Performance Analysis: A Coal Seam Gas Case Study

Fahd Saghir, M.E. Gonzalez Perdomo, and Peter Behrenbruch, University of Adelaide

This paper was prepared for presentation at the SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology Conference held in Brisbane, Australia, 18 – 19 November 2019.

Abstract

Limited research work and publications are available to examine the performance of Progressive Cavity Pumps (PCP) based on machine learning methods, especially in Coal Seam Gas (CSG) operations. Previous work done in this space either focuses on exception-based surveillance on time-series data [1], or the use of machine learning to optimize completion design [2] and production [3].

This paper will discuss how data approximation and unsupervised machine learning methods can be applied to time-series data-sets, using data gathered from automation systems, to help analyze PCP performance and detect anomalous pump behavior.

Introduction

The majority of CSG wells operated in Australia are automated with Remote Telemetry Units (RTUs) and Supervisory Control and Data Acquisition (SCADA) systems [1, 2]. These automation systems gather production, mechanical, and electrical time-series data from PCPs, which allow operators to manage day-to-day production operations. However, SCADA systems are not suited to run advanced analytics and machine learning algorithms that can help determine PCP performance.

To exploit information from SCADA systems, a time-series based image conversion technique is utilized to aid with a better understanding of PCP performance. Machine learning based image classification techniques are applied to these converted images, where they are clustered based on t-Distributed Stochastic Neighbor Embedding (t-SNE) and k-Means algorithms (unsupervised learning).

Results from this study depict how time-series based heatmap conversion, coupled with unsupervised machine learning techniques can provide an innovative method to identify the abnormal behavior of PCPs in CSG wells. The findings discussed in this paper are based on three (3) years' worth of time series PCP data collected from forty-two (42) CSG wells.

Methodology

Time Series Data Approximation

Prior to initiating PCP performance analysis, it is important to extract relevant features from time-series data in order to facilitate work with machine learning methods. After completing the initial pre-processing steps [4], data approximation techniques should be employed to generalize time-series data. An approximated time-series data set reduces dimensionality while capturing key features.

Symbolic Aggregation Approximation (SAX) is a data estimation technique that converts time-series data to symbols based on a gaussian distribution [5]. The SAX symbols represent standard deviation based equidistant regions in a gaussian distribution curve. The number of regions is decided by the pre-selected number of SAX symbols chosen to represent the time series data-set, and the distribution of each region is based on the difference between the measured value and the overall mean. The conversion of time series data to nine (9) SAX symbols is shown in Figure 1.

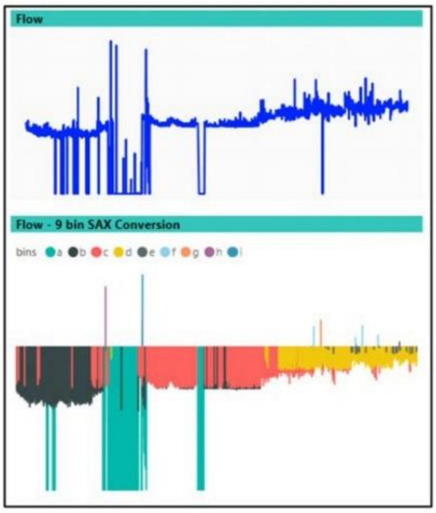


Figure 1—(Top) Time Series Flow Data, (Bottom) Time Series data converted to 9-character bins

Once time series data is converted into SAX representation, the symbol distribution for each measured variable (flow, speed, torque) can then be transformed into a Heatmap representation of the PCP performance [6]. A sample Heatmap transformation for a twenty-four (24) hour period is shown in Figure 2 below.

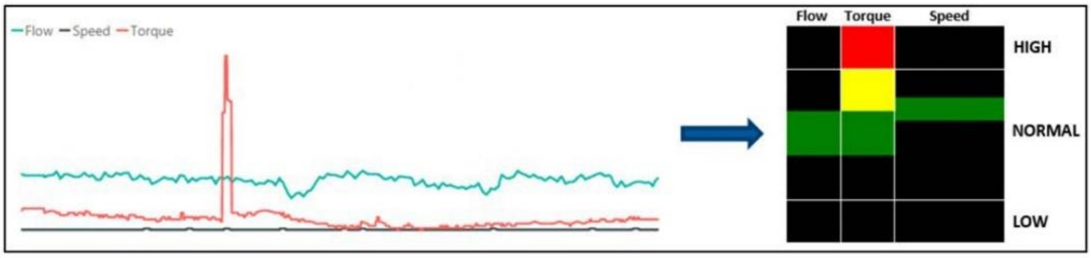


Figure 2—Time Series Data Conversion into Heatmap Image Representation

Heatmap Image Analysis

Time Series Heatmaps allow for improved analysis of PCP performance. The Heatmap conversion, shown in Figure 2, depicts different performance profile areas (high, normal, low) for flow, torque, and speed. The color code in each cell corresponds to the time-period for which the PCP operated in a particular performance profile.

The image shown in Figure 3 depicts a PCP Performance Heatmap generated for a twenty-four (24) hour period. This Heatmap provides a performance overview where flow, speed, and torque have remained in the normal range for a time-period of seven (7) hours or more during a single day. This time-period is indicated by the color green and signifies *Majority Performance*. Color yellow shows a performance *time-period* in the range of two (2) and seven (7) hours, and this is depicted in the Heatmap when the torque is in the higher tier of the normal profile. This Heatmap also depicts high torque for a time-period of two (2) hours or less; hence, this event is indicated with the color red.

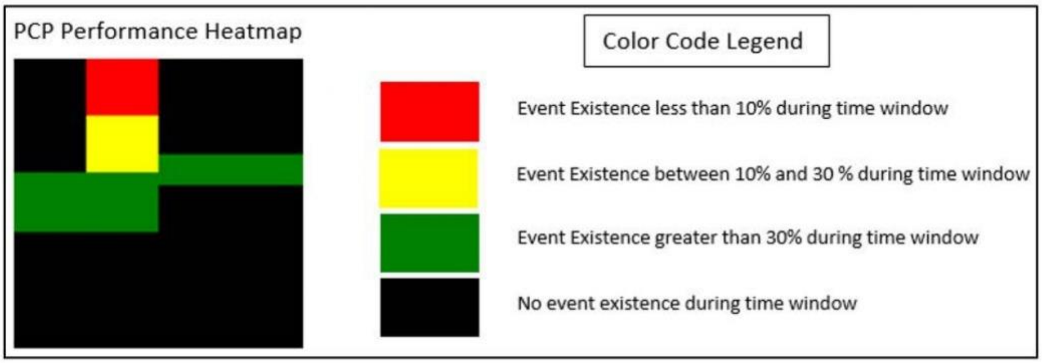

PCP Performance Heatmap
Color Code Legend
Event Existence less than 10% during time window
Event Existence between 10% and 30 % during time window
Event Existence greater than 30% during time window
No event existence during time window

Figure 3—PCP Performance Heatmap Color Code Overview

As shown in Figure 4, a single PCP Performance Heatmap can be further decomposed to *Majority Performance* (green time-period) and *Anomaly Event* (red time-period) Heatmap images. The yellow time-periods will be ignored as part of this study as they represent acceptable performance deviations during normal PCP operations.

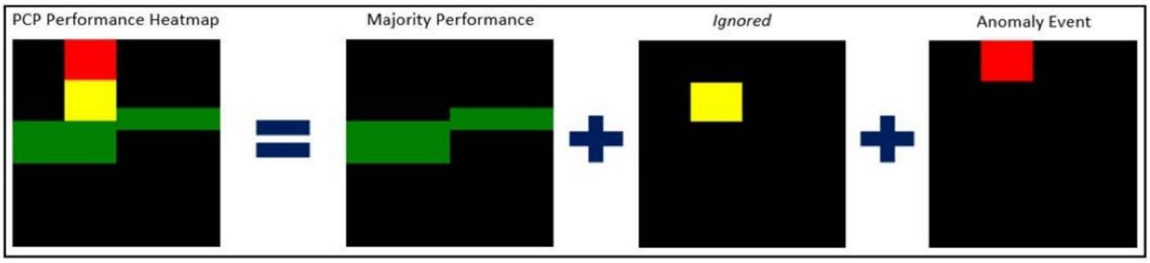


Figure 4—Time Series Heatmap Decomposition to *Majority Performance* and *Anomaly Event*

Difference Heatmaps. *Majority Performance* Heatmaps can aid with the understanding of PCP performance over the life of the pump. This can be achieved by creating *Difference Heatmaps*, where two (2) sequential Heatmap images can be subtracted from each other, to detect a change in the PCP performance profile.

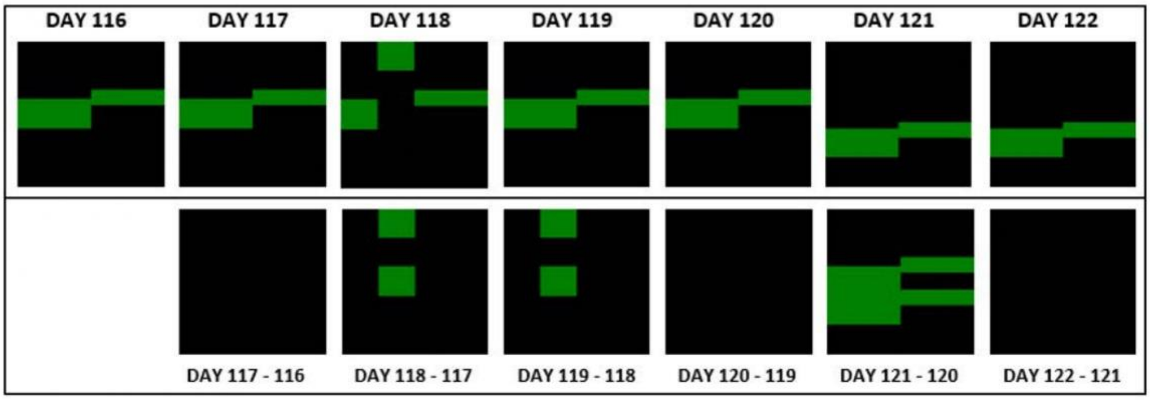


Figure 5—(TOP) Majority Performance Heatmap Images, (BOTTOM) Difference Heatmap Images

In Figure 5 above, the top row depicts *Majority Performance* Heatmaps over a period of seven (7) days. The bottom row shows *Difference Heatmaps* produced by subtraction of two (2) consecutive *Majority Performance* Heatmaps. Empty, or entirely black, *Difference Heatmaps* indicate no change in PCP performance between two (2) consecutive Heatmaps. *Difference Heatmaps* which show highlighted green cells indicate a change in PCP performance.

An interesting observation in the figure above is the change in PCP behavior over days 117, 118, and 119. The *Difference Heatmap* captures the first performance change on day 118 and then on day 119. The two (2) consecutive *Difference Heatmaps* (118-117 and 119-117) are identical, and this depicts that PCP performance temporarily changed between days 117 and 119. On the other hand, a permanent change in PCP performance is observed on day 121.

Machine Learning for PCP Heatmap Image Analysis

Once *Anomaly Event* and *Difference Heatmap* images are derived from time-series data, they can then be clustered via machine learning methods to identify PCP performance and capture anomalous events. Figure 6 outlines the process used to convert SAX images to k-Means clusters.

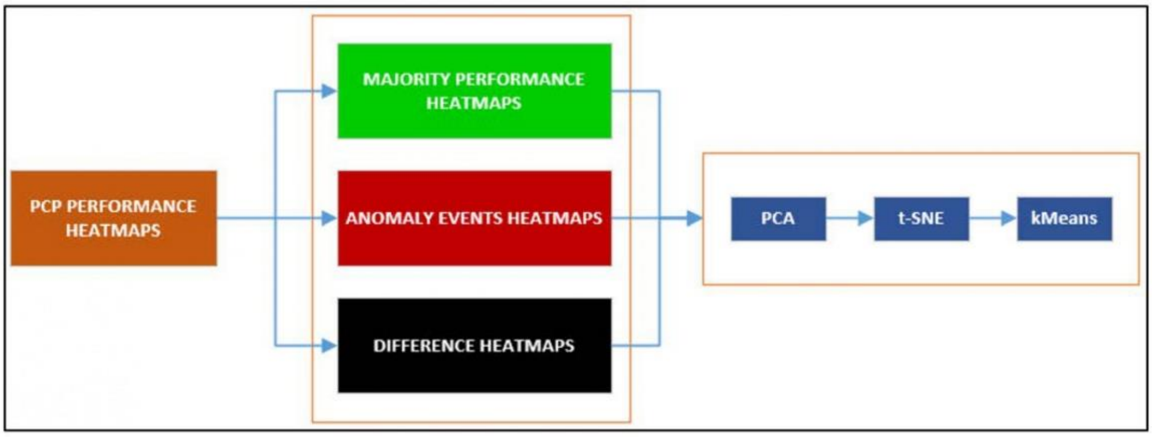


Figure 6—Overview of Unsupervised Machine Learning Process

Dimensionality Reduction through Principal Component Analysis (PCA). As the Heatmap images produced are *112 × 112 × 3* in dimension, they represent a total of *37,632* pixels or sample points. The first step to reducing the number of dimensions is to run a Principal Component Analysis of the images. Based on the plots in Figure 7, it is observed that twenty (20) components capture almost a hundred-percent (100%) variance for all three (3) Heatmap image types.

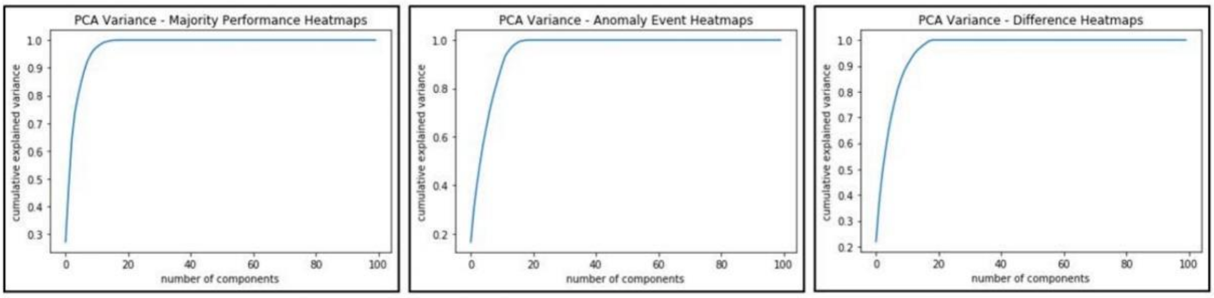


Figure 7—PCA Variance Plots for all three (3) Heatmap Image Types

t-Distributed Stochastic Neighbor Embedding (t-SNE). t-SNE allows for improved visualization of high-dimensional data [7], and based on the PCA results above, the twenty (20) components extracted from the Heatmap images can be visualized in two-dimensional space. Figure 8 illustrates the 2-D t-SNE visualization of twenty (20) components of each Heatmap type captured by the PCA process. This 2-D representation can further be used for k-Means clustering.

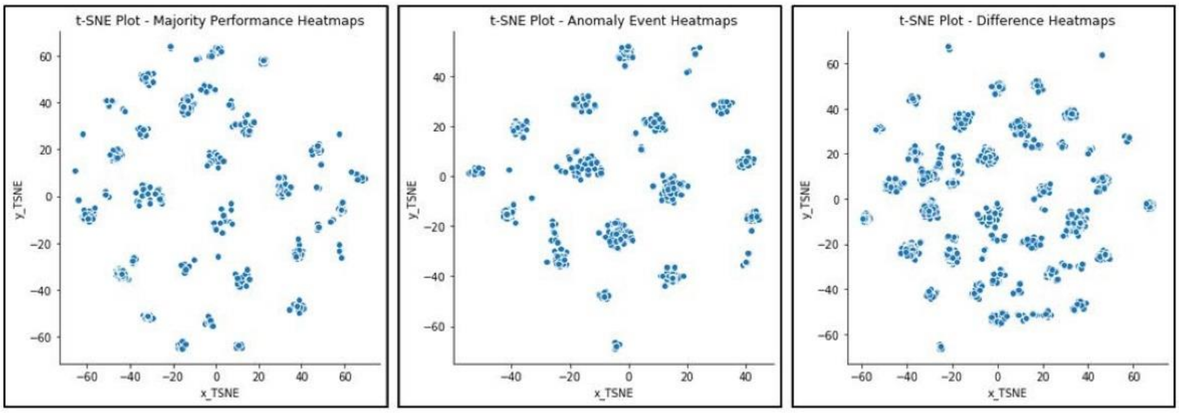


Figure 8—t-SNE 2-D Plots for all three (3) Heatmap Image Types

k-Means for Unsupervised Machine Learning. Prior to conducting k-Means clustering, it is important to determine the optimal number of clusters for the 2-D t-SNE representation of the Heatmap images. This can be achieved by using the elbow method [8] where a range of *k* values are used to compute distortion for each cluster number. The optimal number of clusters or *k* value is selected when the distortions are closer to zero. In Figure 9, the *k* value is determined from the line chart, based on the point of inflection on the curve.

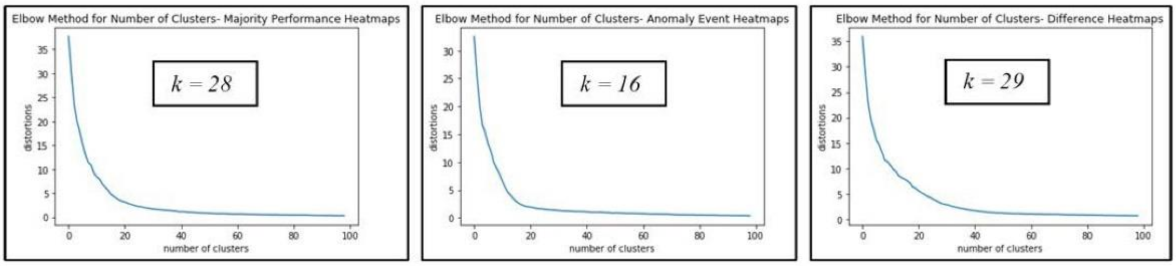


Figure 9—Elbow Method Curve to Determine Optimal *k* Value

Once the optimal number of clusters is determined, the k-Means algorithm is used to identify clusters within the 2-D t-SNE representation. Figure 10 shows the categorized clusters for each Heatmap image type. It is observed that both *Majority Performance Heatmaps* and *Difference Heatmaps* have a similar number of clusters based on the t-SNE plots. Comparatively, the cluster grouping for *Anomaly Event Heatmaps* is more compact compared to other Heatmap types. These clusters are based on historical time-series data gathered from forty-two (42) wells.

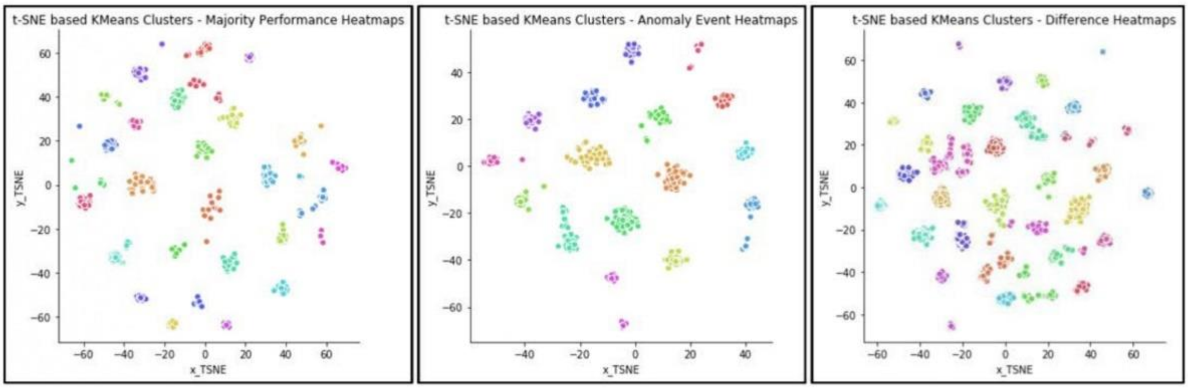


Figure 10—k-Means Clusters for the three (3) Heatmap Image Types

The demarcation of cluster groups for *Majority Performance Heatmaps* and *Difference Heatmaps* could be improved by choosing a higher number of clusters, but for this study, we will choose the clusters based on the elbow method. Also, the k-Means centroids generated during the unsupervised learning stage are recorded to label Heatmaps that will be generated by new time-series data gathered from automation systems.

Results

Image Sequence Visualization

SAX based Heatmaps improve visualization of time-series data, allowing for improved analysis of data in near real-time. As shown in Figure 11, plotting time-series trends in tandem with Heatmap images aid with the analysis of PCP performance and anomalous events.

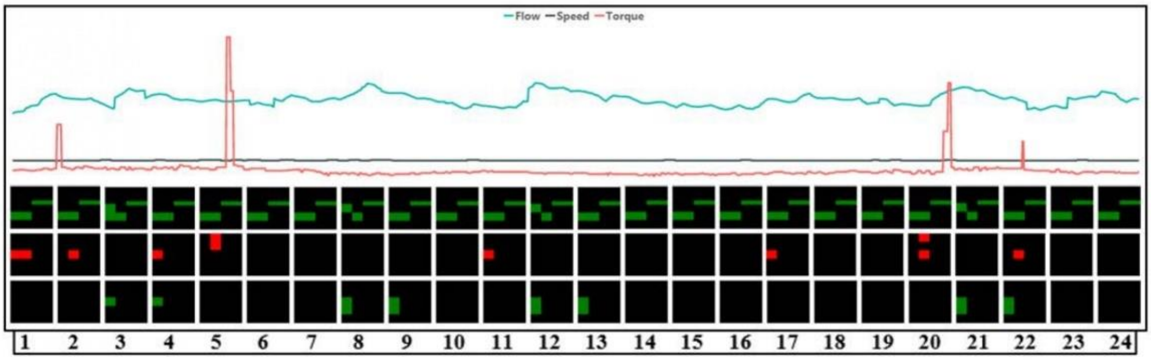


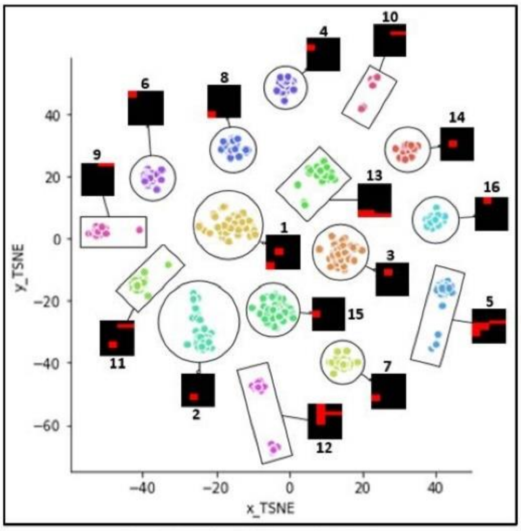
Figure 11—Time Series Trend Plotted with various Heatmaps. (TOP) Time Series Trend, (ROW 1) ***Majority Performance Heatmaps*****, (ROW 2)** ***Anomaly Event Heatmaps*****, (ROW 3)** ***Difference Heatmaps***

Figure 11 shows a twenty-four (24) hour trend, where the Heatmaps are generated on an hourly basis. *Anomaly Event* Heatmaps, which are shown in *ROW 2* (Figure 11), are able to pick up anomalous torque peaks (hours 2, 5, 20, 22), and also flow related anomalous events (hours 11, 12). A torque-flow anomalous event is also detected for this time-series data (hour 1).

Other than the anomalous events, the overall PCP performance is stable for the twenty-four (24) hour period shown in Figure 11. Although the *Difference Heatmaps* show deviations (hours 3+4, 8+9, 12+13, 21+22), these deviations occur in pairs, which indicates that the PCP performance divergence was temporary.

Image Clustering for PCP Performance Analysis

For each Heatmap type, the cluster groups can be labeled based on their image characteristics. Figure 12 shows different *Anomaly Heatmap* images as represented by the cluster numbers. The figure also shows that it is simple to identify which anomaly is represented by each cluster.



Identified by k-Means Clustering

Heatmap Image Clusters

Figure 12—Anomaly Heatmaps

Clusters 4, 6, 7, and 8 can be labeled as "F" as they represent flow anomaly. Clusters 2, 3, 14, and 16 can be labeled as "T" as they represent torque anomaly. Cluster 1 can be labeled as "T|F" as it represents torque-flow anomaly. Clusters 5, 9, 10, 11, 12, and 13 are not labeled, as these Heatmaps represent a speed change, and are not considered as anomalies. Speed changes are either controlled manually (by an operator) or autonomously (through SCADA control logic); hence, they do not represent an anomaly.

k-Means centroids recorded during the unsupervised learning phase are used to label new Heatmaps generated via streaming time-series data. As new data is processed and Heatmap images are generated, the recorded centroids are matched against these images to assign them a cluster number. These cluster numbers are allotted a label as described in the previous paragraph. Figure 13 shows how this cyclical process assists with the autonomous tagging of anomalous events on streaming time-series data.

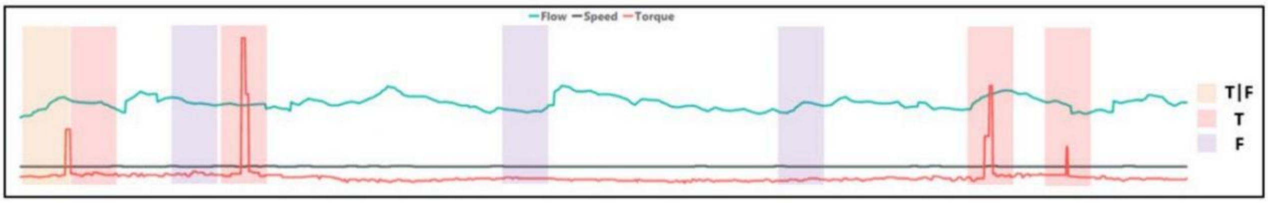


Figure 13—Autonomous Tagging of Streaming Time-Series Data

Conclusion

In this paper, we have described how machine learning methods, when applied to time-series based Heatmap images, can improve PCP performance analysis. In particular, the methods described in this paper can assist with detecting anomalous events on streaming time-series data and assist CSG operators with managing their wells.

Although examples in this paper are based on Heatmaps generated on an hourly basis; the process of autonomous tagging through machine learning methods can also be applied to Heatmaps generated over shorter time periods.

References

1. Hoday, J.P., et al., Diagnosing PCP Failure Characteristics using Exception Based Surveillance in CSG, in SPE Progressing Cavity Pumps Conference. 2013, Society of Petroleum Engineers: Calgary, Alberta, Canada. p. 13.
2. Prosper, C. and D. West, Case Study Applied Machine Learning to Optimise PCP Completion Design in a CBM Field, in SPE Asia Pacific Oil and Gas Conference and Exhibition. 2018, Society of Petroleum Engineers: Brisbane, Australia. p. 10.
3. Silva, T., et al., Achieving Production Optimization Using Progressive Cavity Pumps, Artificial Neural Networks, and System-Based Monitoring, in SPE Heavy Oil Conference-Canada. 2013, Society of Petroleum Engineers: Calgary, Alberta, Canada. p. 15.
4. Saghir, F., M.E.G. Perdomo, and P. Behrenbruch, Application of Exploratory Data Analytics (EDA) in Coal Seam Gas Wells with Progressive Cavity Pumps (PCPs), in SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition. 2019, Society of Petroleum Engineers: Bali, Indonesia.
5. Lin, J., et al., Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 2007. **15**(2): p. 107–144.
6. Saghir, F., M.E.G. Perdomo, and P. Behrenbruch, Converting Time Series Data into Images: An Innovative Approach to Detect Abnormal Behavior of Progressive Cavity Pumps Deployed

in Coal Seam Gas Wells, in Annual Technical Conference and Exhibition. 2019, Society of Petroleum Engineers: Calgary, Canada.

7. Maaten, L.v.d. and G. Hinton, Visualizing data using t-SNE. *Journal of machine learning research*, 2008. **9**(Nov): p. 2579–2605.
8. Gove, R. *Using the elbow method to determine the optimal number of clusters for k-means clustering*. 2017 [cited 2019; Available from: https://bl.ocks.org/rpgove/0060ff3b656618e9136b.

6. Paper 4: Application of machine learning methods to assess progressive cavity pumps (PCPs) performance in coal seam gas (CSG) wells

This paper advances the application of machine learning methods for assessing the performance of PCPs in CSG wells. The paper incorporates a neural network-based dimensionality reduction method to reduce the Heatmap image representation using Convolutional Auto Encoders (CAE). Moreover, the study employs an enhanced high-density-based clustering method HDBSCAN, to create clusters. Data from 359 PCP wells was used to develop the method and validate results for this paper.

To analyze a large amount of data used for the research, the CAEs were employed instead of the PCA approach discussed in previous papers. The CAEs provided a more compressed latent representation of the SAX heatmap images compared to PCA. Additionally, the k-Means clustering method was replaced with HDBSCAN. This was done to identify clusters without providing a pre-set number of clusters, which is the case with the k-Means approach.

This paper also introduces some key concepts and techniques which are pivotal to real-time analysis of PCP performance. Firstly, it introduces the "expanding window technique," which is fundamental for the analysis of PCP performance over its entire operating life. Second, the visual analytics approach to assess PCP performance is discussed, where multivariate trends are plotted against cluster heatmap labels as bar charts to aid in assessing changing ALS performance.

These innovative additions distinguish this research and underscore its significance in the field of PCP performance analysis using machine learning. This study showcases how utilizing machine learning to automatically identify and plot performance heatmap clusters, along with streaming data, provides engineers with an improved overview of PCP performance. This aids in real-time anomaly detection and automation.

Statement of Authorship

Title of Paper	Application of machine learning methods to assess progressive cavity pumps (PCPs) performance in coal seam gas (CSG) wells
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Saghir F, Gonzalez Perdomo ME, Behrenbruch P (2020) Application of machine learning methods to assess progressive cavity pumps (PCPs) performance in coal seam gas (CSG) wells. APPEA J 60(1):197–214

Principal Author

Name of Principal Author (Candidate)	Fahd Saghir		
Contribution to the Paper	Conduct data analysis, write and record experiments, create test reports, tabulate results and write paper.		
Overall percentage (%)	75%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	19/09/2023

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Mary Gonzalez Perdomo		
Contribution to the Paper	Assisted with paper structure, writing and paper review (20%)		
Signature		Date	18/09/2023

Name of Co-Author	Peter Behrenbruch		
Contribution to the Paper	Assisted with paper structure, writing and paper review (5%)		
Signature		Date	10-09-2023

Please cut and paste additional co-author panels here as required.

The APPEA Journal 2020, **60**, 197–214
https://doi.org/10.1071/AJ19044

Peer reviewed paper

Application of machine learning methods to assess progressive cavity pumps (PCPs) performance in coal seam gas (CSG) wells

Fahd Saghir[A,C], *M. E. Gonzalez Perdomo*[A] *and Peter Behrenbruch*[B]

[A]Australian School of Petroleum and Energy Resources, Santos Petroleum Engineering Building,
University of Adelaide, SA 5005, Australia.
[B]Bear and Brook Consulting, 135 Hilda Street Corinda, Qld 4075, Australia.
[C]Corresponding author. Email: fahd.saghir@adelaide.edu.au

Abstract. In Queensland, progressive cavity pumps (PCPs) are the artificial lift method of choice in coal seam gas (CSG) wells, and this choice of artificial lift production stems from the ability of PCPs to better manage the production of liquids with suspended solids. As with any mechanical pumping system, PCPs are prone to natural wear and tear over their operational life, and with the production of coal fines and inter-burden, the run life of PCPs in CSG wells is significantly reduced. Another factor to consider with the use of PCPs is their reliability. As per the CSG production data available through the Queensland Government Data Portal, there are approximately 6400 wells operational in the state as of December 2018. This number is expected to grow significantly over the next decade to meet both international and domestic gas utilisation requirements. Operators supervising these wells rely on a reactive or exception-based approach to manage well performance. In order to efficiently operate thousands of PCP wells, it is pertinent that a benchmark methodology is devised to autonomously monitor PCP performance and allow operators to manage wells by exception. In this study, we will cover the application of machine learning methods to understand anomalous PCP behaviour and overall pump performance based on the analysis of multivariate time-series data. An innovative time-series data approximation and image conversion technique will be discussed in this paper, along with machine learning methods, which will focus on a scalable and autonomous approach to cluster PCP performance and detection of anomalous pump behaviour in near real-time. Results from this study show that clustering real-time data based on converted time-series images helps to pro-actively detect change in PCP performance. Discovery of anomalous multivariate events is also achieved through time-series image conversion. This study also demonstrates that clustering time-series data noticeably improves the real-time monitoring capabilities of PCP performance through improved visual analytics.

Keywords: artificial lift, data analytics, time-series data, visual analytics.

Received 12 December 2019, accepted 23 January 2020, published online 15 May 2020

Introduction

Historical time-series data from 359 wells was utilised to perform the work undertaken as part of this research. This data was provided by two coal seam gas (CSG) operators and spanned a period of three and a half years. Multivariate data from these wells was processed and analysed to create performance heatmap images (Saghir *et al.* 2019*b*), where the images were used to train time-series classification models. This image-based time-series data classification method serves two purposes; first, it helps in the identification of anomalous progressive cavity pump (PCP) behaviour. Second, it assists with gauging the PCP performance over pump run life.

This study will describe the conversion of multivariate time-series data (flow, torque and speed) into performance heatmaps based on the Symbolic Aggregation Approximation (SAX) technique. We will also examine the use of Convolutional Auto Encoders (CAE) and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) methodologies to characterise multivariate time-series data. The conversion of multivariate time-series data to heatmaps also provides improved visual aid to petroleum or well surveillance engineers that can assess varying pump performance and assist with pro-active well management.

Background

Machine learning methods have been applied towards exploration and production-related problems within the CSG operations in Australia. These efforts have mainly focused on well completion design (Prosper and West 2018), drilling and logging activities (Zhong *et al.* 2019) as well as estimation of well production potential (Biniwale and Trivedi 2012). However, there is no research work available on the use of machine learning methods for identifying PCP performance and anomalous behaviour, especially in the case of CSG wells. This is true for both historical and real-time datasets.

Other approaches have been used that enable well surveillance engineers to manage PCP wells by exception. Methods such as exception-based surveillance (EBS) (Hoday *et al.* 2013) have been utilised to monitor PCP performance in near real-time; however, these exceptions are based on set rules or alarms (derived from operator experience), and they rely on availability of bottom hole pressure for calculating failure indicators (displacement and friction performance indicators). The drawback of the EBS approach is its dependence on bottom hole pressure, which is measured by downhole gauges, and not all operators deploy downhole pressure gauges in their CSG wells. These gauges are expensive to install, have a high probability of failure and are not easy to replace or maintain (Firouzi and Rathnayake 2019). Hence, this study will use multivariate sensor data from surface instruments and electrical equipment, which are less prone to failure and easy to calibrate and manage.

Multivariate data – selection of flow, torque and speed as key performance variables

The approach outlined in this study looks at three multivariate data points: flow, torque and speed. These three key performance parameters are selected based on the PCP mechanical design and correlations derived from historical data analysis.

Based on the mechanical design calculations for PCPs, Eqn 1 (Matthews *et al.* 2007) provides a relationship between speed and flow:

$$s_{min} = \frac{q_a}{\omega E} \tag{1}$$

where s_{min} = minimum required pump displacement, q_a = required flow rate, ω = pump rotational speed and E = volumetric pump efficiency.

Eqn 2 (Matthews *et al.* 2007) provides a relationship between torque and speed:

$$P_{pmo} = \frac{CT_{pr}\omega}{E_{pt}} \tag{2}$$

where P_{pmo} = prime mover power output, C = constant, T_{pr} = polished-rod torque, ω = pump rotational speed and E_{pt} = power transmission system efficiency.

A correlation analysis was also conducted to further support the utilisation of flow, torque and speed as the primary multivariate data points for this study. The correlation matrix shown in Fig. 1 was derived from a set 42 random wells, and it illustrates the high dependencies between flow, torque and speed data points.

Moreover, PCP pump failure in CSG wells is most commonly attributed to high torque due to build-up of coal fines in pump cavities (Matthews *et al.* 2007; Saghir *et al.* 2019*b*). Torque is also considered to be an early indicator of other PCP failures, such as blockage at the pump intake or discharge, parted rods and slug build-up. Hence, as shown through the PCP design equations and data correlation, we will use only three data points (flow, torque and speed) to conduct PCP performance analysis and anomaly detection via conversion of time-series data into performance heatmaps.

Related work

As this study covers the analysis of real-time PCP production data, we will utilise this section to briefly discuss published work that encompass methodologies that describe time-series data clustering. The discussion is divided into two parts: (1) time-series data approximation and (2) time-series data clustering.

Time-series data approximation

Although various techniques have been used for high-level representation of time-series data, which mostly focuses on using linear approximations (Dan *et al.* 2013; Bettaiah 2014; Duvignau *et al.* 2018) or wavelet transformation functions (Chaovalit *et al.* 2011; Shaw *et al.* 2015), the SAX technique (Lin *et al.* 2007) is considered the most practical due to its consideration of the temporal nature of time-series data. The SAX technique uses Gaussian distribution to approximate time-series data into *bins*, where equally distributed regions are assigned symbols or characters, thereby reducing the dimensionality of the time-series dataset to a set character range. Fig. 2 shows the approximation of a univariate time-series dataset to nine equally distributed symbol regions.

Once time-series data is approximated with the SAX technique, it is easy to identify anomalies and patterns in the dataset. To do this, multiple methods have been suggested, such as visualising discrete letter sequences through bitmap images (Kumar *et al.* 2005) or using motif discovery to identify patterns within the time-series datasets (Lin *et al.* 2005; Sivaraks and Ratanamahatana 2015; Guigou *et al.* 2017; Gao and Lin 2018). However, these methods are limited to univariate datasets and they assume that time-series data is uniformly recorded. The PCP performance heatmap method employed in this paper will discuss how these shortcomings can be addressed and make a case for utilising SAX representation when streaming time-series data.

Time-series data clustering

Another critical facet of anomaly detection and performance analytics is the process of time-series data clustering. Clustering helps with identifying varying patterns in time-series datasets. However, due to the temporal nature of time-series data, applying classical clustering algorithms will produce inaccurate results (Vidaurre *et al.* 2014). Therefore, neural network based clustering methods (Hatami *et al.* 2017; Ali *et al.* 2019) have become popular in recent times. Using a neural network based clustering method involves dimensionality

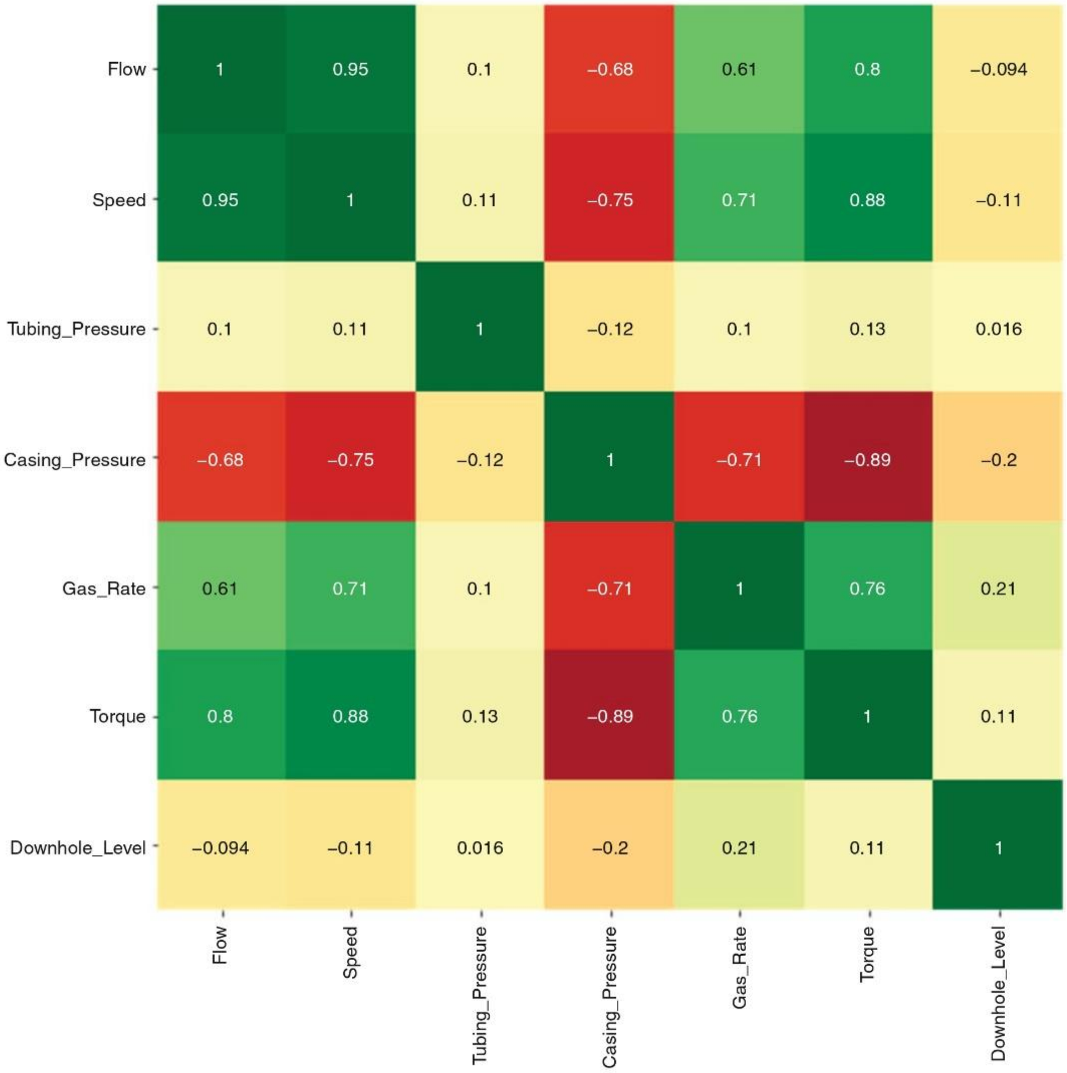


Fig. 1. Correlation matrix depicting high dependencies between flow, torque and speed multivariate data points.

reduction and feature extraction of time-series datasets. This can be achieved in two ways: (1) by feeding time-series data tables directly into a neural network model or (2) by converting time-series data in images.

There are three methodologies that describe the conversion of time-series data into images: recurrence plots (RP) (Wang and Oates 2015), Gramian Angular Fields (GAF) and Markov Transition Fields (MTF) (Hatami *et al.* 2017). Each method works towards creating a matrix representation of time-series data, which is then converted to images.

Although it is not the purpose of this study to compare the results of the image conversion techniques, it is worth mentioning the limitations of the aforementioned methods. Similar to the limitations of data approximation techniques, these image conversion methods have no inherent ability to manage multivariate time-series data. They require a multi-channel neural network approach to overcome this limitation. The requirement for uniformly recorded time-series data is also valid for these methods.

Methodology

Data transformation and preparation

Data transformation and preparation is compulsory when working with time-series datasets and is typically the most tedious task. Sensor data, gathered from industrial control systems, is stored in a hierarchical configuration where streaming data is stored in the order it is recorded (Saghir

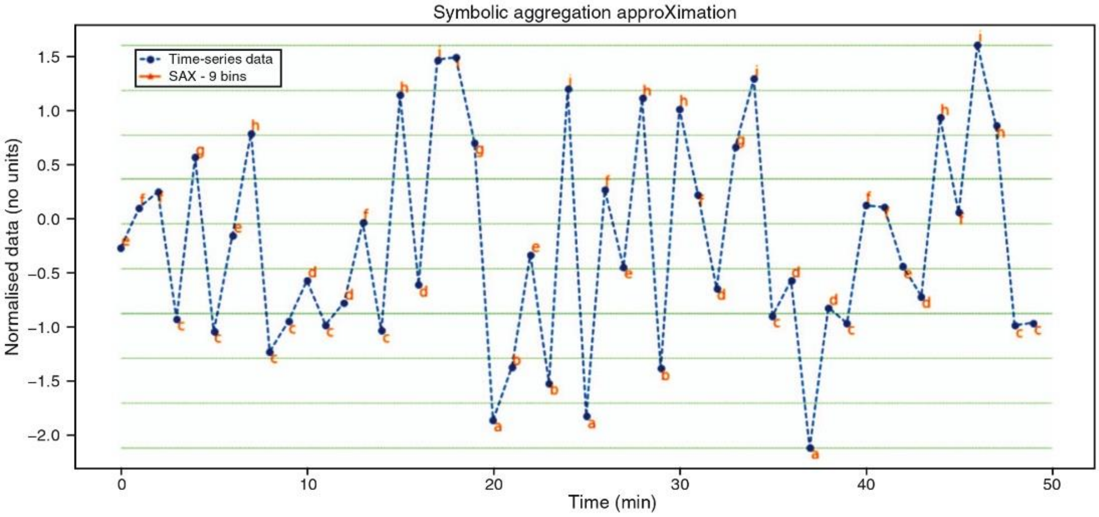


Fig. 2. An example of the SAX technique, where univariate time-series data is approximated to nine symbols.

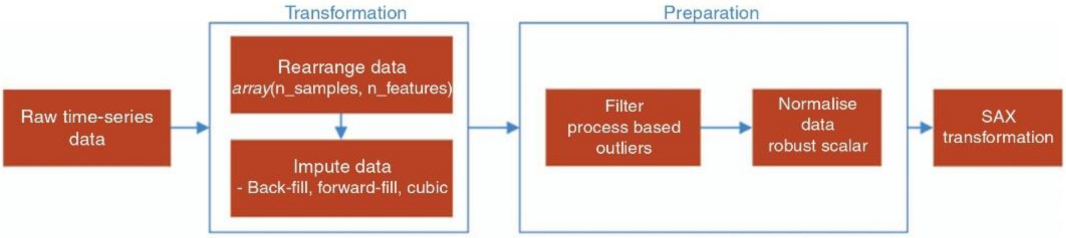


Fig. 3. Flow chart depicting data transformation and preparation steps before conducting SAX transformation.

et al. 2019*a*). Popular data science programming languages process and store data as arrays or vectors (McKinney 2013); hence, the transformation of streaming time-series data is necessary before conducting any detailed analytics (Fig. 3).

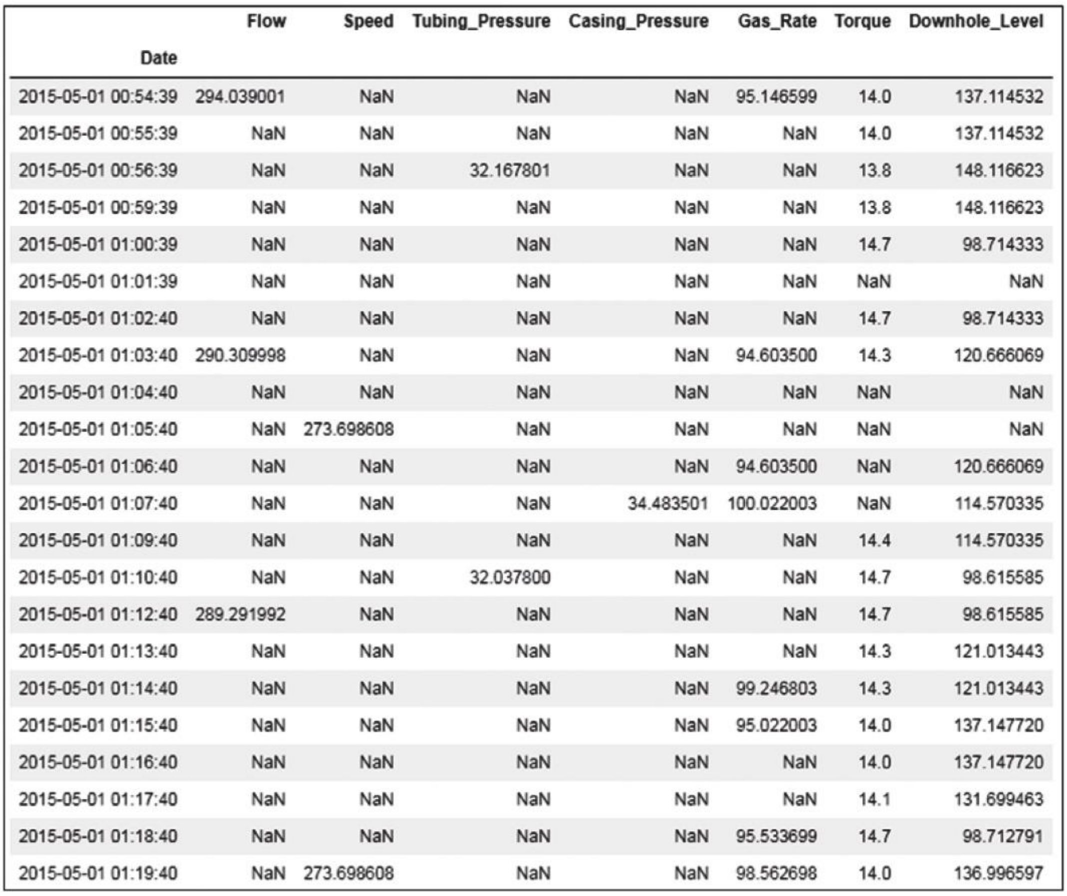
As part of the transformation process, raw time-series data is first rearranged into an array format. Traditionally, time-series data is non-uniform, i.e. not all sensor values are recorded at each time step (or sampling rate) as some values are measured faster than others. Hence, the conversion of raw time-series data to an array format will produce time-step gaps where all values are not recorded at the same interval. A gap laden array is shown in Fig. 4, where it is evident that some values are measured at a higher frequency than others.

The gaps in transformed arrays are filled by using imputation or fill methods that are readily available in machine learning software libraries (Pedregosa *et al.* 2011). An essential factor to consider while filling missing (or NaN) entries is to know the measured variable characteristics. Analog variables (such as flow and torque) should be filled using a cubic imputation method, as this method matches the measurement characteristics. Digital variables (such as speed) should be filled using a forward-fill method, as this characterises step-wise change, i.e. variables change only when a command is initiated to modify them. A filled array is shown in Fig. 5.

Once the time-series array is transformed, data is then filtered and normalised. Process-based outliers, where sensor measurements are out-of-bounds, are deleted from the array based on set limits. The normalisation of time-series data is achieved through adjusting sensor values to a standard scale and removing variability in measurement range across sensors. This ensures identical feature extraction across different wells during the SAX transformation process.

