PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

UNIVERSITY OF KASDI MERBAH OUARGLA

Faculty of New Technologies of Information and Communication

PhD Dissertation

Submitted in fulfillment of the requirements for Doctorate degree in Electrical Engineering

Option: **Automation and Systems Engineering**

Presented by **Aymen Djamel Eddine HARROUZ**

Theme

Enhanced Neural Network Architectures for Data-Scarce Environments and Multi-Parameter Prediction in Oil and Gas Operations

Jury:

President	Bilal BENARABI	MCA	University of Kasdi Merbah Ouargla
Director	Redouane KAFI	MCA	University of Kasdi Merbah Ouargla
Co-Director	Houari TOUBAKH	MCA	University of Kasdi Merbah Ouargla
Examiner	Abdelhai LATI	MCA	University of Kasdi Merbah Ouargla
Examiner	Zakaria LAMMOUCHI	MCA	University of El Oued
Examiner	Bilal MAGNI	MCA	University of Annaba

February 2024

Acknowledgements

Praise be to the Almighty God who has given me faith, courage, and patience to carry out this work.

I thank Dr. Redouane KAFI from Kasdi Merbah university for supervising me, for his orientation, availability, listening, and patience during the realization of this job.

I want to express my deep gratitude to my co-supervisor Dr. Houari TOUBAKH from Kasdi Merbah university, for the confidence he has placed in me, through his presence always with me, by his direction, his modesty, his advice, and constructive remarks for the good progress of this work.

I would like to thank everyone who helped me to improve my work. and who gave me any remark that helped me to perfect this manuscript.

I express my deep gratitude to my parents, and my brother for their encouragement and prayers that allowed me to achieve this modest job. I am very grateful for the confidence they have placed in me.

Finally, I express my gratitude to all those who have contributed in one way or another to the development of this work.

Dedication

I dedicate this work to my parents:

May they find here the testimony of my deep gratitude and acknowledgment

To all those who have helped me - directly or indirectly - and those who shared with me the emotional moments during the accomplishment of this work and who warmly supported and encouraged me throughout my journey.

To all my friends who have always encouraged me, and to whom I wish more success.

Thanks!

Aymen Djamel Eddine HARROUZ

Abstract

Neural networks are a crucial component of modern artificial intelligence, demonstrating impressive abilities in understanding complex patterns and relationships in data. However, applying neural networks in real industrial systems, such as Oil and Gas operations, is challenging due to limited historical data availability, especially for new machines, and the high cost of obtaining or producing data. As a result, there is a scarcity of public data for the research community. This PhD thesis proposes innovative neural network architectures tailored to address the critical challenges in this field. To solve the issue of low prediction accuracy in predicting the health state of tools, a novel neural network architecture is proposed to forecast the Remaining Useful Life when limited training data is provided. a feedback mechanism is incorporated into an artificial neural network in a novel manner, using the values of the output layer neurons as inputs. These inputs are utilized as features to generate precise predictions. To validate the effectiveness of the approach, real dataset from oil and gas wells during production is used, this study focuses on a sub dataset of a sub-surface safety valve tool. Additionally, a custom neural network architecture is proposed to create a data-driven digital twin based on multi-target regression to mitigate the time delay that impacts decision-making for drillers during directional drilling operations. The architecture combines Long-short Term Memory and Multi-Layer Perception branches in a single neural network to forecast and predict important drilling parameters, such as inclination and rate of penetration. To validate this approach a real data collected during a directional drilling operation is used. Furthermore, an incremental learning framework is implemented to simulate the performance of the architectures in real-time, where data is continuously received and the regression models are updated concurrently. The proposed architectures demonstrate superior results compared to existing works in the field. The research conducted in this thesis aims to extend the capabilities of neural network models, uncovering their potential in solving complex problems while contributing to the evolving field of intelligent systems.