Time-series data to SAX symbols conversion – expanding window method

Once data preparation and transformation is complete, multivariate datasets are then converted to SAX symbols based on an aggregate window method. Usually, data analysis is done on a sliding window-based method. Analysis based on the sliding window methodology captures events, features and statistical information from a pre-defined time window with a fixed stride movement. A limitation of this approach is that statistical information (mean, average, variance, etc.) is confined to the window length alone. While using SAX feature extraction for multivariate time-series analysis, a sliding window, as



	Flow	Speed	Tubing_Pressure	Casing_Pressure	Gas_Rate	Torque	Downhole_Level
Date							
2015-05-01 00:54:39	294.039001	NaN	NaN	NaN	95.146599	14.0	137.114532
2015-05-01 00:55:39	NaN	NaN	NaN	NaN	NaN	14.0	137.114532
2015-05-01 00:56:39	NaN	NaN	32.167801	NaN	NaN	13.8	148.116623
2015-05-01 00:59:39	NaN	NaN	NaN	NaN	NaN	13.8	148.116623
2015-05-01 01:00:39	NaN	NaN	NaN	NaN	NaN	14.7	98.714333
2015-05-01 01:01:39	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2015-05-01 01:02:40	NaN	NaN	NaN	NaN	NaN	14.7	98.714333
2015-05-01 01:03:40	290.309998	NaN	NaN	NaN	94.603500	14.3	120.666069
2015-05-01 01:04:40	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2015-05-01 01:05:40	NaN	273.698608	NaN	NaN	NaN	NaN	NaN
2015-05-01 01:06:40	NaN	NaN	NaN	NaN	94.603500	NaN	120.666069
2015-05-01 01:07:40	NaN	NaN	NaN	34.483501	100.022003	NaN	114.570335
2015-05-01 01:09:40	NaN	NaN	NaN	NaN	NaN	14.4	114.570335
2015-05-01 01:10:40	NaN	NaN	32.037800	NaN	NaN	14.7	98.615585
2015-05-01 01:12:40	289.291992	NaN	NaN	NaN	NaN	14.7	98.615585
2015-05-01 01:13:40	NaN	NaN	NaN	NaN	NaN	14.3	121.013443
2015-05-01 01:14:40	NaN	NaN	NaN	NaN	99.246803	14.3	121.013443
2015-05-01 01:15:40	NaN	NaN	NaN	NaN	95.022003	14.0	137.147720
2015-05-01 01:16:40	NaN	NaN	NaN	NaN	NaN	14.0	137.147720
2015-05-01 01:17:40	NaN	NaN	NaN	NaN	NaN	14.1	131.699463
2015-05-01 01:18:40	NaN	NaN	NaN	NaN	95.533699	14.7	98.712791
2015-05-01 01:19:40	NaN	273.698608	NaN	NaN	98.562698	14.0	136.996597

Fig. 4. Raw time-series data transformed into an array, where values are missing due to the measurement of sensor data at varying time steps.

shown in Fig. 6, will only capture approximation for a time window without considering the effect of overall PCP performance since time zero (t_0). This will lead to limited representation of the overall PCP performance.

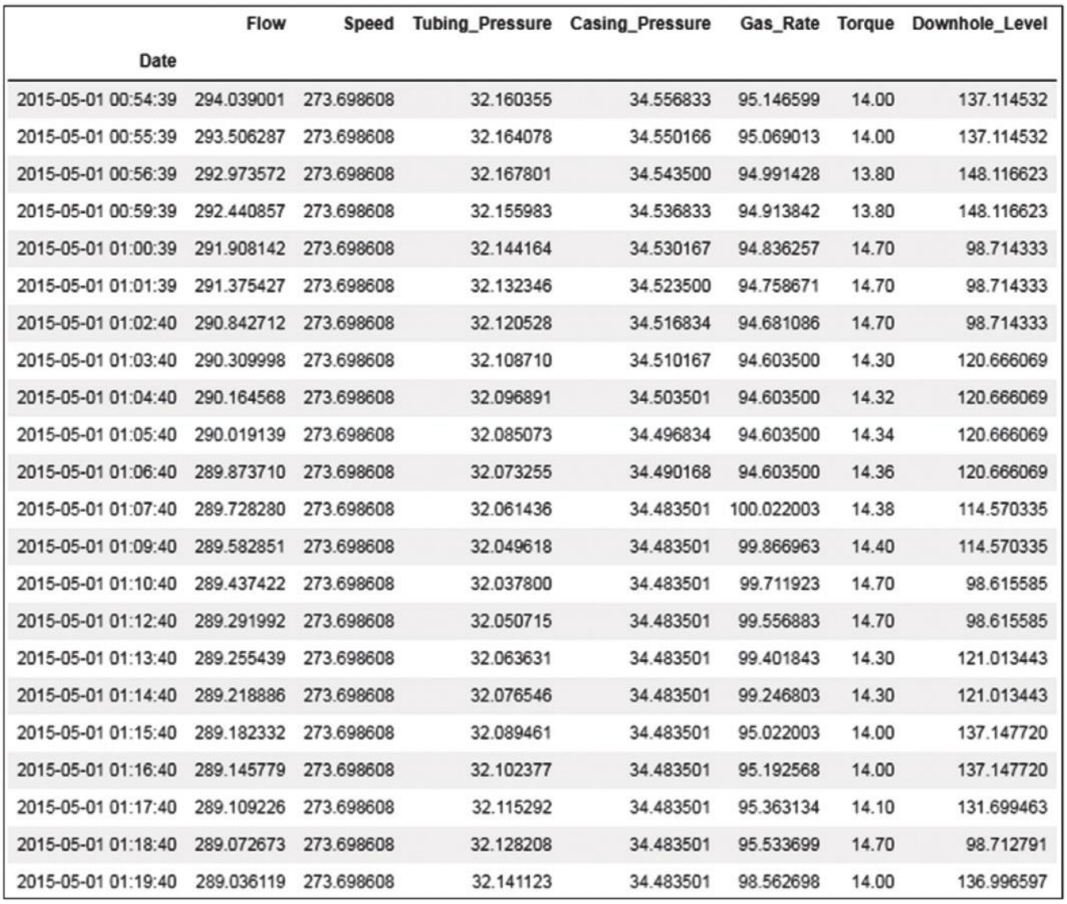
This study proposes the use of the aggregating window technique to overcome the shortfalls of the sliding window method when applied to multivariate time-series analysis. As the name suggests, the aggregating window method takes into consideration data starting from time zero (t_0) and normalises data over the entire aggregate of time-series data at each expansion stride. The process is shown in Fig. 7, where the mean of time-series data is effected every time new measurements are captured in expansion stride. This technique ensures that extracted features are representative of PCP performance from t_0 onwards.

Fig. 8 further demonstrates how SAX transformation is applied with an expanding window technique. In this example, the expansion stride is 20 time-series samples, i.e. normalisation and SAX transformation is executed every 20 min based on a 1 min sampling rate. Fig. 8*a* depicts the first time-series sample window of 20 data points starting at t_0. These data points are normalised, after which SAX transformation is applied to the equally distributed *bin* regions. The SAX symbols for this window are recorded and stored for the purpose of PCP performance heatmap conversion.

Fig. 8*b* depicts the next 20 streaming data points. Once the second expansion stride is completed, all data points starting from t_0 are normalised and SAX transformation is applied to the complete dataset. It can be observed that the SAX symbols in the first window were adjusted as per the normalisation with the new data window. However, we record the SAX symbols for the new window only and discard the symbols of the previous windows. By doing so, the SAX symbols created in the window are representative of the change in PCP performance. The previous windows that have adjusted symbols are of no particular use in terms of machine learning and analytics as they do not represent streaming time-series data.

SAX symbols to performance heatmaps

Converting SAX symbols to performance heatmaps is a four-step process where symbols are first arranged in a matrix, each



Date	Flow	Speed	Tubing_Pressure	Casing_Pressure	Gas_Rate	Torque	Downhole_Level
2015-05-01 00:54:39	294.039001	273.698608	32.160355	34.556833	95.146599	14.00	137.114532
2015-05-01 00:55:39	293.506287	273.698608	32.164078	34.550166	95.069013	14.00	137.114532
2015-05-01 00:56:39	292.973572	273.698608	32.167801	34.543500	94.991428	13.80	148.116623
2015-05-01 00:59:39	292.440857	273.698608	32.155983	34.536833	94.913842	13.80	148.116623
2015-05-01 01:00:39	291.908142	273.698608	32.144164	34.530167	94.836257	14.70	98.714333
2015-05-01 01:01:39	291.375427	273.698608	32.132346	34.523500	94.758671	14.70	98.714333
2015-05-01 01:02:40	290.842712	273.698608	32.120528	34.516834	94.681086	14.70	98.714333
2015-05-01 01:03:40	290.309998	273.698608	32.108710	34.510167	94.603500	14.30	120.666069
2015-05-01 01:04:40	290.164568	273.698608	32.096891	34.503501	94.603500	14.32	120.666069
2015-05-01 01:05:40	290.019139	273.698608	32.085073	34.496834	94.603500	14.34	120.666069
2015-05-01 01:06:40	289.873710	273.698608	32.073255	34.490168	94.603500	14.36	120.666069
2015-05-01 01:07:40	289.728280	273.698608	32.061436	34.483501	100.022003	14.38	114.570335
2015-05-01 01:09:40	289.582851	273.698608	32.049618	34.483501	99.866963	14.40	114.570335
2015-05-01 01:10:40	289.437422	273.698608	32.037800	34.483501	99.711923	14.70	98.615585
2015-05-01 01:12:40	289.291992	273.698608	32.050715	34.483501	99.556883	14.70	98.615585
2015-05-01 01:13:40	289.255439	273.698608	32.063631	34.483501	99.401843	14.30	121.013443
2015-05-01 01:14:40	289.218886	273.698608	32.076546	34.483501	99.246803	14.30	121.013443
2015-05-01 01:15:40	289.182332	273.698608	32.089461	34.483501	95.022003	14.00	137.147720
2015-05-01 01:16:40	289.145779	273.698608	32.102377	34.483501	95.192568	14.00	137.147720
2015-05-01 01:17:40	289.109226	273.698608	32.115292	34.483501	95.363134	14.10	131.699463
2015-05-01 01:18:40	289.072673	273.698608	32.128208	34.483501	95.533699	14.70	98.712791
2015-05-01 01:19:40	289.036119	273.698608	32.141123	34.483501	98.562698	14.00	136.996597

Fig. 5. Transformed array after imputation and fill methods are applied to various measurements and sensor readings. Gaps in the speed columns are filled using a 'forward-fill' method. Remaining variables are filled through cubic imputation.

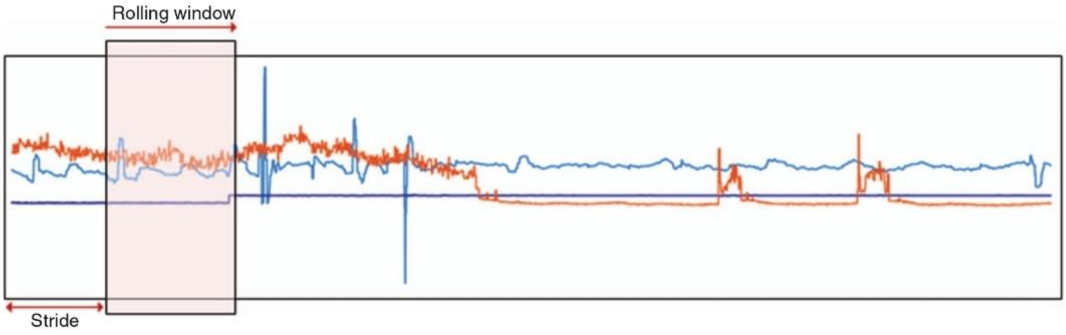


Fig. 6. Example of a rolling window method with a fixed stride.

matrix cell is assigned a timed colour code, a combined heatmap is produced for a single time window, and finally, the heatmaps are segregated by masking colour codes. These steps are described in detail below.

Matrix conversion

SAX symbols from each of the three variables (flow, torque and speed) are converted into matrices. Flow and torque SAX symbols are each converted into a 5 × 1 matrix, where the

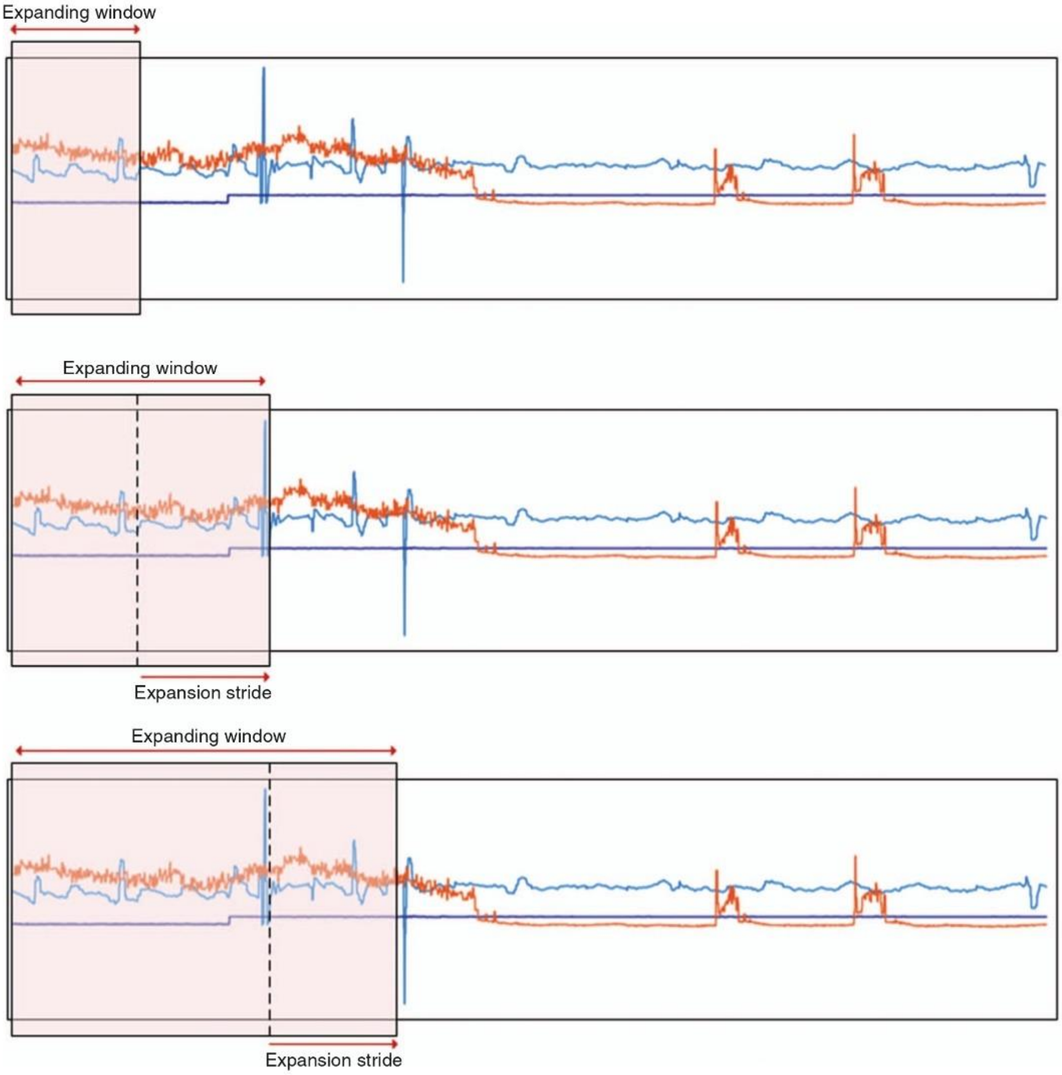


Fig. 7. Example of an Expanding Window method with an expansion stride.

cells denote a low-, normal- and high-operational range. Each cell in the matrix holds either a single SAX symbol or an aggregate of SAX symbols. For flow and torque, symbol (*i*) denotes a high-performance range, and the symbols are placed on the top cell of their respective matrices. Symbol aggregates (*f*+*g*+*h*) and (*c*+*d*+*e*) denote a normal-performance range and are placed in the mid-cells. Similarly, symbols (*a*) and (*b*) represent the low-performance range and are, therefore, placed in the two bottom cells of the matrices. Aggregation of mid-range symbols is performed to suppress their high rate of occurrence within a time window, which can inaccurately be characterised as fluctuating behaviour.

Speed SAX symbols are converted into a 9 × 1 matrix, where each symbol has its own designated cell. Symbols for this time-series data are not aggregated as speed does not have a highly dynamic behaviour versus flow and torque. The matrix conversion for all three variables is shown in Fig. 9.

Timed colour-code conversion

To better visualise the occurrence of SAX symbols in a time window, we need to transform matrix cells as per a set timed behaviour. Each matrix cell represents a count of its respective symbol(s) within a pre-selected time window. If we run SAX symbol conversion on a pre-selected 1-day window and the sampling time for each data point is 1 min, we will then have 1440 SAX symbols for each of the selected variables. In order to capture the occurrence or count of symbols, the colour-code scheme shown in Table 1 is applied to each matrix cell.

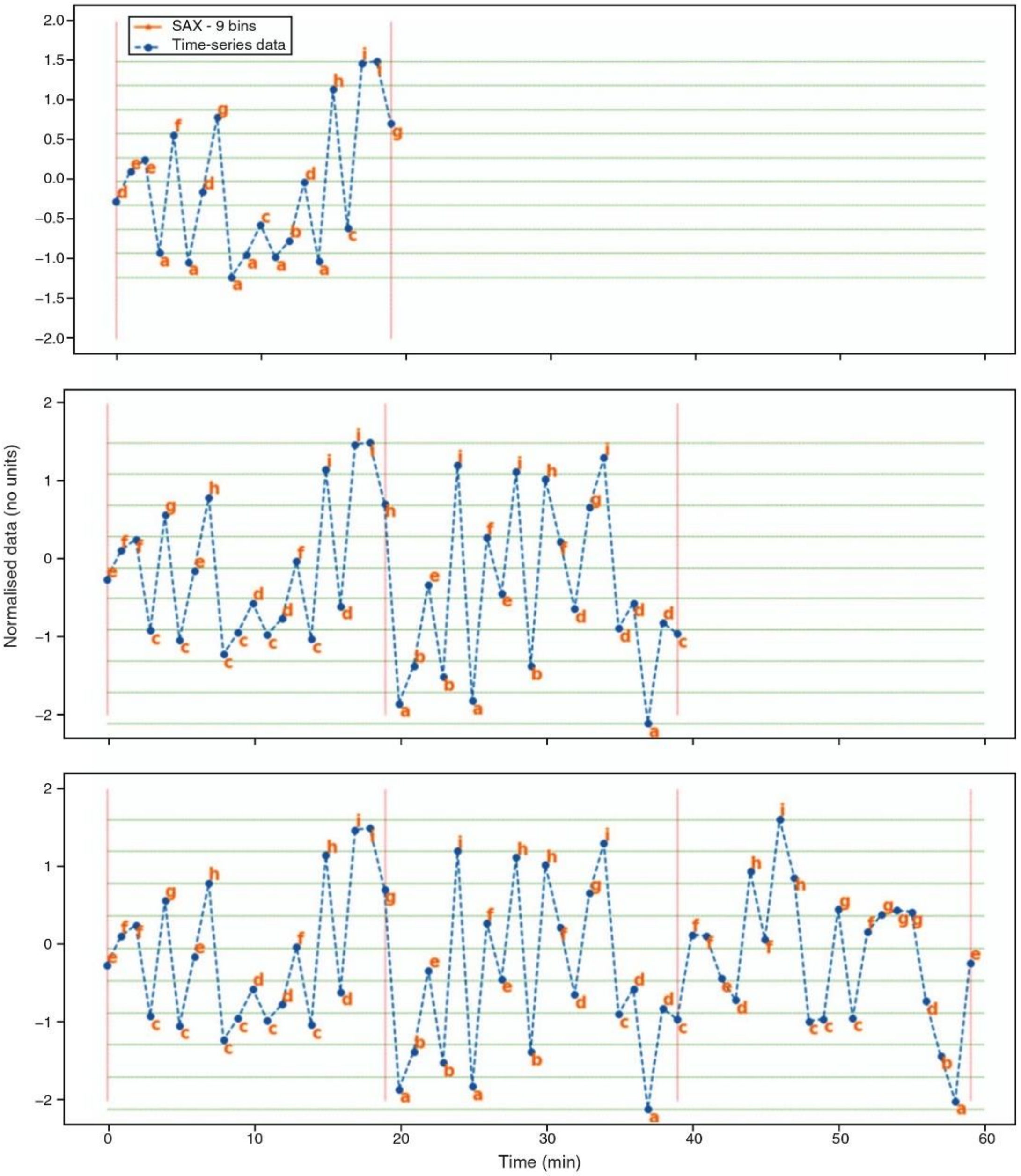


Fig. 8. SAX transformation of time-series data using the Expanding Window Technique.

Combined heatmap transformation

Fig. 10 provides an overview of how a 1-day window of flow time-series data is converted into a heatmap. For this instance, symbol (*i*) is prevalent where it has occurred more than 720 min (or half a day). Hence the matrix cell for the symbol (*i*) is colour coded green. The occurrence of the symbol (*e*) is higher than 60 min but less than 720 min, hence the matrix cell (*c*+*d*+*e*) is colour coded. The occurrence of symbols (*f*) and (*h*), when accumulated together, is also higher than 60 min but less than 720 min. Hence the matrix cell (*f*+*g*+*h*) is colour coded yellow. The same process is shown for torque and speeds SAX symbols in Figs 11 and 12, respectively.

After data variables are converted into colour-coded heatmaps, they are then merged together to form a multivariate heatmap. Fig. 13, shows a multivariate heatmap

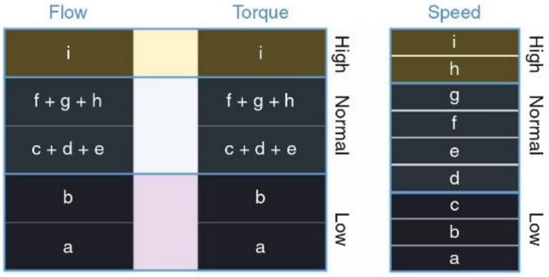


Fig. 9. HEATMAP matrices for flow, torque and speed.

Table 1. Colour code based on SAX symbol count during a 1-day window

SAX symbol count in one day window (1440 min)	Colour code	Performance type
720 < count < 1440	Green	Majority performance
60 < count < 720	Yellow	Variation performance
00 < count < 60	Red	Anomaly event
Count = 00	Black	Nil

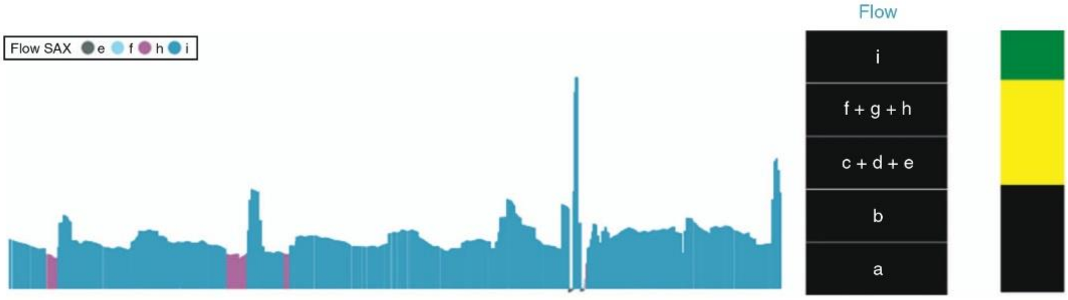


Fig. 10. Flow time-series data conversion to SAX HEATMAP. Symbol (i) is colour coded green and the remaining yellow.

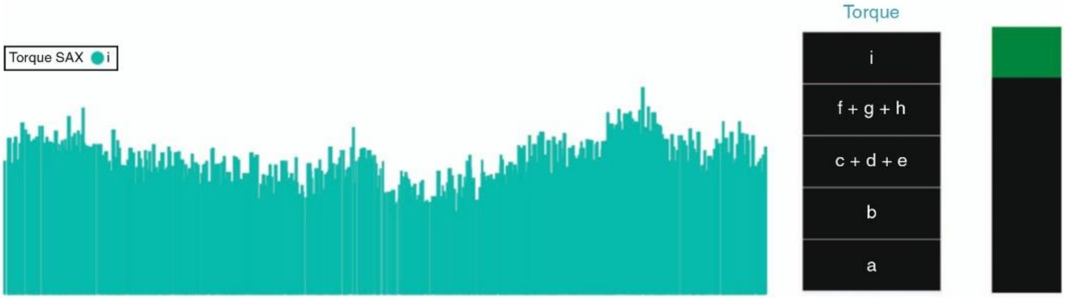


Fig. 11. Torque time-series data conversion to SAX HEATMAP. Symbol (i) is colour coded green.

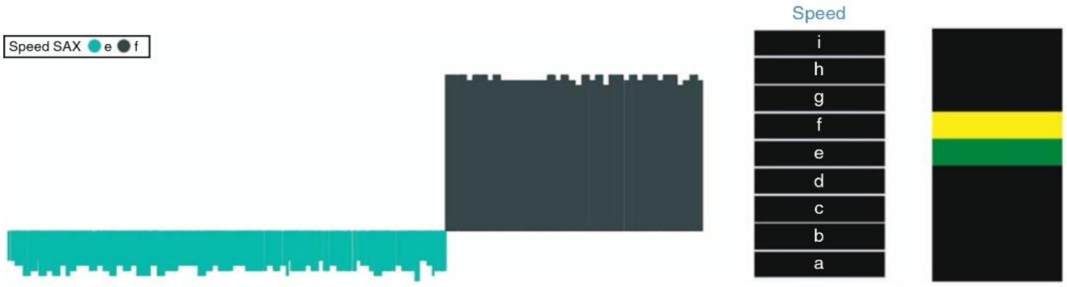


Fig. 12. Speed time-series data conversion to SAX HEATMAP. Symbol (e) is colour coded green and (f) yellow.

for a 1-day time-series data of flow, torque and speed. Two events are observed during this 24-h period; the first event is a speed change, and the second event is unsteady flow. As per the colour-code SAX symbol count, both these events are captured by the multivariate heatmap. However, this heatmap does not recognise if flow fluctuation occurred before or after the speed change. This drawback can be avoided if a smaller time window is used for SAX symbol conversion and heatmap generation.

Table 2 shows the colour-code conversion scheme for a PCP performance heatmap generated for a 1-h window. For a 1-h window, the maximum symbol count for each variable is 60 symbols. The limits for each performance type are set based on the count of SAX symbols, where a count in the range of 15–60 represents majority performance. Variation performance is set between five and 15 symbols. Moreover, an anomaly event is recognised when the symbol count range is between zero and five.

Table 2. Colour code based on SAX symbol count during a 1-h window

SAX symbol count in 1-h window (maximum of 60 SAX symbols)	Colour code	Performance type
15 < count < 60	Green	Majority performance
5 < count < 15	Yellow	Variation performance
00 < count < 5	Red	Anomaly event
Count = 00	Black	Nil

In Fig. 14 performance heatmap images are generated with a 1-h window. With the improved granularity provided by a smaller window analysis, change in performance can be seen for each hour. The change in flow behaviour is picked up by the performance heatmaps on either side of the 6:00 p.m. mark.

Splitting performance heatmaps

To further aid with PCP performance analysis, the performance heatmaps are split by masking the individual colours shown in the colour code Table 2 (Saghir *et al.* 2019*c*). Three heatmaps are extracted by this method, and they are shown in Fig. 15. By splitting the heatmaps, it becomes easier to train the machine learning models to detect performance and anomalous related events individually. Majority performance heatmaps represent the stable PCP performance recorded for a particular time window. Anomaly event heatmaps are representative of behaviour that has occurred for a brief period, as illustrated in Table 2.

The variation performance heatmaps are ignored, as they represent passing variation in sensor data and characterise acceptable performance deviation when physical measurements are in transition. Fig. 16 shows a sample of split heatmaps, which are created for training the auto encoder and the machine learning model.

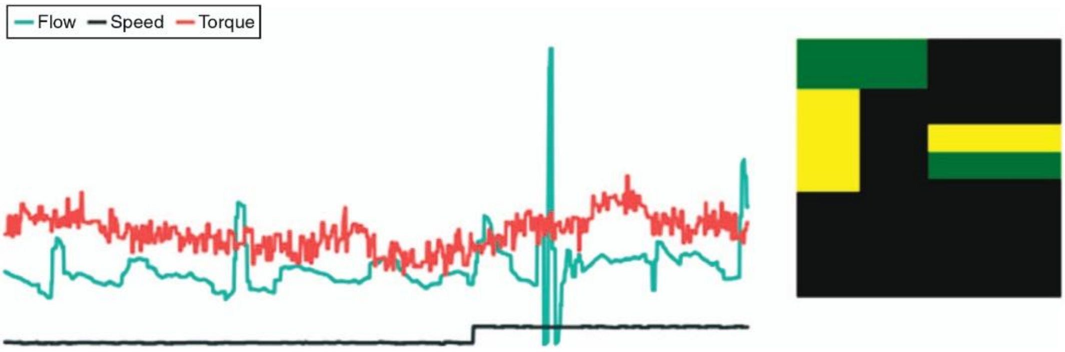


Fig. 13. Multivariate Performance Heatmap for a 1-day time window

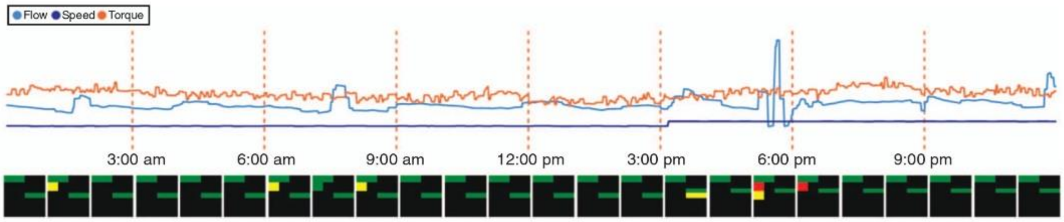


Fig. 14. Multivariate Performance Heatmaps generated on an hourly basis.

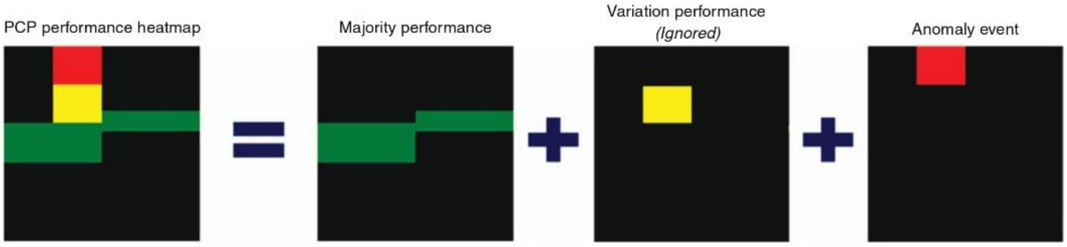


Fig. 15. Multivariate Performance Heatmap split into masked components.

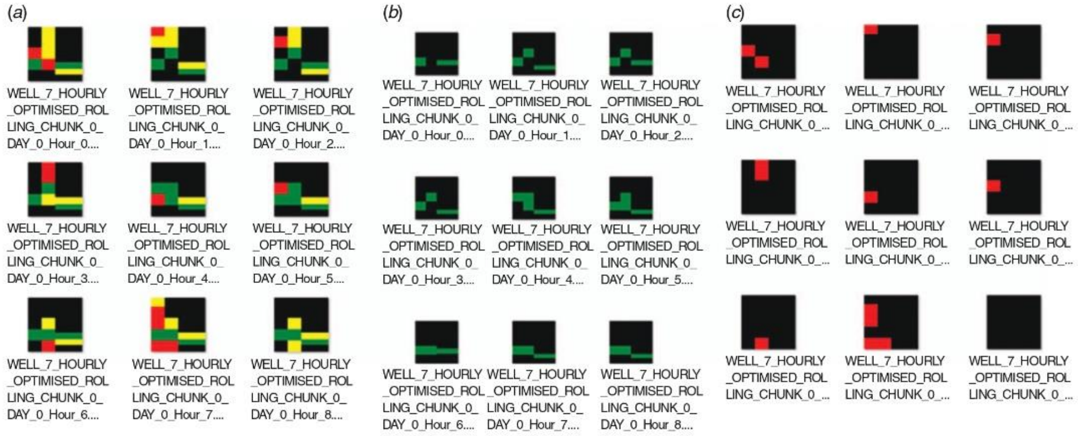

(a)
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_0....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_1....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_2....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_3....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_4....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_5....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_6....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_7....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_8....
(b)
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_0....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_1....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_2....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_3....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_4....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_5....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_6....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_7....
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_ DAY_0_Hour_8....
(c)
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_...
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_...
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_...
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_...
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_...
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_...
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_...
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_...
WELL_7_HOURLY _OPTIMISED_ROL LING_CHUNK_0_...

Fig. 16. Split Heatmaps used for training Auto Encoders and Machine Learning Models. (*a*) Original PCP Performance Heatmaps. (*b*) Majority Performance Heatmaps. (*c*) Anomaly Heatmaps.

Convolutional neural network based auto encoder

Once performance heatmaps are created and masked, a CAE is employed to reduce the dimensionality of the generated images. The performance heatmaps generated for this study have a dimension of 48 × 48 × 3 (6912 pixels). With the CAE, we reduce the dimensionality of the image to latent space, which helps to improve the model training time.

As shown in Fig. 17, we are using a seven-layer neural network to reduce the dimension of the image to 1 × 8 representation of the original 48 × 48 × 3 image. All the layers in the CAE are fully connected. The top section of the network represents the encoder, which reduces the dimension of the original image to a simplified code. The decoder takes the code and converts it back to the original image dimension. Although the encoder would suffice by itself for reducing the image dimensionality, the decoded image is required to compare the loss between the original and reconstructed image to validate the accuracy of the CAE. The accuracy of the CAE will be discussed in the results section.

All PCP performance heatmaps are converted to an encoded representation before the application of the clustering methodology.

Time-series data clustering based on PCP performance heatmaps

Fig. 18 summarises the end-to-end process for converting performance heatmaps into clusters. All heatmaps are encoded via the CAE before being processed by the clustering algorithm. A further dimensionality reduction step is performed with the singular value decomposition (SVD) to convert the encoded images into two-dimensional representation. This is done to visualise the results of the clusters on a single plane (X-Y) plot. The two-dimensional representation also helps with faster computation analysis to determine the time-series clusters.

As per the second step of this process, we pass the SVD data through the HDBSCAN clustering algorithm. Although there are multiple benefits of using a density-based clustering algorithm over partitional clustering algorithms (K-means as an example), there are two benefits that make HDBSCAN suitable for our application (Campello *et al.* 2013). First, unlike partitional clustering algorithms where the number of clusters must be pre-defined, HDBSCAN only requires identification of a minimum number of clusters to work out the optimum cluster number based on a hierarchical cluster tree. Second, the HDBSCAN algorithm can identify outliers,

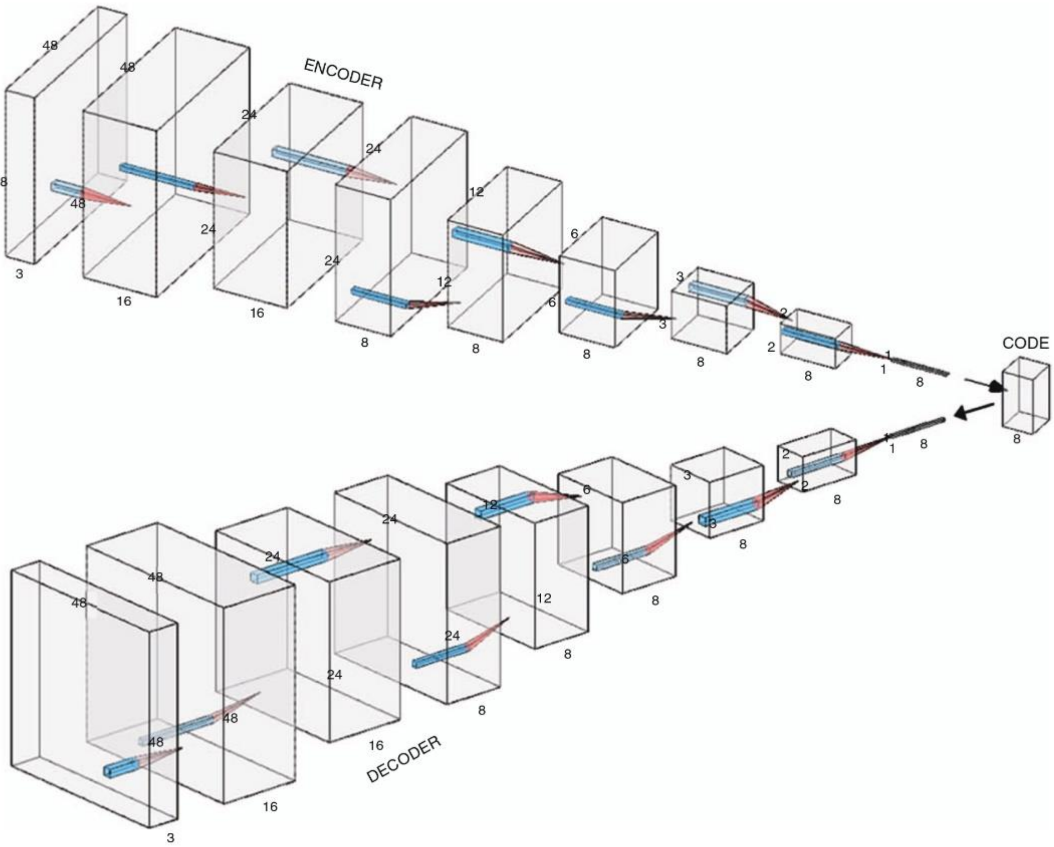


Fig. 17. CAE used for reducing the dimension of the PCP Performance Heatmaps.

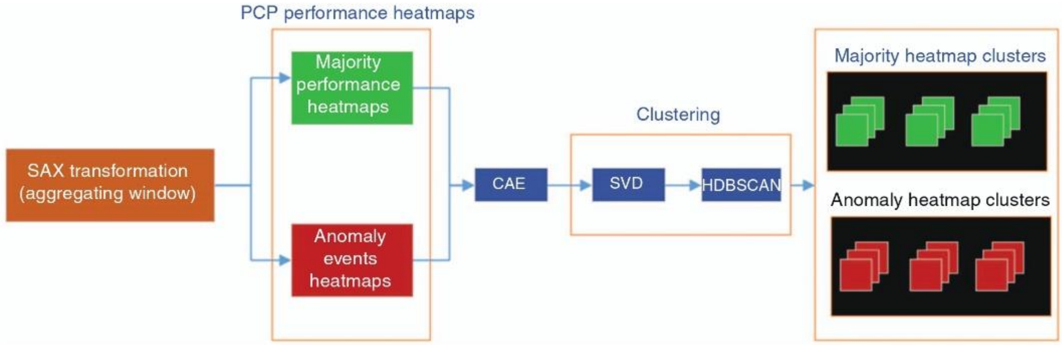


Fig. 18. Summary of the process applied to convert PCP Performance Heatmaps to clusters.

which can further aid with analysis of PCP Performance Heatmaps.

Results

CAE performance

We first describe the performance of the CAE model, as it forms the basis of the clustering and the visual analysis results of this study. Fig. 19 features the performance of the CAE, where the top row (Fig. 19*a*)shows 20 random anomaly heatmaps. Fig 19*b* shows the encoded 1 × 8 images, but they are displayed in a 2 × 4 configuration. Fig 19*c* shows the decoded anomaly heatmaps, and it is observed that the majority of the original images closely match their original counterparts. For anomaly heatmaps that do not have enough training samples, a loss is observed in their decoded images.

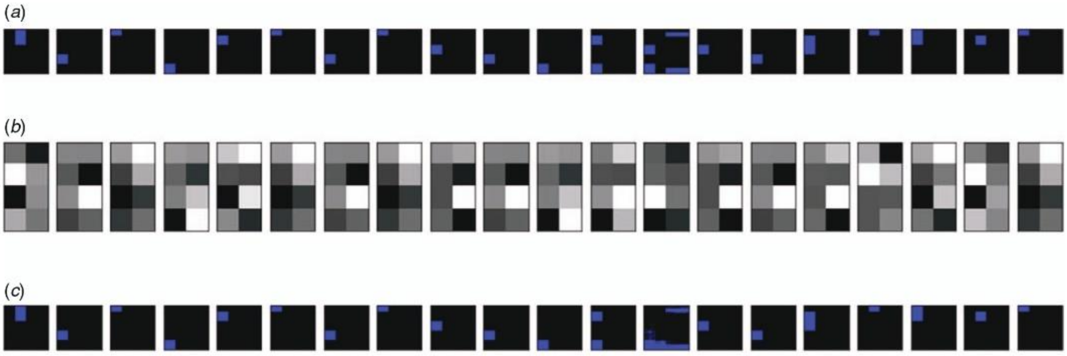


Fig. 19. CAE Encoder and Decoder overview for Anomaly Heatmaps, where (*a*) shows the original Anomaly Heatmap images; (*b*) shows the encoded image (with an 8 × 1 dimension but displayed as a 2 × 4 image); and (*c*) shows the decoded images.

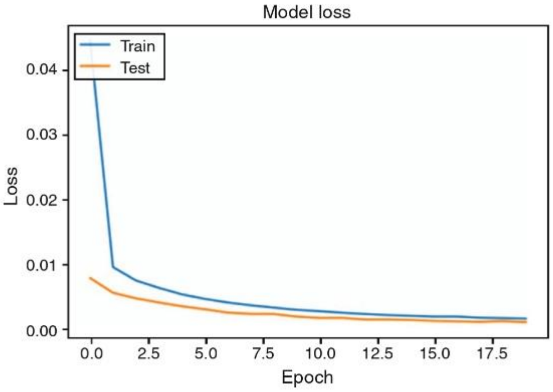


Fig. 20. CAE model loss for Anomaly Heatmaps.

Anomaly heatmaps from 200 wells were used to develop the CAE model. A total of 800 000 images were split into 60/40 configuration to create training and test samples, respectively. The CAE was trained over 20 epochs, and the model results are shown in Fig. 20. The training and validation loss decreased over the epoch cycles and the two lines closely follow each other. This means the model neither under or overperformed and functioned effectively with the anomaly heatmaps.

Figs 21 and 22 show the CAE image conversion overview and model loss, respectively, for the majority performance heatmaps. The CAE model for the majority performance heatmaps was also trained with images from 200 wells and had the same training parameters as the anomaly heatmap CAE. The training and validation loss also decreased in this case, with validation loss slightly higher than the training loss. This indicates marginal overfitting of the majority heatmap CAE model, which falls within an acceptable tolerance range.

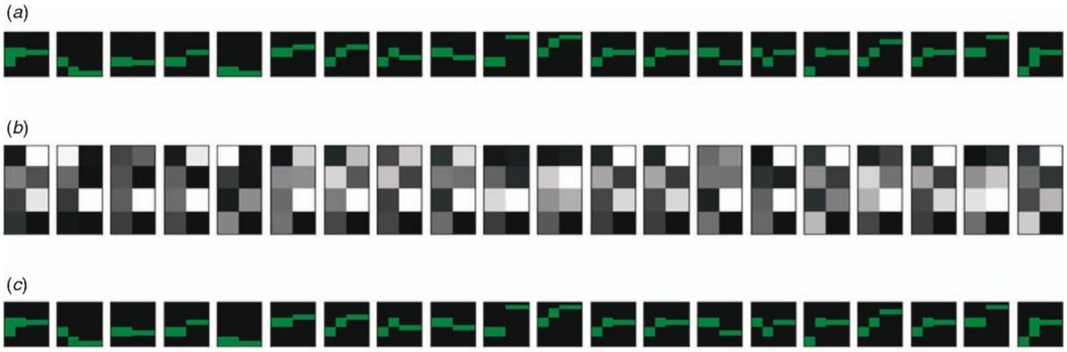


Fig. 21. CAE encoder and decoder overview for Majority Heatmaps, where (*a*) shows the original Majority Heatmap images; (*b*) shows the encoded image (with an 8 × 1 dimension but displayed as a 2 × 4 image); and (*c*) shows the decoded images.

Heatmap clustering

Once the heatmap images were encoded, they were clustered as per the process depicted in Fig. 18. It is noted that the HDBSCAN created a high number of clusters for both the majority performance and anomaly event heatmaps. For the majority performance heatmaps, the HDBSCAN algorithm produced 1000 plus clusters. Moreover, for the anomalous events heatmaps, approximately 300 clusters were produced.

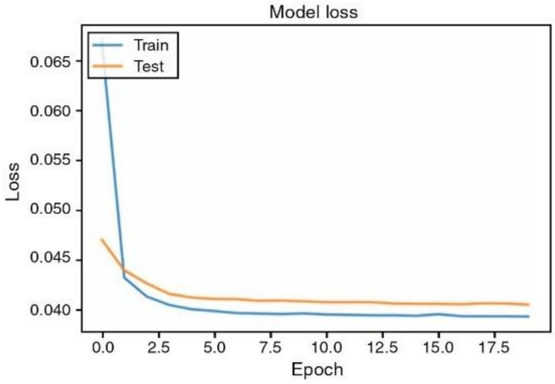


Fig. 22. CAE model loss for Majority Heatmaps.

There are two reasons for such a high cluster count. First, the clusters were created based on time-series heatmaps generated from a large pool of CSG wells. Second, HDBSCAN applied limited generalisation to the identified clusters compared to other methods.

Although the number of clusters looks significantly high, they do not apply to each well. Figs 23 and 24 show the number of majority performance and anomalous event clusters across 17 randomly selected wells. Wells with a higher number of clusters attributed to the significantly high overall cluster count for the time-series images analysed as per the HDBSCAN algorithm. Furthermore, the high cluster numbers also helped to identify CSG wells with highly abnormal PCP performance.

Visual analytics – PCP performance analysis

Although the number of heatmap clusters detected was relatively high, they presented an advantage when used in conjunction with visual analytics tools. Fig. 25 shows how clusters can be visualised in combination with streaming time-series data. The visual analytics interface is a 30-day moving window that helps the observer gauge PCP performance as streaming data is captured and analysed.

Fig. 25*a* shows a stacked bar chart of the majority performance clusters. Each stack is populated on an hourly basis and represents one day of the performance. Different colours in each stack represent a cluster count for that particular day. In this visual analysis, the cluster number is

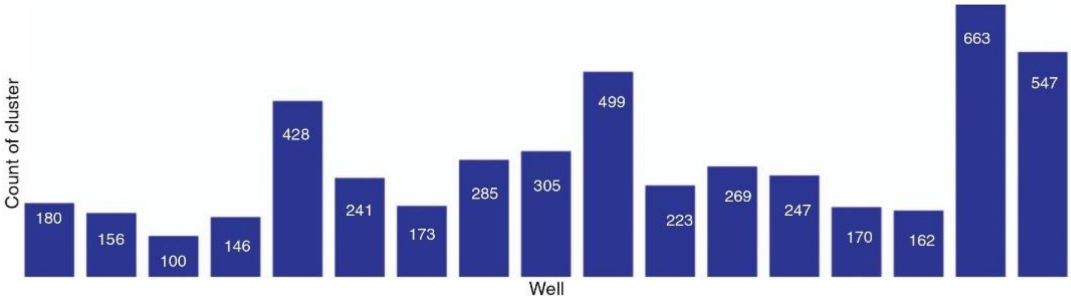


Fig. 23. Count of majority performance clusters across 17 randomly selected wells.

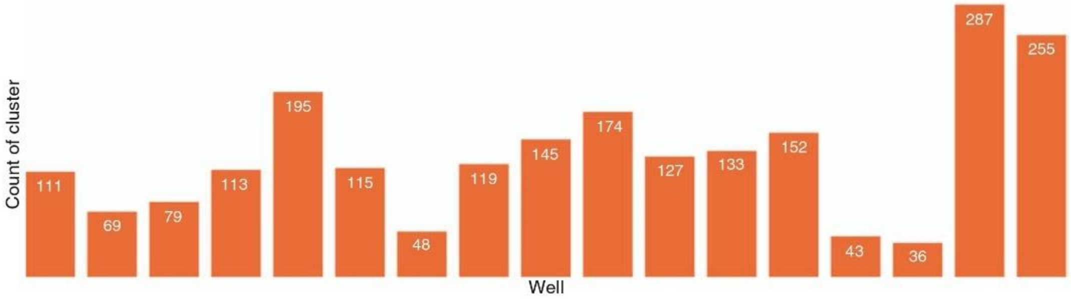


Fig. 24. Count of Anomalous Event Clusters across 17 randomly selected wells.

not essential, but the variation in cluster count over time is of significance. For the majority performance clusters bar charts, the variance in cluster count is related to the change in PCP performance. This is obvious by the change in colour of the stacked columns. While the pump is running, the stacked bar charts will usually contain two to five colours daily. However, during pump shutdown and startup, multi-colour stacks were observed, which indicated the fluctuating performance of the pump during such stages.

Similarly, Fig. 25*c* shows a stacked bar chart of the anomaly event clusters. A single colour stack (blue in this case) represents a 24-h period, where no anomaly was detected. As anomalies were detected and accumulated for a particular day, the bar chart identified periods with atypical performance. Change in the number of anomaly clusters aids the operator in selecting days or periods to conduct an in-depth PCP performance analysis.

To further aid with visual analysis, a cluster variance trend was plotted over the stacked bar chart to provide an overview of changing cluster behaviour. This visualisation is shown in Fig. 26. Furthermore, an in-depth analysis was performed on a daily trend by displaying majority performance and anomalous heatmap images for a selected 24-h period. By focusing on a particular day, operators and well surveillance engineers can

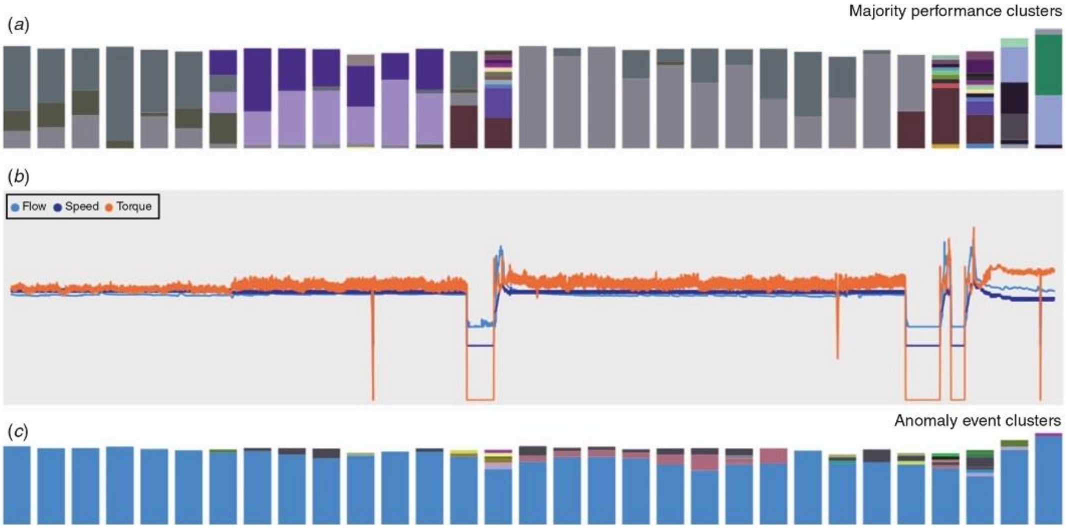


Fig. 25. PCP visual analytics for 30-day rolling period. (*a*) Stacked colour bar chart depicting the count of daily majority performance clusters. (*b*) Time-series trend. (*c*) Stacked colour bar chart depicting the count of daily anomalous event clusters.

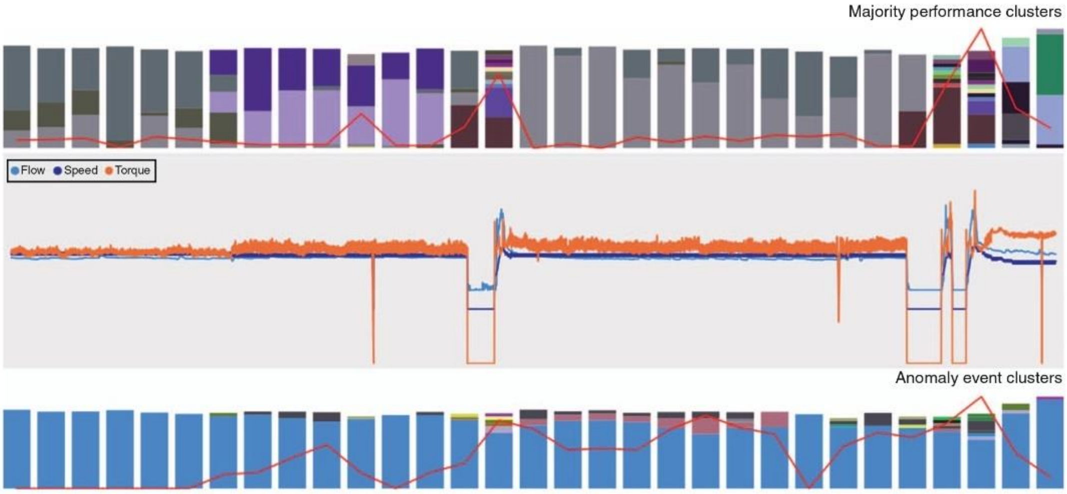


Fig. 26. PCP visual analytics for 30-day rolling period with a cluster variation line plot.

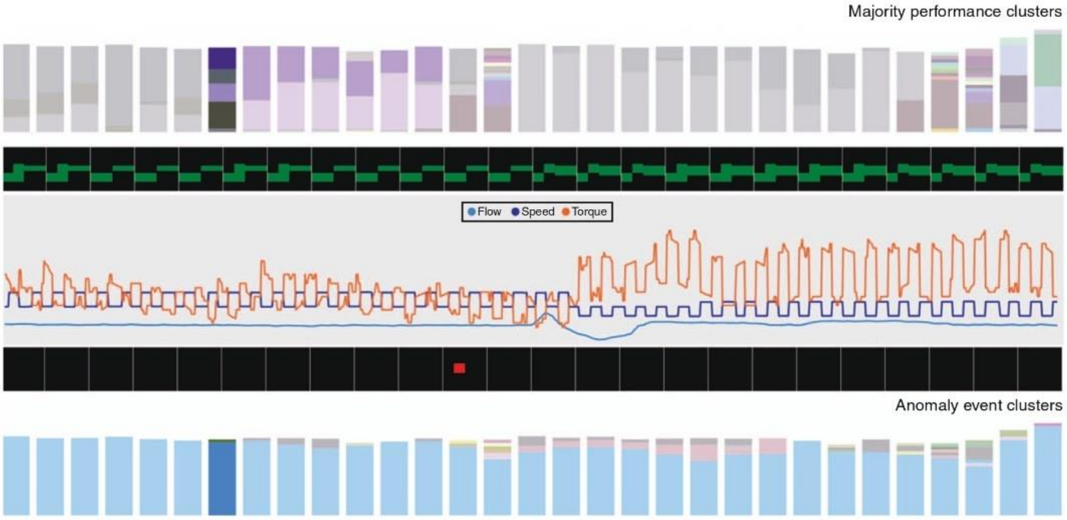


Fig. 27. PCP visual analytics with Heatmap image clusters for a 24-h, in-depth analysis.

correlate anomalous behaviour to overall PCP performance. An example of in-depth daily analysis is shown in Fig. 27, where time-series plot, stacked bar charts and heatmap images are displayed together.

Conclusion

Work described in this study shows that heatmap image-based analysis of time-series data can effectively aid with PCP performance analysis. Although only three multivariate data points (flow, torque and speed) were used in this study, this unique methodology can work with any number of data points.

This study has also shown that visual analysis capability provided by heatmap images can improve operator workload and aid with in-depth PCP performance analysis.

Future work

Although the methodology covered in this study provides an improved method to conduct PCP performance analysis, further improvements are envisaged as part of continuing research. Future publications will cover the following work, some of which are already underway:

- Analysis of CSG well meta-data (well orientation and pump setting depth) to improve cluster generalisation,
- Understanding the progressive behaviour of heatmap clusters to help with sequential prediction analysis, and
- A comparative study to understand clustering behaviour between the image analysis technique conducted in this study and other time-series data based image conversion techniques (RP, GAF and MTF).

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgments

The authors would like to thank the University of Adelaide for permission to submit this paper. This research did not receive any specific funding.

References

Ali, M., Jones, M., Xie, X., and Williams, M. (2019). TimeCluster: dimension reduction applied to temporal data for visual analytics. *The Visual Computer* **35**, 1013–1026. doi:10.1007/s00371-019-01673-y

Bettaiah, V. (2014). The hierarchical piecewise linear approximation of time series data. PhD Thesis, Department of Computer Science, University of Alabama in Hunstville. Available at https://cdm16608.contentdm.oclc.org/digital/collection/p16608coll23/id/2485 [Verified 30 January 2020]

Biniwale, S. S., and Trivedi, R. (2012). Managing LNG Deliverability: An Innovative Approach Using Neural Network and Proxy Modeling for Australian CSG Assets. In 'Abu Dhabi International Petroleum Conference and Exhibition, 11-14 November, Abu Dhabi, UAE.' (Society of Petroleum Engineers.) doi:10.2118/160445-MS

Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In: 'Advances in Knowledge Discovery and Data Mining'. (Eds J. Pei, V. S. Tseng, L. Cao, H. Motoda and G. Xu.) pp. 160–172. (Springer: Berlin Heidelberg)

Chaovalit, P., Gangopadhyay, A., Karabatis, G., and Chen, Z. (2011). Discrete wavelet transform-based time series analysis and mining. *ACM Computing Surveys* **43**, 1–37. doi:10.1145/1883612.1883613

Dan, J., Shi, W., Dong, F., and Hirota, K. (2013). Piecewise Trend Approximation: A Ratio-Based Time Series Representation. *Abstract and Applied Analysis* **2013**, 603629. doi:10.1155/2013/603629

Duvignau, R., Gulisano, V., Papatriantafilou, M., and Savic, V. (2018). Piecewise Linear Approximation in Data Streaming: Algorithmic Implementations and Experimental Analysis. Available at https://arxiv.org/abs/1808.08877 [Verified 30 January 2020]

Firouzi, M., and Rathnayake, S. (2019). Prediction of the Flowing Bottom-Hole Pressure Using Advanced Data Analytics. In 'SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology Conference, 18-19 November, Brisbane, Australia.'. (Unconventional Resources Technology Conference). doi:10.15530/AP-URTEC-2019-198240

Gao, Y., and Lin, J. (2018). Efficient Discovery of Variable-length Time Series Motifs with Large Length Range in Million Scale Time Series. Available at https://arxiv.org/abs/1802.04883 [verified 30 January 2020]

Guigou, F., Collet, P., and Parrend, P. (2017). Anomaly detection and motif discovery in symbolic representations of time series. Technical Report No 69427/2. (Complex System Digital Campus, UNITWIN UNESCO). Available at https://hal.archives-ouvertes.fr/hal-01507517/document [Verified 30 January 2020]

Hatami, N., Gavet, Y., and Debayle, J. (2017). Classification of Time-Series Images Using Deep Convolutional Neural Networks. Available at https://arxiv.org/abs/1710.00886 [Verified 30 January 2020]

Hoday, J. P., Knafl, M., Prosper, C., and Braas, M. (2013). Diagnosing PCP Failure Characteristics using Exception Based Surveillance in CSG. In 'SPE Progressing Cavity Pumps Conference. Calgary, Alberta, Canada.' (Society of Petroleum Engineers.) doi:10.2118/165655-MS

Kumar, N., Lolla, V. N., Keogh, E. J., Lonardi, S., Ratanamahatana, C., and Wei, L. (2005). Time-series Bitmaps: a Practical Visualization Tool for Working with Large Time Series Databases. In 'Proceedings of the 2005 SIAM International Conference on Data Mining'. Available at https://doi.org/10.1137/1.9781611972757.55 [Verified 30 January 2020]

Lin, J., Keogh, E., and Lonardi, S. (2005). Visualizing and discovering non-trivial patterns in large time series databases. *Information Visualization* **4**, 61–82. doi:10.1057/palgrave.ivs.9500089

Lin, J., Keogh, E., Wei, L., and Lonardi, S. (2007). Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery* **15**, 107–144. doi:10.1007/s10618-007-0064-z

Matthews, C. M., Zahacy, T. A., Alhanati, F. J. S., Skoczylas, P., and Dunn, L. J. (2007). 'Petroleum Engineering Handbook. Production Operations Engineering'. (Society of Petroleum Engineers: Richardson, Texas).

McKinney, W. (2013). 'Python for Data Analysis'. (O'Reilly Media: Sebastopol, California).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830.

Prosper, C., and West, D. (2018). Case Study Applied Machine Learning to Optimise PCP Completion Design in a CBM Field. In 'SPE Asia Pacific Oil and Gas Conference and Exhibition, 23–25 October, Brisbane, Australia.' (Society of Petroleum Engineers.) doi:10.2118/192002-MS

Saghir, F., Gonzalez Perdomo, M. E., and Behrenbruch, P. (2019*a*). Application of Exploratory Data Analytics (EDA) in Coal Seam Gas Wells with Progressive Cavity Pumps (PCPs). In ' SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, 29–31 October, Bali, Indonesia.' (Society of Petroleum Engineers.) doi:10.2118/196528-MS

Saghir, F., Perdomo, M. E. G., and Behrenbruch, P. (2019*b*). Converting Time Series Data into Images: An Innovative Approach to Detect Abnormal Behavior of Progressive Cavity Pumps Deployed in Coal Seam Gas Wells. In 'SPE Annual Technical Conference and Exhibition, 30 September–2 October, Calgary, Canada: (Society of Petroleum Engineers.) doi:10.2118/195905-MS

Saghir, F., Gonzalez Perdomo, M. E., and Behrenbruch, P. (2019*c*). Machine Learning for Progressive Cavity Pump Performance Analysis: A Coal Seam Gas Case Study. In 'SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology Conference, 18–19 November,. Brisbane, Australia.' (Unconventional Resources Technology Conference.) doi:10.15530/AP-URTEC-2019-198281

Shaw, P. K., Saha, D., Ghosh, S., Janaki, M. S., and Iyengar, A. N. S. (2015). Investigation of coherent modes in the chaotic time series using empirical mode decomposition and discrete wavelet transform analysis. *Chaos, Solitons and Fractals* **78**, 285–296. doi:10.1016/j.chaos.2015.08.012

Sivaraks, H., and Ratanamahatana, C. A. (2015). Robust and Accurate Anomaly Detection in ECG Artifacts Using Time Series Motif Discovery. *Computational and Mathematical Methods in Medicine* **2015**, 453214. doi:10.1155/2015/453214

Vidaurre, D., Iead, R., Harrison, S., Smith, S., and Woolrich, M. (2014). Dimensionality reduction for time series data. Available at https://arxiv.org/abs/1406.3711 [Verified 30 January 2020]

Wang, Z., and Oates, T. (2015). Imaging Time-Series to Improve Classification and Imputation. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015). Available at https://www.ijcai.org/Proceedings/15/Papers/553.pdf [Verified 30 January 2020]

Zhong, R., Johnson, R. L., and Chen, Z. (2019). Using Machine Learning Methods to Identify Coals from Drilling and Logging-While-Drilling LWD Data. In 'SPE/AAPG/SEG.) Asia Pacific Unconventional Resources Technology Conference, 18–19 November,. Brisbane, Australia.' (Unconventional Resources Technology Conference.) doi:10.15530/AP-URTEC-2019-198288

The authors

Fahd Saghir is an Automation Engineer with 10+ years of experience in the digital oilfield domain. Since completing his BSc in Electrical Engineering from the University of Houston in 2006, Fahd has been involved in creating digital solutions for the production and operations verticals within the oil and gas sector, based on innovative hardware and software technologies. Fahd is currently pursuing a PhD in Petroleum Engineering from the University of Adelaide. His research work focuses on the use of Machine Learning methods to classify and detect abnormal PCP performance in CSG wells. This research investigates an innovative approach where time-series data is transformed into heatmap images, and the images are then used to classify PCP performance in near real-time. Fahd is also an active SPE volunteer and has participated as a speaker and moderator at multiple SPE conferences and webinars. He is currently a member of the Digital Solutions Committee, which falls under SPE's Digital Energy Technical Section.

Mary Gonzalez is a senior lecturer at the Australian School of Petroleum and Energy Resources (ASPER) at the University of Adelaide. Her research and teaching focus is on reservoir and production engineering, particularly production enhancement and optimisation. She joined the ASP in 2009 after several years of experience in the oil and gas industry, where she provided practical petroleum engineering, consultancy services and solutions in the areas of subsurface and production engineering. Mary has published several articles in peer-reviewed journals and presented at international conferences. She has served as a reviewer for different journals and as a mentor for young professionals, and she is the Community Education Chair and the ASPER Faculty Officer for the SPE.

Professor Peter Behrenbruch is currently the Managing Director of Bear and Brook Consulting Pty Ltd (since 2003). He is also an Adjunct Professor at the Ho Chi Minh University of Technology, Faculty of Geology and Petroleum Engineering, Vietnam. Behrenbruch's last full-time industry position (2008–2009) was Chief Operating Officer/Managing Director for East Puffin (SINOPEC) for the Puffin offshore development project, Timor Sea. He held a similar position (2007–2008) for AED Oil Ltd on the same project. He was also the inaugural Head of the School of Petroleum Engineering and Management (2001–2003) and full-time Professor at the University of Adelaide (2001–2007), with tenure since 2004. More recently, he taught as a Visiting Professor at the University of Western Australia (2014), Curtin University (2014), Stanford University (2000) and several other institutions.

7. Paper 5: Application of streaming analytics for Artificial Lift systems: a human-in-the-loop approach for analyzing clustered time-series data from progressive cavity pumps

The paper provides a comprehensive methodology for replicating the clustering process for multivariate time-series data. This begins with the creation of a performance heatmap-specific autoencoder designed to reduce the size of heatmap images, thus improving computational efficiency. The paper also explores various dimensionality reduction methods for visualizing the performance heatmaps in a two-dimensional space.

Moreover, the paper discusses the pivotal role of petroleum engineers in the human-in-the-loop approach for labelling performance heatmaps. Engineers are entrusted with cluster labeling to accurately categorize different PCP performance states. The paper presents a workflow for human-assisted labelling of streaming time-series data using Major and Anomaly clusters.

Petroleum and surveillance Engineers narrow down events of interest for real-time alerts by pairing relevant Major and Anomaly clusters. This empowers them to enhance their management of ALS through machine learning-supported exception-based surveillance. Based on the labelled Major and Anomaly pairs, the method identifies ten performance-related events and five anomalous events when analyzing the heatmap images, demonstrating its real-world effectiveness.

This collaborative effort between engineers and machine learning algorithms significantly enhances the accuracy and the overall efficacy of the streaming analytics system, hence assisting with improved management of a large fleet of CSG wells.

Statement of Authorship

Title of Paper	Application of streaming analytics for Artificial Lift systems: a human-in-the-loop approach for analysing clustered time-series data from progressive cavity pumps
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Saghir, F., M.E. Gonzalez Perdomo, and P. Behrenbruch, Application of streaming analytics for Artificial Lift systems: a human-in-the-loop approach for analysing clustered time-series data from progressive cavity pumps. Neural Computing and Applications, 2022. 35(2): p. 1247-1277.

Principal Author

Name of Principal Author (Candidate)	Fahd Saghir		
Contribution to the Paper	Conduct data analysis, write and record experiments, create test reports, tabulate results and write paper.		
Overall percentage (%)	75%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	19/09/2023

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Mary Gonzalez Perdomo		
Contribution to the Paper	Assisted with paper structure, writing and paper review (20%)		
Signature		Date	18/09/2023

Name of Co-Author	Peter Behrenbruch		
Contribution to the Paper	Assisted with paper structure, writing and paper review (5%)		
Signature		Date	10-09-2023

Please cut and paste additional co-author panels here as required.

Neural Computing and Applications (2023) 35:1247–1277
https://doi.org/10.1007/s00521-022-07995-8

REVIEW

Application of streaming analytics for Artificial Lift systems: a human-in-the-loop approach for analysing clustered time-series data from progressive cavity pumps

Fahd Saghir[1] · M. E. Gonzalez Perdomo[1] · Peter Behrenbruch[2]

Received: 6 December 2021 / Accepted: 25 October 2022 / Published online: 11 November 2022

Abstract

Assessing real-time performance of Artificial Lift Pumps is a prevalent time-series problem to tackle for natural gas operators in Eastern Australia. Multiple physics, data-driven, and hybrid approaches have been investigated to analyse or predict pump performance. However, these methods present a challenge in running compute-heavy algorithms on streaming time-series data. As there is limited research on novel approaches to tackle multivariate time-series analytics for Artificial Lift systems, this paper introduces a human-in-the-loop approach, where petroleum engineers label clustered time-series data to aid in streaming analytics. We rely on our recently developed novel approach of converting streaming time-series data into heatmap images to assist with real-time pump performance analytics. During this study, we were able to automate the labelling of streaming time-series data, which helped petroleum and well surveillance engineers better manage Artificial Lift Pumps through machine learning supported exception-based surveillance. The streaming analytics system developed as part of this research used historical time-series data from three hundred and fifty-nine (359) coal seam gas wells. The developed method is currently used by two natural gas operators, where the operators can accurately detect ten (10) performance-related events and five (5) anomalous events. This paper serves a two-fold purpose; first, we describe a step-by-step methodology that readers can use to reproduce the clustering method for multivariate time-series data. Second, we demonstrate how a human-in-the-loop approach adds value to the proposed method and achieves real-world results.

Keywords Multivariate time-series data · Unsupervised clustering · Machine Learning · Anomaly detection · Artificial lift · Progressive cavity pump · Coal seam gas

1 Introduction

The State of Queensland is home to approximately nine thousand natural gas wells [1], where energy operators depend on positive displacement pumps to produce hydrocarbons from these geographically dispersed Coal Seam Gas (CSG) assets. As the natural gas supplied from these wells is critical to sustaining energy demand in domestic and international markets, operators need to avoid unplanned downtime caused by pump failures. To monitor pump performance, data acquisition and control systems are deployed across the entire fleet of CSG wells, where they gather and transmit time-series data from pumps and well sensors. Depending on the natural gas operating company, a petroleum engineer may be assigned to manage anywhere from fifty to a hundred wells. They monitor streaming time-series data to evaluate pump performance and anticipate any failure that may disrupt gas production. However, monitoring, analysing, and mitigating issues on a well-by-well basis is a tedious task, and most often, critical pump events are either missed or ignored [2]. Most importantly, CSG producers are looking to add several hundred wells in the coming years to sustain global energy

✉ Fahd Saghir
fahd.saghir@adelaide.edu.au

[1] Australian School of Petroleum and Energy Resources, Santos Petroleum Engineering Building, University of Adelaide, Adelaide, SA 5005, Australia

[2] Bear and Brook Consulting, 135 Hilda Street, Corinda, QLD 4075, Australia

demand, which will only exacerbate the real-time pump performance analysis issue. This is where machine learning-assisted pump performance analysis can improve pump life.

1.1 Drawbacks of time-series analysis methods used for artificial lift systems

Generally, time-series analysis of Artificial Lift systems is based on either fuzzy logic [3, 4], physics-based models [5] or machine learning [6–9] based pattern recognition methods. However, such methods present a drawback when assessing an Artificial Lift system's performance as they identify events without context, which may or may not impact the pump performance. Moreover, these methods rely on labelled or known events, and any new or outlier events are not detected. Furthermore, it is rare to find labelled datasets for Artificial Lift applications. In most cases, the assistance of subject matter experts (SMEs) is required to label data sets for improved failure prediction results [10]. However, labelling patterns in raw time-series data is challenging for SMEs, as each pump presents a different data performance profile where the same anomaly or event may have very different characteristics, such as amplitude and length of an event.

1.2 Limitations of time-series clustering methods

In a recently published paper, where the authors benchmarked eight (8) well-known time-series clustering methods [11], they set limitations for their evaluation methods which are mentioned below:

1. Uniform length time-series: The benchmarked methods mentioned in the paper above were tested on time-series data of uniform length for a pre-defined time-window length. However, time-series data from industrial sensors mostly have non-uniform lengths.
2. Known number of clusters: The datasets tested to benchmark the clustering methods had a known number of clusters (or k values). As our previous publications have demonstrated [12–14], it is impossible to pre-define a set number of clusters for industrial time-series data, especially when dealing with data gathered from Artificial Lift Systems.

Another notable research on deep time-series-based clustering [15] mentions similar or related drawbacks. These will be discussed later in the Related Works section.

1.3 A practical approach for streaming time-series analysis of artificial lift systems

To address the drawbacks and limitations mentioned above, we propose a human-assisted approach to labelling clustered time-series data that can be utilized for running streaming performance analytics of positive displacement pumps.

Our research has two unique parts; firstly, we define a streamlined process to cluster multi-variate time-series data. This process is based on our previous research work where we convert multi-variate time-series data into performance heatmap images [14]. These images are then converted to unlabelled clusters based on the methodology defined later in the paper.

Second, to assist with the cluster labelling process, we developed a cluster analysis tool for engineers, where they could apply their petroleum domain expertise to label events of interest. Through this tool, petroleum engineers can combine their expertise with streaming analytics and automate the process of labelling events of interest, allowing them to manage Artificial Lift System proactively. Furthermore, petroleum engineers from two CSG operating companies currently use the cluster analysis tool system developed as part of this research for their daily analysis of approximately five-hundred PCP wells.

2 Overview of Coal Seam Gas production

In eastern Australia, natural gas is predominantly produced through CSG production, where coal seams are depressurized through a dewatering process that allows gas to escape from coal cleats and flow to the surface. Positive displacement pumps are installed in CSG wells, which produce water to the surface and, in the process, depressurize the coal seams. In the oil and gas industry, such pumps are referred to as Artificial Lift Pumps, and a network of these pumps collectively forms an Artificial Lift System. In Fig. 1 (Left), we see how water is displaced from the bottom of the well through the Production Tubing, and gas is produced via the production casing.

A salient characteristic of CSG wells is that they have a shorter production life span, usually ten (10) years, compared to conventional gas-producing wells. This lifespan is shown in Fig. 1 (right), with three (3) distinct stages, where a large quantity of water is produced initially to depressurize the coal seams, followed by a production stage with an increase in gas production. Finally, gas rates decline towards the end of the production lifecycle.

As gas production depletes quickly, CSG producers in Queensland must periodically drill and add new wells to

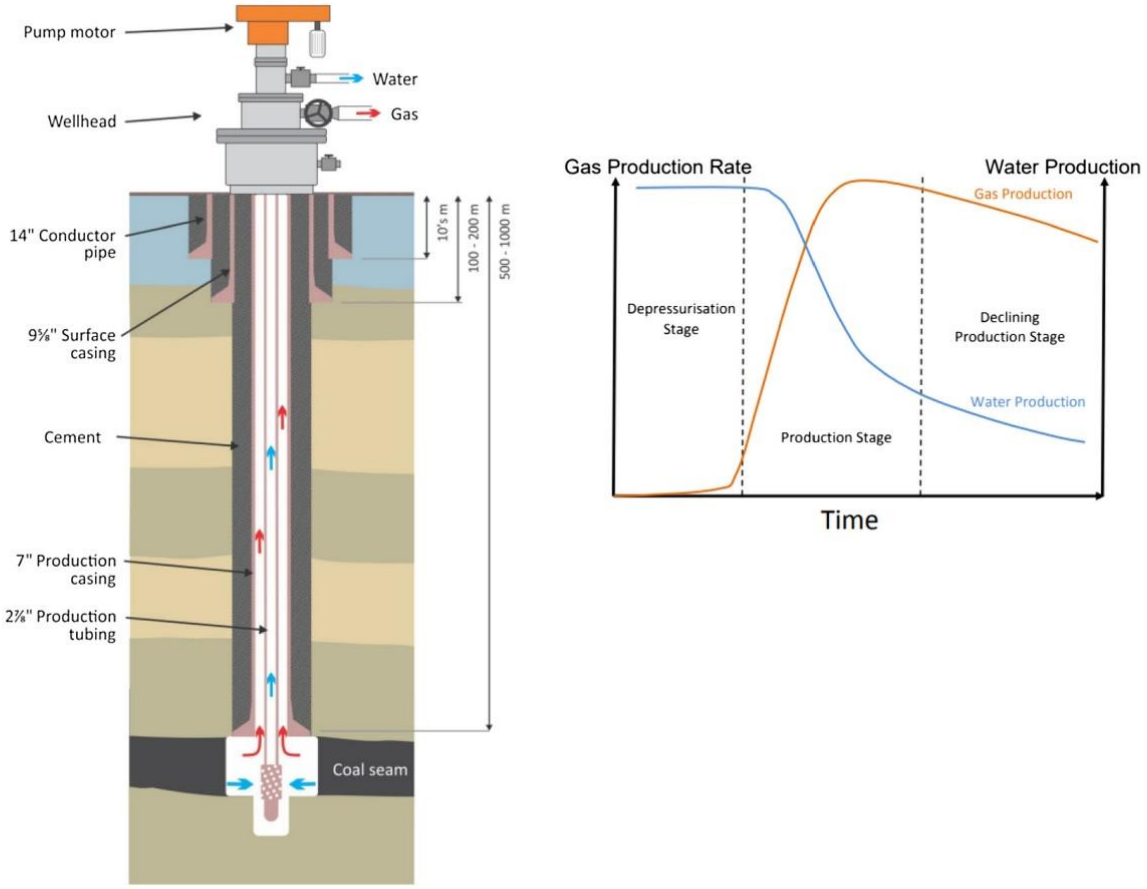


Fig. 1 (Left) Depiction of natural gas production from coal seam gas (CSG) wells [16], (right) production stages of a coal seam gas well [17]

maintain natural gas supplies. Hence, many CSG wells are dotted across Queensland, and this geographical spread and density are shown in Fig. 2.

2.1 Progressive Cavity Pumps

Like any positive displacement pump, a rotor and a stator work in tandem to push the liquid through to achieve vertical hydraulic lift. Figure 3 shows various components of a PCP assembly installed in a producing well. The rotor and elastomer assembly are designed such that the cavities between them push the fluid through when the rotor is operational.

Equations (1) and (2) show the correlation between speed, flow and torque. Time-series trends of these three parameters provide the necessary operational details of PCPs over their lifetime. Hence, the multivariate time-series analysis of our study will focus on these three parameters.

The correlation between flow and speed is shown in Eq. (1) [18].

$$q_{th} = s\omega \tag{1}$$

where q_{th} = theoretical flow, s = pump displacement, ω = rotational speed.

The correlation between torque and speed is shown in Eq. (2) [18].

$$T_{\mathrm{pr}} = \frac{P_{\mathrm{pmo}} E_{\mathrm{pt}}}{C\omega} \tag{2}$$

where T_{pr} = polished rod torque, P_{pmo} = prime mover power, E_{pt} = power transmission efficiency, C = constant, ω = rotational speed

2.2 Data gathering from CSG wells

As CSG wells are located in remote and geographically dispersed areas, operators must utilize Supervisory Control and Data Acquisition (SCADA) Systems to control wells

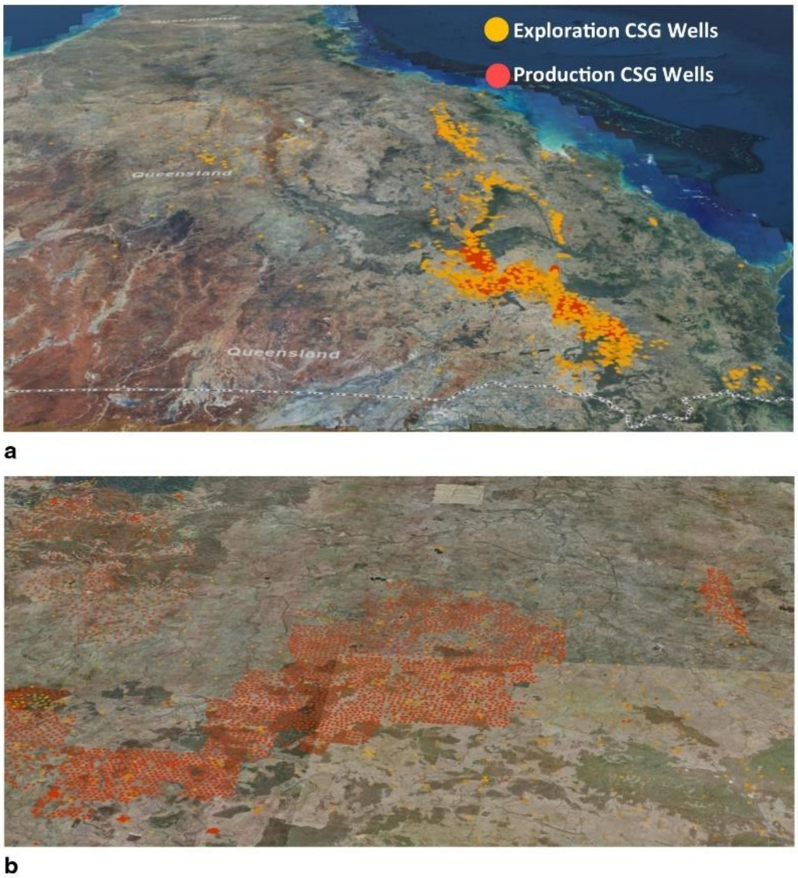


Fig. 2 a Geographical Spread of CSG Wells in Queensland. **b** High-Density Well Population of CSG Wells (Source: Queensland Globe)

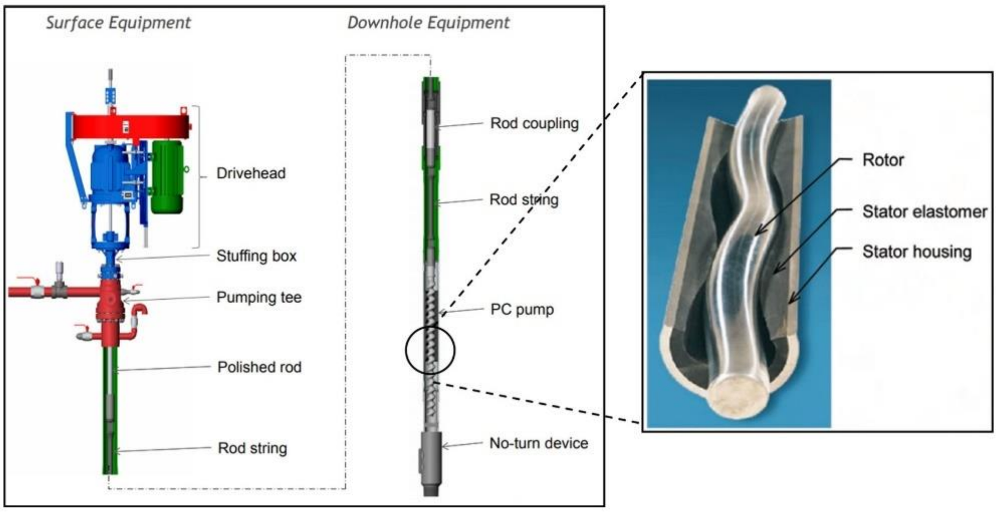


Fig. 3 (Left) Main components of a PCP system. (Right) Cut-out view of PCP rotor and stator [18]

through Wireless Telemetry. Ultra-high frequency (UHF) or microwave radio transmit data from CSG wells to a central control room. Figure 4 shows a layout of a typical CSG well site. The Remote Telemetry Unit (RTU) installed at each well site is responsible for recording data from multiple sensors and forwarding it to a central SCADA system. The data are stored and historized in data servers and delivered onwards to a corporate Historian database. It is important to note that data transferred via SCADA systems may not always have a fixed transmit rate; hence, data reporting time in most cases is asynchronous where time windows are not of identical length. Some SCADA systems use a report-by-exception approach, where data are only transmitted when a critical data point changes based on a pre-set percentage change. The report-by-exception method also produces data of unequal time windows.

3 Related work

Unlike univariate time-series data, applying anomaly detection and clustering methods to multivariate time-series data are a complex task which requires additional interpretation and insights [19]. In this section, we will further shed light on research gaps in multivariate time-series based anomaly detection and clustering methods. Furthermore, our previous work on Symbolic Aggregation Approximation (SAX)-based performance heatmap conversion [14] will be discussed to demonstrate why this novel approach provides a better basis for a human-in-the-loop approach when clustering multivariate time-series data. Finally, we will discuss why a human-in-the-loop approach adds value to time-series analysis process proposed in this paper.

3.1 Neural net-based anomaly detection

Neural nets have become a popular choice to solve anomaly detection problems in time-series data. One approach proposes using a fully connected convolutional network, U-Net, to identify anomalies in multivariate time-series data [20]. This method treats a fixed-length multivariate time-series snapshot as a multi-channel image. A U-Net segmentation technique is applied to obtain a final convolution layer corresponding to an augmentation map. The last layer includes the anomaly identification classes for the time-series snapshot, and each anomaly class is considered a mutually exclusive event. However, there are two drawbacks to this anomaly detection approach. Firstly, as the U-Net architecture accepts a fixed number of samples as input in a time window, the time-series data must be resized based on up-sampling or down-sampling techniques. Second, as each anomaly is a mutually exclusive event, it is difficult to segregate anomalies of interest from a routine change in process behaviour.

Another neural net-based anomaly detection approach proposes a Multi-Scale Convolutional Recurrent Encoder–Decoder (MSCRED) method [21]. This method converts multivariate time-series data into signature matrices based on the pairwise inner-product of time-series data streams. The matrices are encoded using a fully connected convolutional encoder. A Convolutional Long Short-Term

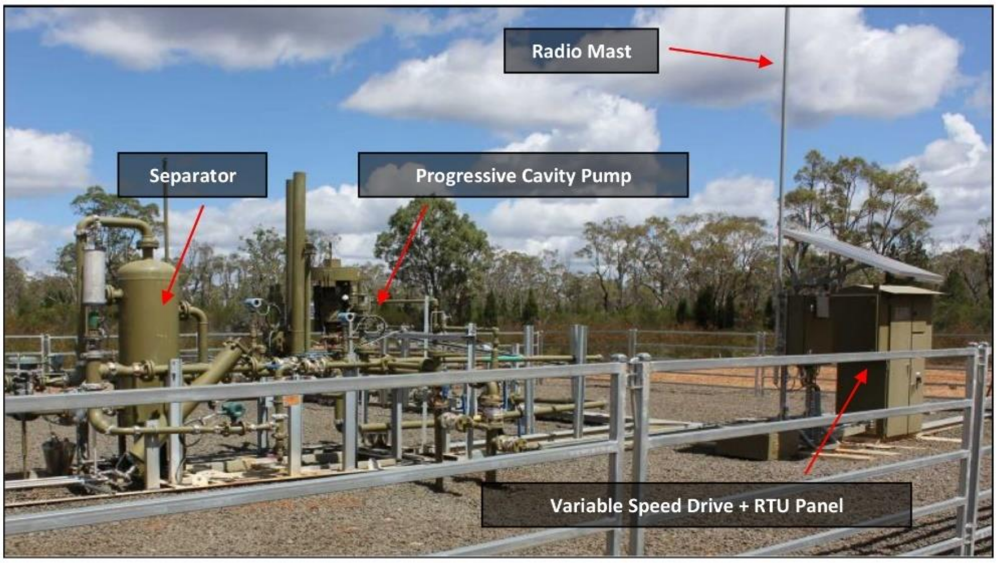


Fig. 4 Layout and key components of a CSG well

Memory (ConvLSTM) network is used to extract the hidden layer of each encoder stage, which is added to a convolutional decoder to produce a reconstructed signature matrix. The difference between the original signature and the reconstructed matrix is labelled as the residual signature matrix. This matrix defines a loss function that helps the model detect anomalies in multivariate time-series data. The residual signature matrix also helps determine the duration of anomaly events in time-series data based on small, medium, and large time-window duration.

Although the MSCRED methodology is novel in its approach, there are three limitations to using this approach for multivariate time-series analysis. Firstly, identifying anomaly events depends on the time-window duration. Therefore, the duration of the small, medium and large time windows will have to be tuned based on the properties of the time-series data and the application where it will be applied. Secondly, this approach does not consider the state of the process from time zero (t_0), when the process was initiated for the first time. This restriction, therefore, fails to observe any changes in pump mechanical degradation, which can provide additional insights into time-series-based performance analysis.

3.2 Neural net-based time-series clustering

Multiple research papers have recently been published on the use of neural net based time-series clustering methods [15, 22–24], both for univariate and multivariate data sets. These novel research methods extract feature matrices which are fed to a neural net architecture to extract low-dimensional embedding. The embeddings are then used to cluster the time-series data with a known clustering method, primarily the *k*-means method, which means the number of clusters must be known beforehand.

Although our approach is similar, we do not need to know the number of clusters beforehand. Most importantly, our low-dimensional embeddings are based on the novel approach of SAX derived time-series performance heatmap images.

3.3 Converting time-series data into performance heatmap images

This section provides an overview of how the SAX-based performance heatmaps are created for improved understanding of Artificial Lift Performance analysis and, more importantly, how these images provide contextual clustering of multivariate time-series data.

3.3.1 Expanding window technique

To understand how PCPs operate in CSG operations, it is essential to look at their performance from the day they are initiated into operation for dewatering wells. For this purpose, we use the expanding window technique shown in Fig. 5, which evaluates the multivariate data in the expansion stride based on the elapsed pump performance. By doing so, the exploratory data analysis methods utilized for performance analysis can capture the varying mechanical dynamics in the PCP through the pump's life.

3.3.2 Symbolic aggregation approximation (SAX)-based performance heatmaps for PCPs

Performance heatmaps help capture the temporal variation and time-window-based impact of multiple sensor readings in a single image [12]. By converting time-series data into performance heatmaps, it is possible to visualize the sequential variation in sensor readings while understanding the influence of change in sensor values between time windows. Furthermore, the performance heatmap approach is exempt from some of the shortcomings of other time-series-based image conversion techniques.

While other time-series to image conversion methods require a fixed sampling rate for each analysed time window to produce consistent images, the performance heatmap technique overcomes this deficiency by converting sensor values into Symbolic Aggregation Approximation (SAX) symbols [27]. The SAX symbols obtained through the conversion of time-series data are transformed into a symbol matrix and then converted to a performance heatmap—an example of SAX-based time-series image conversion [12]. Figure 6 shows a 1-h time-series trend of flow, speed and torque converted to a performance heatmap.

Moreover, most image conversion techniques [28] are developed for univariate time-series data. Although some techniques convert multivariate data into images [29], they mostly rely on converting univariate data into images and then either stack them horizontally or vertically to create a single 2D image.

3.3.3 Majority and anomaly heatmap images

Once the performance heatmaps are created, they can be split into majority and anomaly event images. Table 1 shows the time-based colour code used to label major, variation and anomaly event in a performance heatmap. In this study, we will only focus on majority and anomaly events, as the variation events are events in transition that are not significant in deducing any abnormal behaviour of the PCPs.

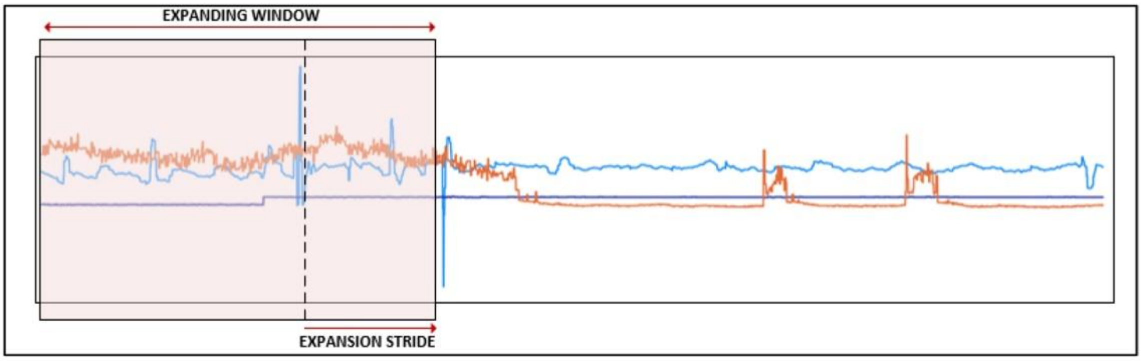


Fig. 5 Depiction of the Expanding Window technique captures PCP performance variation from time t_0

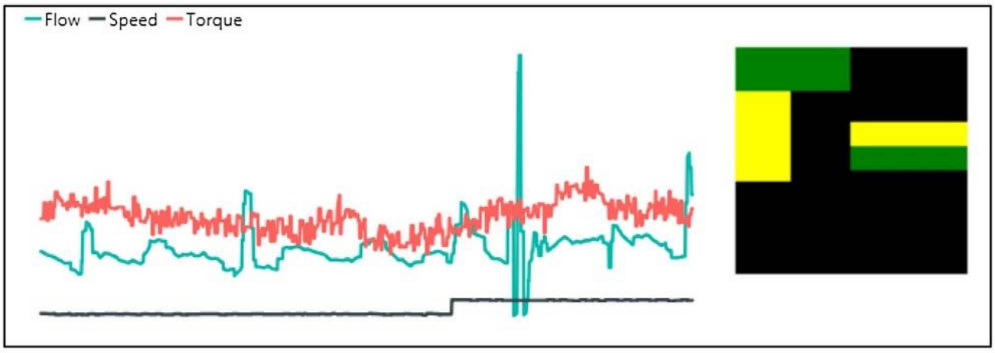


Fig. 6 An example of SAX based time-series image conversion [12]

Table 1 Color code for performance heatmaps based on the counts of SAX symbols in a 1-day window

SAX Symbol Count in One (1) Day Window (1440 MINUTES)	Colour Code	Performance Type
720 < Count < 1440	🟩	Majority Event
60 < Count < 720	🟨	Variation Event
00 < Count < 60	🟥	Anomaly Event
Count = 00	⬛	Nil

Figure 7 shows how a performance heatmap is split into majority and anomaly event images. We will use these images to create our unsupervised image cluster library for time-series data. Each image has a *48* × *48x3 (6912)* pixel dimension.

Figure 8 provides a breakdown of the majority and anomaly heatmaps where we have three (3) parameters (flow, torque and speed) represented in their respective columns. The position of the coloured boxes provides the state of each parameter, which can either be low, medium or high. These states can be used in cluster labelling processing to group various time-series clusters into similar performance and anomalous event categories.

3.4 Advantages of a human-in-the-loop approach for data labelling

Multivariate time-series data collected from industrial processes are seldom labelled, and hence, extracting any meaningful information from such data requires input from domain experts [19, 25]. This holds true for CSG operations, where the geophysical dynamics between coal seams and the pump require inference from domain experts for correct interpretation of normal and abnormal behaviour. By adding human inference to unlabelled data sets, it becomes easier for domain experts to accept the results generated by machine learning solutions [26].

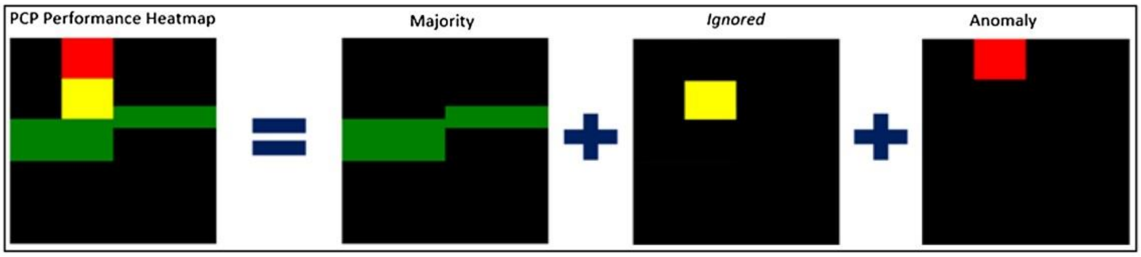


Fig. 7 Splitting a performance heatmap to majority and anomaly event images [12]

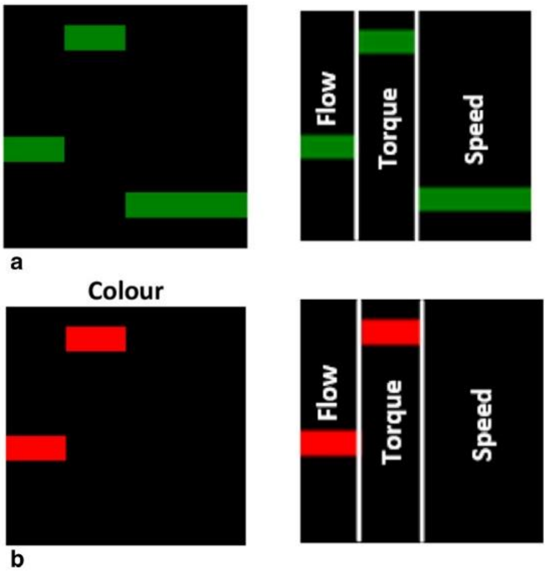


Fig. 8 **a** Majority heatmap depicting medium flow, high torque and low speed. **b** Anomaly heatmap depicting medium flow, high torque and no speed event

In our methodology, we rely on cluster labelling from petroleum engineers to correctly identify various performance states of PCPs used in Artificial Lift Systems. The SAX performance heatmaps provide petroleum engineers with a visual context as to why different clusters are identified based on the majority and anomaly heatmaps. In the next section, we will cover in detail the methodology to cluster SAX-derived performance heatmaps and how petroleum engineers label these clusters via a cluster analysis tool.

4 Methodology

In this section, we will discuss the methodology shown in Fig. 9, which is used to cluster un-labelled time-series data:

1. *Develop a Performance Heatmap-specific Auto-Encoder:* This step will be used as the first dimensionality reduction method to reduce the size of the images, which will lower memory usage and improve calculation times on the computer used for conducting the experiments.
2. *Embedding-based Dimensionality Reduction:* In this step, we will experiment with various dimensionality reduction methods (DRMs) and reduce the dimension of auto-encoded images to a 2-dimensional plane. Doing so will better visualize the performance heatmaps grouping in an XY plot.
3. *Hierarchical Density-Based Spatial Clustering (HDBSCAN): We* will use HDBSCAN to identify various clusters within the 2-dimensional plane of various DRMs in step 2.
4. *Cluster Labelling:* Once the images have been clustered, we will assign a label to known historical events from a selected number of wells using a cluster analysis tool. Once these events are labelled, we will use an automated cluster labelling pipeline to identify events on real-time data.

Assumptions Our methodology works under the following assumptions:

1. *Availability of Key Data Variables:* To analyse PCP performance, the time-series data should have flow, toque and speed variables. These three variables are needed to produce SAX performance heatmap that is required for the clustering process. For other multivariate time-series application, key variables should be defined based on the processed being analysed.
2. *Domain Expertise:* Petroleum engineers using the cluster analysis tool should have relevant experience in their field to properly label SAX performance heatmap clusters.
3. *Data Completeness: The* multivariate data set used for the clustering process must cover the entire operation cycle of a PCP in CSG operation, i.e. the time-series data from beginning to end-of-life of PCPs. This will

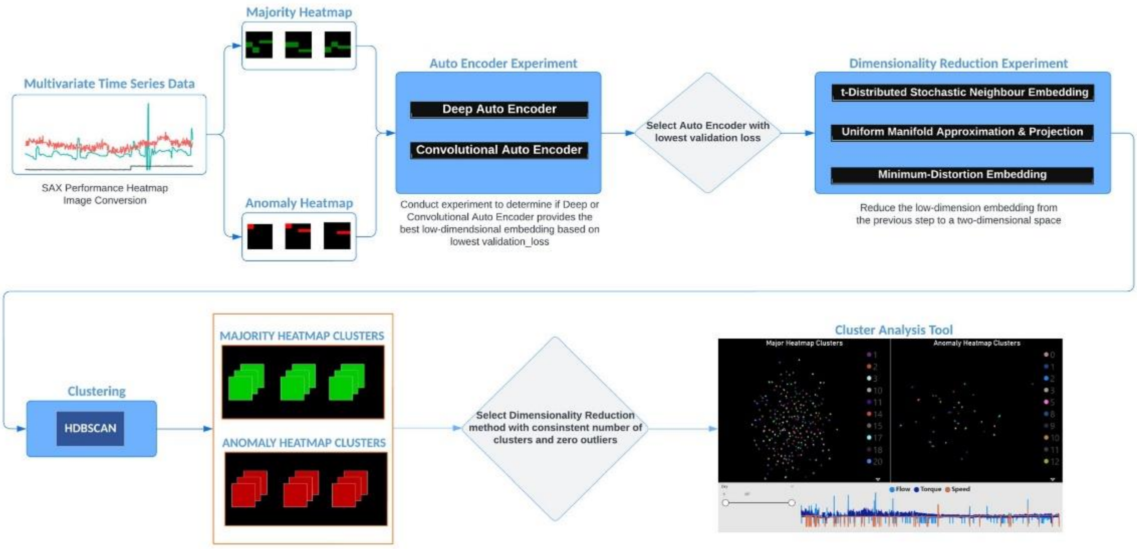

Majority Heatmap
Multivariate Time Series Data
SAX Performance Heatmap
Image Conversion
Anomaly Heatmap
Auto Encoder Experiment
Deep Auto Encoder
Convolutional Auto Encoder
Conduct experiment to determine if Deep or Convolutional Auto Encoder provides the best low-dimendsional embedding based on lowest validation_loss
Select Auto Encoder with lowest validation loss
Dimensionality Reduction Experiment
t-Distributed Stochastic Neighbour Embedding
Uniform Manifold Approximation & Projection
Minimum-Distortion Embedding
Reduce the low-dimension embedding from the previous step to a two-dimensional space
Clustering
HDBSCAN
MAJORITY HEATMAP CLUSTERS
ANOMALY HEATMAP CLUSTERS
Select Dimensionality Reduction method with consinstent number of clusters and zero outliers
Cluster Analysis Tool
Major Heatmap Clusters
Anomaly Heatmap Clusters
Flow
Torque
Speed

Fig. 9 An overview of the methodology steps

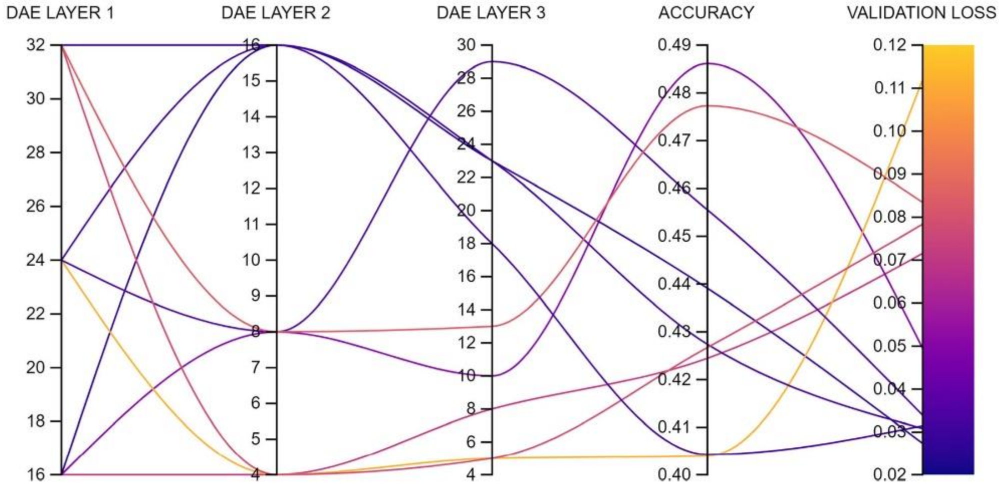


Fig. 10 An example of Weights and Biases (WANDB.) Sweep to analyse the effect of various deep auto-encoder (DAE) layers on validation loss

help capture various performance heatmaps over the life cycle of CSG wells.

Experiment tracking setup We used Weights & Biases [30] for experiment tracking and visualizations to develop insights for this paper. The Weights & Biases application allows automated tracking of machine learning experiments through Code Sweeps. Through this, multiple combinations of model training, hyperparameter tuning and clustering results can be captured and visualized to obtain the best results for machine learning projects. Figure 10 provides an overview of how multiple sweep experiments can be recorded and visualized to provide actionable insights into the effect of different layer properties for a deep auto-encoder (DAE). All coding for these experiments was done using Python 3.7 and necessary statistics, computer vision and machine learning libraries suited for this Python version.

4.1 I. Auto-encoder-based dimensionality reduction

This section will look at selecting the most optimum auto-encoder (AE) to reduce the latent representation of the

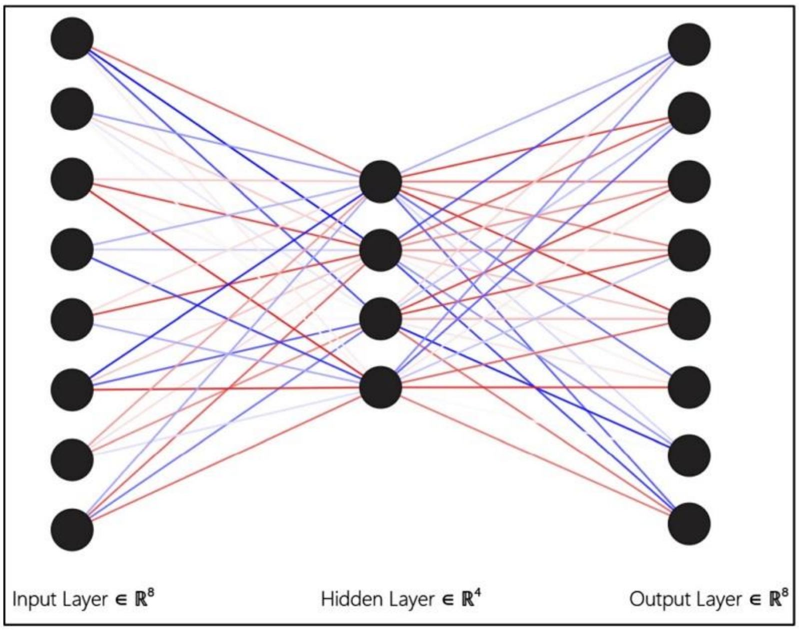

Input Layer ∈ $\mathbb{R}^8$
Hidden Layer ∈ $\mathbb{R}^4$
Output Layer ∈ $\mathbb{R}^8$

Fig. 11 An overview of a fully connected neural network with an input, hidden and output layer

performance heatmaps. The SAX-based performance heatmaps have a dimension of 48 × 48 ± 3 pixels (6912 pixels in total). Any data clustering problem must represent data in a two-dimensional space to evaluate results through an *X*–*Y* scatter plot. We will utilize an AE approach to minimize pixel dimensionality. Furthermore, reducing dimensionality allows us to examine many images due to reduced processing memory requirement, which improves the overall clustering analysis of the Performance Heatmaps.

4.1.1 i Deep auto-encoder (DAE)

We will start the experiment by developing a DAE that reduces the performance heatmap to fewer dimensions. Figure 11 shows a fully connected neural network with an input, hidden and output layer. These layers form a DAE, where the hidden layer is the reduced latent representation of the input layer. The output layer is the input layer reconstruction based on the hidden layer's interpretation. To gauge the performance of the DAE, we track the validation loss (*val_loss*), where the lowest value determines the best performing DAE architecture.

Table 2 Parameter settings for a 2-layer DAE sweep experiment

Parameter	Settings
Layer 1	[16, 32]
Layer 2	[4, 8, 16]
Train images	134, 346
Test images	33, 587
Loss function	Binary cross entropy
Epochs	100

Step 1–2-Layer DAE sweep run

In Step 1, we begin the experiment by evaluating a two-layer DAE to gauge the performance of *val_loss* over different channel sizes. The parameters for the first Sweep Run are as follows:

The settings shown in Table 2 run six (6) sweep experiments and measure the *val_loss* for different layer combinations. Figure 12 shows that for a two-layer deep neural network, a 16-channel *layer2* produces the minimum *val_loss* compared to an eight or four-channel *layer2*. As shown in Fig. 13, the decoded image for a *16* × *8*-channel DAE configuration does not accurately represent the original image. However, as shown in Fig. 14, the decoded image for a *16* × *16*-channel DAE configuration is a more accurate copy of the original image. Table 3 confirms that *sweep-4*, where *layer1* and *layer2* are sixteen-channel each, produces the best *val_loss* for a two-layer DAE.

Step 2–3-Layer DAE sweep run

In this step, we will add a third layer to the DAE and try further dimensionality reduction. As per Table 4, we will try dimensions *8*, *4* and *2* for the third layer in the DAE Table 3 shows the setup for the Sweep experiment used in this step.

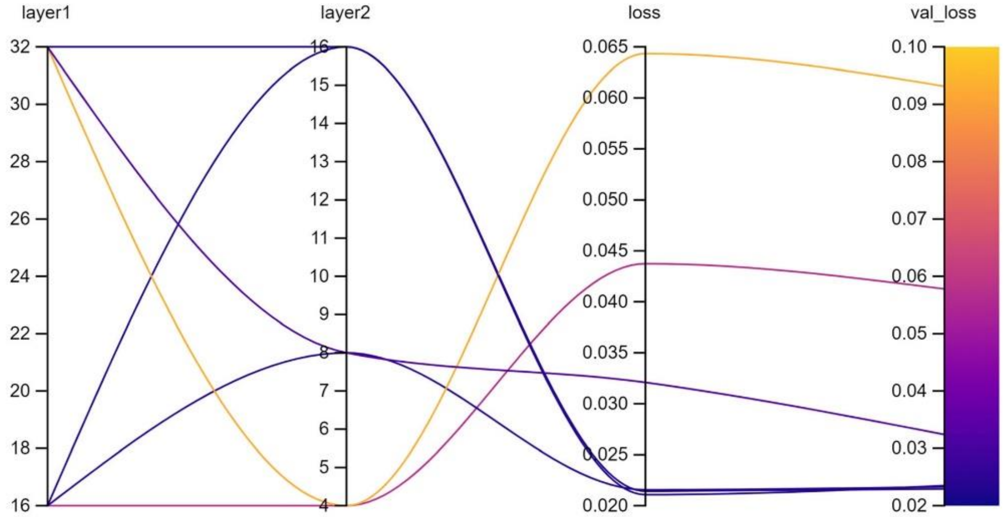


Fig. 12 Results showing the *val_loss* from a 2 Layer DAE Sweep Run

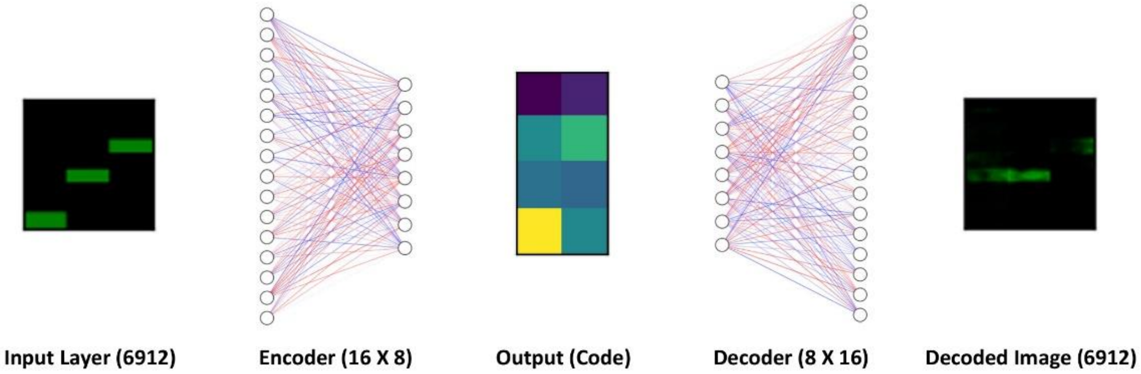


Fig. 13 Result of *16* × *8* DAE showing the Decoded Image versus the Input Image

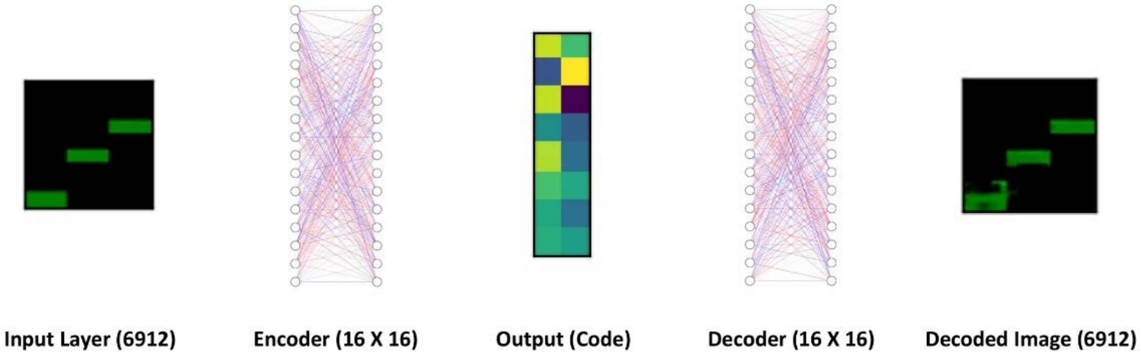


Fig. 14 Result of *16* × *16* DAE showing the decoded image versus the input image

Table 3 *val_loss* results from the 2 Layer DAE sweep run

Name	layer1	layer2	Loss	val_loss
sweep-6	16	4	0.043735	0.057785
sweep-5	16	8	0.021559	0.023302
sweep-4	**16**	**16**	**0.021424**	**0.02292**
sweep-3	32	4	0.064341	0.093155
sweep-2	32	8	0.032088	0.032394
sweep-1	32	16	0.021083	0.023502

Table 4 Parameter settings for a 3-layer DAE sweep experiment

Parameter	Settings
Layer 1	[16]
Layer 2	[16]
Layer 3	[2, 4, 8]
Train images	134, 346
Test images	33, 587
Loss function	Binary cross entropy
Epochs	100

Figure 15 provides an overview of the three-layer DAE Sweep experiment. A *16* × *16x8* DAE produces the minimum *val_loss*, and the results of this layer configuration are shown in Fig. 16. Results from all three (3) sweep runs are summarized in Table 5.

Step 3—4-Layer DAE sweep run

In this step, we will experiment with dimensions *8, 4* and *2* in the four-layer of the DAE. Table 6 shows the setup for this sweep experiment.

Figure 17 shows that further dimensionality reduction to *2* or *4* channels increases the *val_loss*; hence, further reduction from *8* channels is not feasible. However, an eight *(8)* channel fourth-layer does improve the overall val_loss of the DAE from ***0.02292*** (Table 5) to ***0.022079*** (Table 7). Results from the four-layer DAE are shown in Fig. 18, which validates that reducing dimensionality below *8* channels is not feasible with a D.A.E. Hence, our final DAE configuration is *16* × *16x8* × *8* for reducing the time-series heatmaps from 6912 pixels (*48* × *48x3*) to 8 dimensions.

4.1.2 ii. Convolutional auto-encoder

To see if the *val_loss* and dimensions can be reduced further, we will use a four-layer convolutional auto-encoder (CAE) architecture. Table 8 shows the Sweep experiment parameters that are investigated using a four-layer architecture to see if the CAE can reduce the image to *8* or fewer dimensions while improving val_*loss*.As shown in Fig. 19, CAE, *val_loss* for less than *8* dimensions in the fourth layer is relatively high versus *4* or *2* dimensions. However, the *16* × *16x8* × *8* CAE configuration further reduces the *val_loss* compared to the DAE. Table 9 shows the comparison between the DAE and CAE *val_loss*. Based on this result, we will use a *16* × *16x8* × *8* CAE to encode the *48* × *48x3* major and anomaly event images to *8* dimensions. The final CAE architecture to encode the images is shown in Fig. 20.

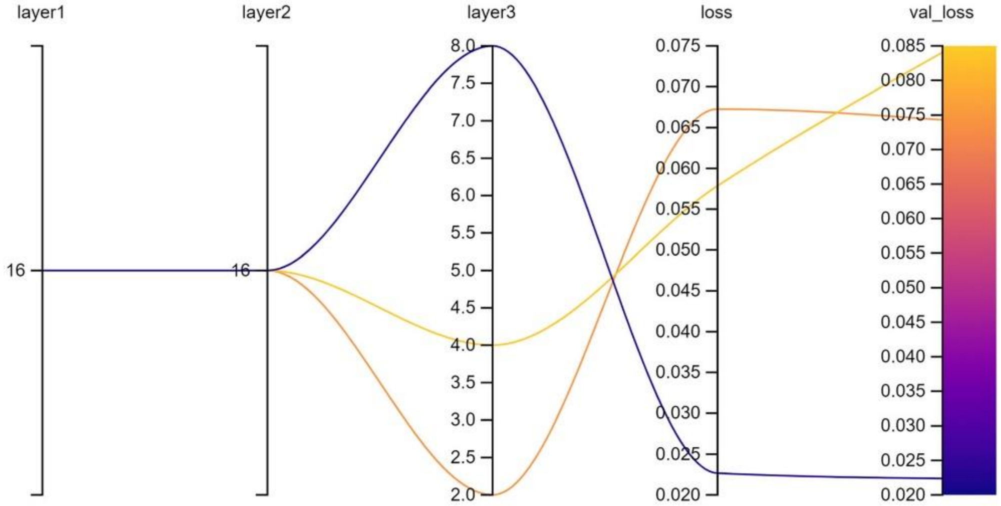


Fig. 15 Results showing the *val_loss* from a 3-layer DAE sweep run

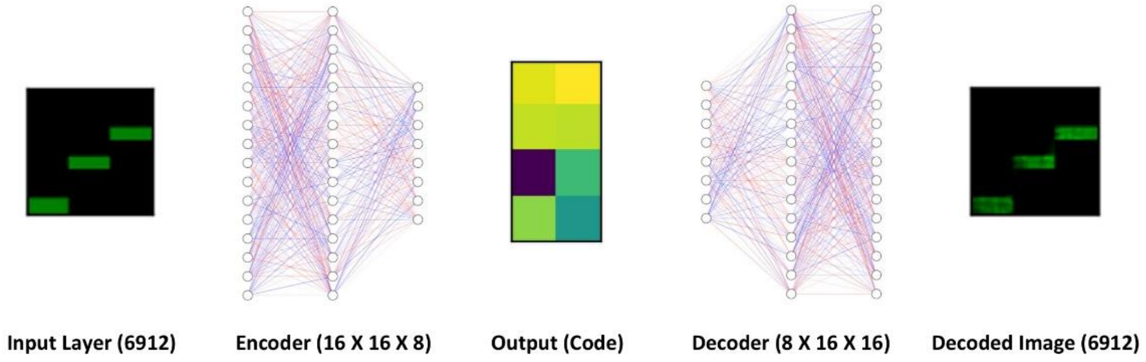


Fig. 16 Result of *16* × *16x8* DAE showing the decoded image versus the input image

Table 5 *val_loss* results from the 3-layer DAE sweep run

Name	layer1	layer2	layer3	loss	val_loss
sweep-3	16	16	2	0.067277	0.074296
sweep-2	16	16	4	0.057862	0.084062
sweep-1	**16**	**16**	**8**	**0.022665**	**0.022379**

Table 6 Parameter settings for a 4-layer DAE sweep experiment

Parameter	Settings
Layer 1	[16]
Layer 2	[16]
Layer 3	[8]
Layer 4	[2, 4, 8]
Train images	134,346
Test images	33,587
Loss function	Binary cross entropy
Epochs	100

4.2 II. High-density dimensionality reduction

We have demonstrated that the time-series-based images can be reduced to a latent size of eight (8) dimensions with a convolutional autoencoder. However, to provide a visual distribution context to time-series image clustering, we need to reduce the number of dimensions to two (2), and this can be achieved by utilizing high-density dimensionality reduction techniques. For this paper, we will experiment with three (3) methods which are *t*-distributed stochastic neighbour embedding (*t*-SNE) [31], uniform manifold approximation and project (UMAP) [32], and the minimum-distortion embedding (MDE) [33] method. These methods take high-density multi-dimensional points and assign them to a two-dimensional map.

4.2.1 i *t*-Distributed stochastic neighbour embedding (*t*-SNE)

t-SNE determines the conditional probability of high-dimensional data points by computing the Euclidean distance between the points. The probability represents similarities between two points and determines if these points could be picked as neighbours [31]. The probability $p_{j|i}$ is represented by Eq. (3) [31], where x_i and x_j are the data points being compared for similarity. Figure 21 depicts the t-SNE distribution for various numbers of major heatmap images. This distribution provides abstract localization with no recognizable high-density areas for the images.

$$p_{(ji)} = \frac{\exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum \exp\left(\frac{-\|x_i - x_k\|^2}{2\sigma_i^2}\right)} \tag{3}$$

where $p_{j|i}$ = conditional probability.

4.2.2 ii. Uniform manifold approximation and projection (UMAP)

Although *t*-SNE and UMAP share some clustering similarities [32], UMAP differentiates itself by creating high- and low-dimensional similarities for the distances between two points. Equations (4) and (5) [32] provide an overview of how these dimensionalities are calculated.

$$v_{j|i} = \exp\left[\frac{\left(-d\left(x_i, x_j\right) - \rho_i\right)}{\sigma_i}\right] \tag{4}$$

where $v_{j|i}$= high dimensional similarities, σ_i = normalizing factor, ρ_i = distance to the nearest neighbour

$$w_{ij} = \left(1 + a \parallel y_i - y_j \parallel_2^{2b}\right)^{-1} \tag{5}$$

where W_{ij} = low-dimensional similarities.

Figure 22 shows the UMAP distribution of various number of heatmap images. As highlighted in Fig. 23, the

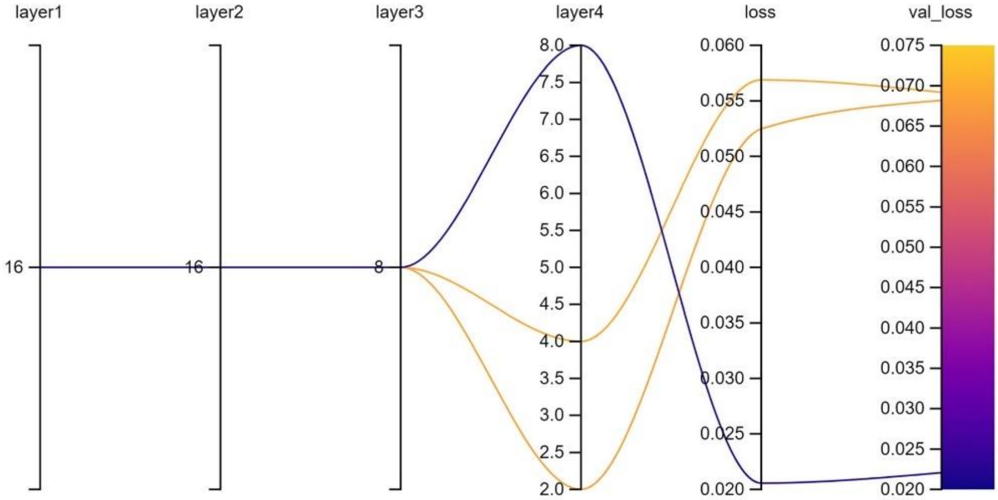


Fig. 17 Results showing the *val_loss* from a 4-layer DAE sweep run

Table 7 *val_loss* results from the 4-layer DAE sweep run

Name	layer1	layer2	layer3	layer4	loss	val_loss
sweep-3	16	16	8	2	0.05247	0.068188
sweep-2	16	16	8	4	0.056899	0.069199
sweep-1	**16**	**16**	**8**	**8**	**0.020568**	**0.022079**

high-density groupings are visible within the overall high-dimensional structure.

4.2.3 iii. Minimum-distortion embedding (MDE)

As the name suggests, the MDE dimensionality reduction method (DRM) pairs items based on vectors calculated from distortion functions. Equation (6) [32] shows the equation to calculate the embedding for vectors, aiming to minimize the average distortion. Like UMAP, similar items will have vectors near each other, and different items will have far apart vectors.

$$(X) = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} f_{ij}(d_{ij}) \tag{6}$$

where E = embedding, $d_{ij} = \|x_i - x_j\|_2$ = set of allowable embeddings, f_{ij} = distortion functions, $\mathcal{E}$ = set of vector pairs.

Figure 24 depicts the distribution of the embeddings for various number of major heatmap images. Again, a concentrated mass in the centre represents similar vectors, and dissimilar vectors are spread around the concentrated group.

Figure 25b shows a zoomed-in view of the concentrated mass of the similar vectors, and Fig. 25c shows how this mass further consists of neighbourhoods of high-density areas.

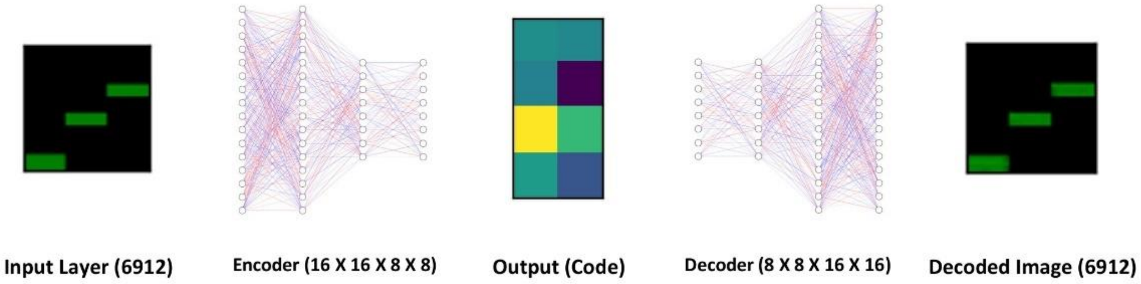


Fig. 18 Result of *16* × *16x8* × *8* DAE showing the decoded image versus the input image

Table 8 Parameter settings for a 4-layer CAE sweep experiment

Parameter	Settings
Layer 1	[16]
Layer 2	[16]
Layer 3	[8]
Layer 4	[2, 4, 8]
Train Images	134, 346
Test images	33, 587
Loss function	Binary cross entropy
Epochs	100

4.3 III. Hierarchical density-based spatial clustering (HDBSCAN)

In this step, we will use HDBSCAN to conduct unsupervised clustering of the major heatmap images. The HDBSCAN algorithm is a density-based clustering method, where a simplified cluster tree is produced from which significant clusters are extracted [34].

Based on the experiment run, we see that t-SNE 2-d dimensions produced increasing clusters as the image numbers increased. However, UMAP and MDE have very

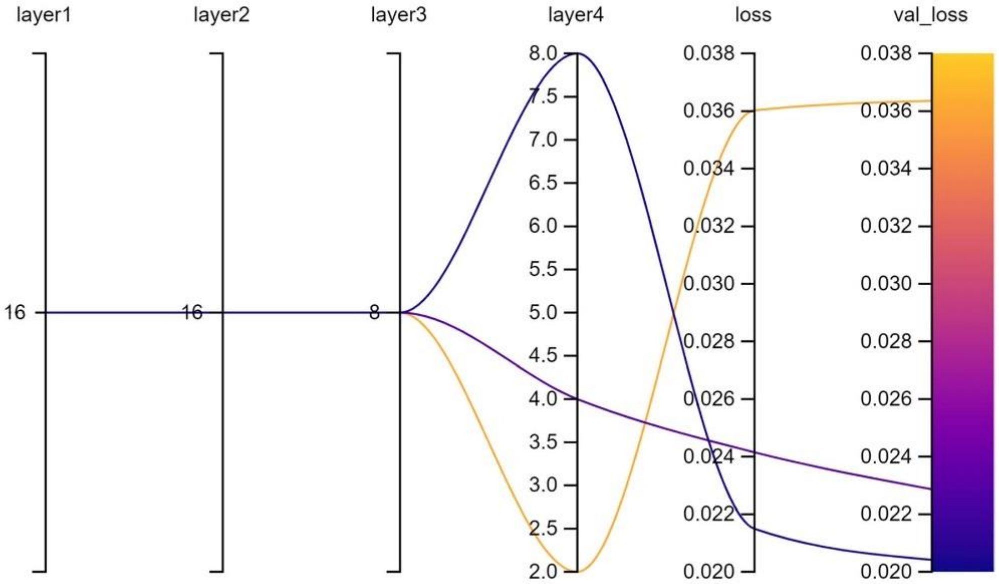


Fig. 19 Sweep run confirming that *16* × *16x8* × *8* CAE architecture produces the best *val_loss* for a 4 Layer CAE

Table 9 Parameter settings for a 4-layer CAE sweep experiment

Encoder type	layer1	layer2	layer3	layer4	loss	val_loss
DAE	16	16	8	8	0.020568	0.022079
CAE	**16**	**16**	**8**	**8**	**0.0215**	**0.02042**

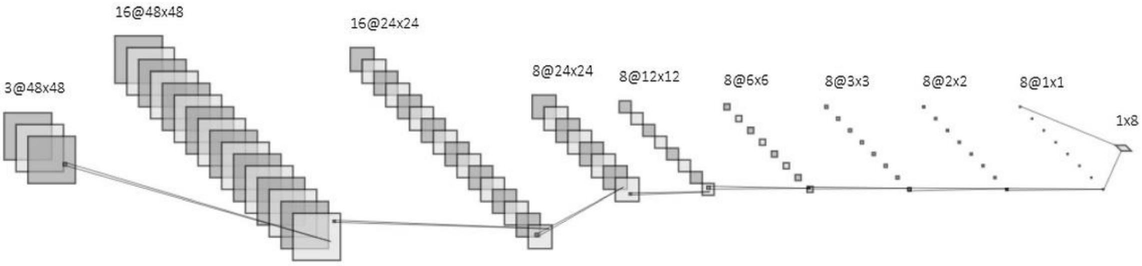


Fig. 20 *16* × *16x8* × *8* Convolutional auto-encoder

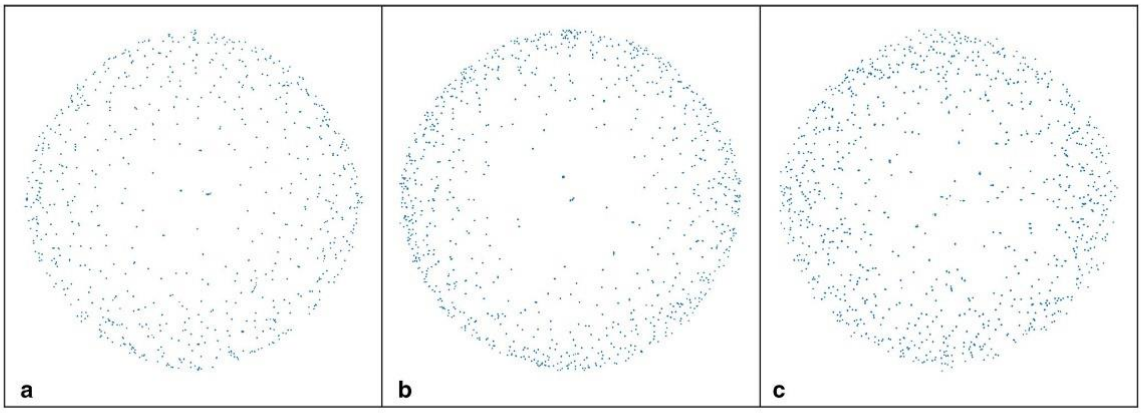


Fig. 21 a *t*-SNE distribution for 172,193 major heatmap images, **b** *t*-SNE distribution for 554,436 major heatmap images, **c** *t*-SNE distribution for 817,475 major heatmap images

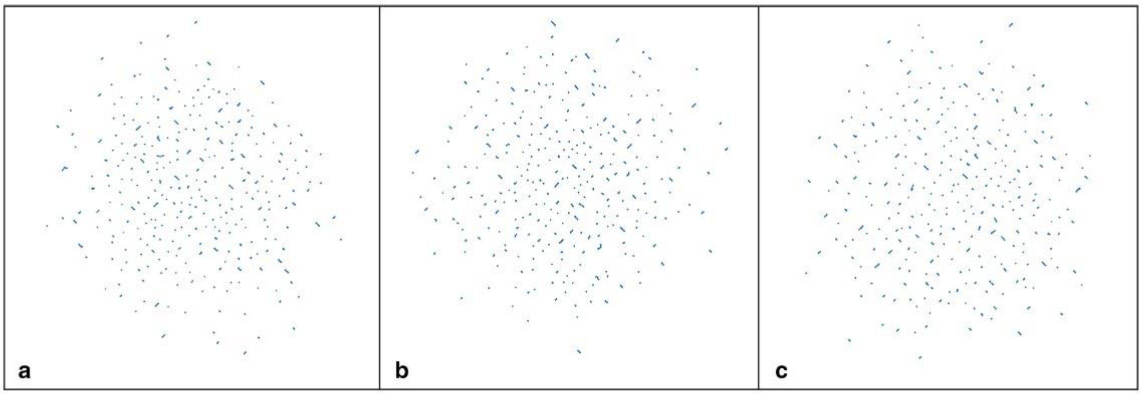


Fig. 22 **a** UMAP distribution for 172,193 major heatmap images, **b** UMAP distribution for 554,436 major heatmap images, c UMAP distribution for 817,475 major heatmap images

similar cluster numbers with the increase in image numbers. These results are shown in Fig. 26.

The experiment parameters are set as per Table 10. The cluster size determines the minimum samples in a cluster for it to be considered unique, and the sample size determines the critical mass within the cluster neighbourhood [35]. As per Table 10, our cluster size is *5,* and the sample size is *200*, which means that our clusters should have a minimum of *5* points, and if that cluster grows beyond *200* points, then that cluster becomes a core point.

Figure 27 shows the minimum and maximum clusters produced by each DRM method. Based on this experiment run, we will discard *t*-SNE for any further assessment. Also, the UMAP clustering with HDBSCAN provides the narrowest distribution of clusters across the number of images used in this experiment. Based on this information, we will have an in-depth look at UMAP and MDE clusters to understand how the images are sorted in the 2D plane and confirm which DRM methods provide us with a useable cluster distribution.

4.3.1 i. Clustering analysis

To understand the cluster formation in the MDE and UMAP reduction methods, we need to look at how the number of clusters and outliers varies with different combinations of cluster size and sample size parameters in HDBSCAN. To do this, we will run a clustering experiment with the parameters shown in Table 11.

Figure 28 shows how cluster size and sample size impact the cluster and outlier count in the MDE and UMAP reduction methods. Clustering the UMAP distribution provides consistent cluster counts, with zero outliers in most cases. Moreover, running an independent UMAP clustering experiment as per Table 12 shows that the sample size of less than 30 produces the most consistent

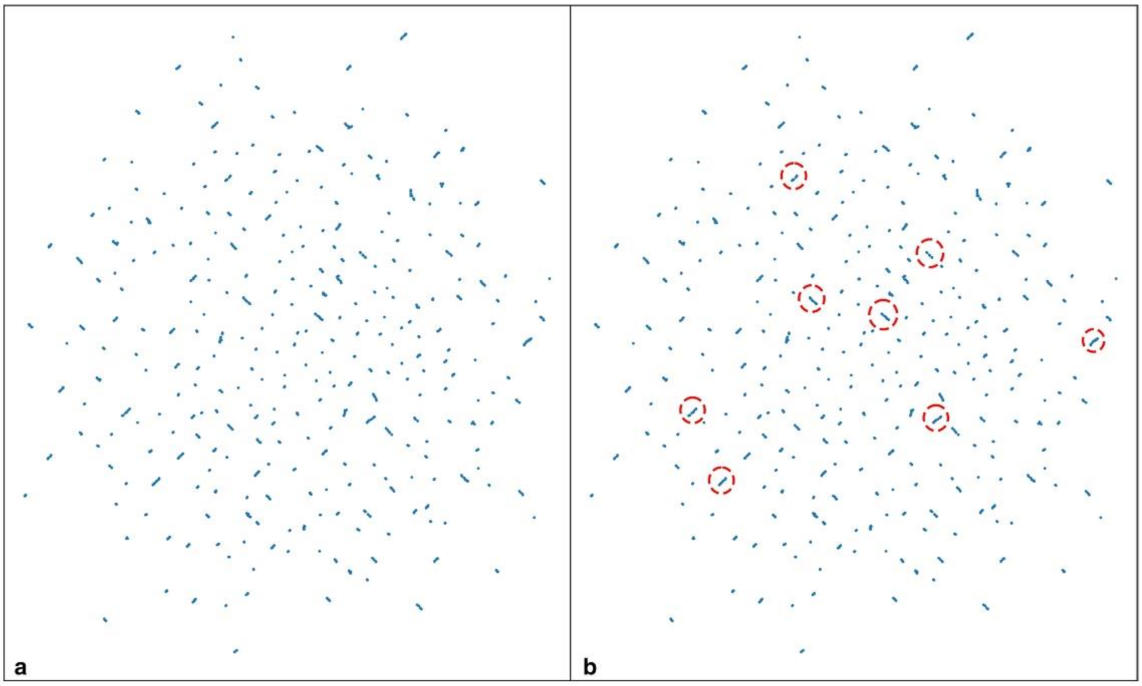


Fig. 23 a UMAP distribution of major heatmaps (554,436 Images), **b** observations of high-density areas in the UMAP distributions, which are separated from low-density areas

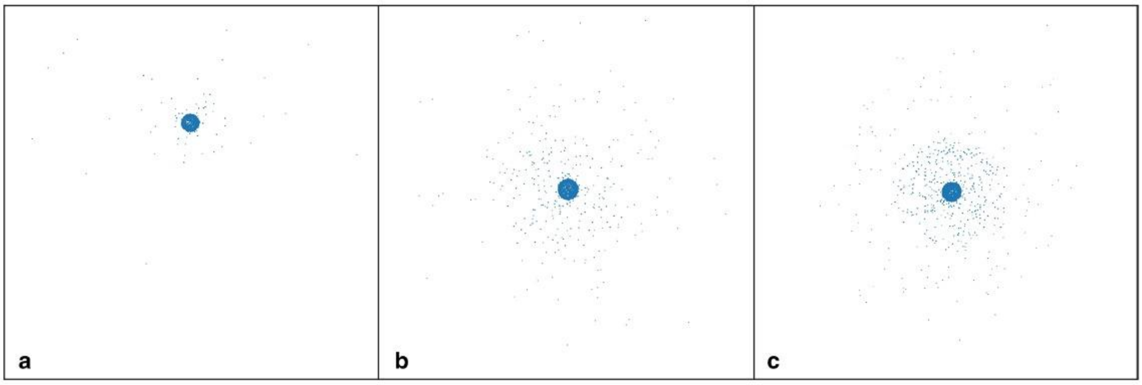


Fig. 24 a MDE distribution for 172,193 major heatmap images, **b** MDE distribution for 554,436 major heatmap images, **c** MDE distribution for 817,475 major heatmap images

results where the outliers are minimised and the number of clusters is below 1000. The details of the independent UMAP clustering experiment are shown in Table 13 and Fig. 29. In Fig. 30a, we see that the MDE method produces a large spread of outliers beyond the core cluster area when the sample size is 5 and the cluster size is 2. However, in Fig. 30b we observe that the UMAP method produces 996 clusters with 0 outliers with the same sample and cluster size settings. Hence, it is clear that the UMAP DRM produces the most consistent number of HDBSCAN-derived clusters when the sample size is set to 5.

The 996 clusters and 0 outliers from the UMAP clusters will be used to identify time-series events. Although we have 996 clusters identified in the UMAP distribution, we will use the time-series labelling methodology to generalize the cluster grouping.

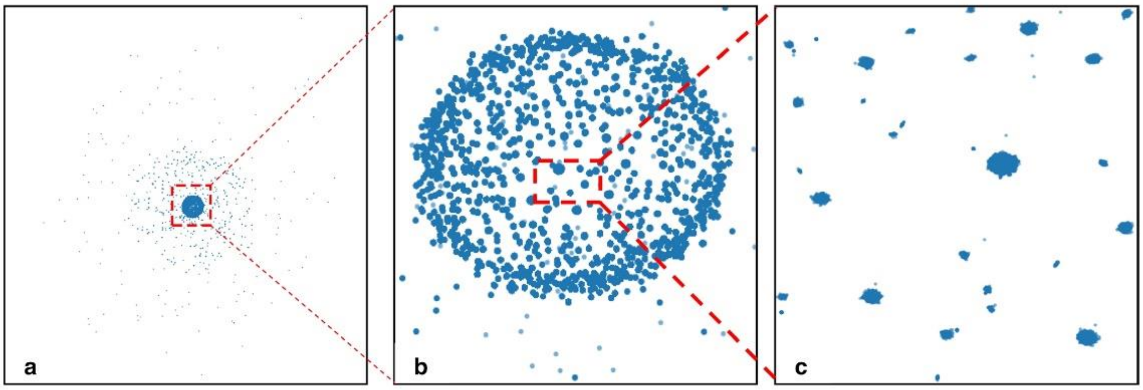


Fig. 25 **a** MDE distribution for 554,436 major heatmap images, **b** zoomed view of the high-density area, c zoomed view of the neighbourhoods within the high-density MDE area

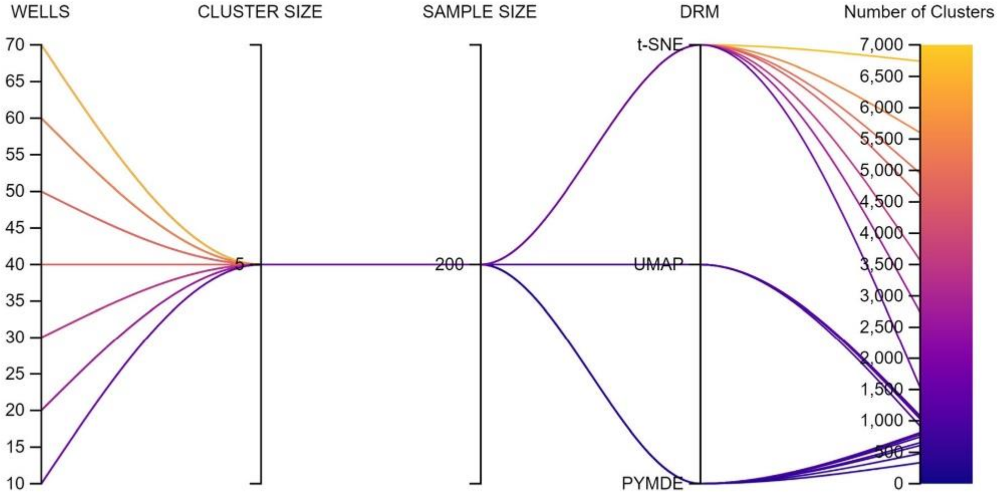


Fig. 26 Experiment results highlighting the progression of the number of clusters when running HDBSCAN clustering on different DRM methods

Table 10 Parameter settings for the HDBSCAN unsupervised clustering experiment

Parameter	Settings
Wells	[10, 20, 30, 40, 50, 60, 70]
Cluster size	[5]
Sample size	[200]
Dimensionality Reduction	[*t*-SNE, UMAP, MDE]

4.3.2 ii. Analysing the UMAP and HDBSCAN clusters for Performance Heatmap grouping

To understand the cluster formation in the UMAP reduction methods, we will investigate two cluster areas, as shown in Fig. 31. The performance image grouping for major heatmaps, as shown in Fig. 32a and b, depicts that similar images are grouped in their respective high-density areas. We will use the assigned cluster numbers to label PCP performance events and identify any cluster repetition patterns.

Using the experiment steps explained in the previous sections, we get a UMAP and HDBSCAN cluster layout for anomaly heatmaps, as shown in Fig. 33. For the anomaly heatmaps, we get 98 clusters and 0 outliers. Investigating Cluster Area 1, we see the groupings created in the identified dense area. Like major heatmaps, we will use these cluster numbers to identify abnormal and anomalous PCP performance events.

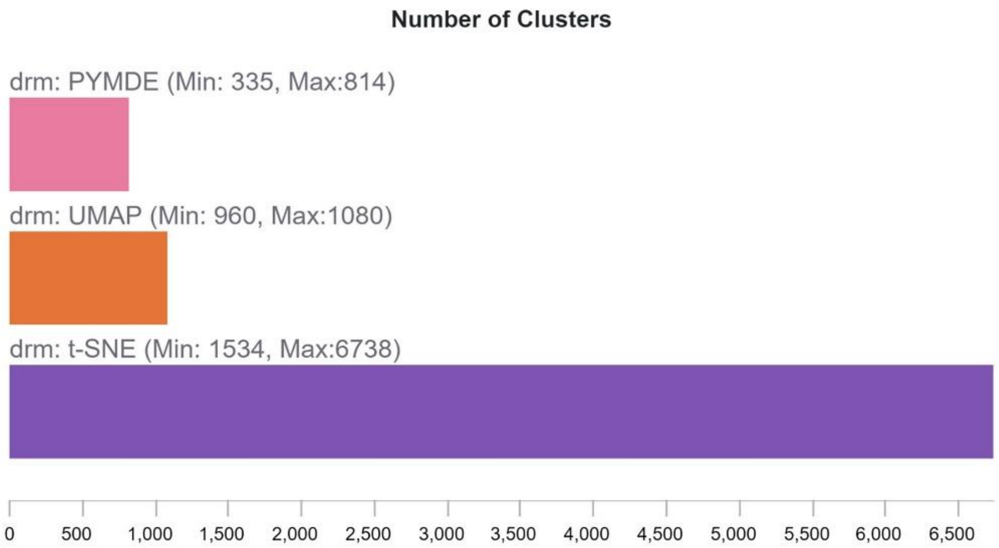


Fig. 27 Number of minimum and maximum clusters per DRM method. The minimum cluster number corresponds to images from 10 wells, and the maximum cluster number corresponds to images from 70 wells

Table 11 Parameter Settings for the HDBSCAN unsupervised clustering experiment

Parameter	Settings
Wells	[70]
Cluster size	[2, 5, 10, 15, 25]
Sample size	[5, 10, 25, 50, 100, 200]
Dimensionality reduction	[UMAP, MDE]

Table 12 Parameter settings for the HDBSCAN unsupervised clustering experiment

Parameter	Settings
Wells	[70]
Cluster size	[2, 5, 10, 15, 25]
Sample size	[5, 10, 25, 50, 100, 200]
Dimensionality reduction	[UMAP]

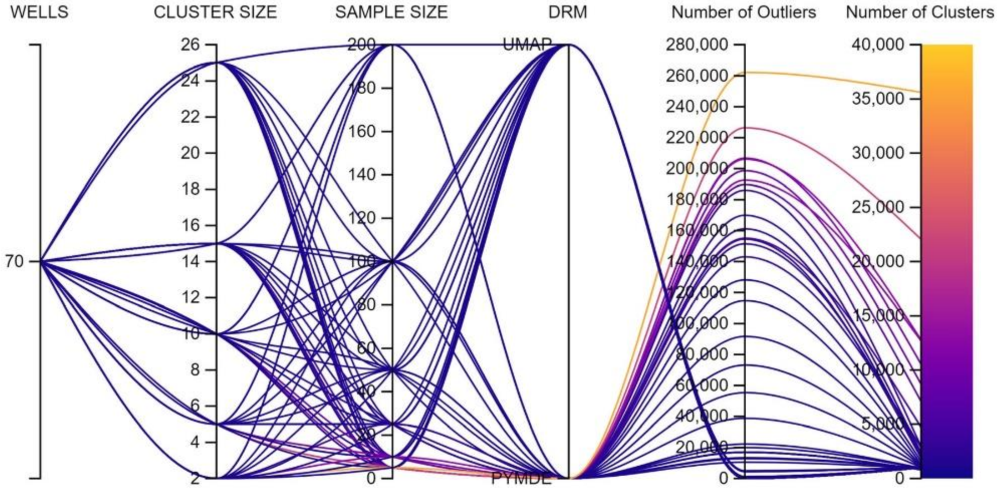


Fig. 28 Experiment run showing the effect of Cluster Size and Sample Size on UMAP and MDE cluster count

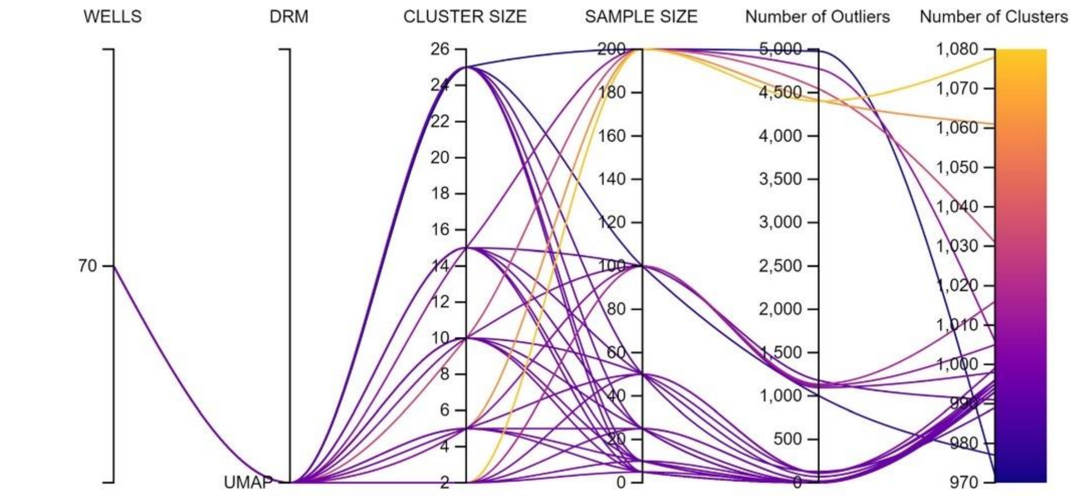


Fig. 29 Experiment run showing the effect of cluster size and sample size on UMAP cluster and outlier distribution

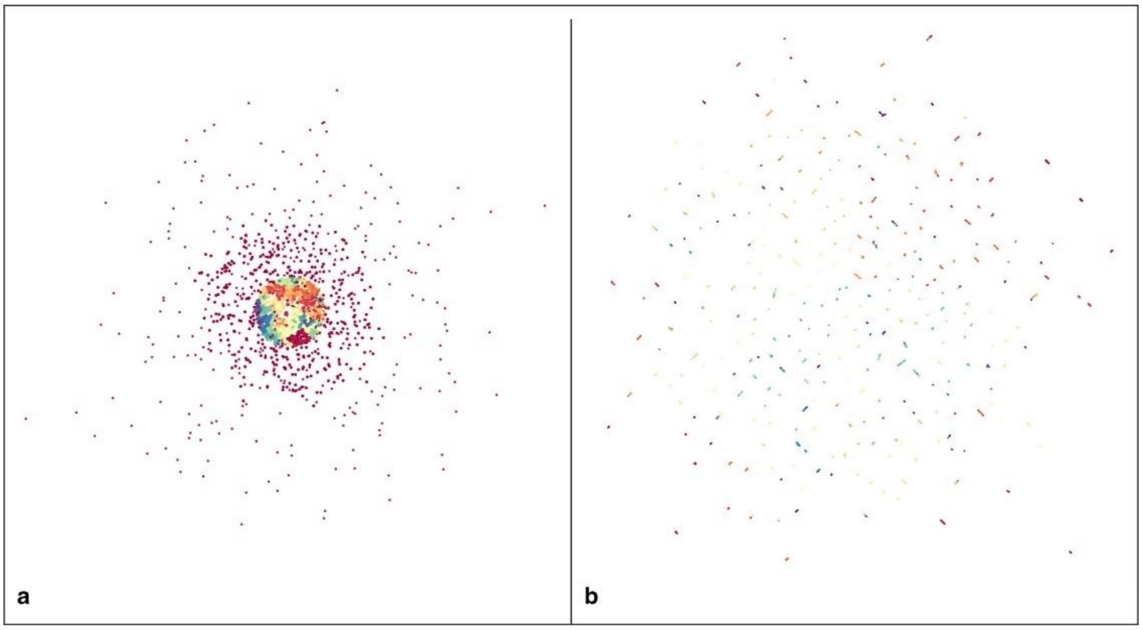


Fig. 30 **a** MDE clusters depicting 1055 clusters and 10082 outliers, **b** UMAP clusters representing 996 clusters and 0 outliers

4.4 IV Cluster labelling

After numbering the major and anomaly heatmap clusters in the previous step, we will now use a cluster labelling tool to add context to cluster numbers. The cluster labelling tool, developed using PowerBI, provides an intuitive approach to identifying clusters. As part of the cluster labelling process, as depicted in Fig. 34, we will use pre-identified event dates marked by production and Artificial Lift engineers to label events of interest. Furthermore, we will also discuss how cluster labelling can help identify the progression of the majority of heatmaps over the lifetime of a well and depict the degradation of a PCP.

4.4.1 i. Cluster labelling tool

The cluster labelling tool has three (3) areas shown in Fig. 35. The major heatmap clusters and anomaly heatmap

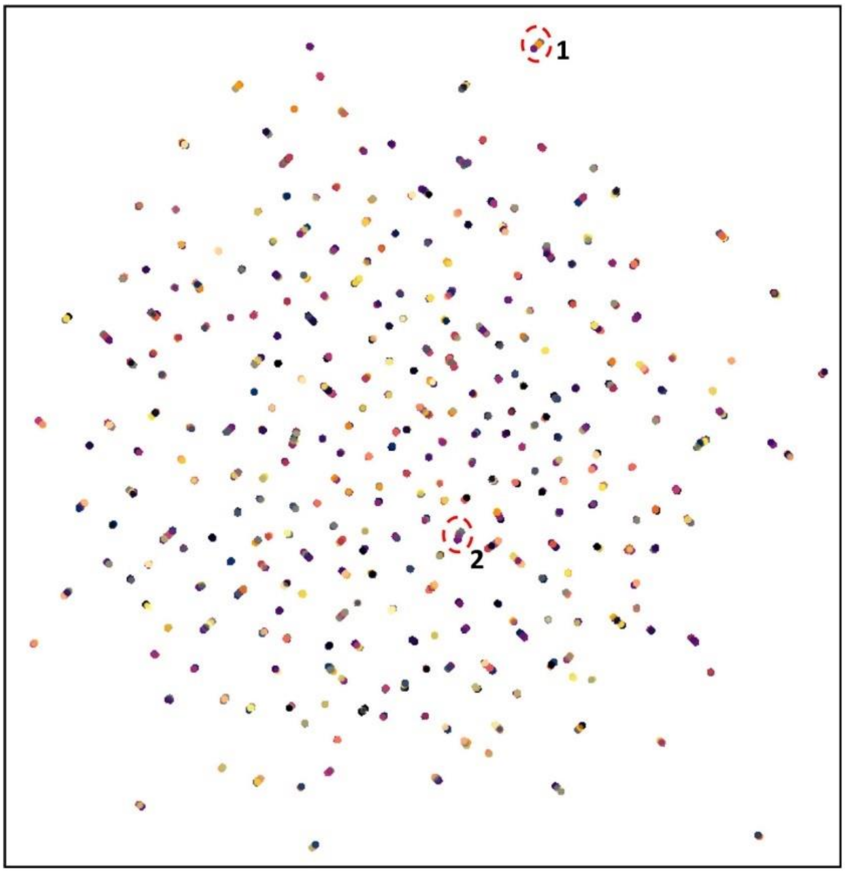


Fig. 31 Area 1 and Cluster Area 2 are investigated to understand the performance heatmap grouping

clusters areas, shown in Fig. 35a and b, respectively, show the UMAP distribution of clusters for a particular well being analysed for labelling the time-series data. Fig. 35c shows the time-series trend with a days filter to browse various periods where abnormal or activity of interest may have occurred during PCP operations. Such periods can then be used to place clusters in categories that identify abnormal or anomalous PCP behaviour.

In Fig. 36, we look at a flow disturbance event on Day 84 of PCP operation. Two major heatmap clusters (*44*, *240*) and two anomaly heatmap clusters (*87*, *90*) were observed on Day 83. Upon selecting the area of flow disturbance on the time-series trend, we observed that both anomaly heatmap clusters, *87* & *90*, are prevalent and relate to the flow column, as shown in Fig. 8. Furthermore, when the petroleum engineer selects abnormal behaviour period (red dotted area in Fig. 37), only major heatmap cluster *240* is visible, which depicts that the pump is in a high flow and high torque state. Hence, by looking at this grouping, we can state that on Day 84, the PCP saw flow anomaly events while in a high flow, high torque state. Using this methodology, we can group the major and anomaly heatmaps clusters into various states of PCP operations.

5 Results

This study aimed to demonstrate a streamlined and reproducible method of labelling time-series data gathered from CSG wells, so it may aid production engineers with identifying PCP performance profiles and abnormal production events. Our end-to-end approach is shown in Fig. 38, where saved cluster weights and labels help the streaming analytics process and allow operators to manage PCP wells by exception.

We will highlight in this section how our methodology produces meaningful labelled cluster groups that can be visualized as a coloured sequential bar chart against time-series data. Moreover, the anomalous events detected by our method were consistent among the two operators, specifically the solids and gas through pump events which are detrimental to PCP operational life. The streaming

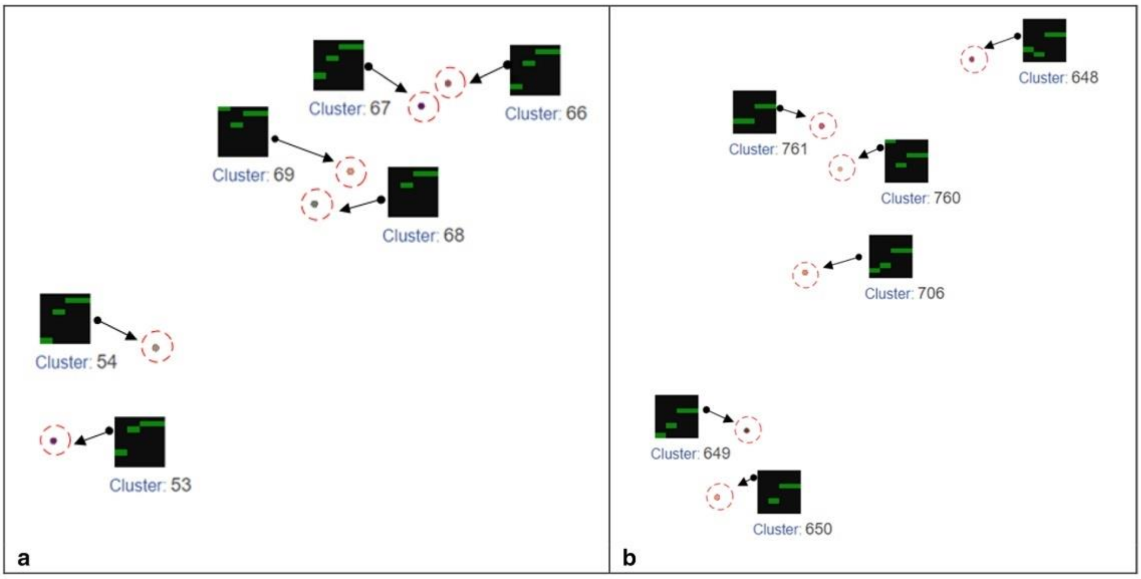


Fig. 32 **a** Major heatmap grouping in Cluster Area 1, **b** major grouping in Cluster Area 2

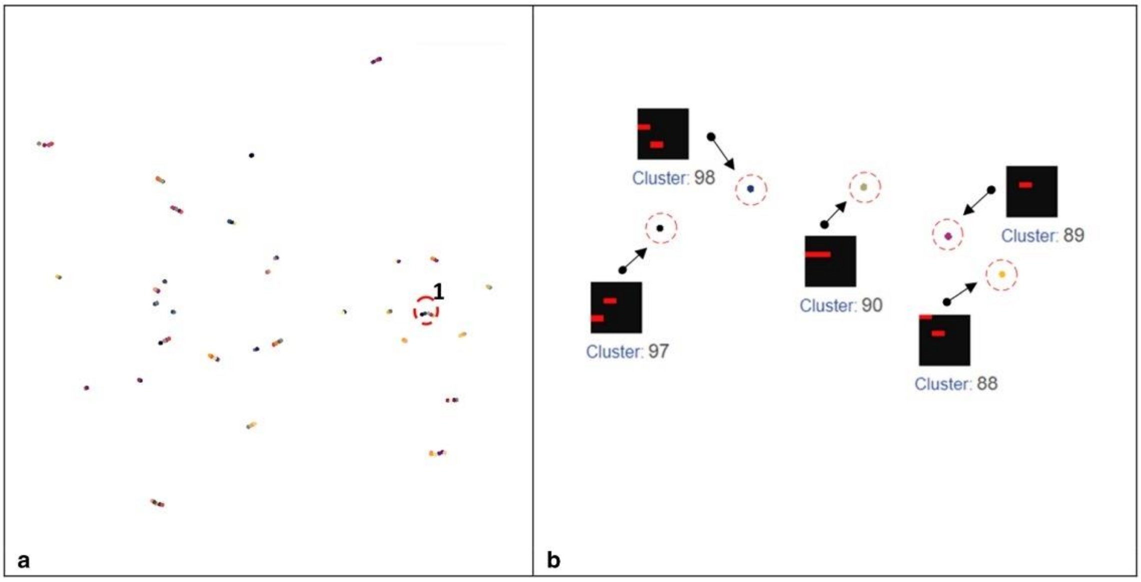


Fig. 33 **a** Cluster grouping for anomaly heatmaps, **b** anomaly heatmap grouping in Cluster Area 1

analytics approach was not only able to capture the abnormal event amplitude but also the longevity of the event.

5.1 I. Grouping cluster labels

Based on the observations made with the cluster labelling tool on *20* wells, the *996* major heatmap clusters and *98* anomaly heatmaps were segregated into groups, as shown in Table 14. The group labels were defined based on the experience of production and well surveillance engineers.

Table 13 HDBSCAN clustering sweep results showing the effect of sample size and cluster size on UMAP clusters

Cluster size	Sample size	Number of clusters	Number of outliers
2	**5**	**996**	**0**
15	**5**	**995**	**0**
25	**5**	**994**	**0**
5	**5**	**996**	**0**
10	**5**	**996**	**0**
5	**10**	**996**	**0**
2	**10**	**996**	**0**
10	**10**	**996**	**0**
25	**10**	**994**	**0**
15	**10**	**995**	**0**
10	**25**	**994**	**0**
25	**25**	**993**	**0**
5	25	996	18
15	**25**	**993**	**0**
2	25	996	18
5	50	996	110
2	50	999	116
25	50	989	128
15	50	993	69
10	50	995	115
15	100	991	1179
25	100	977	1002
5	100	1005	1118
10	100	998	1096
2	100	1016	1136
15	200	1006	4774
10	200	1031	4541
5	200	1061	4414
25	200	972	4976
2	200	1078	4402

In Table 15, we see the groups with a sample set of images they represent. For example, the major heatmap group labelled erratic torque shows that images within this group did not have a stable torque profile as no green box was recorded in the centre column. This depicts that the torque fluctuated significantly in this period; hence, no SAX character existed long enough to record any symbol count as a major event as per Table 1. Similarly, the image grouping for anomaly heatmaps in Table 16 describes the state that the PCP is in momentarily or may be considered an abrupt change.

It is important to note that the characteristics of major or the anomaly heatmap groups can provide a performance profile for PCPs independently or in combination.

5.2 II. Cluster sequencing and visual analytics

To understand how heatmap groups define the performance of a PCP, we will use the colour code in Table 17 to represent each image group. These colour codes are then plotted along with the time-series data to understand the cluster sequencing and identify patterns in PCP performance.

In Fig. 39, we see the heatmap groups plotted with the time-series trend for a lifespan of PCP well. Figure 39a represents the major heatmap group, and Fig. 39c represents the anomaly heatmap group. As seen in the progression of the major heatmap groups, it matches the state of the PCP performance during the dewatering, stable-flow, and high-torque pumping regimes.

If we look closer at a one-week PCP performance window, as shown in Fig. 40, the details in the major and anomaly heatmap groups become more apparent. In this case, we see the major heatmap groups clearly marking the areas of solids through the pump where the PCP torque increases. At the same time, anomaly heatmap groups also present markers of change (primarily high torque, high

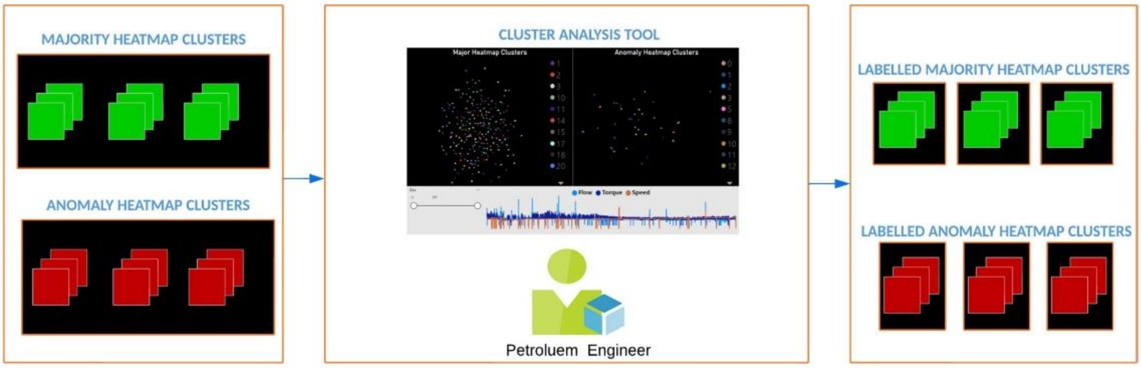


Fig. 34 Cluster labelling process with human-in-the-loop

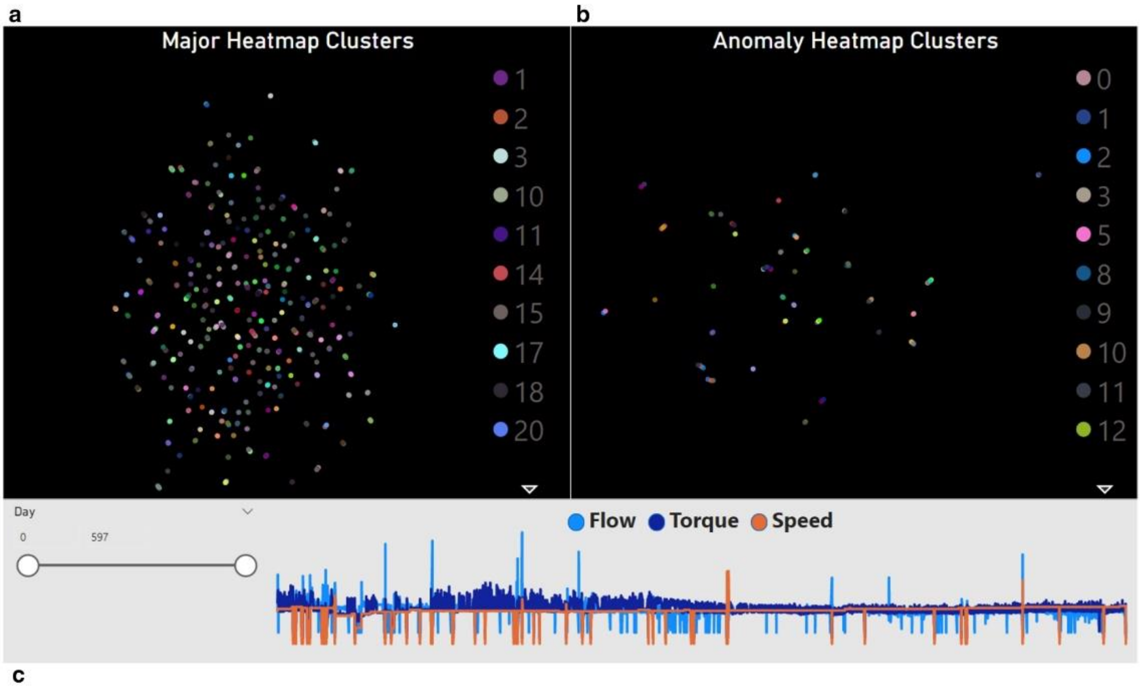


Fig. 35 **a** Major heatmap cluster area, **b** Anomaly heatmap cluster area, **c** Time-series trend area

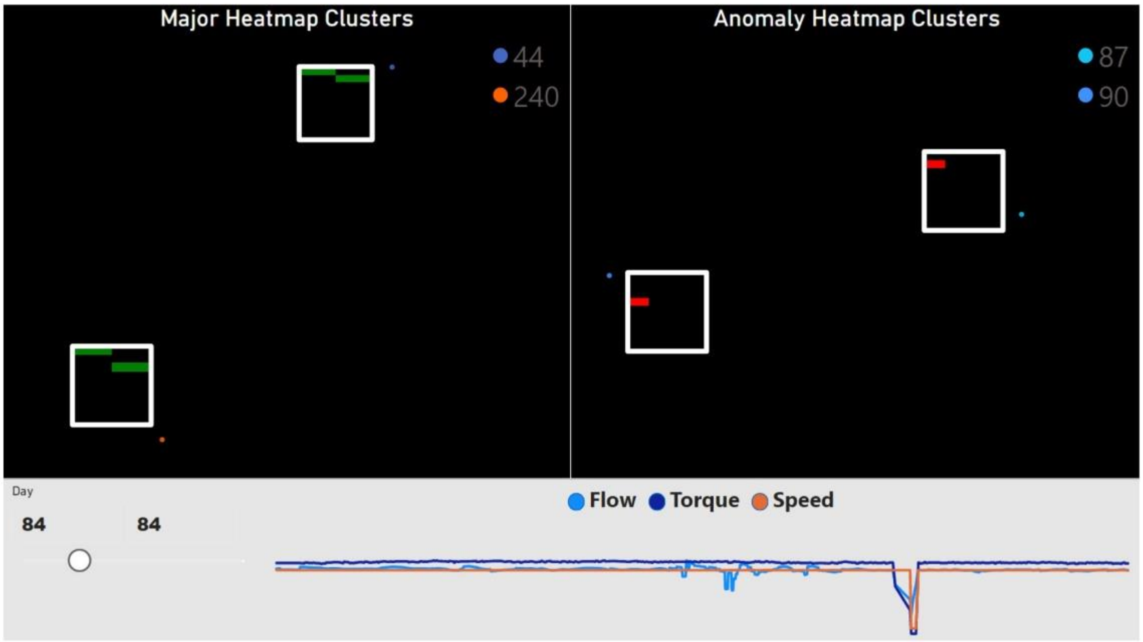


Fig. 36 Day 84 of PCP operations with the respective major and anomaly heatmap clusters

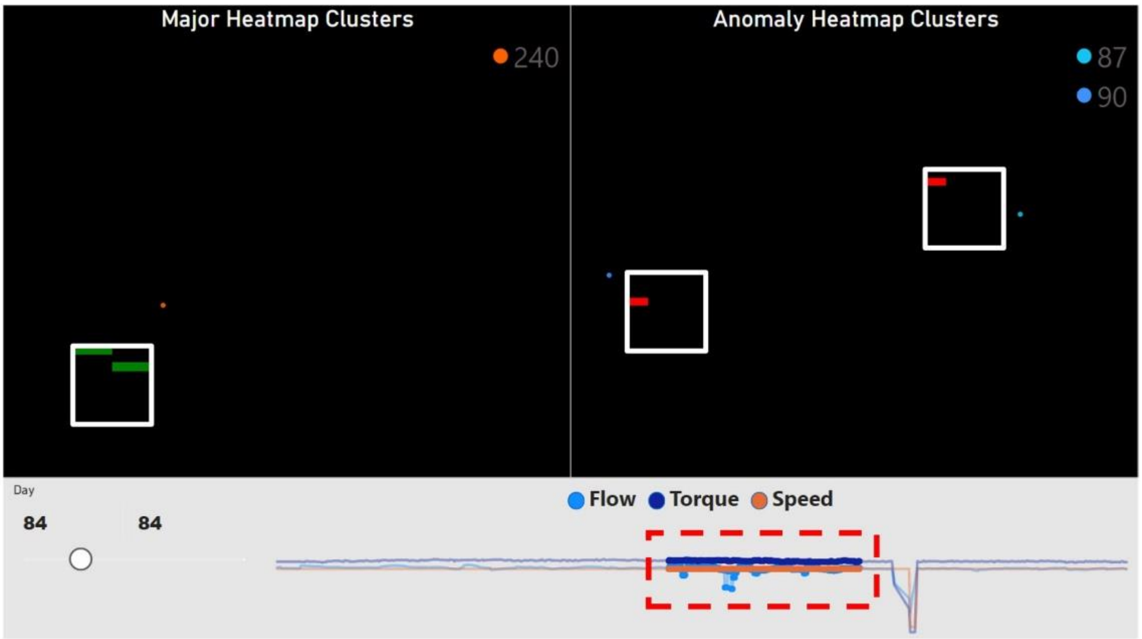


Fig. 37 Day 84 of PCP operations with the selected abnormal behaviour and the respective major and anomaly heatmap clusters

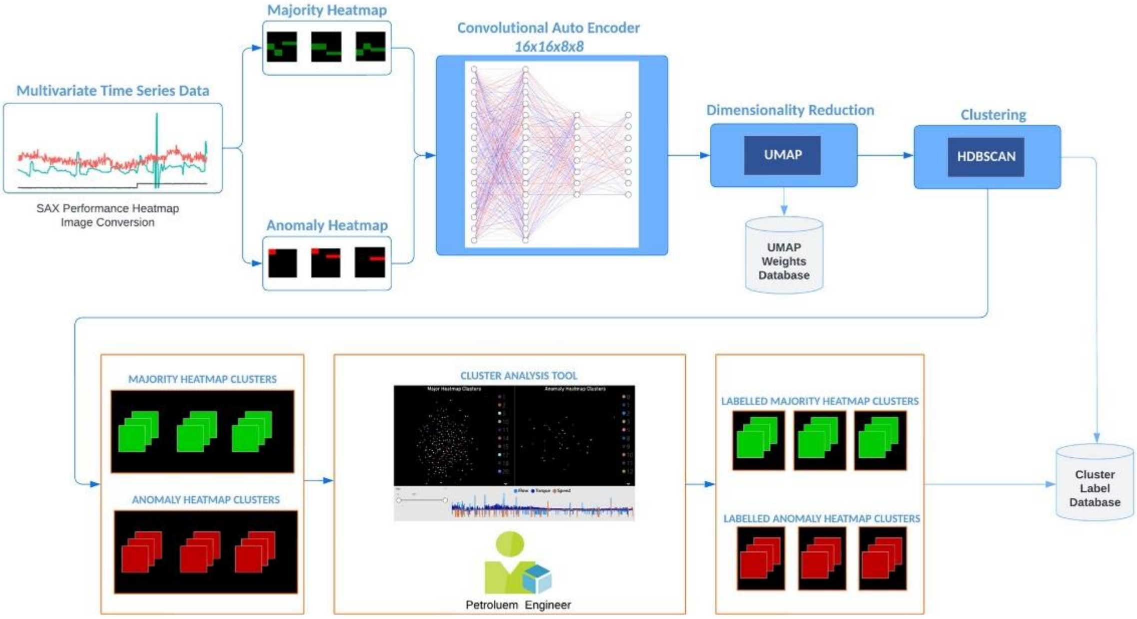


Fig. 38 Final method for labelling time-series data with the human-in-the-loop approach

flow, low flow) either during or before the solids through pump events occur.

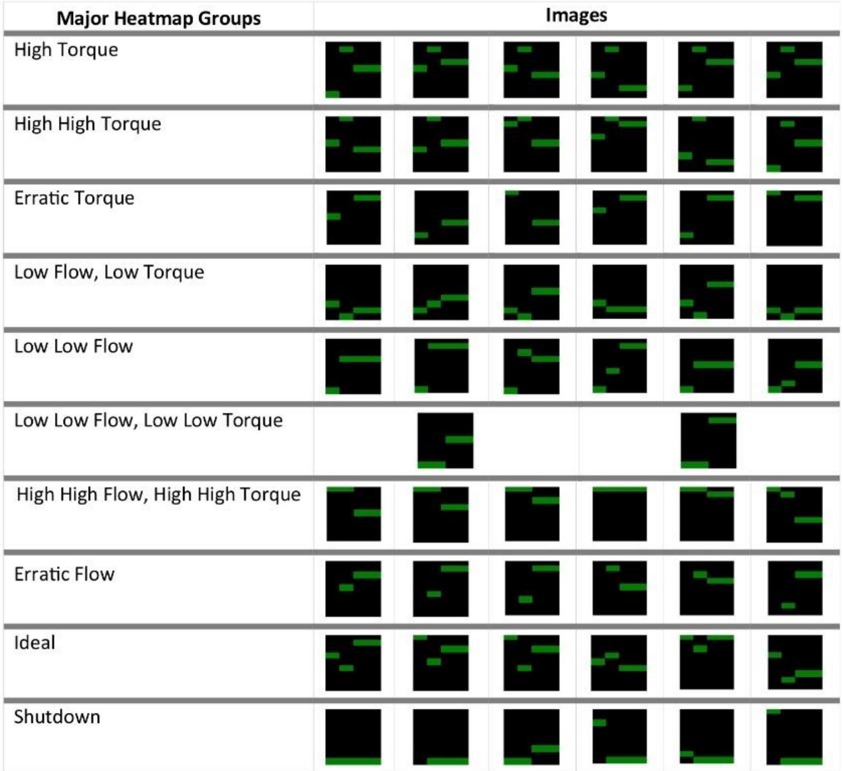
Table 14 Heatmap groups based on the observations made in the cluster labelling tool

Major heatmap groups	Anomaly heatmap groups
High torque	High flow
High high torque	Low flow
Erratic torque	High torque
Low flow, low torque	Low torque
Low low flow	Flow and torque
Low low flow, low low torque	
High high flow, high high torque	
Erratic flow	
Ideal	
Shutdown	

5.3 III. Cluster group consistency for anomalous events

Another finding during this study was the repeatability of major and anomaly heatmap group sequencing for events of interest. For example, in Fig. 41, we see the solids through pump events on multiple wells, where the major heatmap groups diverge from *ideal* to either *high torque* or *high high torque*. The major heatmap sequencing is very similar for all such events, regardless of the event's intensity or duration.

Table 15 Sample images of major heatmap groups



Major Heatmap Groups	Images
High Torque	
High High Torque	
Erratic Torque	
Low Flow, Low Torque	
Low Low Flow	
Low Low Flow, Low Low Torque	
High High Flow, High High Torque	
Erratic Flow	
Ideal	
Shutdown	

Table 16 Sample images of anomaly heatmap groups

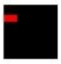
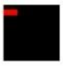
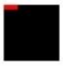
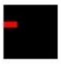
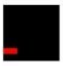
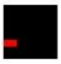
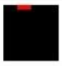
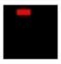
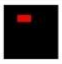
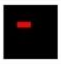
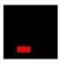
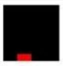
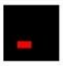
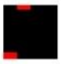
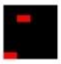
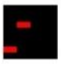
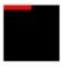
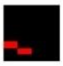
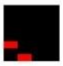
Anomaly Heatmap Groups	Images
High Flow	   
Low Flow	 
High Torque	   
Low Torque	  
Torque and Flow	     

Table 17 Color codes for major and anomaly heatmap groups

Major Heatmap Groups		Anomaly Heatmap Groups	
High Torque	🟥	High Flow	🟦
High High Torque	🟥	Low Flow	🟪
Erratic Torque	🟥	High Torque	🟥
Low Flow, Low Torque	🟧	Low Torque	🟧
Low Low Flow	🟦	Flow and Torque	🟪
Low Low Flow, Low Low Torque	🟧		
High High Flow, High High Torque	🟥		
Erratic Flow	🟦		
Ideal	🟩		
Shutdown	⬛		

Similarly, in Fig. 42, we see gas through pump events. In this case, the major heatmap groups fluctuate between *Ideal* and *Erratic Torque*, whereas the anomaly heatmap groups consistently present with *low flow* and *flow and torque* events.

5.4 IV. Streaming analytics application for PCP performance analysis

Putting the previous steps together, we provided two natural gas operators with a streaming analytics tool, which assists them with identifying early PCP performance issues and alerts when critical anomalous events are detected. An overview of the application is shown in Fig. 43.

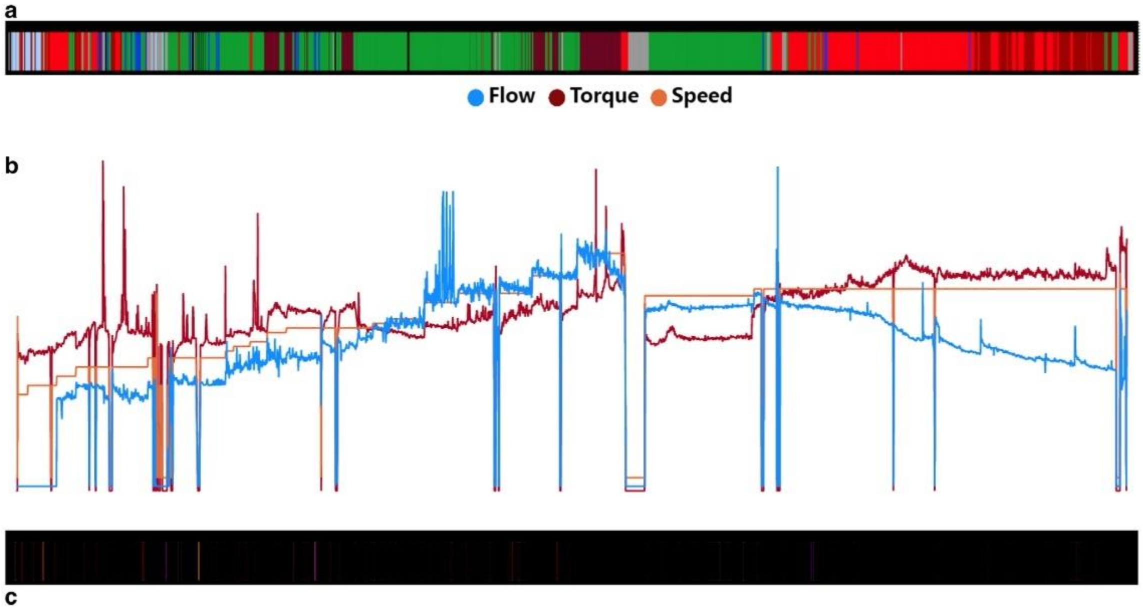


Fig. 39 **a** Major heatmap group plotted with the time-series data, **b** time-series trend for a sample well, **c** anomaly Heatmap group plotted with the time-series data

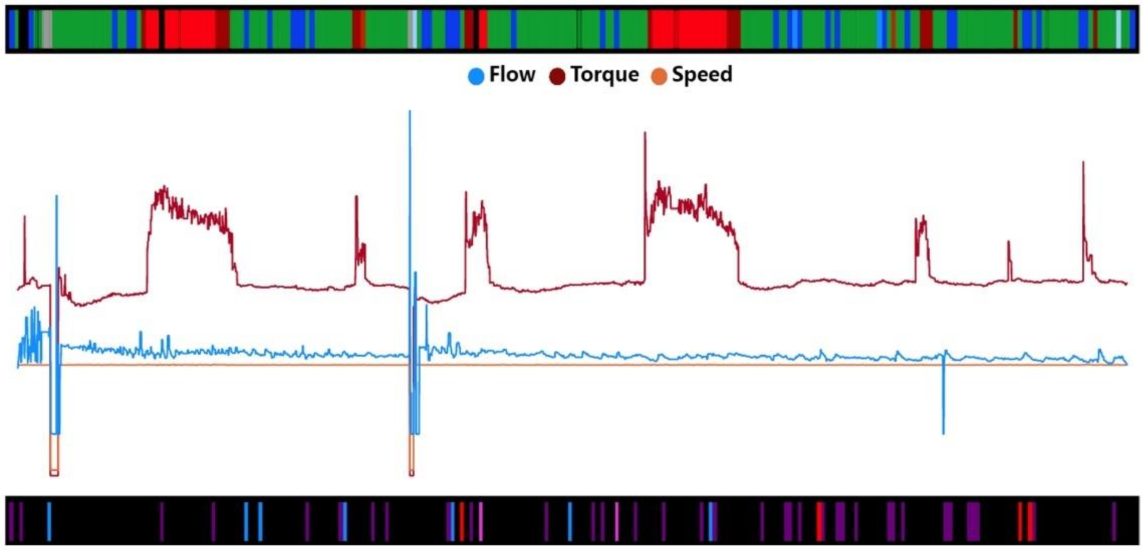


Fig. 40 PCP performance profile over one-week showing regions of abnormal activity

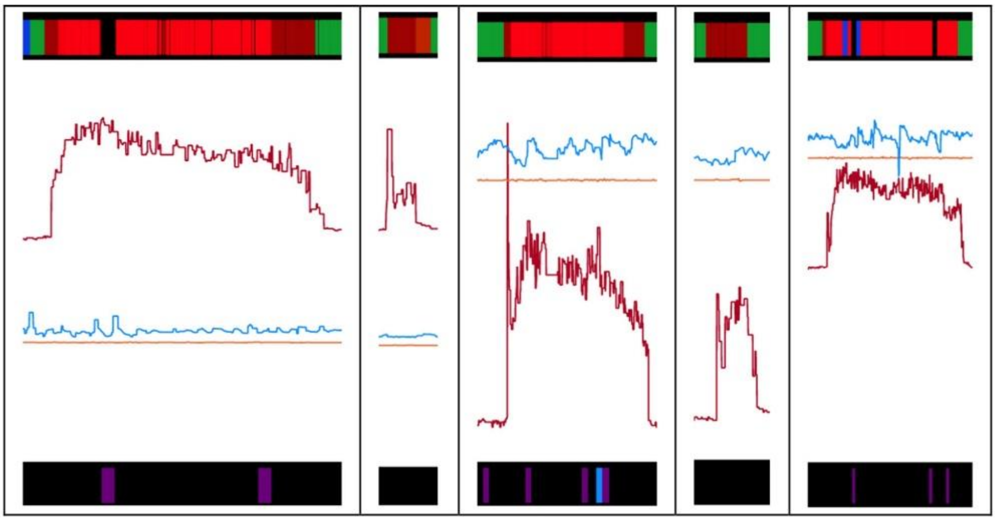


Fig. 41 Solids through pump profile for different wells

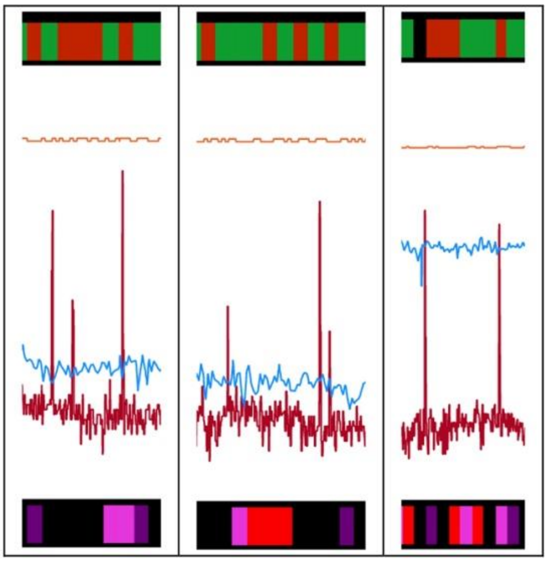


Fig. 42 Gas through pump profile for different wells

6 Conclusion and future works

Based on the above methodology, we demonstrated that the human-in-the-loop cluster labelling method and the streaming analytics tools developed as part of this research provide a reliable and scalable approach to determining and evaluating the performance of PCP-operated wells.

We have shown that various performance patterns can be detected with this approach, and the repeatability of the heatmap patterns provides a better understanding of changing PCP behaviour. Furthermore, notification of changes in performance profile and anomaly markers can be automated, where only events that require immediate attention can be reported in real time. By doing so, production and surveillance engineers can manage their wells by exception, aided by informed insights by the method proposed in this study.

Most importantly, by allowing petroleum engineers to aid with the labelling of time-series data, we could gain their trust in a machine learning-driven approach and, in turn, capture their knowledge of assessing Artificial Lift systems.

During this study, it became evident that the level of granularity to detect performance changes could be improved with smaller expansion stride lengths. In a forthcoming paper, we will present the effect of expansion stride length on cluster groups. Moreover, there is work in progress to apply this method to electric submersible pumps, which are centrifugal pumps used as an Artificial Lift method in conventional oil reservoirs.

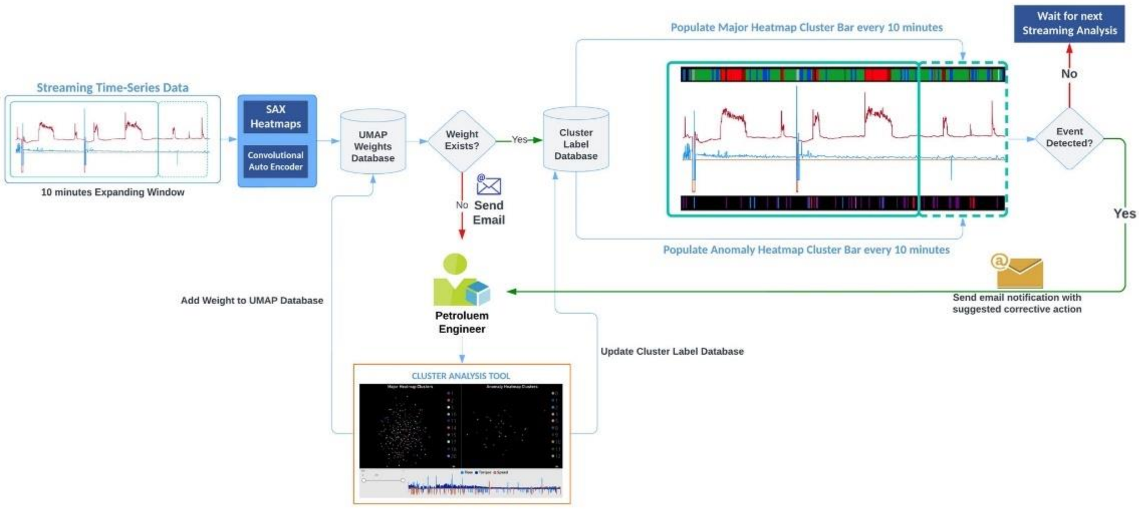

Streaming Time-Series Data
10 minutes Expanding Window
SAX Heatmaps
Convolutional Auto Encoder
UMAP Weights Database
Weight Exists?
Yes
No
Send Email
Cluster Label Database
Populate Major Heatmap Cluster Bar every 10 minutes
Populate Anomaly Heatmap Cluster Bar every 10 minutes
Wait for next Streaming Analysis
No
Event Detected?
Yes
Send email notification with suggested corrective action
Add Weight to UMAP Database
Petroluem Engineer
Update Cluster Label Database
CLUSTER ANALYSIS TOOL

Fig. 43 Streaming analytics application deployment architecture

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability statement The data that support the findings of this study are available from Santos Limited and Westside Corporation Pty Ltd, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of Santos Limited and Westside Corporation Pty Ltd.

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

1. Queensland borehole series metadata record. 2022; Available from: https://www.data.qld.gov.au/dataset/queensland-borehole-series.
2. Hoday JP et al (2013) Diagnosing PCP failure characteristics using exception based surveillance in CSG. In: SPE progressing cavity pumps conference. 2013, Society of Petroleum Engineers: Calgary, Alberta, Canada, p. 13.
3. Awaid A et al (2014) ESP well surveillance using pattern recognition analysis, oil wells, petroleum development Oman. In: International petroleum technology conference. 2014, International Petroleum Technology Conference, Doha, Qatar, p. 22.
4. Thornhill DG, Zhu D (2009) Fuzzy analysis of ESP system performance. In: SPE annual technical conference and exhibition. 2009, Society of Petroleum Engineers: New Orleans, Louisiana, p. 7.
5. Al Sawafi M et al (2021) Intelligent operating envelope integrated with automated well models improves asset wide PCP surveillance and optimization. In: Abu Dhabi international petroleum exhibition & Conference, OnePetro.
6. Abdelaziz M, Lastra R, Xiao JJ (2017) ESP data analytics: predicting failures for improved production performance. In: Abu Dhabi international petroleum exhibition & conference. 2017, Society of Petroleum Engineers: Abu Dhabi, UAE, p. 17.
7. Ocanto L, Rojas A (2001) Artificial-lift systems pattern recognition using neural networks. In: SPE Latin American and Caribbean Petroleum engineering conference. 2001, Society of Petroleum Engineers: Buenos Aires, Argentina, p. 6.
8. Liu S et al (2011) Automatic Early Fault Detection for Rod Pump Systems. In: SPE annual technical conference and exhibition. 2011, Society of Petroleum Engineers: Denver, Colorado, USA, p 11.
9. Andrade Marin A et al (2021) Real Time Implementation of ESP predictive analytics—towards value realization from data science. In: Abu Dhabi International Petroleum Exhibition & Conference.
10. Liu Y et al (2011) Semi-supervised failure prediction for oil production wells. In: 2011 IEEE 11th International conference on data mining workshops.
11. Javed A, Lee BS, Rizzo DM (2020) A benchmark study on time series clustering. Mach Learn Appl 1:100001
12. Saghir F, Gonzalez Perdomo ME, Behrenbruch P (2020) Application of machine learning methods to assess progressive cavity pumps (PCPs) performance in coal seam gas (CSG) wells. APPEA J 60(1):197–214.

13. Saghir F, Gonzalez Perdomo ME, Behrenbruch P (2019) Application of exploratory data analytics EDA in coal seam gas wells with progressive cavity pumps PCPs. In: SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition. 2019, Society of Petroleum Engineers: Bali, Indonesia, p. 10.
14. Saghir F, Gonzalez Perdomo ME, Behrenbruch P (2019) Converting time series data into images: An innovative approach to detect abnormal behavior of progressive cavity pumps deployed in coal seam gas wells. In: SPE Annual Technical Conference and Exhibition. 2019, Society of Petroleum Engineers: Calgary, Alberta, Canada, p. 14.
15. Alqahtani A et al (2021) Deep time-series clustering: a review. Electronics 10(23):3001
16. Huddlestone-Holmes CA, Elaheh KJ (2018) Decommissioning coal seam gas wells—Final Report of GISERA Project S.9: Decommissioning CSG wells. CSIRO.
17. Commonwealth of Australia 2014, Coal seam gas extraction: modelling groundwater impacts. 2014, Department of the Environment.
18. Matthews CM et al (2007) Petroleum Engineering Handbook. In: Production Operations Engineering. 2007, Society of Petroleum Engineers.
19. Choi K et al (2021) Deep learning for anomaly detection in time-series data: review, analysis, and guidelines. IEEE Access 9:120043–120065
20. Wen T, Keyes R (2019) Time series anomaly detection using convolutional neural networks and transfer learning. ArXiv, 2019. arXiv:1905.13628.
21. Zhang C et al (2018) A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data.
22. Tadayon M, Iwashita Y (2020) A clustering approach to time series forecasting using neural networks: a comparative study on distance-based vs. feature-based clustering methods. arXiv preprint arXiv:2001.09547..
23. Ienco D, Interdonato R (2020) Deep multivariate time series embedding clustering via attentive-gated autoencoder. Springer International Publishing, Cham
24. Xu C, Huang H, Yoo S (2021) A deep neural network for multivariate time series clustering with result interpretation. In: 2021 International Joint Conference on Neural Networks (IJCNN)
25. Freeman, C, Beaver I (2019) Human-in-the-Loop Selection of Optimal Time Series Anomaly Detection Methods. In 7th AAAI Conf Hum Comput Crowdsourcing (HCOMP)
26. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D et al (2022) Human-in-the-loop machine learning: a state of the art. Artif Intell Rev. https://doi.org/10.1007/s10462-022-10246-w
27. Lin J et al (2007) Experiencing SAX: a novel symbolic representation of time series. Data Min Knowl Disc 15(2):107–144
28. Wang Z, Oates T (2015) Imaging time-series to improve classification and imputation. arXiv preprint arXiv:1506.00327.
29. Yang C et al Multivariate time series data transformation for convolutional neural network. In: 2019 IEEE/SICE international symposium on system integration (SII). 2019.
30. Biewald L (2020) Experiment racking with weights and biases, 2020. Available from: https://www.wandb.com/.
31. van der Maaten L, Hinton G (2008) visualizing data using t-SNE. J Mach Learn Res 9:2579–2605
32. McInnes L, Healy J, Melville J (2018) Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.
33. Agrawal A, Ali A, Boyd S (2021) Minimum-distortion embedding. Found Trends Mach Learn 14(3):211–378
34. Campello RJGB, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. Springer, Heidelberg, pp 160–172.
35. Malzer C, Baum M (2020) A hybrid approach to hierarchical density-based cluster selection. In 2020 IEEE In Conf Multisensor Fusion Integr Intell Syst (MFI). IEEE, pp. 223–228

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

8. Paper 6: Performance analysis of artificial lift systems deployed in natural gas wells: A time-series analytics approach

The culmination of this research is presented in the final paper, which consolidates all the methods and procedures detailed in the preceding publications. It introduces the Artificial Lift Systems Analytics Application (ALSAA), a pivotal development for addressing the CSG industry's real-time monitoring and data annotation process. ALSAA exemplifies how multiple wells can be efficiently monitored through exception-based surveillance. Additionally, this paper illustrates the versatility of performance heatmap images, demonstrating their application to other forms of Artificial Lift Systems (ALS), including Electrical Submersible Pumps (ESPs) and Electrical Submersible Progressive Cavity Pumps (ESPCPs).

The paper offers a comprehensive view of how the ALSAA is harnessed by petroleum and surveillance engineers for labelling ALS performance using a data annotation toolbar and applying these results to streaming data. It details the complete workflow involved in the data annotation process. Additionally, the paper delves into the concepts of events and sequences. Events signify notable changes or anomalies within the ALS performance, such as variations in downhole pressure or the occurrence of specific operational issues. Conversely, sequences are combinations of events that transpire in a particular order or pattern. Through the identification and analysis of these events and sequences, engineers can gain valuable insights into the performance of the ALS and take proactive measures to address potential issues or failures.

The paper demonstrates how live monitoring dashboards and ALS analysis tools can detect early signs of issues, such as gas intake, and end-of-life for an ALS. Overall, the paper addresses two crucial research gaps in the field of artificial lift system management. Firstly, it focuses on the accurate labeling of ALS data and, secondly, on the capacity to monitor a significant number of wells on an exception basis. ALSAA overcomes these challenges and showcases how they are addressed, significantly contributing to the broader field of ALS operations in CSG wells.

Statement of Authorship

Title of Paper	Performance analysis of artificial lift systems deployed in natural gas wells: A time-series analytics approach
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Saghir, F., Perdomo, M. G., & Behrenbruch, P. (2023). Performance analysis of artificial lift systems deployed in natural gas wells: A time-series analytics approach. Geoenergy Science and Engineering, 230, 212238.

Principal Author

Name of Principal Author (Candidate)	Fahd Saghir
Contribution to the Paper	Conduct data analysis, write and record experiments, create test reports, tabulate results and write paper.
Overall percentage (%)	75%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.

Signature		Date	19/09/2023

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Mary Gonzalez Perdomo
Contribution to the Paper	Assisted with paper structure, writing and paper review (20%)

Signature		Date	18/09/2023

Name of Co-Author	Peter Behrenbruch
Contribution to the Paper	Assisted with paper structure, writing and paper review (5%)

Signature		Date	10-09-2023

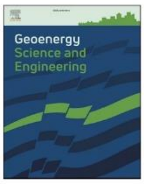
Please cut and paste additional co-author panels here as required.

230 (2023) 212238

Contents lists available at ScienceDirect

Geoenergy Science and Engineering

journal homepage: www.sciencedirect.com/journal/geoenergy-science-and-engineering



Performance analysis of artificial lift systems deployed in natural gas wells: A time-series analytics approach

Fahd Saghir [a,*], Maria Elena Gonzalez Perdomo [a], Peter Behrenbruch [b]

[a] *University of Adelaide, Australia, Australia*
[b] *Bear and Brook Consulting, Australia, Australia*

ARTICLE INFO

Keywords:
Natural gas production
Time-series analytics
Artificial lift analytics
Coal seam gas
Automated workflow
ESP
ESPCP
PCP

ABSTRACT

In Coal Seam Gas (CSG) production, Artificial Lift (AL) systems comprising various downhole pumps produce coal-fine-laden water. Over time, the accumulation of coal fines results in mechanical failures of such pumps, which leads to a loss in natural gas production. This problem is exacerbated by CSG operators having to manage thousands of wells in their day-to-day operations. The reason for having such an exuberant number of natural gas wells is due to the fact that CSG reservoirs have a short production cycle when compared to conventional reservoirs. Therefore, thousands of CSG wells must operate simultaneously to sustain natural gas volumes that satisfy domestic energy demand, and help meet contractual export obligations.

To manage a large fleet of CSG wells, operators rely on Petroleum and Surveillance Engineers to mitigate production issues and advise timely corrective actions. Current methods of monitoring CSG wells involve observing real-time trends and alarms from Supervisory Control and Data Acquisition (SCADA) systems. However, managing a large fleet of wells with a monitoring-only approach can be detrimental to AL operations. Methods such as Exception Based Surveillance have been used to improve engineers' workload, but such methods are informative at best. In addition, they do not provide early indications of changed downhole pump performance. To improve how AL systems are monitored in CSG operation, we propose a novel well surveillance method that can help mitigate pump failure and provide engineers with insightful information to take corrective actions.

In this work, we will present an innovative time-series analytics approach that examines the performance of AL systems in near real-time and helps Well Surveillance and Production Engineers make timely decisions that aid in mitigating pump failures. We used time-series data from 448 CSG wells to build the streaming analytics methodology that autonomously detects various performance states of downhole pumps during CSG production. Furthermore, we successfully detected the onset of AL systems failures, such as solids build-up, gas intake, high torque and pump degradation. We validated our solution on a separate set of 428 wells and were able to show reliable detection of detrimental events on live data from producing natural gas wells. This solution is currently deployed by 2 CSG operators where real-time notifications aid Petroleum and Well Surveillance Engineers with proactively managing AL systems across multiple CSG assets.

1. Introduction

The state of Queensland in Eastern Australia is responsible for producing 5513 TJ of natural gas derived energy per day (Queensland borehole series metadata record, 2022), which is utilized for domestic utilization and international export. To maintain an uninterrupted natural gas supply, energy companies rely on AL systems to deliver reliable performance for displacing water from coal cleats. Typically, Progressive Cavity Pumps (PCPs) are used for displacing water from coal seams. However, with the recent introduction of horizontal wells, traditional Electric Submersible Pumps (ESPs) and Electric Submersible Progressive Cavity Pumps (ESPCPs) have become popular artificial lift pump choices amongst natural gas producers in Queensland (Rajora et al., 2019).

The process of producing natural gas from CSG wells involves the depressurization of coal cleats through the production of reservoir

[*] Corresponding author.
E-mail addresses: fahd.saghir@adelaide.edu.au (F. Saghir), maria.gonzalezperdomo@adelaide.edu.au (M.E. Gonzalez Perdomo), peter.behrenbruch@bearnbrook.com (P. Behrenbruch).

https://doi.org/10.1016/j.geoen.2023.212238
Received 18 June 2023; Received in revised form 27 July 2023; Accepted 6 August 2023
Available online 9 August 2023

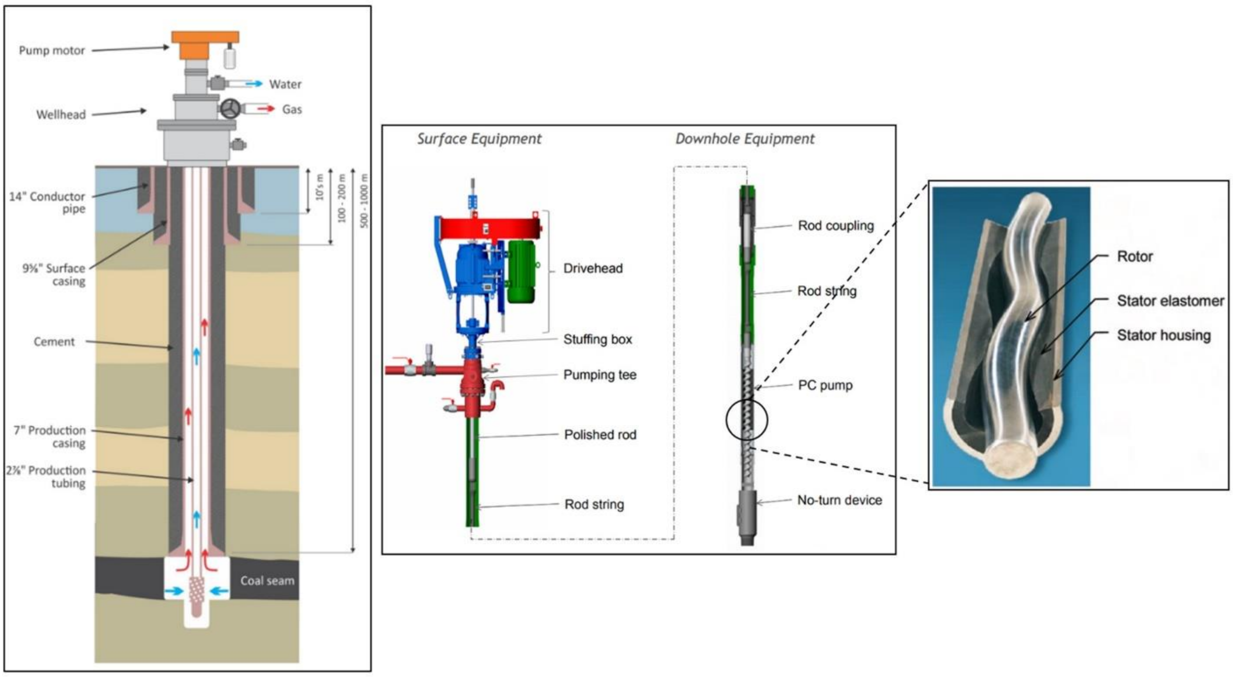


Fig. 1. (Left) Natural Gas Production from Coal Seam Gas (CSG) wells (Commonwealth of Australia, 2014, 2014). (Centre) (Main Components of a PCP system. (Right) Cut-out view of PCP Rotor and Stator (Matthews et al., 2007).

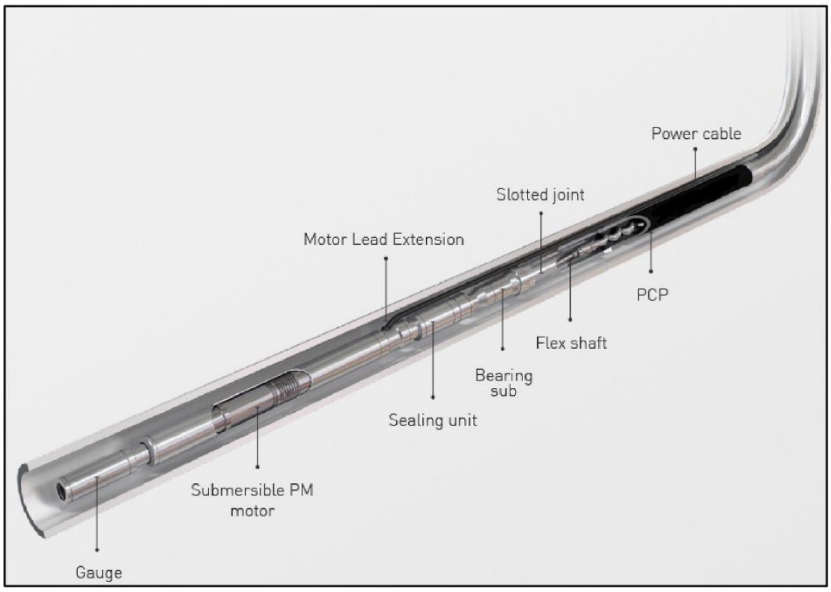


Fig. 2. Components of an electric submersible progressive cavity pump (ESPCP).

water, which causes an intake of disintegrated solids and coal fines into the Artificial Lift Systems. Fig. 1 shows an overview of a CSG well, along with the surface and downhole equipment of a PCP (Commonwealth of Australia, 2014, 2014; Matthews et al., 2007). The components for the ESPCP and ESP systems are shown in Figs. 2 and 3, respectively.

PCP and ESPCP are positive displacement pumps that use a metallic rotor and elastomer stator to move fluids. The main difference between the two is the location of the motor - PCP has the motor at the surface and uses rod strings to transfer energy to the downhole pump. In contrast, ESPCP has the entire motor and pump assembly downhole, eliminating the need for rod strings. As a result, ESPCPs are more efficient, but the choice between the two depends on the economics and well-completion strategy of the operator. ESPs are a type of centrifugal pump that, like an ESPCP, consists of a motor and a pump, which are placed downhole and are used to lift fluids from the wellbore to the surface. The main advantage of ESPs is that they can pump large

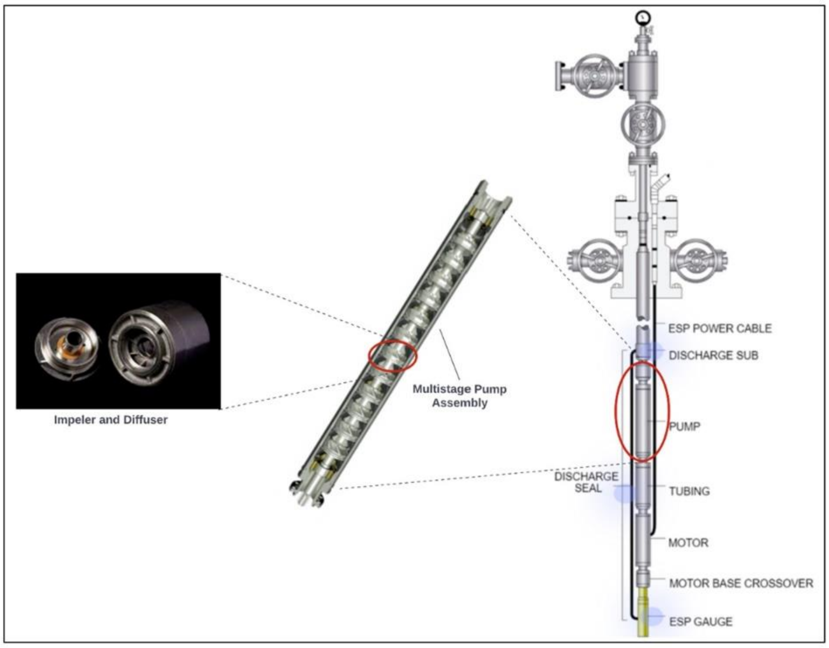


Fig. 3. Components of an electric submersible pump (ESP).

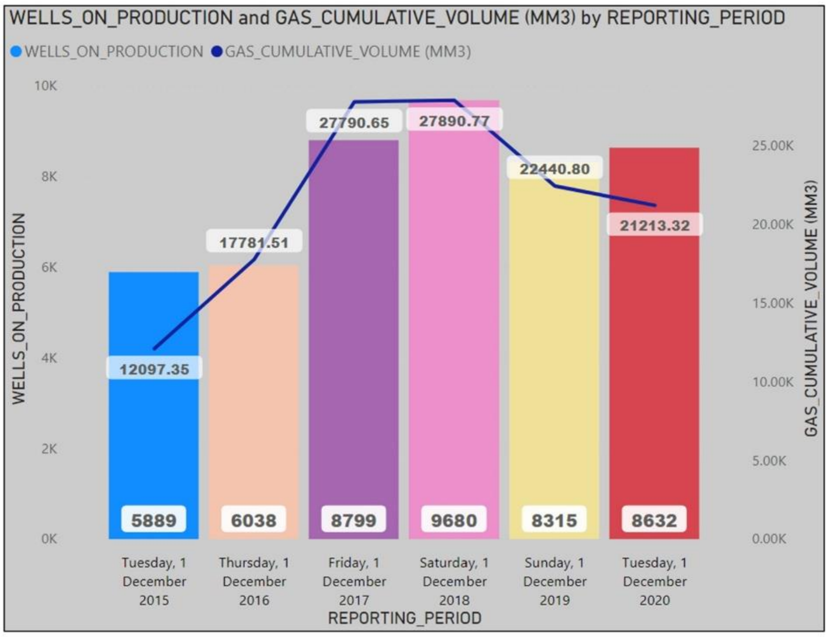


Fig. 4. Data from the Queensland Government showing number of wells & cumulative gas production from 2015 to 2020 (Queensland borehole series metadata record, 2022).

volumes of fluids and are relatively efficient, as they do not require rod strings similar to PCPs. The choice between ESPs and other types of pumps depends on the water flow rate CSG operators try to achieve.

For each AL system, intake of produced water mixed with solids can be detrimental to the performance of pumps, often resulting in unwanted shutdowns and, in some cases, pump failure. In addition, each pump failure initiates a well workover, which is expensive and results in deferred production. Hence, assessing the real-time performance of AL systems is critical to managing and efficiently operating CSG wells.

To effectively manage CSG production and address any potential problems related to the downhole pumps, real-time data from CSG wells is collected using Supervisory Control and Data Acquisition (SCADA) systems. This data is then made available to central control rooms and company headquarters, where Well Surveillance and Petroleum

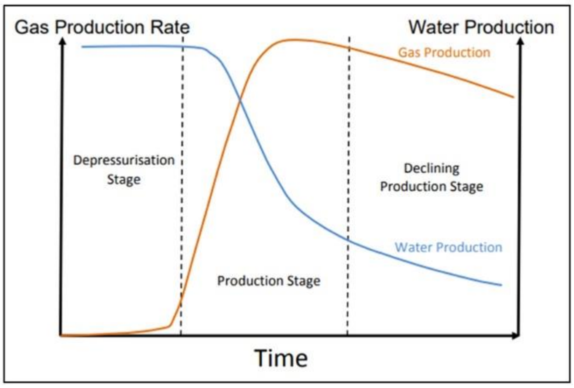


Fig. 5. Production stages of a coal seam gas well (Commonwealth of Australia, 2014, 2014).

Engineers analyze it. These professionals are responsible for monitoring the performance of AL systems and providing recommendations for corrective actions that will help extend the production life of the wells. Companies can ensure that their CSG operations are running smoothly and efficiently by using SCADA and having dedicated experts to review the data.

However, the CSG industry faces a unique challenge in continuously monitoring the performance of its wells, due in large part to the sheer number of wells in operation. As shown in Fig. 4, the number of wells in production has remained consistently high since 2015, with thousands of new wells brought online in 2017 to meet domestic and international energy demands. Although official figures for 2021 have yet to be published, the number of production wells is expected to surpass 10,000 due to new drilling programs initiated by CSG operators in Queensland.

Fig. 5 illustrates the different stages of production for a typical CSG well, which can produce natural gas for up to 10 years but typically reaches peak production within the first 2–3 years. To maintain a steady supply of natural gas, it is necessary to drill a large number of wells and bring them online quickly. However, managing a large number of wells can take time and effort, especially when providing timely input to address performance issues and optimize production.

While improved methods for monitoring Artificial Lift Systems have been discussed and demonstrated extensively through publications, there is minimal work discussing real-time performance analysis of AL pumps deployed in CSG wells. For CSG wells, Knafl et al. (2013) defined an Exception Based Surveillance (EBS) methodology for diagnosing PCP failure characteristics. The method used in the EBS approach calculates displacement and friction variables (derived from existing sensors) to analyze PCP behaviour. Of the multiple sensors used to calculate the displacement and friction variables, the downhole pressure sensor is essential to derive both these measurements. Unfortunately, downhole pressure sensors are prone to consistent failure in CSG operations (Rathnayake et al., 2022); hence, relying on this EBS approach is impractical.

A recent study by Rathnayake et al. (Rathnayake and Firouzi, 2021) proposes a statistical method to detect early PCP failures; however, the authors indicate that this method produces extensive periods of false alarms. If these periods are removed from the training data, the overall accuracy of the predictive model depreciates significantly. They suggest that the model can be improved by providing guidelines to select appropriate training data and including longer observation times prior to pump failure. The limitation of this method also renders it unfeasible to conduct real-time performance analysis of Artificial Lift Systems. Additionally, our research found no existing publications on proactive surveillance methods for ESPs and ESPCPs in CSG operations. This lack of research is likely because these advanced lifting systems were only recently introduced to Australian CSG operations in 2019, and operators are still gaining experience in their use and maintenance (Rajora et al., 2019). As a result, there is a gap in knowledge and understanding when it comes to proactive surveillance methods for these systems in the CSG industry.

In summary, our study proposes using a novel time-series analytics method in combination with real-time monitoring as a way for CSG operators to effectively monitor and analyze the performance of their Artificial Lift (AL) systems. Our approach uses human-labelled time-series data, offering a practical and insightful way to determine real-time AL system performance. We offer examples and case studies demonstrating how our system can assist Well Surveillance and Petroleum Engineers in proactively identifying and resolving potential issues that may result in AL system failure. By combining our time-series analytics method with real-time monitoring, we aim to provide CSG operators with an effective and practical Artificial Lift surveillance system

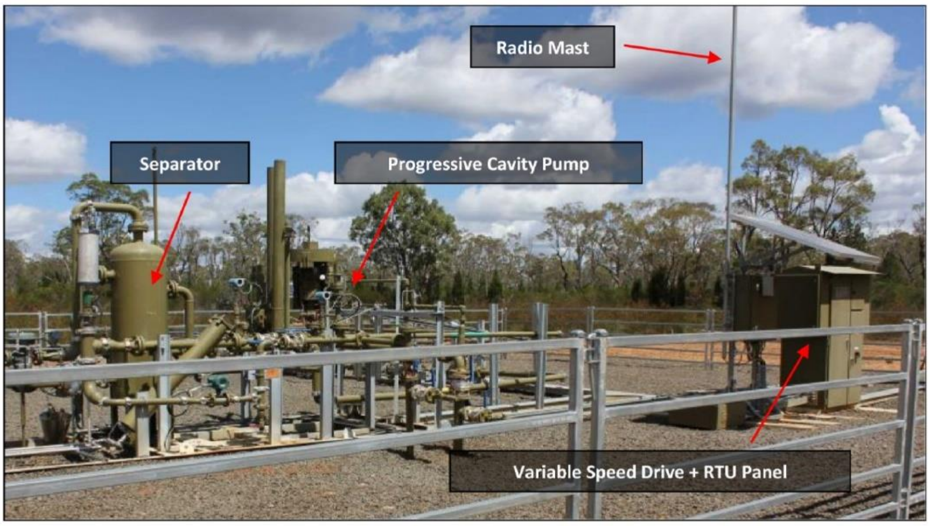


Fig. 6. Typical layout of a coal seam gas well.

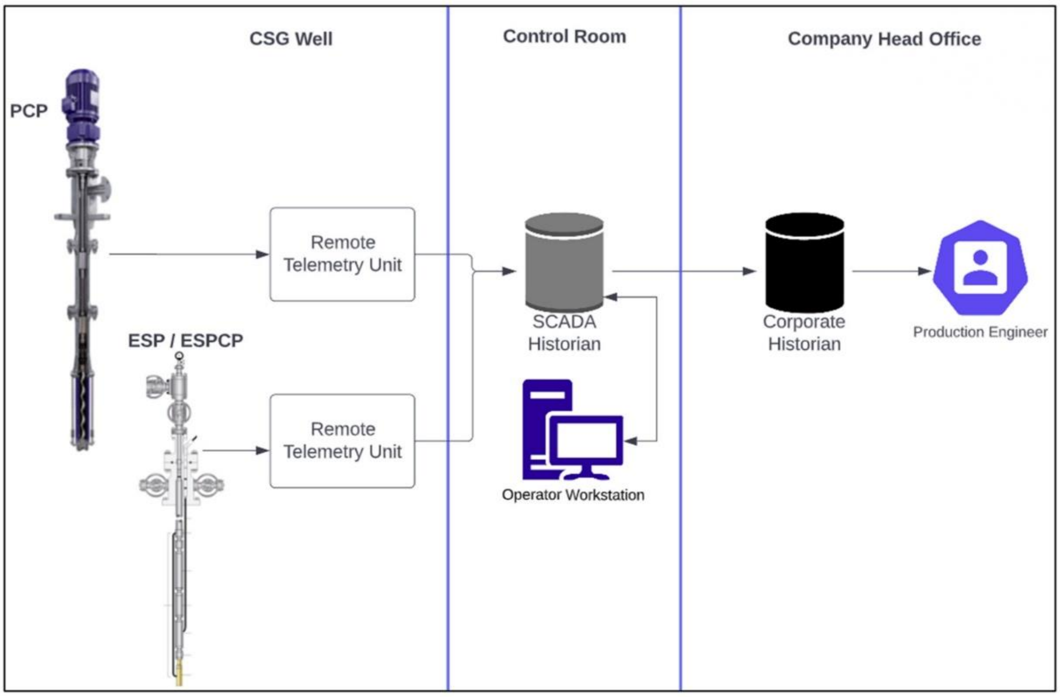


Fig. 7. Data flow from CSG wells to the company head office.

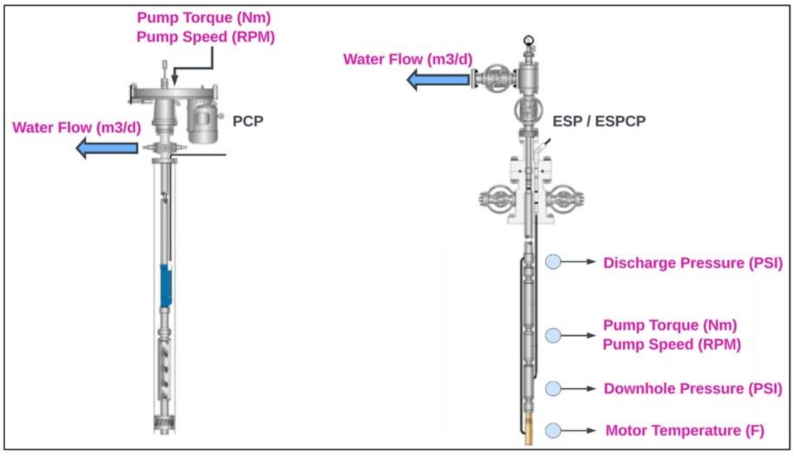


Fig. 8. Data measurement points for the parameters used for each well type.

for determining real-time AL system performance and mitigating issues that may lead to failure.

2. Literature review

2.1. *Performance analytics for artificial lift systems*

Over the past decade, several papers have discussed time-series analytics related to Artificial Lift Systems. Especially in the past three (3) years, we have seen an influx of such papers published in notable journals or presented at conferences. During our research, we conducted a comprehensive literature review of the papers published in the past decade, and identified two (2) distinct categories to classify the papers. These categories include machine-learning-based and hybrid-based analytics models (combining machine learning and physics-based approaches). However, most of these papers are dedicated to ESPs (Silvia et al., 2023; Cardona et al., 2023; Brasil et al., 2023; Al-Ballam et al., 2023; Sharma et al., 2022; Iranzi et al., 2022; Abdalla et al., 2022; Andrade Marin et al., 2021; Ambade et al., 2021; Alamu et al., 2020; Takacs and Takacs, 2018; Abdelaziz et al., 2017; Awaid et al., 2014; Camilleri, 2013; Thornhill and Zhu, 2009; Ocanto and Rojas, 2001), and very limited work is presented on using Machine Learning models for PCPs (Knafl et al., 2013; Rathnayake and Firouzi, 2021; Tan et al., 2021). We have discussed the shortcomings of these approaches in our

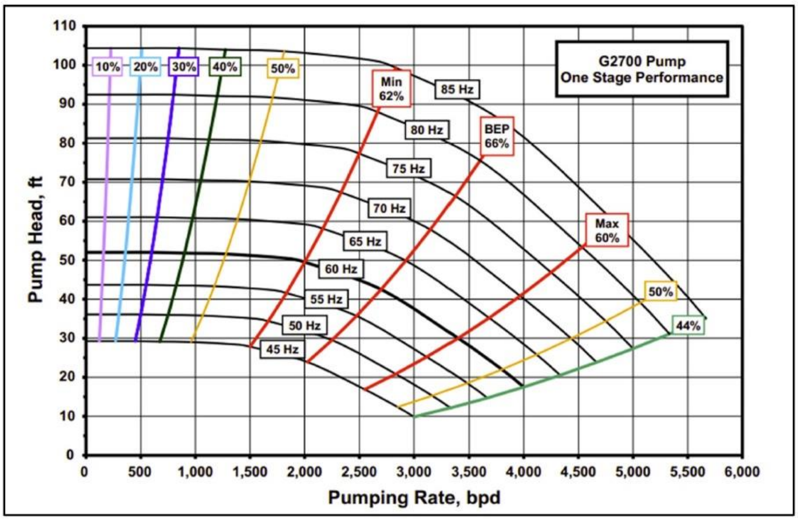


Fig. 9. A typical Pump Performance Curve for an ESP at various Frequencies (Pump Speed) (Takacs and Takacs, 2018).

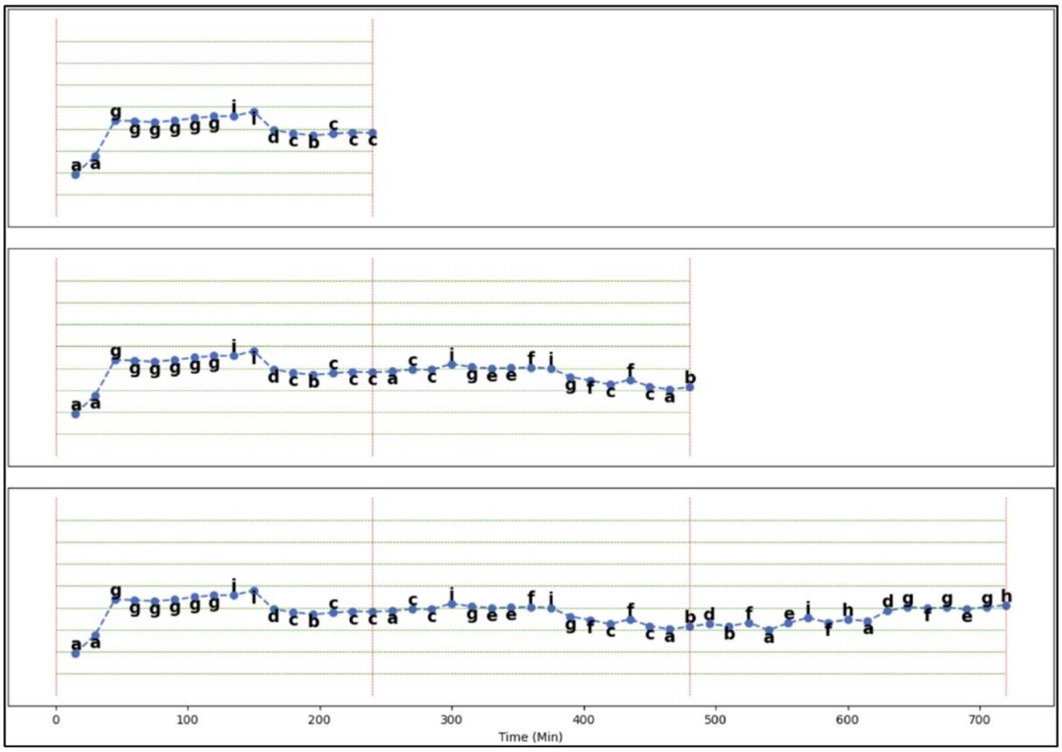


Fig. 10. Effect of Sliding Window Method on SAX symbols.

past papers (Saghir et al., 2019a, 2019b, 2020, 2023), and below, we will briefly discuss limitations in each of the categories mentioned above.

2.1.1. Limitations with machine learning based models for artificial lift systems

In the papers analysed during our research, we noticed that the machine learning models work well within the dataset they have been trained for. Moreover, machine learning models have an extensive data management pipeline which entails data pre-processing and quality checks. The process of data pipeline is quite intricate. Any error that occurs during the preprocessing steps is carried forward to the inference of the machine learning model that has already been trained. We also discovered that the machine learning approach has a noticeable

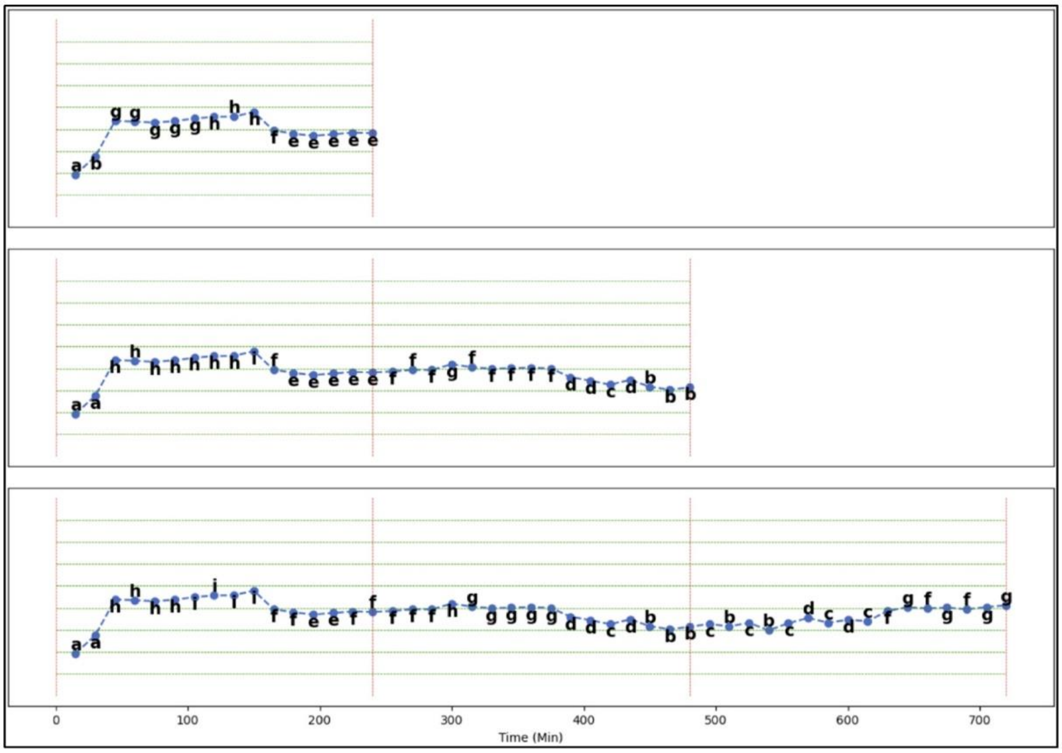


Fig. 11. Effect of Expanding Window Method on SAX symbols.

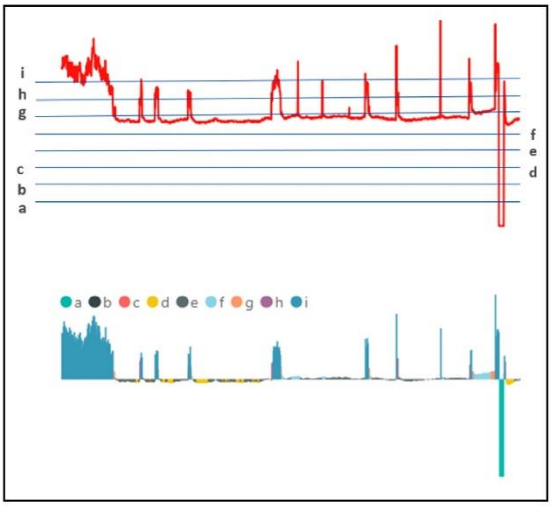


Fig. 12. An Example of converting univariate time-series data to SAX symbols.

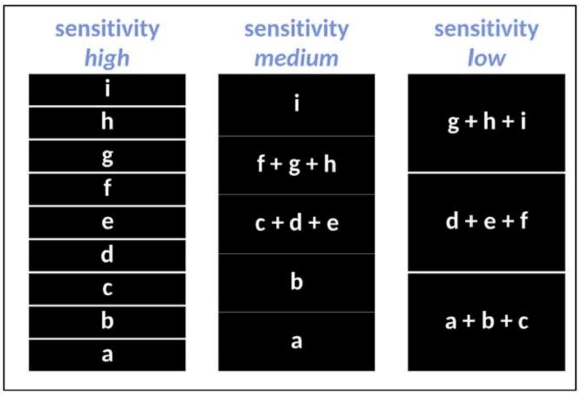


Fig. 13. High sensitivity SAX array (left), medium sensitivity SAX array (centre), low sensitivity SAX array (right).

limitation: the models need to be tailored for every Artificial Lift type. In some cases, the models even have to be specific to the operator, reservoir, and pump type. One of the most apparent restrictions of machine learning models is their dependence on data frequency. If a model was trained on a 10-s frequency, any deviation from this frequency would necessitate the model being retrained. In SCADA systems that rely on radio telemetry, maintaining a consistent data transmission rate can be difficult due to fluctuating connectivity rates and potential failures. To summarise, Machine Learning models work well with Artificial Lift Systems but require effort to maintain. They are highly dependent on data quality and need regular re-training to adapt to varying scenarios related to data collection.

2.1.2. Limitations with hybrid based models for artificial lift systems

Although hybrid models are more accurate than machine learning models, they share the same limitations as mentioned in the previous section. Additionally, setting up physics-based simulations to act as error correction for machine learning models requires significant effort

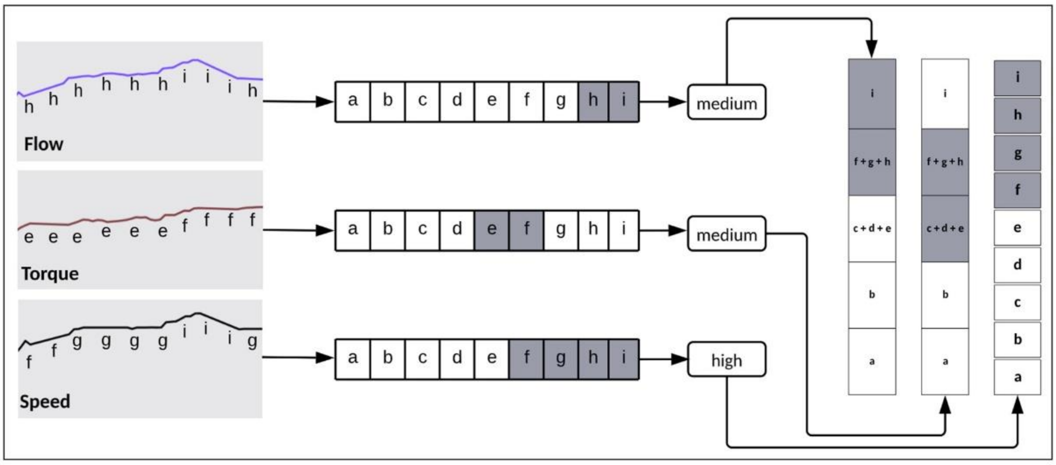
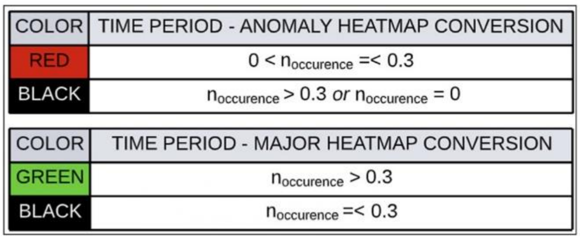


Fig. 14. Method of converting multivariate data to SAX Sensitivity Arrays.



COLOR	TIME PERIOD - ANOMALY HEATMAP CONVERSION
RED	0 < $n_{occurence}$ =< 0.3
BLACK	$n_{occurence}$ > 0.3 *or* $n_{occurence}$ = 0

COLOR	TIME PERIOD - MAJOR HEATMAP CONVERSION
GREEN	$n_{occurence}$ > 0.3
BLACK	$n_{occurence}$ =< 0.3

Fig. 15. Method of converting multivariate data to SAX sensitivity arrays.

and time as these simulations are computationally intensive. Moreover, creating hybrid models requires end-users to rely on expensive software licenses for simulation software provided by technology service companies.

2.2. Symbolic approximation aggregation and time-series performance heatmaps

After thorough analysis and careful consideration of the key points highlighted in the preceding sub-section, it became abundantly clear that machine learning models developed for AL systems were constrained by certain limitations. In order to overcome the challenges posed by different types of Artificial Lift systems and varying frequencies of data, we needed to find an innovative and all-encompassing solution. After rigorous investigation and extensive research efforts, we decided to proceed with the Symbolic Aggregation Approximation (SAX) method — a breakthrough approach renowned for its ability in analyzing time-series data (Lin et al., 2003). The SAX methodology encompasses two primary steps: symbolic representation and approximation. In the symbolic representation step, time-series data is discretized into a set of symbols, effectively reducing dimensionality and capturing essential features. The approximation step then involves matching symbol sequences with predefined distributions to extract valuable patterns and trends (Keogh et al., 2005).

The inherent variability and complexity introduced by distinct types of AL systems demanded a technique capable of seamlessly integrating

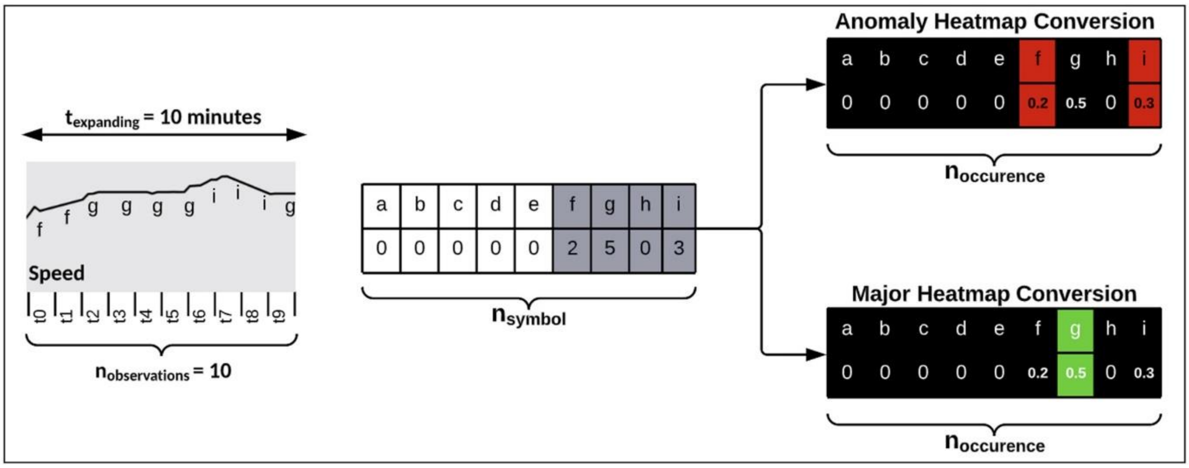


Fig. 16. Applying the Heatmap Colour Conversion method to the SPEED parameter from Fig. 14.

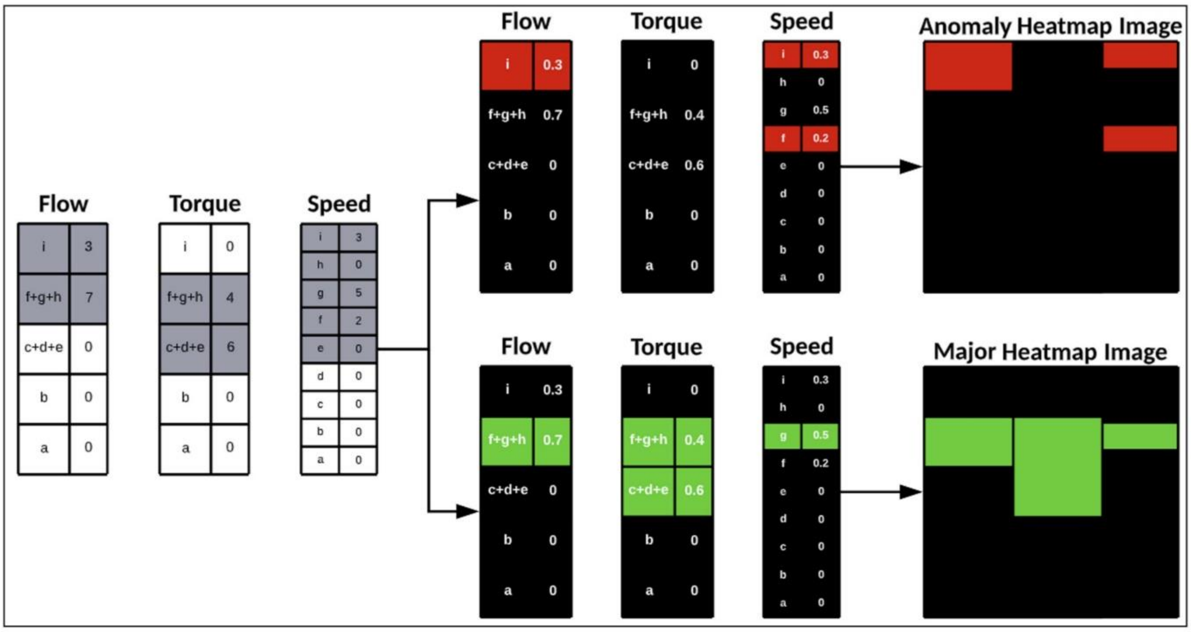


Fig. 17. Applying the Heatmap Colour Conversion method to the Speed parameter.

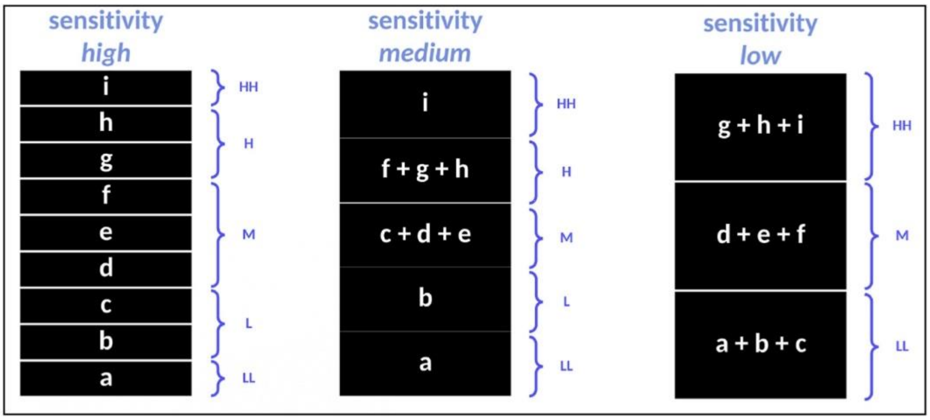


Fig. 18. Label Coding schema based on the position of colour markers.

and adapting to these dissimilarities. Furthermore, the importance of handling data with varying frequencies cannot be overstated. In oil and gas production, data collection intervals often differ, making it arduous for machine learning models to maintain consistent performance across all datasets. In our earlier research we identified that disparity in data frequency was a major roadblock, hindering our ability to deploy uniform and efficient solutions across the board.

Hence our approach of using expanding window with SAX derived time-series performance heatmaps offered a unique advantage in handling data of various intervals simultanouelsy (Saghir et al., 2019b). By transforming time-series data into performance heatmaps, SAX also allowed us to abstract the essential characteristics of each dataset, effectively mitigating the disparities arising from different Artificial Lift systems (Saghir et al., 2020, 2023). In the later section of this paper, we discuss SAX based performance heatmaps in details, and how they provide a usefule tool for AL system analysis.

3. Methodology

3.1. Data

We developed our analytics method using raw time-series data from 428 CSG wells. The data comprised 8 years of AL operations, included all CSG production stages and covered typical pump degradation behaviour and failures. A typical layout of a CSG well is shown in Fig. 6.

3.1.1. Data aggregation from CSG wells

To gather data from CSG wells, natural gas operators utilize Remote Telemetry Units (RTUs) in tandem with SCADA systems to push the data to the Company's head office, and this data flow is depicted in Fig. 7. An RTU serves two purposes at the CSG well site; firstly, it gathers data from local sensors to provide primary control of the AL system. Secondly, it stores and forwards the data to the SCADA Historian in the local control

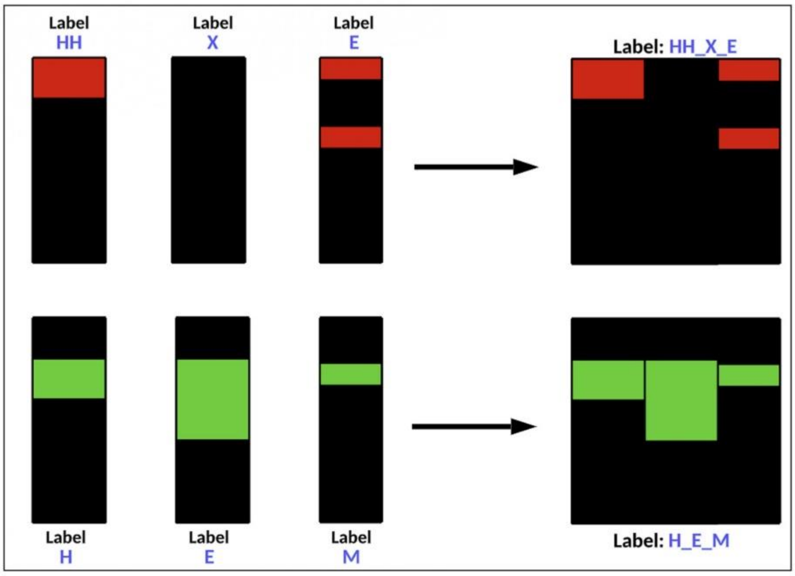


Fig. 19. Example of Heatmap Image labelling based on the Label Coding Schema.

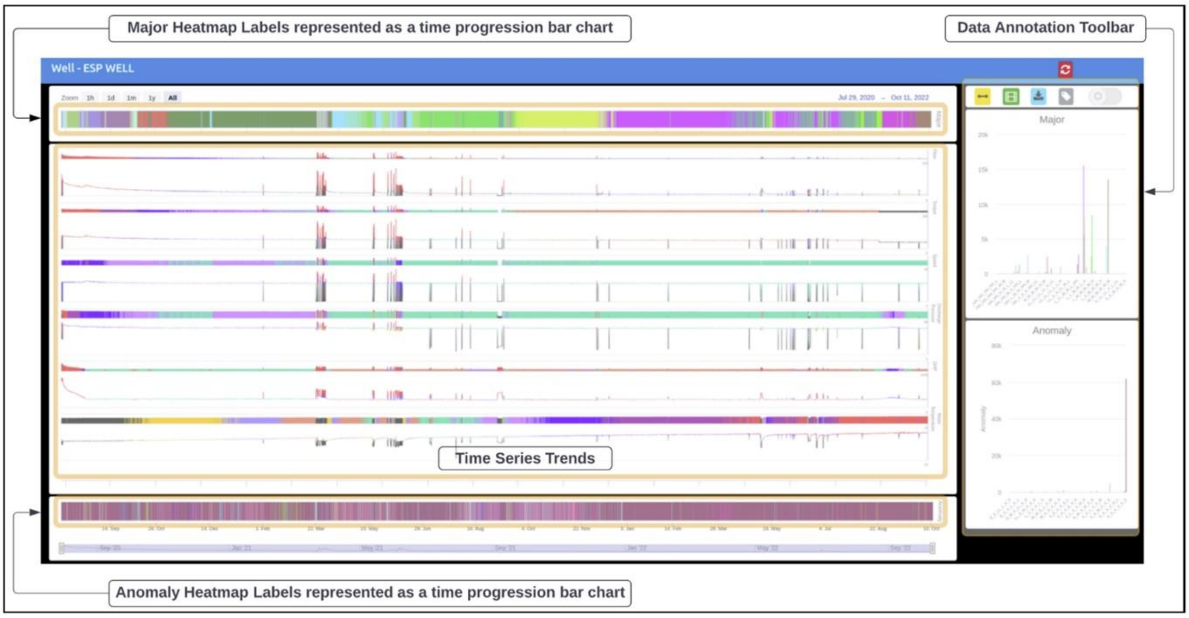

Major Heatmap Labels represented as a time progression bar chart
Data Annotation Toolbar
Well - ESP WELL
Zoom 1h 1d 1m 1y All
Jul 29, 2020 → Oct 11, 2022
Major
Anomaly
Time Series Trends
Anomaly Heatmap Labels represented as a time progression bar chart

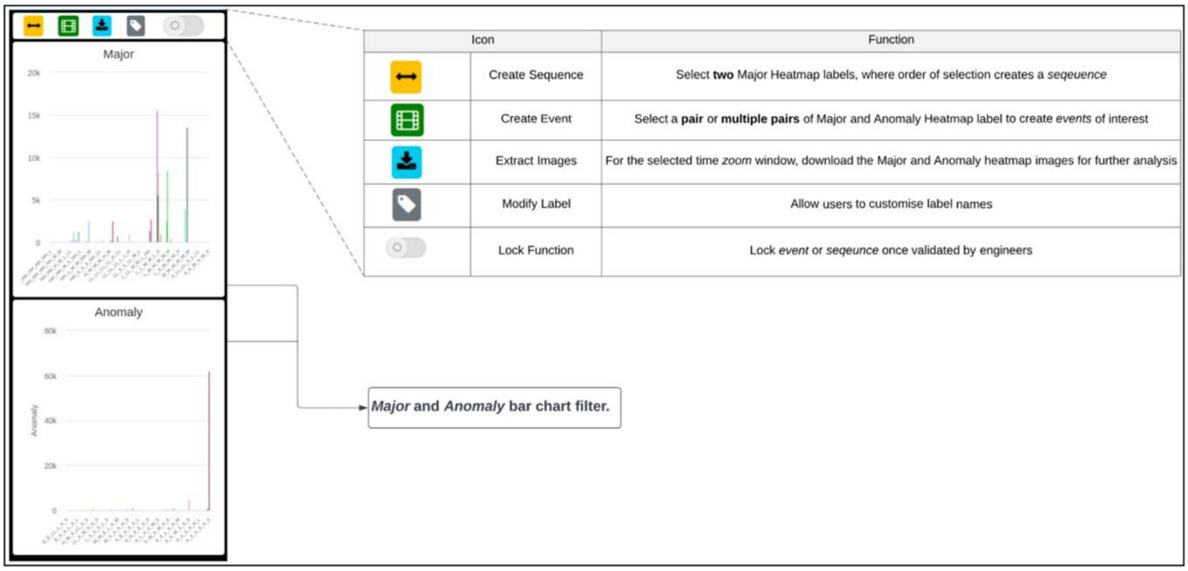
Fig. 20. User Interface created for Data Annotation.

room. The SCADA Historian then pushes the data to the Corporate Historian, making it available to end-users, including Well Surveillance and Petroleum Engineers.

For our study, we utilized unfiltered data from the Corporate Historian. We built our real-time analysis methodology on unfiltered data, which assisted us with addressing conditions where data from CSG wells would show drift due to sensor calibration issues. Furthermore, as data collected from CSG wells is primarily asynchronous (i.e., data not measured at a fixed interval), we spent considerable time ensuring that the data ingestion pipeline could handle streaming data of various lengths.

3.1.2. Data parameters

Our study used three parameters, namely, **Pump Speed**, **Pump Torque** and **Water Flow Rate**, to analyze PCP wells. From a mechanical performance and reservoir behaviour perspective, these three parameters have a high co-relation across PCP-operated wells (Rathnayake et al., 2022; Saghir et al., 2019c). For ESPCPs and ESPs, we used six parameters: **Pump Speed**, **Pump Torque**, **Water Flow Rate, Downhole Pressure, Discharge Pressure** and **Motor Temperature**. The



Icon		Function
⟷	Create Sequence	Select **two** Major Heatmap labels, where order of selection creates a *seqeuence*
🎞	Create Event	Select a **pair** or **multiple pairs** of Major and Anomaly Heatmap label to create *events* of interest
⭳	Extract Images	For the selected time *zoom* window, download the Major and Anomaly heatmap images for further analysis
🏷	Modify Label	Allow users to customise label names
○	Lock Function	Lock *event* or *seqeunce* once validated by engineers

***Major* and *Anomaly* bar chart filter.**

Fig. 21. Detailed view of the data annotation toolbar.

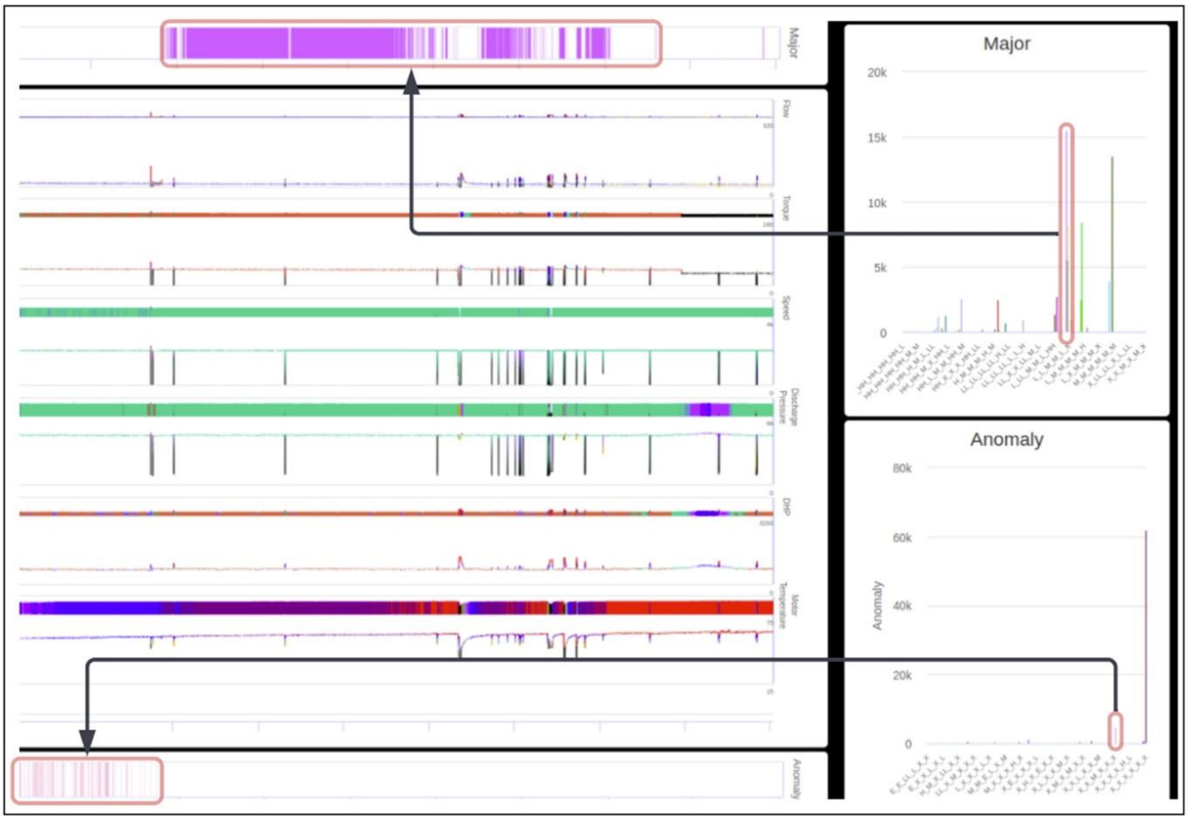


Fig. 22. Using the Major and Anomaly bar charts as filters for the time progression bar chart.

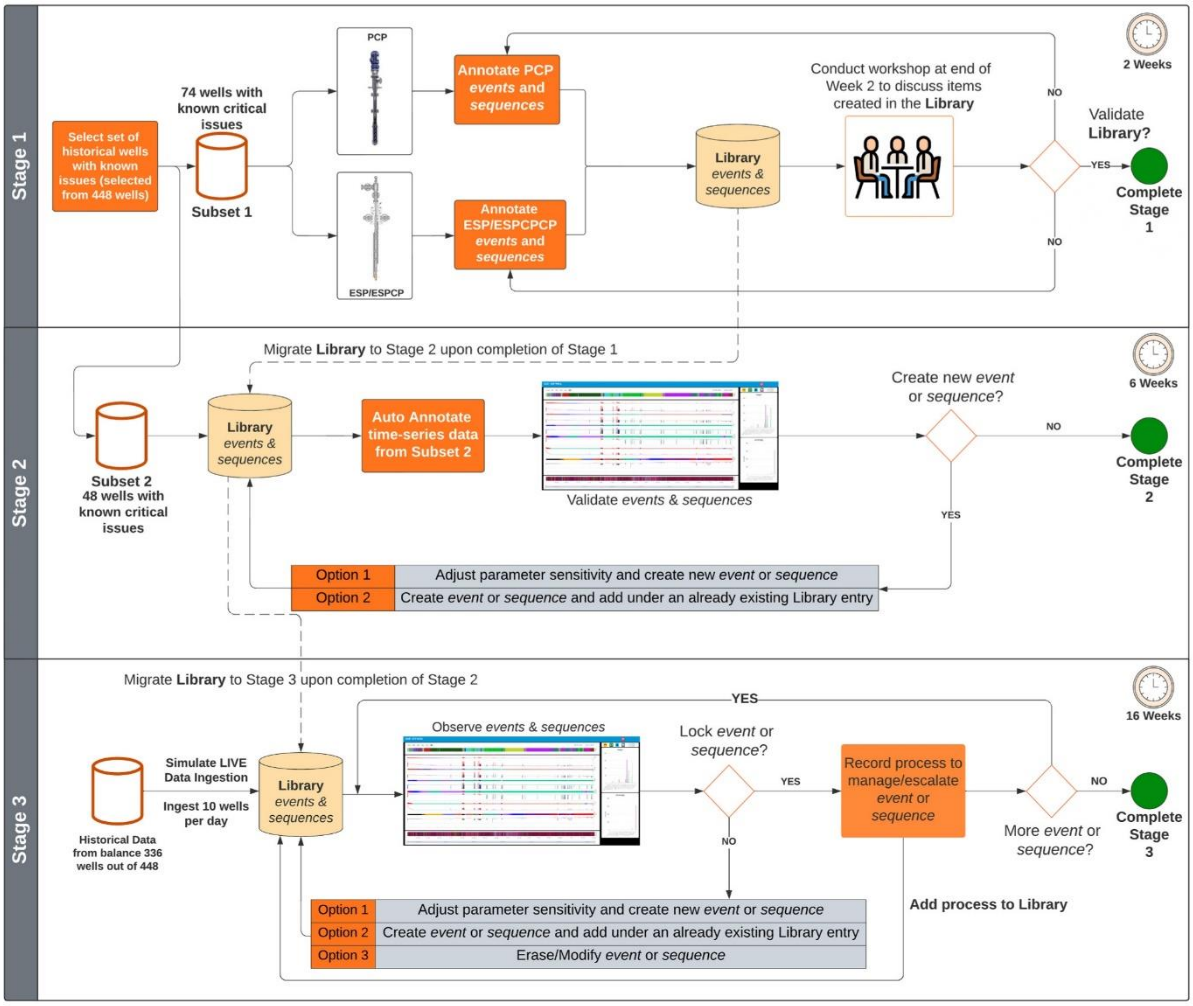

Stage 1
Select set of historical wells with known issues (selected from 448 wells)
74 wells with known critical issues
Subset 1
PCP
ESP/ESPCP
Annotate PCP events and sequences
Annotate ESP/ESPCPCP events and sequences
Library events & sequences
Conduct workshop at end of Week 2 to discuss items created in the Library
NO
Validate Library?
YES
NO
2 Weeks
Complete Stage 1
Stage 2
Migrate Library to Stage 2 upon completion of Stage 1
Subset 2 48 wells with known critical issues
Library events & sequences
Auto Annotate time-series data from Subset 2
Validate events & sequences
Create new event or sequence?
NO
YES
6 Weeks
Complete Stage 2
Option 1 Adjust parameter sensitivity and create new event or sequence
Option 2 Create event or sequence and add under an already existing Library entry
Stage 3
Migrate Library to Stage 3 upon completion of Stage 2
YES
16 Weeks
Historical Data from balance 336 wells out of 448
Simulate LIVE Data Ingestion
Ingest 10 wells per day
Library events & sequences
Observe events & sequences
Lock event or sequence?
YES
NO
Record process to manage/escalate event or sequence
More event or sequence?
NO
Complete Stage 3
Add process to Library
Option 1 Adjust parameter sensitivity and create new event or sequence
Option 2 Create event or sequence and add under an already existing Library entry
Option 3 Erase/Modify event or sequence

Fig. 23. A detailed flow chart showing the 3 stages that are undertaken as part of the data annotation process.

location of these data parameters is shown in Fig. 8.

Equations (1) and (2) demonstrate the correlation between the three PCP parameters, **Pump Speed**, **Pump Torque** and **Water Flow Rate.**

$$q_{th} = s\,\omega \tag{1}$$

Where,

Q_{th} = theoretical flow (m^3/d), s = pump displacement ($m^3/d/rpm$), ω = rotational speed (rpm)

$$T_{pr} = \frac{P_{pmo}E_{pt}}{C\omega} \tag{2}$$

Where,

T_{pr} = polished rod torque (Nm), P_{pmo} = prime mover power (kW),
E_{pt} = power transmission efficiency (unitless),
C = constant (unitless), ω = rotational speed (rpm).

Equations (1) and (2) are also valid for ESPCPs. However, as ESPCPs and ESPs have downhole sensors that measure **Downhole Pressure, Discharge Pressure** and **Motor Temperature**, we added these parameters to our list based on previous failure diagnostic studies of downhole motor-driven pumps.

Equation (3) shows the relationship between torque and speed for ESP wells.

The correlation between torque and speed for ESPs is shown in equation (3) (Takacs and Takacs, 2018).

$$T_{np} = \frac{5252 \times P_{np}}{N} \tag{3}$$

Where,

T_{np} = motor nameplate torque (Nm), P_{np} = mechanical power (kW), N = motor speed (rpm).

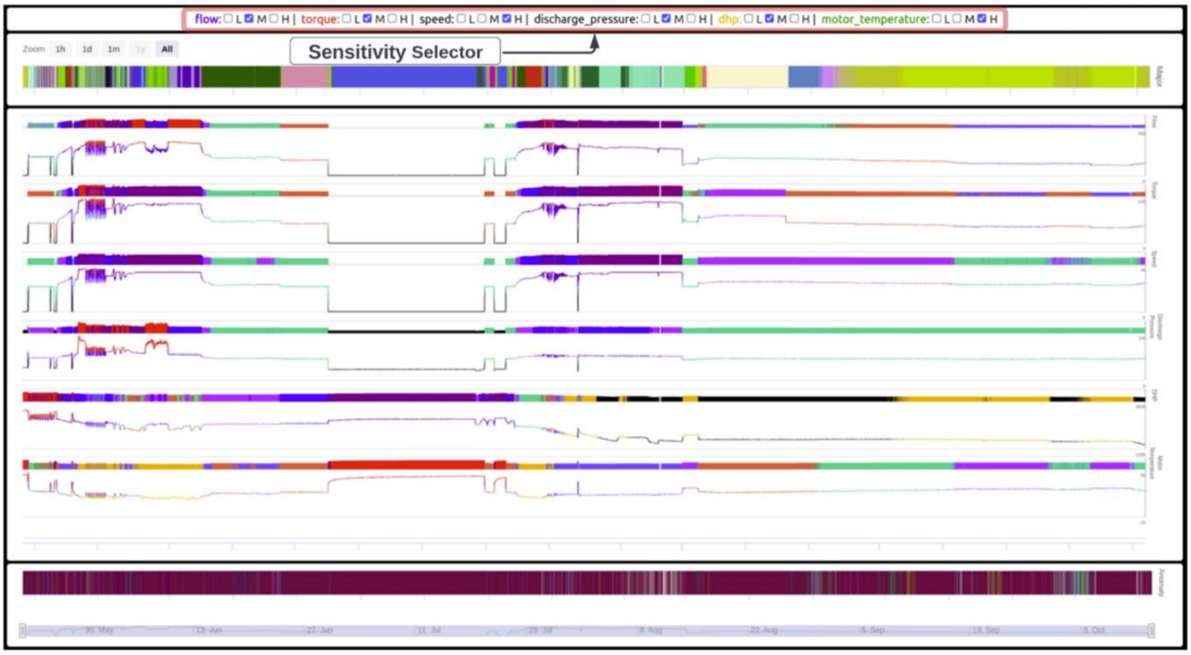
The correlation between motor current, motor voltage (both of which affect speed proportionally) and flow for ESPs is shown in equation (4) (Camilleri, 2013). The pump differential pressure measurement is the difference between the pump discharge and downhole pressure.

$$Q_p = \frac{V_m \times I \times PF \times \eta_m \times \sqrt{3}}{746} \times \frac{1}{\Delta P} \times \eta_p \times 58847 \tag{4}$$

Where,

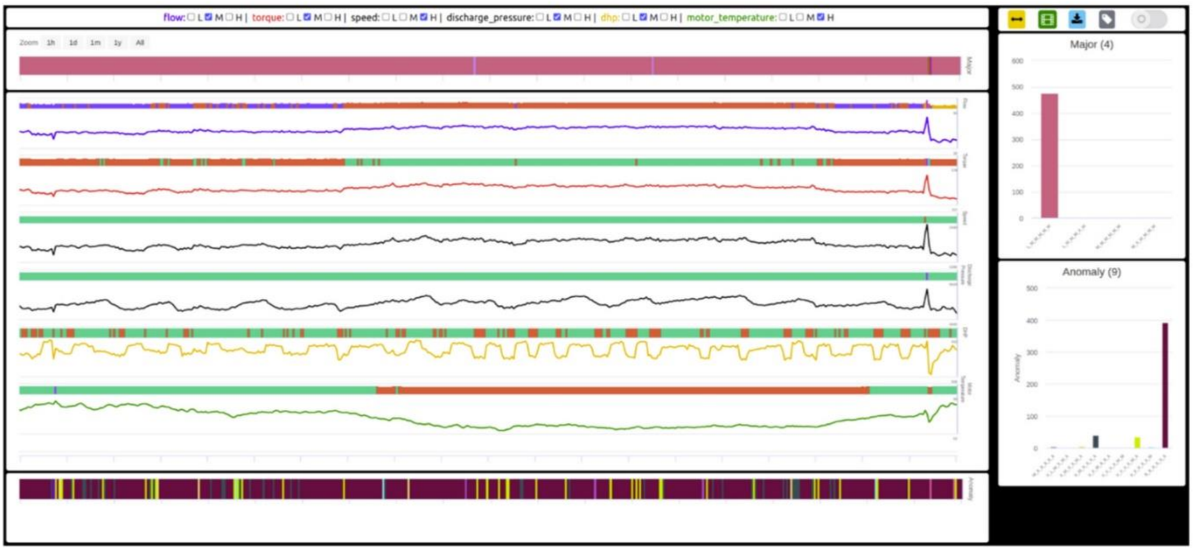
Q_p = flow rate (m^3/d), V_m = motor voltage (V), I = motor current (A),

η_m = motor efficiency (unitless),



flow: □ L ☑ M □ H | torque: □ L ☑ M □ H | speed: □ L □ M ☑ H | discharge_pressure: □ L ☑ M □ H | dhp: □ L ☑ M □ H | motor_temperature: □ L □ M ☑ H

Fig. 24. Sensitivity Selector feature in DAT.



flow: □ L ☑ M □ H | torque: □ L ☑ M □ H | speed: □ L □ M ☑ H | discharge_pressure: □ L ☑ M □ H | dhp: □ L ☑ M □ H | motor_temperature: □ L □ M ☑ H

Fig. 25. ESP parameters plotted with the SAX Sensitivity of Medium (Flow), Medium (Torque), High (Speed), Medium (Discharge Pressure), Medium (Downhole Pressure), and High (Motor Temperature).

η_p = pump efficiency (unitless), ΔP = pump differential pressure (PSI).

Moreover, pump manufacturers provide pump performance curves, shown in Fig. 9, which can be used to deduce flow rate based on pump operating frequency (speed).

3.2. Symbolic Aggregation Approximation (SAX) based Performance Images

In this section, we will expand on the previously developed SAX Performance Images method (Saghir et al., 2019a, 2019b, 2020) and introduce additional steps that help with the streaming analysis of time-series data. Most importantly, we will shed light on clustering SAX Performance Images by utilizing the expertise of Well Surveillance and Petroleum Engineers.

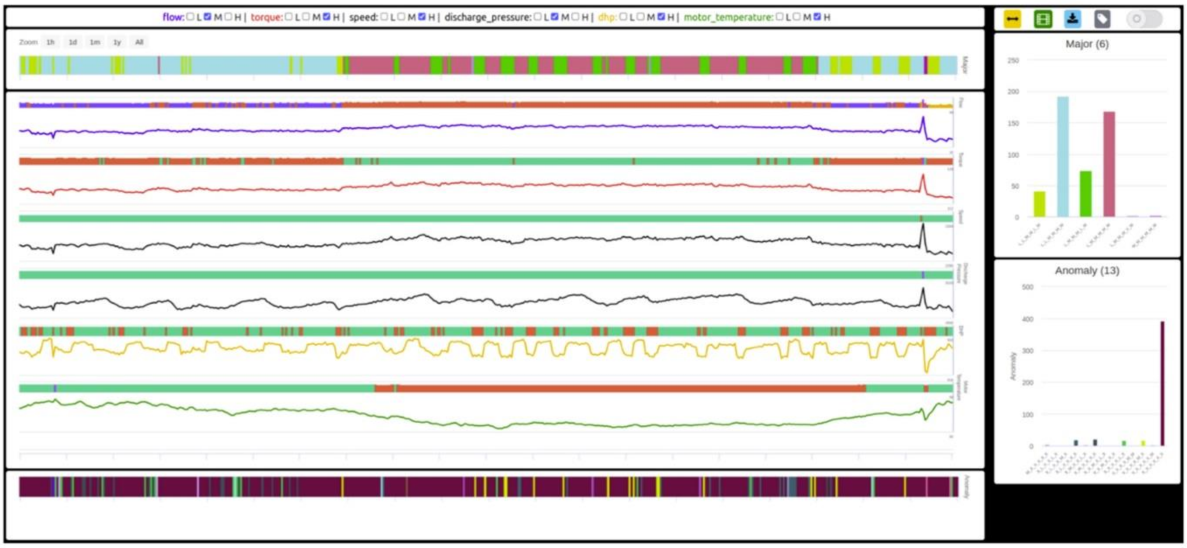

flow: □ L ☑ M □ H | torque: □ L □ M ☑ H | speed: □ L □ M ☑ H | discharge_pressure: □ L ☑ M □ H | dhp: □ L □ M ☑ H | motor_temperature: □ L □ M ☑ H

Fig. 26. ESP parameters plotted with the SAX Sensitivity of Medium (Flow), High (Torque), High (Speed), Medium (Discharge Pressure), High (Downhole Pressure), and High (Motor Temperature).

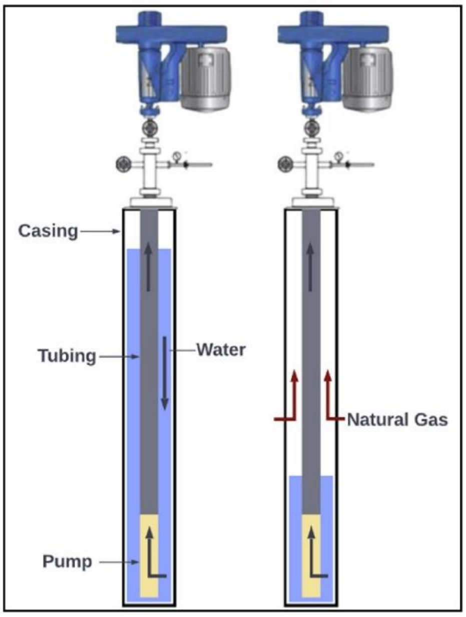


Fig. 27. Csg well drawdown process.

3.2.1. Expanding window technique for time-series analysis

Time-series analytics predominantly relies on sliding window or moving average techniques to examine changes in temporal behaviour (Braverman and Kao, 2016). Our earlier work found two significant drawbacks of this technique when applied to time-series data gathered from PCP wells. Firstly, a sliding window does not analyze the start-to-end performance of a PCP well; hence changes induced by mechanical and reservoir behaviour are missed. Second, as the Pump Speed is manually controlled by operators or adjusted using control algorithms, a sliding window of any time dimension provides only immediate information regarding changing pump behaviour due to speed control. Fig. 10 depicts the shortcomings of a sliding window method. This figure shows three time-series steps where SAX conversion is applied to each step in isolation. Once the SAX conversion is applied to the first window, the SAX symbols do not change when new data comes through. Hence, the temporal information of preceding data is lost with this method.

To overcome these drawbacks, we utilized the expanding window technique. This method allows us to account for the effect of past observations on current and future observations. In this technique, a fixed-size window is used to define a subset of the time series data, which is gradually expanded over time as more data becomes available. The expanding window technique allows for incorporating additional information as it becomes available, leading to more accurate predictions and a better understanding of the underlying trends and patterns in the data. In Fig. 11, we can see how the prior SAX symbols adjust based on new information. Although we do not use the preceding information, it is essential to note new data's effect on previous SAX symbols. As new data adjusts based on historical data, we can easily observe the effect of mechanical degradation and reservoir behaviour via the adjusting SAX symbols. This will be discussed in detail in the forthcoming sections. Additionally, when a new pump was installed in a CSG well, we restarted the expanding window analysis, which forced the SAX symbols to readjust based on the new pump behaviour.

3.2.2. Multivariate time-series conversion to SAX sensitivity arrays

SAX was first introduced in 2003 b y Lin et al. as a dimensionality reduction method to represent univariate time-series data as discrete symbols (Lin et al., 2003), an example of which is shown in Fig. 12. Since then, multiple papers have discussed using SAX for anomaly detection in time-series data (Keogh et al., 2005; Lin et al., 2007; Kumar et al., 2005; Keogh and Lin, 2005).

However, most of these papers have focused on its application to univariate time-series data and did not provide a comprehensive method to analyze multivariate time-series data.

Our approach to converting multivariate data to SAX involved transforming each time-series parameter into nine (9) SAX symbols using the expanding window technique (Saghir et al., 2020). With these nine (9) symbols, we can individually set the sensitivity of each

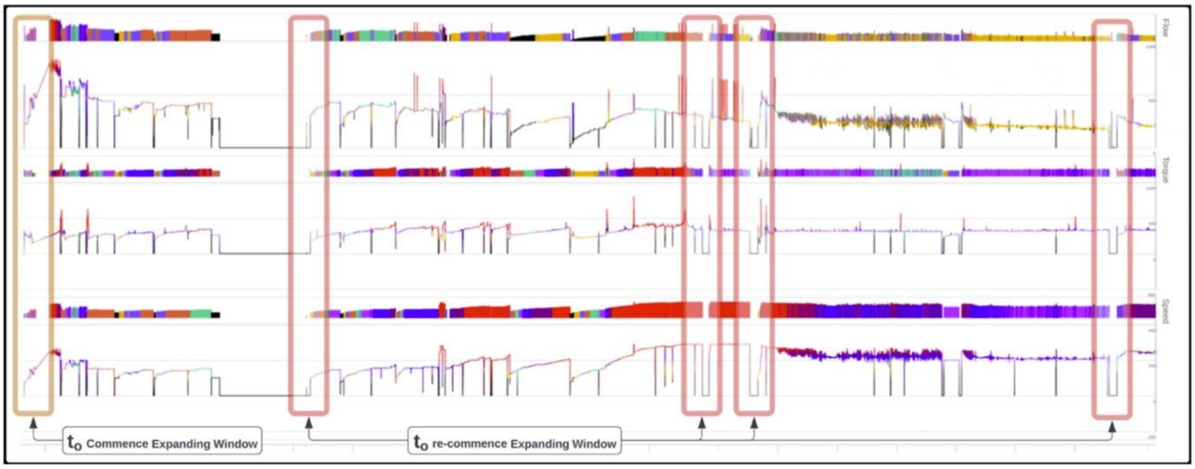


Fig. 28. Diagram depicting shutdown periods after which expanding window was reset for calculating SAX symbols.

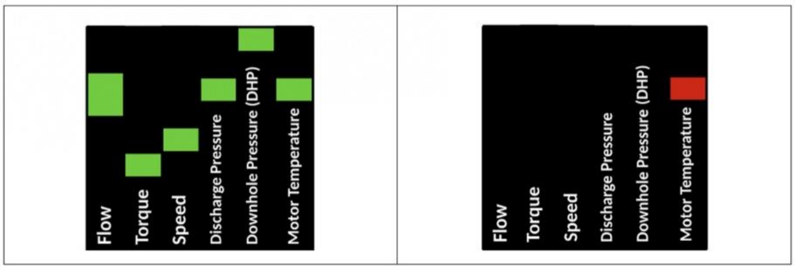


Fig. 29. Major and anomaly heatmap for ESP/ESPCP wells.

parameter, hence accounting for fast or slow-changing parameters. Our research determined that three sensitivities adequately define the SAX bins for the time-series parameters. Namely, these sensitivities are *low*, *medium*, *high*, and their SAX symbol distribution are set as *9x1*, *5x1* and *3x1* arrays, respectively. These sensitivities are depicted in Fig. 13.

Furthermore, in Fig. 14, we show how multivariate time-series data is converted to SAX sensitivity arrays. The selection of SAX sensitivity depends on a parameter's influence on the performance of the AL system. For example, the pump speed is set either manually or through an automated control algorithm in the RTU; hence setting the speed to a *high* sensitivity ensures that the model picks up any speed change in real-time. On the other hand, flow and torque parameters are set to *medium* as they do not change as abruptly as the speed parameter. It is important to note that the Petroleum or Surveillance Engineers can dynamically set the parameter sensitivity while analyzing the pump performance. We will show this feature in the upcoming sections.

3.2.3. *Converting SAX sensitivity arrays to Performance Heatmap Images*

Our previous publications (Iranzi et al., 2022; Awaid et al., 2014) provided a Performance Image Conversion methodology based on a 1-h expanding window. However, once we engaged with other CSG operators, we realized that having a more robust heatmap conversion methodology was required to cater for an expanding window of any size. Fig. 15 shows how we colour-code Anomaly and Major heatmaps based on the SAX sensitivity arrays.

The colour is dependent on the $n_{occurrence}$ factor, which can be calculated using the following equation:

$$n_{occurence} = \frac{n_{symbol}}{n_{expanding}} \quad (5)$$

Where,

N_{symbol} = count of SAX symbol in the expanding window,

$N_{observations}$ = number of recorded observations in the expanding window.

Fig. 16 shows how the colour code is applied to the speed parameter. In this observation, we have assumed that there are ten readings ($n_{observations}$) in the $t_{expanding}$ window of size 10 min. Based on the number of SAX occurrences, the SAX sensitivity array is coloured accordingly. Anomaly Heatmaps identify abrupt changes that occur in the expanding window period. Likewise, Major Heatmaps identify the change in the overall performance of the AL system. In the later sections of this paper, we will discuss in detail how these heatmaps aid in creating *events* of interest and assist Petroleum and Surveillance Engineers in labelling time-series data. Finally, in Fig. 17, we elaborate on how the colour code is applied to individual parameters to create the SAX Performance Heatmap Images.

3.3. *Time series data annotation with the aid of SAX heatmap images*

As stated earlier in the paper, there is a significant lack of labelled time-series data for AL systems operated in CSG wells. This presents a considerable challenge in assessing the performance of these systems methodically. To have any meaningful impact on how the performance of these systems can be evaluated, it is essential to have a thorough data annotation process in place.

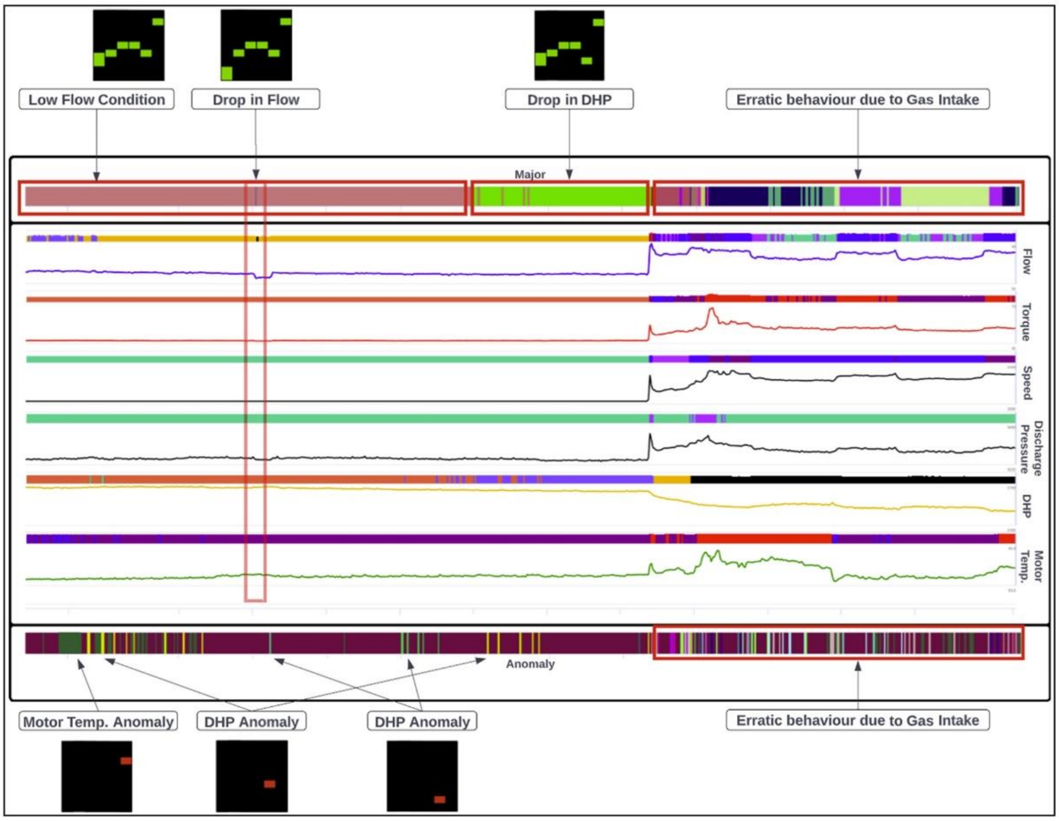


Fig. 30. – Changes in Major and Anomaly Labels before Gas Intake behaviour in ESP.

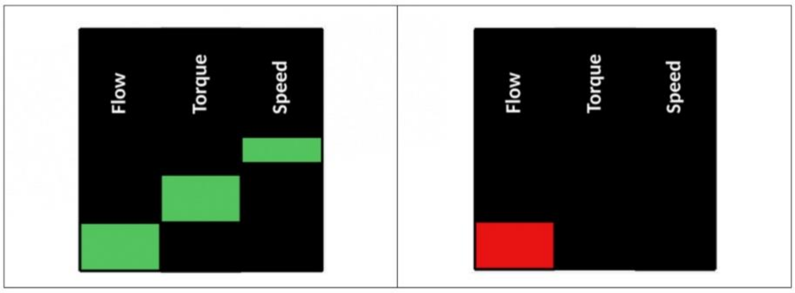


Fig. 31. Major and anomaly heatmap for PCP wells.

The SAX Heatmap Images are a powerful tool that can significantly reduce the effort and time required for annotating time series data. In addition, as the heatmap images are much more compact than the original time series data, they significantly reduce the amount of data that needs to be annotated.

During our research, we relied on the expertise of experienced Well Surveillance and Petroleum engineers to help with the time-series labelling process. These engineers were able to provide valuable insights into the data and were able to identify patterns and anomalies that would have been difficult to detect otherwise. This helped us ensure that the data annotation process was accurate and reliable.

The data annotation process was carried out in multiple stages. In the first stage, the engineers were asked to combine the Major and Anomaly heatmap images based on their expertise and experience and create *events* and *sequences* that define the performance of the AL system. In the second stage, the engineers were asked to review and validate the *events* and *sequences* of historical data not used in the first stage. In the final stage, the engineers were asked to evaluate the *events* and *sequences* on pseudo-live data (historical data ingested as a live data feed) to understand how they would respond to real-time alerts. A methodology to respond to alerts was also formalized during this stage.

Once the data annotation phase was completed, a real-time data ingestion engine was created to automate the detection of *events* via a machine learning model. The end-to-end process from data annotation to solution deployment was completed in six months. In the coming subsections, we will provide an overview of the data annotation process and

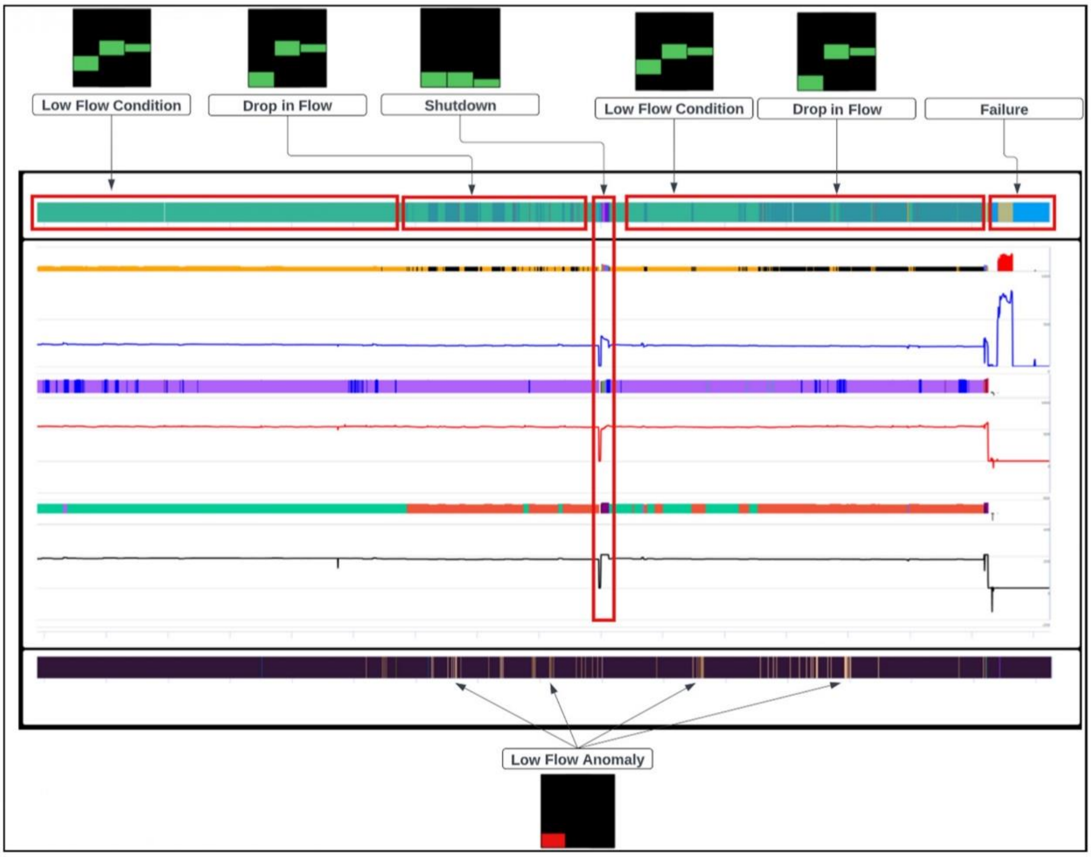
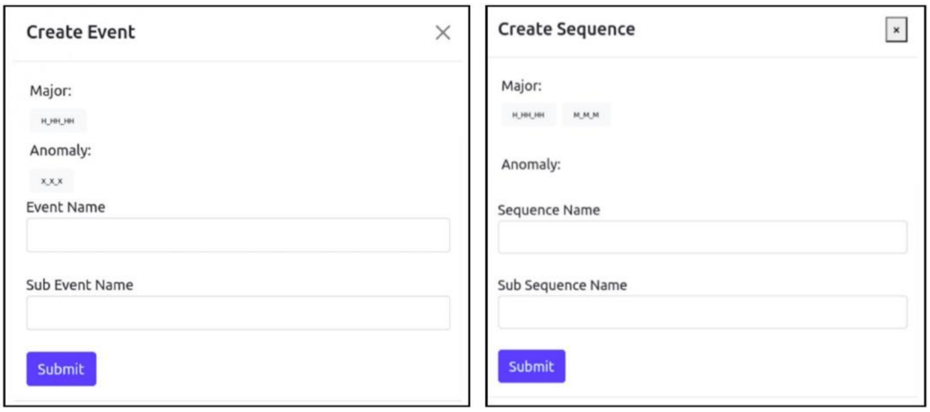


Fig. 32. Changes in major and anomaly labels before PCP failure due to end of life.



Create Event

Major:
H_HH_HH

Anomaly:
X_X_X

Event Name

Sub Event Name

Create Sequence

Major:
H_HH_HH M_M_M

Anomaly:

Sequence Name

Sub Sequence Name

Fig. 33. (Left) User Interface to add sub-events under *events*, (Right) User Interface to add sub-sequence under sequences.

show a User Interface was specifically developed to aid engineers with labelling time series data.

3.3.1. Labelling SAX images

Before initiating the annotation process, it is essential to establish a logical grouping of heatmaps as the foundation for event labelling. In our past research, we employed a Convolutional Auto-Encoder (CAE) based clustering technique to label SAX images (Saghir et al., 2020, 2022). Our previous work focused only on PCPs with one set of SAX Array sensitivity.

However, the pre-labelling process utilizing CAE-based clustering became increasingly complex as we collaborated with additional CSG operators who utilized ESPs, ESPCPs, and conventional PCPs. The increased complexity was due to the additional parameters and SAX

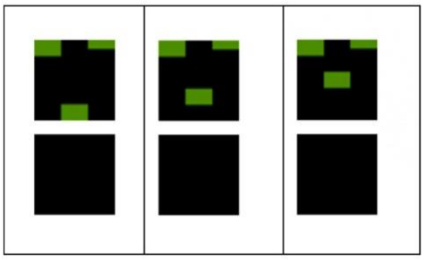


Fig. 34. –Multiple sub-events combined to create the Drawdown event.

array sensitivities used for ESPs and ESPCPs.

Hence, we developed a simple yet effective method to automatically pre-label SAX images to overcome these additional complexities. With this new approach, the SAX images were automatically labelled based on the position of the colour markers within the SAX sensitivity arrays. The label coding schema is presented in Fig. 18, and the application of this schema is shown in Fig. 19. An empty SAX sensitivity array is labelled X, and an array with more than one coloured block is labelled E.

3.3.2. *User Interface for time series data annotation*

Our initial implementation of the data annotation User Interface (UI) was based on PowerBI and was designed to cater solely to PCP wells (Saghir et al., 2019c). However, this design had a limitation: it could only function with a single set of sensitivity levels for PCP parameters. As the requirement expanded to include ESP and ESPCP wells, and the need for adjustable sensitivities arose, it became necessary to rebuild the data annotation tool. In the new iteration of the UI, we used JavaScript to create a customized dashboard, which allowed for a simplified data annotation process involving the three types of Artificial Lift systems. The UI design was based on feedback received from Petroleum and Well Surveillance Engineers.

The resulting interface greatly simplified the data annotation process, allowing for more efficient use of resources, as the annotation process could be completed with fewer personnel or in a shorter amount of time. Also, the re-designed UI reduced the risk of errors and inconsistencies, resulting in a more accurate and reliable data annotation process.

The resulting User Interface (UI) for Data Annotation is shown in Fig. 20. Breaking down the interface, two areas above and below the Time Series Trends show the progression of Major and Anomaly Heatmaps. The different colours correspond to the Major and Anomaly Heatmap labels as processed by the label coding schema. On the right, we have the Data Annotation Toolbar (DAT) that is used by Production and Well Surveillance engineers to examine and annotate the time-series data. The various selectable options in the DAT are shown in Fig. 21. Each bar in the Major and Anomaly bar chart window is designed as a filter. Clicking on any of the labels on the bar chart will only show that particular label in the respective progression bar chart, as shown in Fig. 22. This filter allows users to easily select areas of interest and use this information to create *events* and *sequences* that can be used to assess AL system performance to look at early failure trends and identify abnormal system behaviour.

3.3.3. *Time series data annotation process*

The data annotation process aims to capture *events* & *sequences* deemed necessary to Petroleum and Well Surveillance Engineers. In addition, the DAT helps facilitate the knowledge capture process from these experts to improve the *events* and *sequences* database over time. Each stage of the data annotation process is covered in detail in Fig. 23.

The research team was involved daily with the engineers during Stage 1 of the data annotation process. This level of collaboration allowed for real-time feedback and adjustments to be made to the data annotation process as it was being carried out. The research team was able to provide guidance and address any issues that arose during the annotation process, ensuring that the data annotation process remained on track. In Stages 2 and 3 of the data annotation process, the level of collaboration between the research team and engineers was reduced. However, meetings were still held to discuss procedural matters and make minor adjustments to the DAT. These meetings were essential to ensure that the data annotation process continued to be improved and optimized.

4. Results

4.1. *Observations made during the data annotation process*

The six months spent annotating the data gave us insights into how engineers engaged with the labelled Major and Anomaly heatmaps. This section covers valuable observations that helped us develop a useful time-series analysis tool for monitoring AL system performance with streaming data.

4.1.1. *Observation 1: dynamically adjusting SAX array sensitivities*

During Stage 2, it was observed that the real-time parameters for horizontal ESP and ESPCP wells were not changing as dynamically as they were in vertically installed PCP wells, which led to a need for adjusting the SAX array sensitivities. In response to this issue, we developed a feature in the DAT that allowed engineers to adjust the SAX array sensitivities dynamically. By doing so, the engineers had higher confidence when creating *events* and *sequences*.

Fig. 24 shows the Sensitivity Selector feature, where engineers can select the sensitivity of any plotted parameters. This feature adjusts labels based on the pre-computed SAX symbols. By adjusting the sensitivity dynamically, engineers could understand which parameters impacted the Major and Anomaly heatmap labels most. Figs. 25 and 26 show ESP trends with two different SAX sensitivity settings.

In Fig. 25, the parameters for an ESP well are plotted with the SAX sensitivity of Medium (Flow), Medium (Torque), High (Speed), Medium (Discharge Pressure), Medium (Downhole Pressure), and High (Motor Temperature). In Fig. 26, the same parameters are plotted with the SAX sensitivity of Medium (Flow), High (Torque), High (Speed), Medium (Discharge Pressure), High (Downhole Pressure), and High (Motor Temperature). Over the plotted period, Fig. 25 has four (4) Major labels and nine (9) Anomaly labels.

Whereas Fig. 26 has six (6) Major labels and thirteen (13) Anomaly labels. The adjusted SAX sensitivity in Fig. 26 can distinctly identify the effect of changing Downhole Pressure. By adjusting the SAX sensitivities, engineers could ascertain which parameters impacted the SAX images most, providing labels that could easily be used to create meaningful *events* and *sequences*.

4.1.2. *Observation 2: reset expanding window after PCP shutdown*

When a vertical PCP commences de-watering of a CSG well, the water column in the casing is pumped out via the tubing to expose the coal formations that release gas into the casing. This process is known as water drawdown and is shown in Fig. 27. However, when a pump shuts down, the water column rises back into the casing, so the drawdown process must be commenced again.

During the data annotation process, it was observed that re-commencing the drawdown produced SAX images that were not representative of the pump performance. This happened due to the pump efficiency loss caused by the stator's wear and tear. Hence, a modification was made to the SAX pipeline, where the expanding window was reset whenever a pump restarted after a period of 24 h. By doing so, the SAX images would adjust to represent the drawdown process based on the lower pump efficiency. For example, in Fig. 28, we show the shutdown periods that are greater than 24 h, after which the expanding window was reset to calculate SAX symbols based on adjusted pump

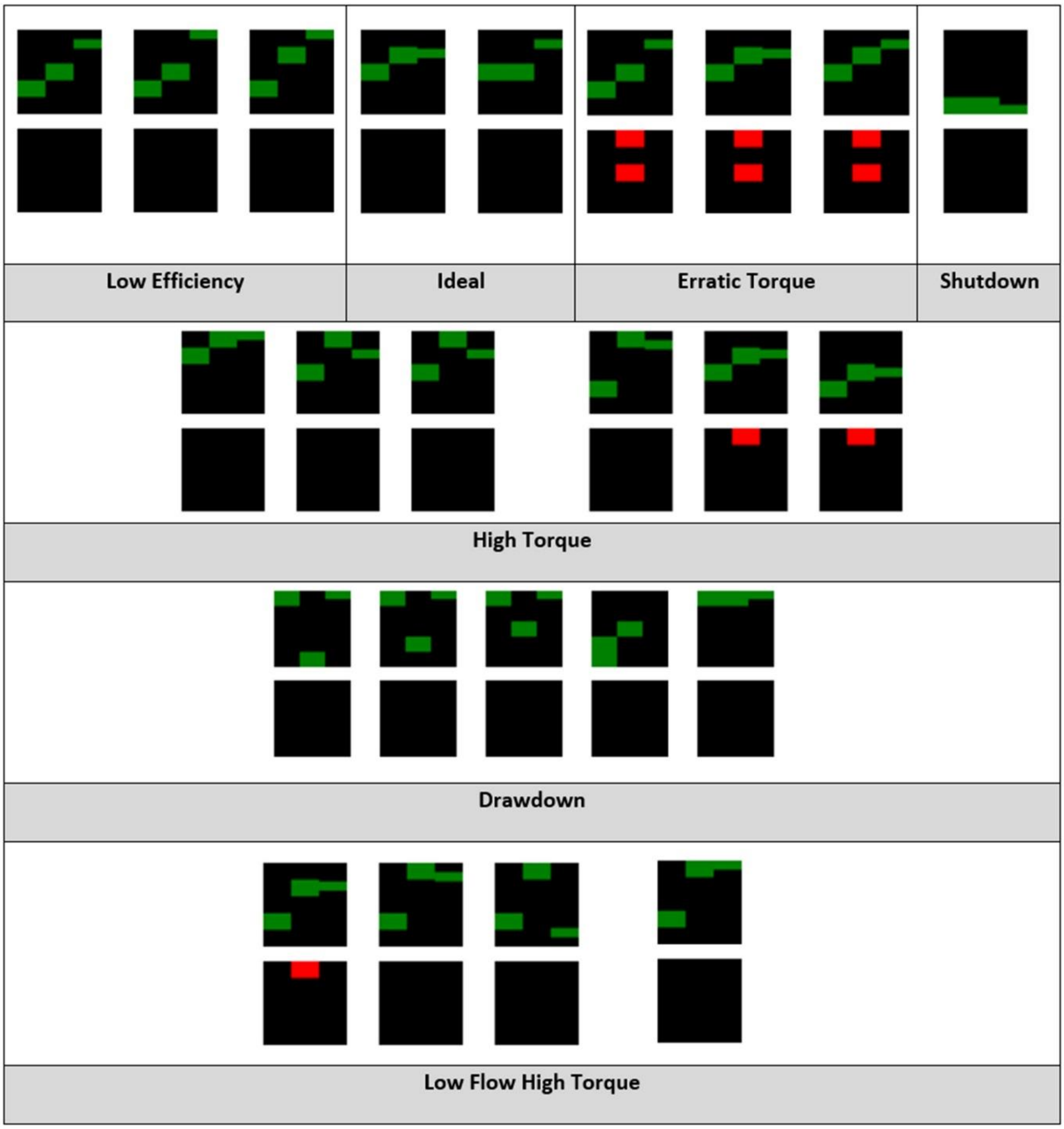


Fig. 35. PCP Events stored in the Events and Sequence Library.

efficiency.

4.1.3. Observation 3: detecting early onset of abnormal AL behaviour

To gauge changes in AL performance, the engineers tracked how the Major and Anomaly heatmap labels changed prior to a particular abnormal AL behaviour. The change in labels was beneficial in identifying changes in both the mechanical and reservoir behaviour associated with AL-operated wells. Fig. 29 shows the SAX Major and Anomaly Heatmaps for an ESP well with six parameters. Fig. 30 shows how Major and Anomaly label changes progress to a Gas Intake event. It can be seen that before the erratic behaviour of the ESP, there was a drop in downhole pressure (shown in the Major Heatmap labels), and there was also an increase in the frequency of downhole pressure anomalies (as shown in the Anomaly Heatmap labels).

Similarly, before a PCP failure due to pump end-of-life, it can be observed in Fig. 31 that there is a drop in flow as depicted by the changing Major Heatmaps. Simultaneously, there is an increase in the number of Low Flow Anomaly Heatmaps. Hence, to pre-empt failure due to end-of-life behaviour, it is vital to observe the behaviour of the PCP when it enters the Low Flow condition Fig. 32. In the upcoming section, we will cover critical *events* and *sequences* that are considered early indicators of abnormal AL behaviour.

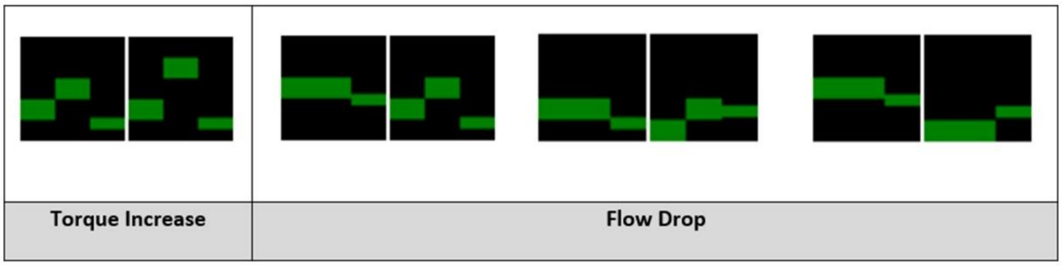


Fig. 36. PCP Sequences stored in the Events and Sequence Library.

4.1.4. Observation 4: grouping similar performance characteristics

During the data annotation process, engineers recognized that various Major and Anomaly pairs could be grouped collectively to form the same event or sequence. By doing so, similar performance characteristics could be grouped to reduce the overall number of *events* and *sequences*.

Fig. 33shows the User Interface developed for grouping multiple combinations of sub-events and sub-sequences into single *events* and *sequences*. In Fig. 34, we show an example where three combinations of sub-events (Major and Anomaly pairs) are combined to detect a drawdown event. If any Major and Anomaly sub-event pairs occur in the time-series data, it is designated as a drawdown event.

4.1.5. Observation 5: events and Sequences Library

The *events* and *sequences* library had 18 entries after Stage 3 of the data annotation process. The *events* and *sequences* shown in Fig. 35, Fig. 36, Figs. 37 and 38 were classified as must-have by Petroleum and Surveillance engineers to assist with exception-based surveillance.

4.2. Artificial Lift System Analytics Application

To understand the effectiveness of the data annotation process and how the resulting *events* and *sequences* impact the performance analysis of the AL systems, we created an Artificial Lift System Analytics Application (ALSAA) to facilitate the work for engineers. In this section, we will showcase how ALSAA, facilitated by the *events* and *sequences* library, improved the management of AL systems. In addition, the application also provided the operators with previously unseen insights, which helped improve the overall performance of these systems.

4.2.1. Artificial Lift System Analytics Application (ALSAA)

ALSAA comprises of three main components: data ingestion, SAX processing and the application dashboard. The data flow architecture for ALSAA is shown in Fig. 39. The data ingestion and SAX processing components are invoked every 5 min, and they update the time series database. Anomaly and Major labels are passed to the *events* and *sequences* library, where they are checked against entries in the library. If a match exists, the alert database is updated, and notifications are pushed to the application dashboard. As part of the application, we allowed specific users to edit the *events* and *sequences* library through the dashboard. This feature was added in-case engineers wanted to add a new event or sequence of interest that may not have been recorded or observed during the data annotation process.

4.2.2. Exception-based surveillance dashboard

The exception-based surveillance dashboard is the home page of ALSAA. The dashboard is split into four (4) areas shown in Fig. 40.

4.2.2.1. Dashboard area 1. The dashboard Area 1, shows various menu options for ALSAA. The detail of each menu option is shown in Fig. 41. These options will be discussed in more detail in the forthcoming sections.

4.2.2.2. Dashboard area 2. As shown in Fig. 42, Dashboard Area 2 depicts live cards which provide a quick summary to the engineers for various wells. Most importantly, the number of wells under observation, low-efficiency wells and drawdown wells are critical in tracking abnormal AL behaviour.

4.2.2.3. Dashboard area 3. Dashboard Area 3 summarises the wells covered in Dashboard Area 2. The summary aims to provide engineers with the ability to track wells of interest proactively. For example, Fig. 43 shows an overview of Under Observation wells. This summary provides an overview of the event being tracked for each well, their Activity trend, and which cause is being tracked. Wells are added to the Under Observation card through the Live Alerts page, details of which are covered in the next section.

4.2.2.4. Dashboard area 4. The Live Map gives engineers an overview of the well conditions based on various filters in Dashboard Area 4. As shown in Fig. 44, users can filter wells based on labels, *events* and *sequences*. They can further see the behaviour of the screened wells based on the time-period selection. The selected filter can be applied to all dashboard areas. The map-based filter selection is a quick way for engineers to find wells with issues they may be interested in investigating further.

4.2.3. Automated real-time alerts

In our previous approach, where we assessed only PCPs with performance heatmaps (Saghir et al., 2022), we used an image clustering approach to label the images and produce necessary alerts. However, this approach depended highly on engineers supervising cluster labelling and ensuring efficient output of automated alerts. In our current approach, the auto-labelled heatmap images provided a more streamlined and efficient automated alert method, requiring minimum engineers' supervision. Fig. 45 shows how the alerts are processed once the image labels are processed in the *events* and *sequences* Library.

The Alerts Logic Processor (ALP) assesses the streaming *events* and *sequences* and puts them into categories to be displayed in the Application Dashboard. The application has three alert types: Notification, Warning and Critical Alarm. The ALP also calculates the cumulative time of each category and records any necessary action the engineers took. The Alert View on the Application Dashboard is shown in Fig. 46.

4.3. Additional analytics tools based on heatmap labels, events and sequences

As part of ALSAA, five visual analytics tools were developed to assist engineers with further analyzing AL performance. These tools and their functionalities are listed below.

4.3.1. Unique label analysis

This tool aims to monitor Unique Major and Anomaly labels and identify if an AL system behaves differently from similar AL system types. Fig. 47 shows the tracking of unique *events* over three days for a

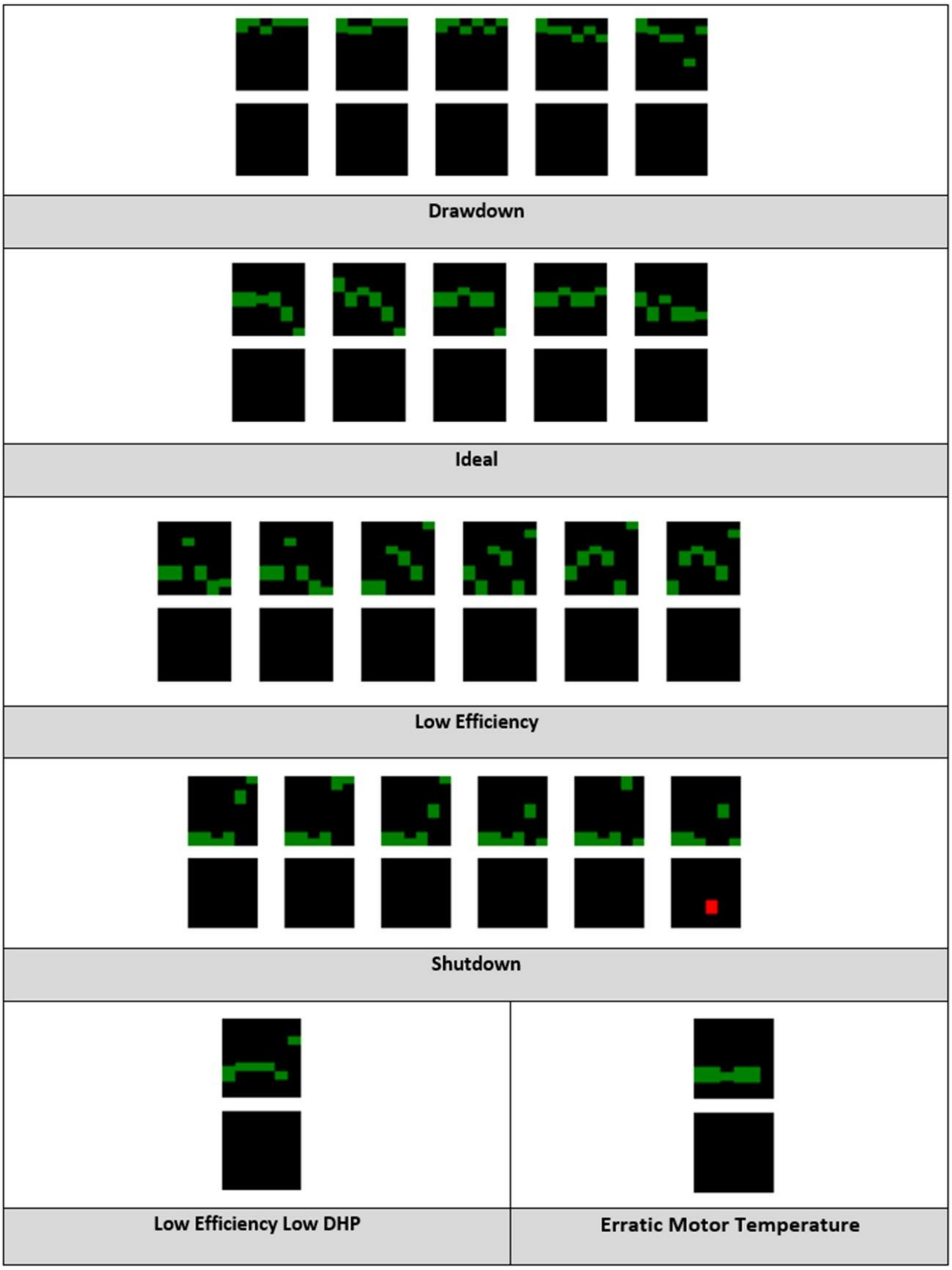


Fig. 37. ESP Events stored in the Events and Sequence Library.

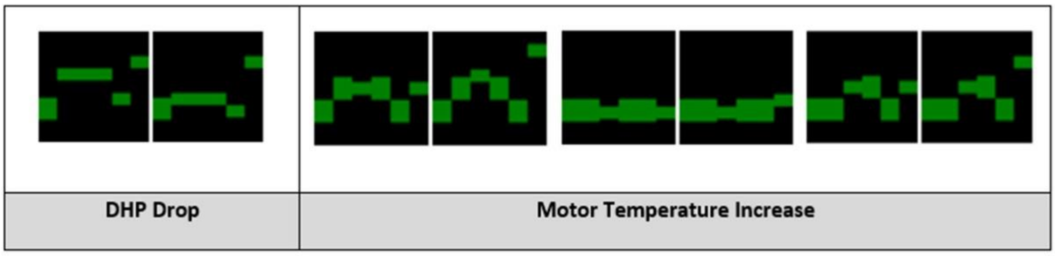


Fig. 38. ESP Sequences stored in the Events and Sequence Library.

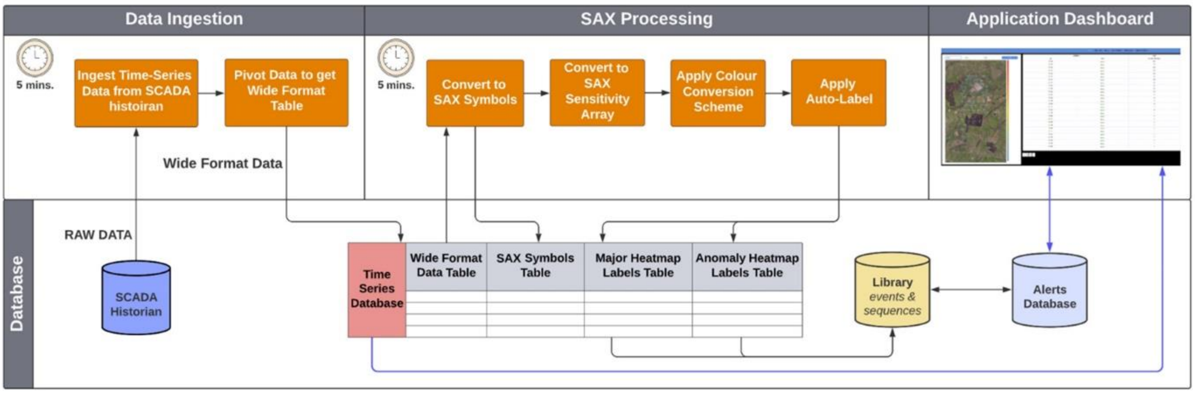
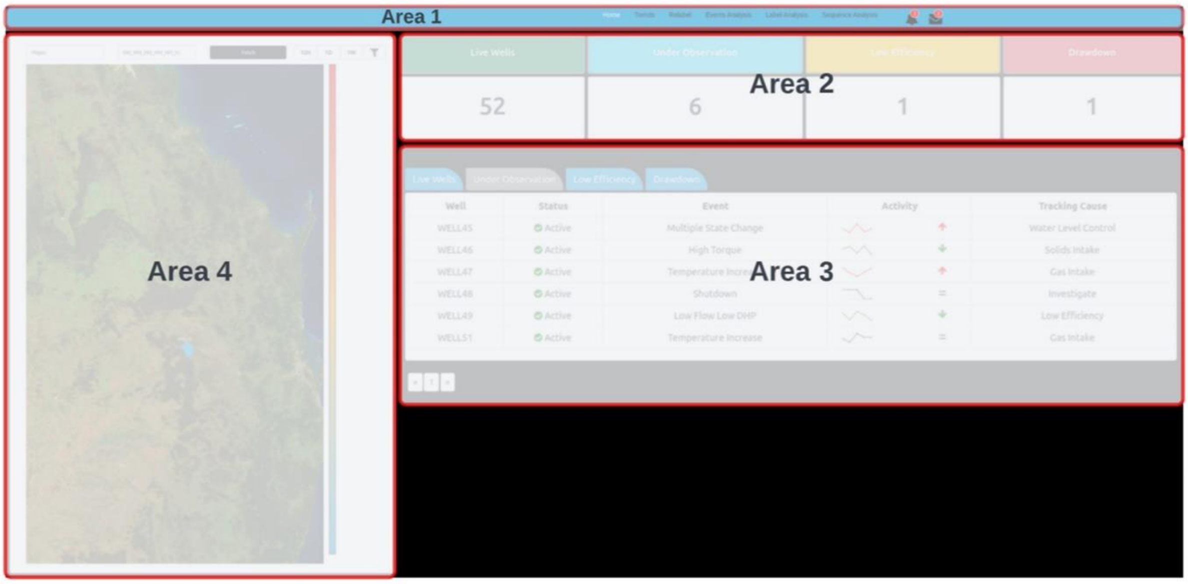


Fig. 39. – Data flow architecture: Artificial lift systems analytics application.



Well	Status	Event	Activity	Tracking Cause
WELL45	Active	Multiple State Change	↑	Water Level Control
WELL46	Active	High Torque	↓	Solids Intake
WELL47	Active	Temperature Increase	↑	Gas Intake
WELL48	Active	Shutdown	=	Investigate
WELL49	Active	Low Flow Low DHP	↓	Low Efficiency
WELL51	Active	Temperature Increase	=	Gas Intake

Fig. 40. Surveillance Dashboard. Area 1: Selection Menu and Alerts Notifications. Area 2: Display Cards for Live Wells, Under Observation Wells, Low-Efficiency Wells and Drawdown Wells. Area 3: Tabs showing details of wells mentioned on the display cards in Area 2. Area 4: Live Well Filter allows users to filter wells based on Labels, Events and Sequences.

particular well. On day 1, the tool recorded one label out of seven which did not exist in the same or similar wells to that date. By monitoring these unique labels, engineers could identify new behaviour of AL systems and identify an opportunity to create new *events* or *sequences.*

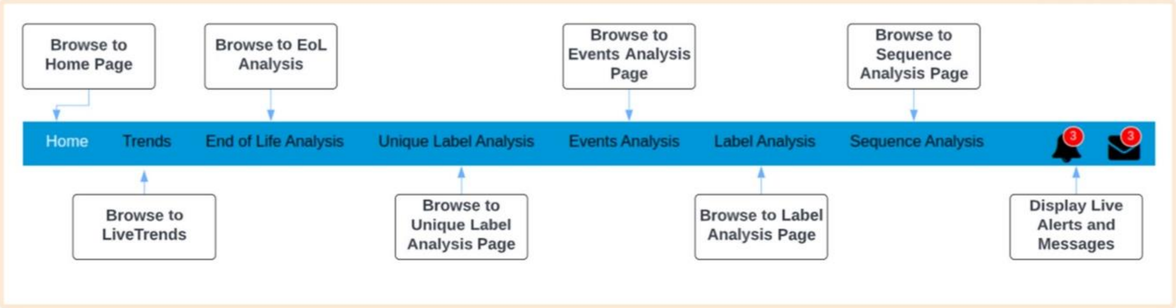
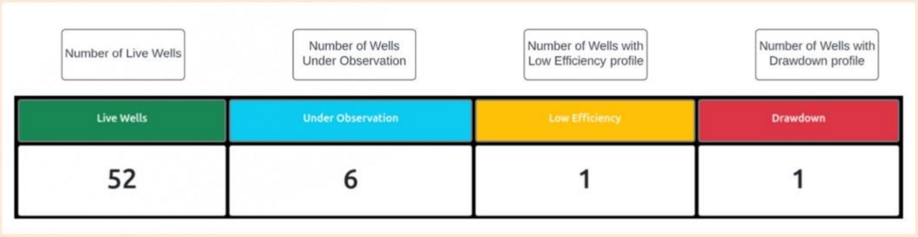
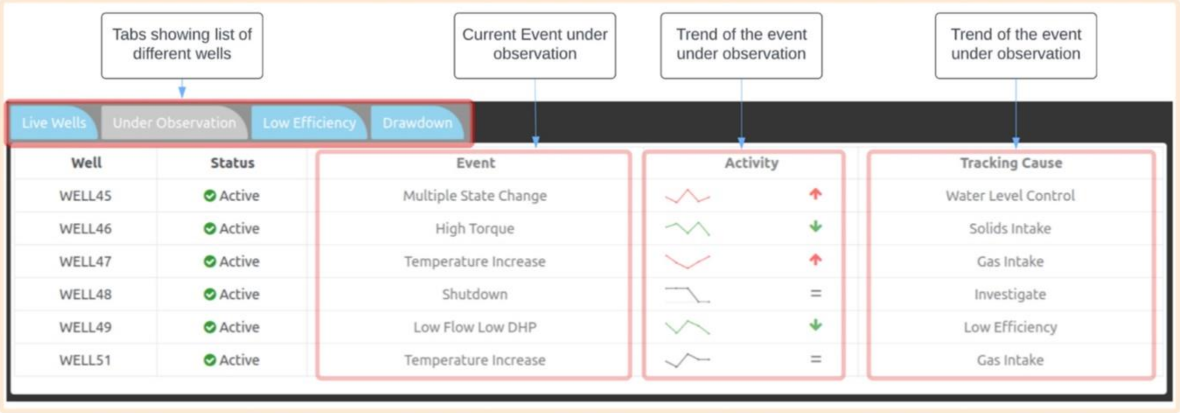


Fig. 41. Dashboard area 1: Menu options in the artificial lift systems analytics application.


Number of Live Wells
Number of Wells Under Observation
Number of Wells with Low Efficiency profile
Number of Wells with Drawdown profile

Live Wells	Under Observation	Low Efficiency	Drawdown
52	6	1	1

Fig. 42. Dashboard Area 2: Live Cards displaying the number of various wells to track Artificial Lift performance.


Tabs showing list of different wells
Current Event under observation
Trend of the event under observation
Trend of the event under observation
Live Wells
Under Observation
Low Efficiency
Drawdown

Well	Status	Event	Activity	Tracking Cause
WELL45	✓ Active	Multiple State Change	↑	Water Level Control
WELL46	✓ Active	High Torque	↓	Solids Intake
WELL47	✓ Active	Temperature Increase	↑	Gas Intake
WELL48	✓ Active	Shutdown	=	Investigate
WELL49	✓ Active	Low Flow Low DHP	↓	Low Efficiency
WELL51	✓ Active	Temperature Increase	=	Gas Intake

Fig. 43. – Dashboard Area 3: Live Cards displaying the number of various wells to track Artificial Lift performance.

4.3.2. End-of-life analysis

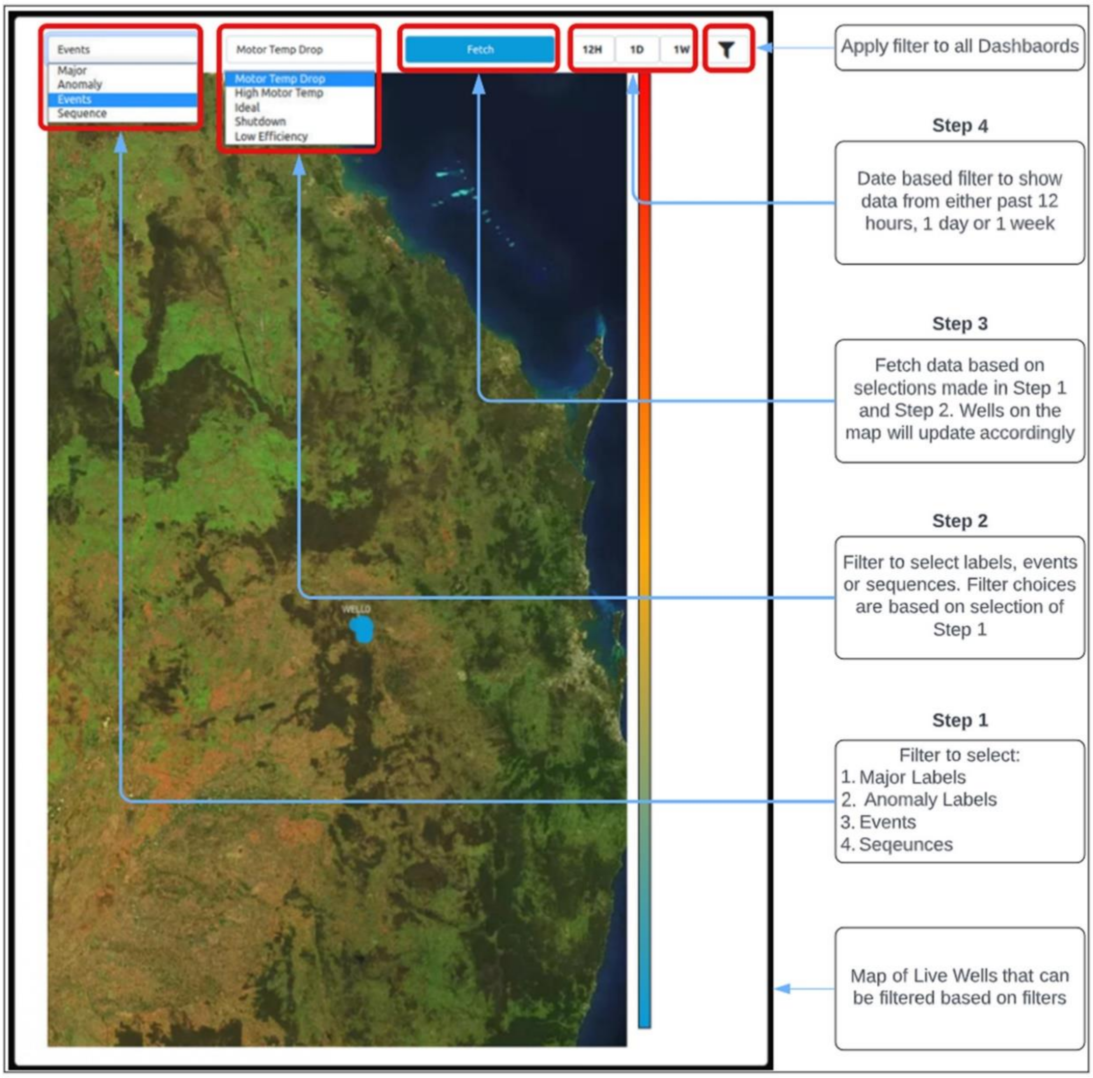
As the pumps age, their efficiency decreases depending on the wear and tear of internal components. Hence, it is important for engineers to understand how efficiency behaves over time based on the Major Heatmap labels. Each Major Heatmap label was given a pre-determined score, and the value was plotted as a heatmap over time. Fig. 48 shows the end-of-life heatmap laid over a live well plot. The various stages during the pump life are shown in the heatmap plot, with a clear indication of when the pump enters the low-efficiency stage, followed by the end-of-life stage.

4.3.3. Label analysis

The Label Analysis tools allow engineers to see trends in Major and Anomaly Labels in a calendar heatmap format. The Label Analysis has two side-by-side calendars; for each calendar, the user can select the Label Type (Major or Anomaly) and Count Type. There are three Count Types a user can choose from.

- Minimum: Display the labels which had a minimum count for a day
- Maximum: Display the labels which had the maximum count for a day
- Difference: Display the number of times a label has changed during a day

In Fig. 49, we see the Label Analysis Tool, where both the heatmap calendars show Major Heatmap Labels. The calendar on the left shows the Major Labels with the maximum count for each calendar day. The calendar on the right shows the difference in Major Labels for each calendar day. The label analysis tool shows the performance trend over the calendar year as seen in the figure. By selecting various



Apply filter to all Dashbaords

Step 4

Date based filter to show data from either past 12 hours, 1 day or 1 week

Step 3

Fetch data based on selections made in Step 1 and Step 2. Wells on the map will update accordingly

Step 2

Filter to select labels, events or sequences. Filter choices are based on selection of Step 1

Step 1

Filter to select:
1. Major Labels
2. Anomaly Labels
3. Events
4. Seqeunces

Map of Live Wells that can be filtered based on filters

Fig. 44. Dashboard Area 4cx: Live map with filters.

combinations of Label and Count Type, engineers can analyze data based on the AL system's behaviour change.

4.3.4. *Events analysis*

Similar to the Label Analysis tool, the Events Analysis Tool is a calendar-based heatmap which shows the occurrence of the *events* count in a calendar year. In addition, this tool allows the user to understand event occurrence trends and enables engineers to see how often an event can impact a particular AL type. Fig. 50 shows the Event Analysis Heatmap for a PCP well with the Erratic Torque event over a calendar year.

4.3.5. *Sequence Analysis Tool*

The Sequence Analysis Tool shares the same layout as the Event Analysis Tool but with the feature of a heatmap that highlights the regions where selected *sequences* of interest have occurred. This assists engineers in analyzing changes in performance behaviour more effectively. Fig. 51 shows the Sequence Analysis Heatmap for a PCP well with the Flow Drop sequence over a calendar year.

5. Conclusion and future works

This paper presented an innovative time-series analytics approach that autonomously detects various performance states of downhole pumps during CSG production. Our proposed solution provides engineers with real-time notifications that aid in proactively managing AL systems across multiple CSG assets. This is a significant improvement over traditional surveillance methods as it allows engineers to take corrective actions before mechanical failures occur, which leads to a loss in natural gas production.

The detailed methodology, results, and observations presented in this paper demonstrate the efficacy of the proposed approach in detecting various performance states of downhole pumps during CSG production. One of the key advantages of our approach is the ability to analyze time-series data from multiple CSG wells in near real-time. In addition, the analytics application developed in this work allows

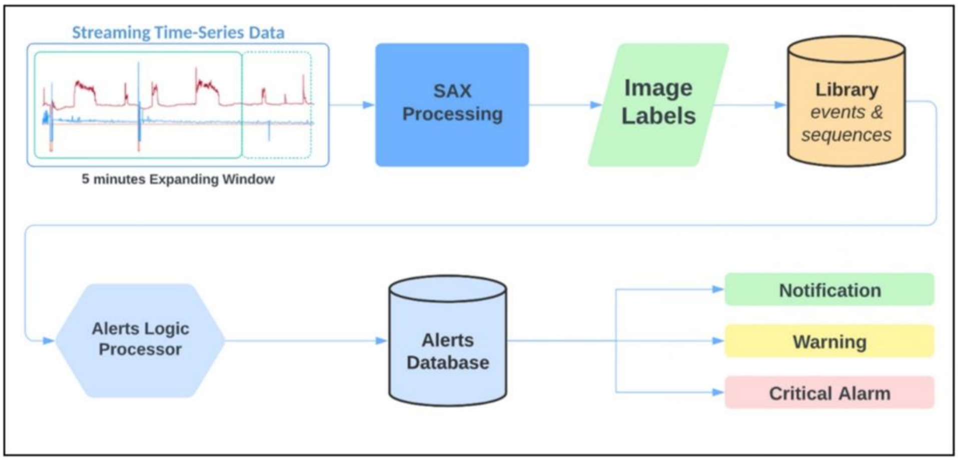
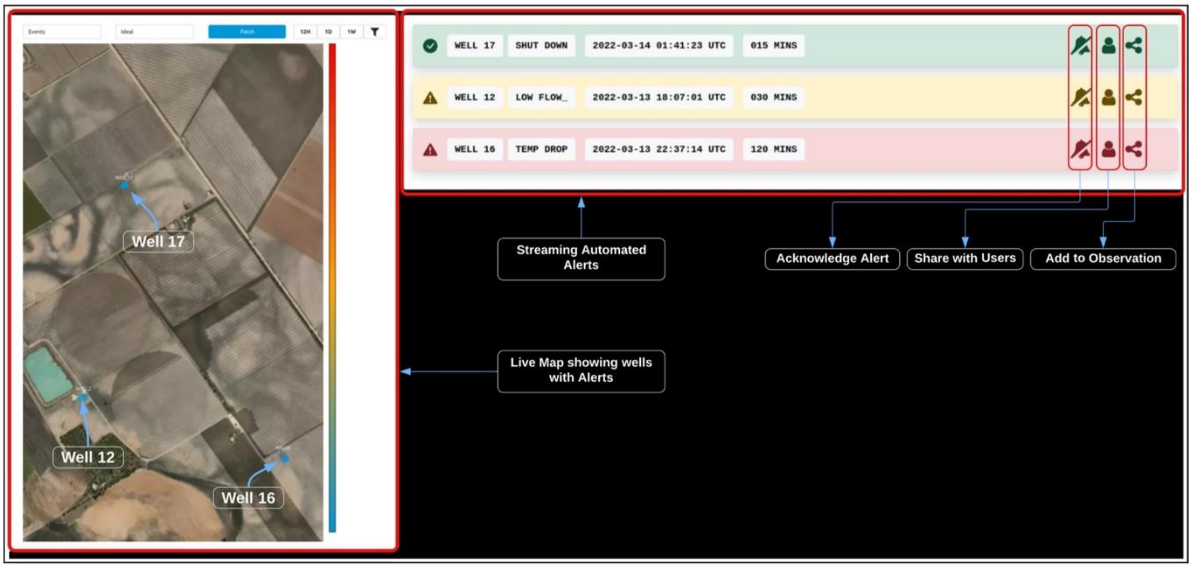


Fig. 45. Automated real-time alert data flow architecture.


WELL 17 SHUT DOWN 2022-03-14 01:41:23 UTC 015 MINS

WELL 12 LOW FLOW_ 2022-03-13 18:07:01 UTC 030 MINS

WELL 16 TEMP DROP 2022-03-13 22:37:14 UTC 120 MINS

Fig. 46. – Alert view in the application dashboard.

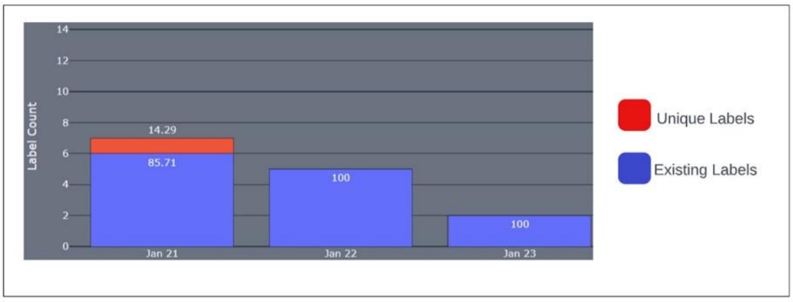


Fig. 47. Unique label tracking.

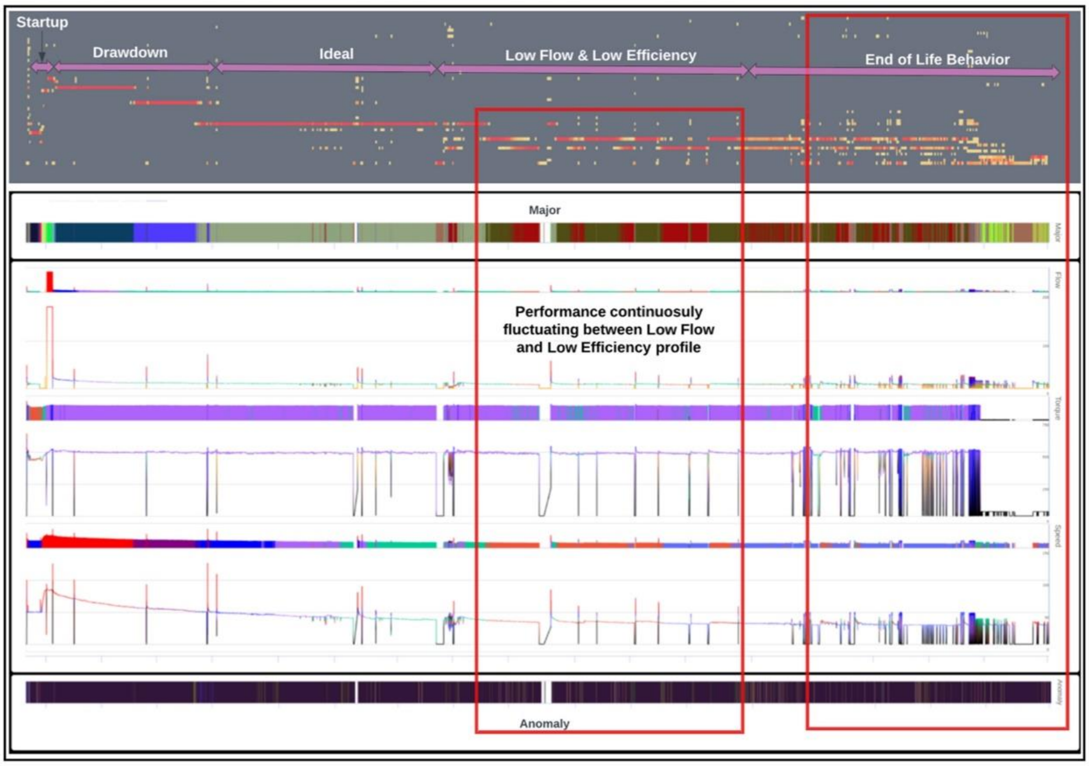


Fig. 48. End-of-life heatmap analysis.

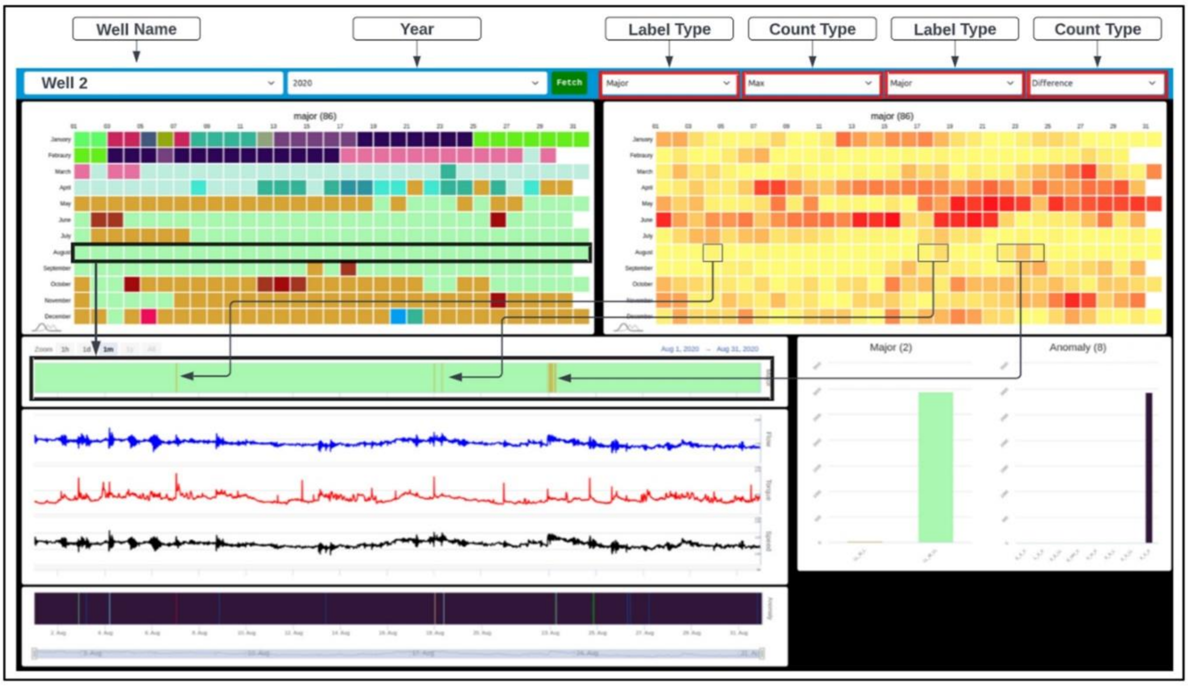


Fig. 49. Label analysis tool.

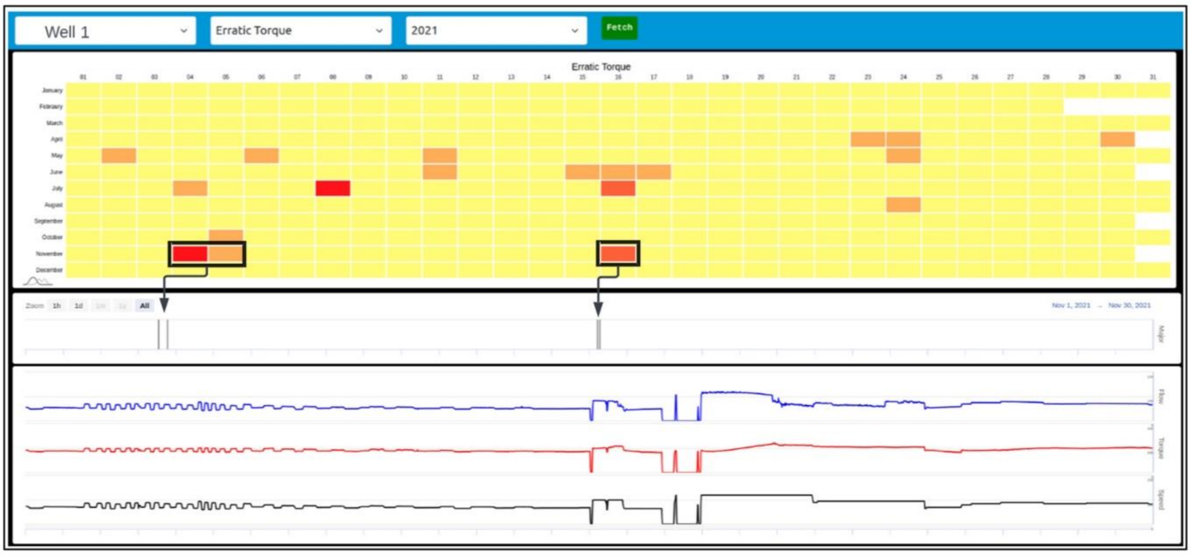


Fig. 50. – Events analysis tool.

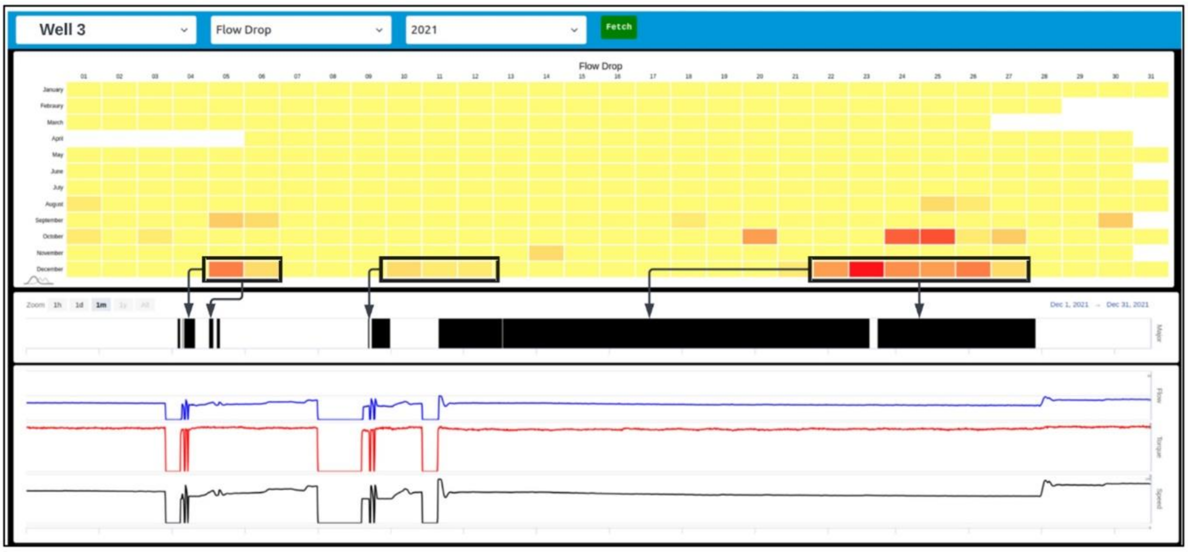


Fig. 51. Sequence analysis tool.

engineers to see real-time results from live wells, providing an automated system to assess any abnormal AL system behaviour and investigate further factors to foresee any impending failure.

The Analysis Tools discussed in this paper have the potential to provide engineers with additional capabilities to conduct detailed investigations of different performance parameters. For example, the Events Analysis Tool can help engineers identify the impact of certain *events* on overall pump life. Furthermore, the *events* and *sequences* library developed during the data annotation stage proved critical for real-time analysis, providing engineers with clear insights into why changes in behaviour occurred during CSG operations.

Although the time-series analytics method was developed using data from CSG-operated wells, we are highly confident that a similar approach can be applied to other rotating equipment operated in Oil and Gas and other industrial applications. For example, this application can be highly beneficial for mud motors used in directional drilling operations within Oil and Gas. Mud motors are essentially PCPs, but rather than having a motor drive the rotor, the force generated by drilling mud is used to drive the rotor, which helps to transfer energy to the drill bit. The mud motor, also commonly known as the power section, is shown in Fig. 52. Also, we are currently discussing with an operator from the Middle East to test the method on AL systems operating in conventional oil and gas reservoirs.

Additionally, to enhance the user experience, we are presently developing a search engine based on time-series data, which will allow users to easily search for specific *events* and *sequences* by typing various states into a search bar. The search engine will then generate a table of wells with the desired *events* and *sequences* listed, making it easier for

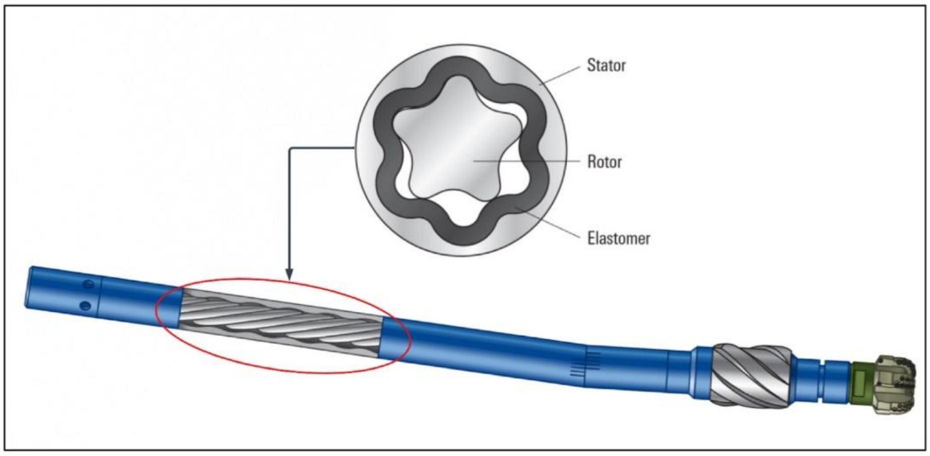


Fig. 52. Various components of the Bottom Hole Assembly along with the mud motor (power section) cut out showing the rotor and stator similar to a PCP. (SLB).

users to locate the information they need quickly and efficiently.

Declaration of competing interest

The authors would like to provide the following *Declaration of Interest* statement.

Two natural gas producers based in Queensland provided financial compensation to the research project's lead author for deploying pre-developed time-series analytics models and the Artificial Lift Systems Analytics Application within their respective cloud platforms.

It is important to note that the lead author had already completed the required research and development work for the time-series models, methods, Data Annotation Tool, and the Artificial Lift Systems Analytics Application before commencing work with these companies.

The provided compensation was used to customize the data ingestion pipeline and improve the Analytics Application dashboards based on each company's specific requirements. However, these customizations are not discussed in the paper, and the necessary data has been anonymized due to the Confidentiality Agreement with the companies. Furthermore, as mentioned in the manuscript, the companies assisted with the data annotation process, which helped create the events & sequences database used for Artificial System performance tracking.

The co-authors of this manuscript have no conflicts of interest to disclose. Furthermore, all authors have approved the final version of the manuscript and agree to its submission to the Engineering Applications of Artificial Intelligence Journal.

Data availability

The authors do not have permission to share data.

References

Abdalla, R., et al., 2022. Machine learning approach for predictive maintenance of the electrical submersible pumps (ESPs). ACS Omega 7 (21), 17641–17651.

Abdelaziz, M., Lastra, R., Xiao, J.J., 2017. ESP data analytics: predicting failures for improved production performance. In: Abu Dhabi International Petroleum Exhibition & Conference. Society of Petroleum Engineers, Abu Dhabi, UAE, p. 17.

Al-Ballam, S., Karami, H., Devegowda, D., 2023. A hybrid physical and machine learning model to diagnose failures in electrical submersible pumps. In: SPE/IADC Middle East Drilling Technology Conference and Exhibition.

Alamu, O.A., et al., 2020. ESP data analytics: use of deep autoencoders for intelligent surveillance of electric submersible pumps. In: Offshore Technology Conference.

Ambade, A., et al., 2021. Electrical submersible pump prognostics and health monitoring using machine learning and natural language processing. In: SPE Symposium: Artificial Intelligence - towards a Resilient and Efficient Energy Industry.

Andrade Marin, A., et al., 2021. Real time implementation of ESP predictive analytics - towards value realization from data science. In: Abu Dhabi International Petroleum Exhibition & Conference.

Awaid, A., et al., 2014. ESP well surveillance using pattern recognition analysis, oil wells, Petroleum development Oman. In: International Petroleum Technology Conference. International Petroleum Technology Conference, Doha, Qatar, p. 22.

Brasil, J., et al., 2023. Diagnosis of operating conditions of the electrical submersible pump via machine learning. Sensors 23 (1), 279.

Braverman, V., 2016. Sliding window algorithms. In: Kao, M.-Y. (Ed.), Encyclopedia of Algorithms. Springer, New York: New York, NY, pp. 2006–2011.

Camilleri, L., 2013. System, Method, and Computer Readable Medium for Calculating Well Flow Rates Produced with Electrical Submersible Pumps. USPTO, United States. Sensia LLC.

Cardona, L.E., Vivas Sanchez, P.J., Joya, B., 2023. Failure prediction methodology for ESP and operational behavior. In: SPE Latin American and Caribbean Petroleum Engineering Conference.

Commonwealth of Australia 2014, 2014. Coal Seam Gas Extraction: Modelling Groundwater Impacts. Department of the Environment.

Iranzi, J., et al., 2022. A nodal analysis based monitoring of an electric submersible pump operation in multiphase flow. Appl. Sci. 12 (6), 2825.

Keogh, E., Lin, J., 2005. Clustering of time-series subsequences is meaningless: implications for previous and future research. Knowl. Inf. Syst. 8 (2), 154–177.

Keogh, E., Lin, J., Fu, A., Hot, S.A.X., 2005. Efficiently finding the most unusual time series subsequence. In: Fifth IEEE International Conference on Data Mining. ICDM'05).

Knafl, M., et al., 2013. Diagnosing PCP failure characteristics using exception based surveillance in CSG. In: SPE Progressing Cavity Pumps Conference.

Kumar, N., et al., 2005. Time-series Bitmaps: a Practical Visualization Tool for Working with Large Time Series Databases.

Lin, J., et al., 2003. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms, pp. 2–11.

Lin, J., et al., 2007. Experiencing SAX: a novel symbolic representation of time series. Data Min. Knowl. Discov. 15 (2), 107–144.

Matthews, C.M., et al., 2007. Petroleum engineering handbook. In: Production Operations Engineering. Society of Petroleum Engineers.

Ocanto, L., Rojas, A., 2001. Artificial-lift systems pattern recognition using neural networks. In: SPE Latin American and Caribbean Petroleum Engineering Conference. Society of Petroleum Engineers, Buenos Aires, Argentina, p. 6.

Queensland borehole series metadata record. Available from: https://www.data.qld.gov.au/dataset/queensland-borehole-series.

Rajora, A., et al., 2019. Deviated pad wells in surat: journey so far. In: SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology Conference.

Rathnayake, S.I., Firouzi, M., 2021. Statistical process control for early detection of progressive cavity pump failures in vertical unconventional gas wells. In: SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology Conference.

Rathnayake, S., Rajora, A., Firouzi, M., 2022. A machine learning-based predictive model for real-time monitoring of flowing bottom-hole pressure of gas wells. Fuel 317, 123524.

Saghir, F., Perdomo, M.G., Behrenbruch, P., 2019a. Machine learning for progressive cavity pump performance analysis: a coal seam gas case study. In: SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology Conference. Society of Petroleum Engineers, Brisbane, Australia.

Saghir, F., Gonzalez Perdomo, M.E., Behrenbruch, P., 2019b. Converting time series data into images: an innovative approach to detect abnormal behavior of progressive cavity pumps deployed in coal seam gas wells. In: SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers, Calgary, Alberta, Canada, p. 14.

Saghir, F., Gonzalez Perdomo, M.E., Behrenbruch, P., 2019c. Application of exploratory data analytics EDA in coal seam gas wells with progressive cavity pumps PCPs. In:

SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition. Society of Petroleum Engineers, Bali, Indonesia, p. 10.

Saghir, F., Gonzalez Perdomo, M.E., Behrenbruch, P., 2020. Application of machine learning methods to assess progressive cavity pumps (PCPs) performance in coal seam gas (CSG) wells. The APPEA Journal 60 (1), 197–214.

Saghir, F., Gonzalez Perdomo, M.E., Behrenbruch, P., 2022. Application of streaming analytics for Artificial Lift systems: a human-in-the-loop approach for analysing clustered time-series data from progressive cavity pumps. Neural Comput. Appl. 35 (2), 1247–1277.

Saghir, F., Gonzalez Perdomo, M.E., Behrenbruch, P., 2023. Application of streaming analytics for Artificial Lift systems: a human-in-the-loop approach for analysing clustered time-series data from progressive cavity pumps. Neural Comput. Appl. 35 (2), 1247–1277.

Sharma, A., Songchitruksa, P., Sinha, R.R., 2022. Integrating domain knowledge with machine learning to optimize electrical submersible pump performance. In: SPE Canadian Energy Technology Conference.

Silvia, S., et al., 2023. Case study: predicting electrical submersible pump failures using artificial intelligence and physics-based hybrid models. In: SPE Symposium Leveraging Artificial Intelligence to Shape the Future of the Energy Industry.

SLB. PowerPak steerable motors. Available from: https://www.slb.com/drilling/bottomhole-assemblies/directional-drilling/powerpak-steerable-motors.

Takacs, G., 2018. Chapter 4 - use of ESP equipment in special conditions. In: Takacs, G. (Ed.), Electrical Submersible Pumps Manual, second ed. Gulf Professional Publishing, pp. 153–240.

Tan, C., et al., 2021. The health index prediction model and application of PCP in CBM wells based on deep learning. Geofluids 2021, 6641395.

Thornhill, D.G., Zhu, D., 2009. Fuzzy analysis of ESP system performance. In: SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers, New Orleans, Louisiana, p. 7.

9. Conclusions and Recommendations

9.1. Conclusion

In conclusion, this thesis presents a groundbreaking approach to analyzing time-series data using SAX-based Heatmap images for the purpose of ALS performance analysis in CSG production. The results demonstrate the effectiveness of this methodology in providing improved visualization of real-time trends and automating the detection of abnormal events in the context of ALS operations. By allowing engineers to focus on wells requiring attention and proactively managing ALS systems, this approach significantly enhances the efficiency of CSG asset management and minimizes natural gas production losses due to mechanical failures.

The detailed methodology, results, and observations presented in this research demonstrate the efficacy of the proposed approach in detecting various performance states of ALS systems during CSG production. One key advantage of the developed approach is the ability to analyze time-series data from multiple CSG wells in near real-time. In addition, the analytics application developed in this work allows engineers to see real-time results from live wells, providing an automated system to assess any abnormal ALS system behavior and investigate further factors to foresee any impending failure.

Moreover, the analysis tools discussed in this research have the potential to provide engineers with additional capabilities to conduct detailed investigations of different performance parameters. For example, the Events Analysis Tool can help engineers identify the impact of certain events on overall pump life. Furthermore, the events and sequences library developed during the data annotation stage proved critical for real-time analysis, providing engineers with clear insights into why changes in behavior occurred during CSG operations.

Chapters 7 and 8 showcase the practical application of the research work by two CSG operators who gained valuable insights into their ALS operations. With the help of real-time alerts, these operators can now manage a large number of wells with ease by identifying exceptions that require immediate attention. Additionally, the analytics application developed as part of this research work has the potential to evolve into a fully autonomous control system, where parameters such as pump speed can be adjusted without human intervention.

9.2. Summary of Findings

The culmination of this research reveals remarkable findings, contributing to the advancement of ALS performance analysis in CSG production. The following key outcomes underscore the significance of the developed methodology:

a. Innovative Application of SAX

The study innovatively applies SAX to transform complex multivariate time-series data into visual performance heatmaps. This novel approach serves to streamline the labeling process for petroleum and surveillance engineers, providing a comprehensive and intuitive representation of ALS dynamics.

b. Efficiency and Accuracy Through Labeling

The assignment of labels, encompassing events and sequences, to time-series images emerges as a pivotal enhancement. This labeling strategy significantly augments the efficiency and accuracy in detecting abnormal ALS performance. The structured labeling schema facilitates swift and effective decision-making processes.

c. Confidence in Result Accuracy

Petroleum and surveillance engineers derive heightened confidence in the accuracy of obtained results through the systematic labeling process. This meticulous organization of pertinent data ensures a foundation for informed decision-making, fostering a robust analytical framework.

d. Analytics Platform for Efficient Well Management

The research presents a user-friendly analytics platform that demonstrates the effectiveness of streaming analytics. This platform empowers CSG operators to adeptly manage a substantial number of wells. By leveraging real-time data, the platform contributes to the efficient monitoring and optimization of ALS performance across diverse operational scenarios.

e. Collaborative Approach for Early Detection

The collaborative approach advocated in this study empowers engineers to create events and sequences. This collaborative effort is instrumental in the early detection of abnormal ALS performance, allowing for proactive corrective actions and the prevention of downtime.

f. Real-time ALS Performance Analysis Framework

The developed real-time ALS performance analysis framework advocates a manage-by-exception approach. This strategic framework streamlines ALS management practices within the CSG industry. By focusing attention on wells requiring intervention, the framework optimizes operational efficiency and minimizes production losses due to mechanical failures.

9.3. Summary of Results

a. Real-time identification of ALS-related issues

Starting in 2021, ALSAA was implemented with two (2) CSG operators who collectively operated close to a thousand wells, and the live alerts enabled these operators to monitor these large number of CSG wells by exception. The analytics tool proved to be efficient in detecting issues with various ALS systems, significantly reducing the time needed to identify and address problems. The tool provided engineers with real-time alerts and valuable insights into specific performance behaviors, enabling them to quickly and accurately address any issues that arose. This mitigated various unnecessary shutdowns, and possibly prevented ALS failures due to undetected abnormal behavior.

b. Actionable Alerts

Operators have identified nine (9) actionable alerts for PCPs and eight (8) actionable alerts for ESPs/ESPCPs. These actionable alerts have enabled the operators to carry out necessary interventions and performance analysis to improve overall production from CSG wells. Furthermore, the data annotation tool provided as part of ALSAA, provides operators with the flexibility to modify and add alerts to meet their specific operational needs.

c. Labelled Time-series Data Repository

As the time-series data was automatically labelled based on the SAX-based heatmap clusters, this provided operators with a host of advantages, including efficient data retrieval and the development of additional analytics applications. In Chapter 8, it is demonstrated how labelled time-series data facilitated the development of five (5) insightful applications. These applications provided operators with valuable insights into PCP end-of-life analysis, tracking and identification of unique labels, and deep-dive analysis tools to understand the patterns of labels, sequences, and events. This immensely improved task efficiency, facilitating root-cause analysis for certain PCP performance issues.

d. Improved ALS workover management

The end-of-life heatmap analysis tool presented in Chapter 8 assists operators with planning pump changeover activities well in advance versus waiting for pumps to fail. This insight helps companies to streamline workover activities, and significantly minimize production downtime. Furthermore, the tool enables operators to efficiently manage pump inventory and assist with supply chain related decisions. By improving ALS workover management, companies reduce operational costs and maintain undisrupted gas production levels, leading to improved profits for CSG operators.

9.4. Recommendations

While the time-series analytics method was developed using data from CSG-operated wells, a similar approach can be applied to other rotating equipment operated in the Oil and Gas industry and other industrial applications. This research paves the way for transformative advancements in real-time monitoring and predictive maintenance across multiple sectors, potentially enhancing operational efficiency and reducing downtime.

Furthermore, the study highlights the versatility of SAX-based performance heatmap analysis, as it can be applied to a wide range of rotating equipment in the Oil and Gas industry and other industrial applications where ALS systems are employed. The innovative time-series analytics tools and event analysis capabilities presented in this research offer engineers valuable insights into performance parameters and the impact of events, making it a valuable asset for various industries beyond CSG production.

The findings showcased in this research not only have implications for the Australian context but also hold the potential to foster valuable collaborations with oil and gas operators in countries like India and the USA, where CSG operations are prevalent.

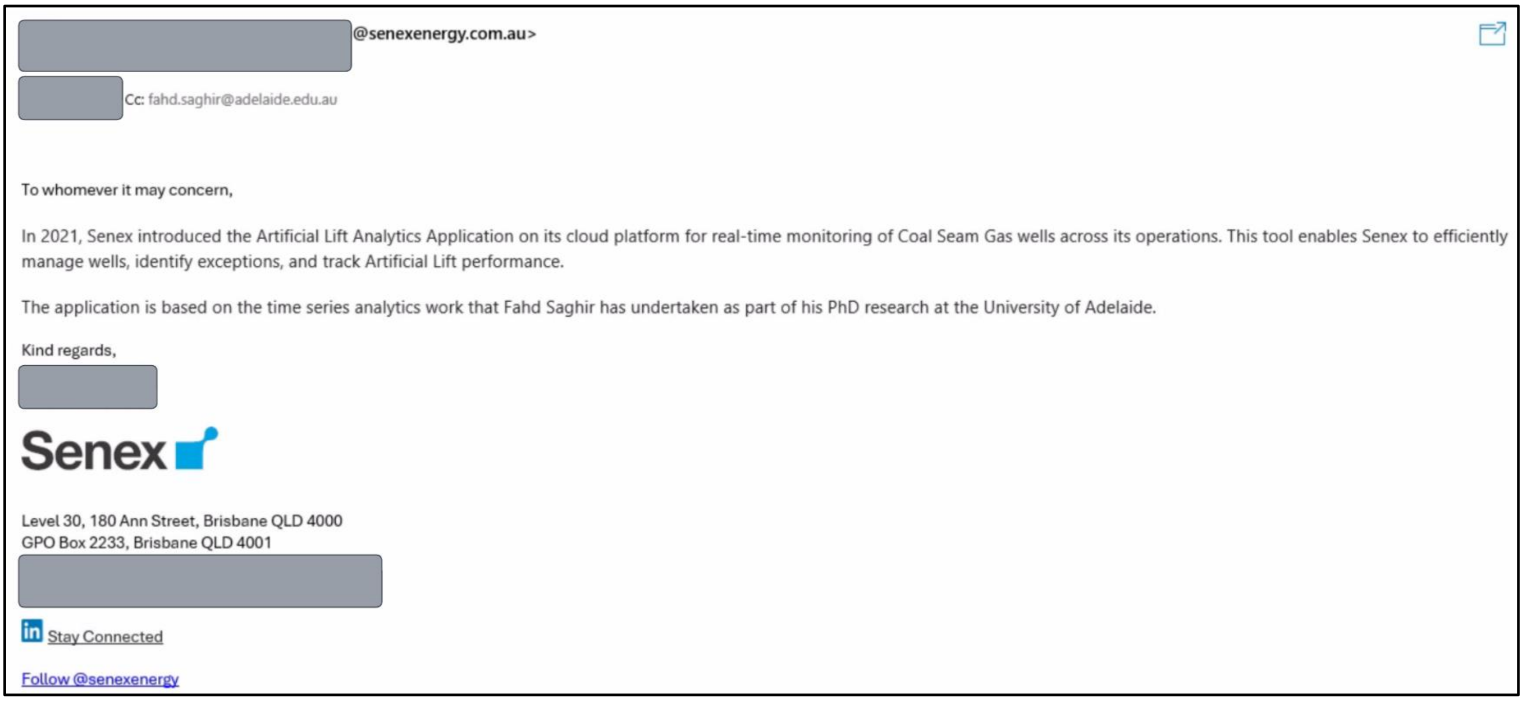
Expanding the scope beyond CSG operations, potential collaborators for advancing the domain of this research extend to any operator heavily reliant on ALS for hydrocarbon production. The inherent adaptability of the presented methodology not only invites collaboration but also signifies its potential for broader refinement, especially in the context of global applicability. This adaptability positions the research as a versatile framework that can be tailored to address the diverse needs and operational nuances of hydrocarbon producers worldwide, thereby fostering collaborative efforts and advancements beyond the realm of CSG operations.

One notable avenue for further improvement is the prospect of leveraging labelled datasets to pave the way for advanced time-series-based search engines. Such engines would revolutionize data retrieval processes by allowing users to access specific information through intuitive search prompts, significantly enhancing the efficiency of data retrieval and analysis. This development could serve as a technological breakthrough, providing a versatile tool that transcends to any time-series based application.

A particularly compelling avenue for further exploration in this research would be the development of a fully autonomous control system. This system would possess the capability to automatically regulate parameters, thereby adjusting the performance of mechanical equipment in real-time to optimize production while concurrently extending equipment run life. The foundation for such an autonomous control system could be established by leveraging insights gained from engineers' responses to alerts generated by the current analytics tool. By learning from and emulating the decision-making processes of engineers in response to specific alerts, this autonomous system could offer a transformative approach to enhancing operational efficiency and prolonging the life of mechanical equipment in dynamic environments.

APPENDIX

Letter from CSG operator acknowledging the use of Artificial Lift Analytics Application



[redacted]@senexenergy.com.au>

[redacted] Cc: fahd.saghir@adelaide.edu.au

To whomever it may concern,

In 2021, Senex introduced the Artificial Lift Analytics Application on its cloud platform for real-time monitoring of Coal Seam Gas wells across its operations. This tool enables Senex to efficiently manage wells, identify exceptions, and track Artificial Lift performance.

The application is based on the time series analytics work that Fahd Saghir has undertaken as part of his PhD research at the University of Adelaide.

Kind regards,

[redacted]

Senex

Level 30, 180 Ann Street, Brisbane QLD 4000
GPO Box 2233, Brisbane QLD 4001

[redacted]

in Stay Connected

Follow @senexenergy