Keywords: *Neural Networks, Fault Prognosis, Digital Twin, Multi-Target Regression, Oil and Gas Operations.*

Résumé

Les réseaux de neurones sont un composant crucial de l'intelligence artificielle moderne, démontrant des capacités impressionnantes à comprendre des schémas complexes et des relations dans les données. Cependant, l'application des réseaux de neurones dans des systèmes industriels réels, tels que les opérations pétrolières et gazières, est difficile en raison de la disponibilité limitée des données historiques, en particulier pour les nouvelles machines, et du coût élevé d'obtention ou de production des données. En conséquence, il existe une rareté de données publiques pour la communauté de recherche. Cette thèse de doctorat propose des architectures de réseaux de neurones innovantes adaptées pour relever les défis critiques dans ce domaine. Pour résoudre le problème de faible précision de prédiction de l'état de santé des outils, une nouvelle architecture de réseau de neurones est proposée pour prédire la durée de vie restante des systèmes avec des données d'entraînement limitées. L'approche proposée intègre un mécanisme de rétroaction dans un réseau de neurones artificiel, en utilisant les valeurs des neurones de la couche de sortie comme entrées. Ces entrées sont utilisées comme caractéristiques pour générer des prédictions précises. L'efficacité de cette approche est validée à l'aide de données réelles provenant de puits de pétrole et de gaz en production, spécifiquement pour prédire l'espérance de vie d'un outil de vanne de sécurité en sous-surface. De plus, une architecture de réseau de neurones personnalisée est introduite pour créer un Digital Twin basé sur une régression multi-cible afin de réduire le délai temporel impactant la prise de décision pour les foreurs pendant les opérations de forage directionnel. L'architecture combine des branches de Mémoire à Court et Long Terme et de Perceptron Multi-Couches dans un seul réseau de neurones pour prévoir et prédire des paramètres de forage importants, tels que l'inclinaison et la taux de pénétration. Cette approche est validée à l'aide de données réelles collectées lors d'une opération de forage directionnel. De plus, une structure d'apprentissage incrémentiel est implémenté pour simuler les performances des architectures en temps réel, où les données sont continuellement reçues et les modèles de régression sont mis à jour simultanément. Les architectures proposées démontrent des résultats supérieurs par rapport aux travaux existants dans état de l'art. La recherche menée dans cette thèse vise à étendre les capacités des modèles de réseaux de neurones, révélant leur potentiel pour résoudre des problèmes complexes tout en contribuant au domaine évolutif des systèmes intelligents.

Mots clés : *Réseaux de neurones, Prognostic des défauts, Digital Twin, Régression multi-cible, Opérations pétrolières et gazières.*

Contents

List of Figures

List of Tables

List of Abbreviations

AI:	Artificial Intelligence
ML:	Machine Learning
CRISP-DM:	Cross-Industry Standard Process for Data Mining
IoT:	Internet of Things
AEM:	Abnormal Event Management
NPT:	Non-Productive Time
CSV:	Comma Separated Values
WITSML:	Well-site Information Transfer Standard Markup Language
DT:	Digital Twin
CBM:	Condition-Based Maintenance
PdM:	Predictive Maintenance
PHM:	Prognosis and Health Management
HI:	Health Indicator
RUL:	Remaining Useful Life
PRUL:	Predicted Remaining Useful Life
OPRUL:	Optimal Predicted Remaining Useful Life
ToF:	Time of Failure
TTF:	Time-To-Failure
MD:	Measured Depth
MWD:	Measurement While Drilling
BHA:	Bottom Hole Assembly
SSSV:	Sub-Surface Safety Valve

BSW:	Basic Sediment and Water
PCK:	Production Choke
PDG:	Permanent Downhole Gauge
CKGL:	Gas Lift Choke
TPT:	Temperature and Pressure Transducer
P-TPT:	Pressure at the Temperature and Pressure Transducer
T-TPT:	Temperature at the Temperature and Pressure Transducer
ROP:	Rate of Penetration
NN:	Neural Networks
ANN:	Artificial Neural Network
MLP:	Multi-Layer Perception
RNN:	Recurrent Neural Networks
LSTM:	Long Short-Term Memory
CNN:	Convolutional Neural Network
GAN:	Generative Adversarial Network
DCNN-Bi-LSTM:	Deep Convolutional Neural Network-Bidirectional LSTM
GRU:	Gated Recurrent Unit
FFBPN:	Feed-Forward Back Propagation Network
SAAN:	Sequence Adaptation Adversarial Network
TLNN:	Tri-Layered Neural Network
BLNN:	Bi-Layered Neural Network
OLNN:	One Layer Neural Network
SVM:	Support Vector Machine
LSVM:	Linear Support Vector Machine
QSVM:	Quadratic Support Vector Machine
SVMR:	Support Vector Machine Regression
KNN:	K-nearest Neighbor

FRN: Fixed Radius Neighbor

DT: Decision Tree

SA: Self-Attention

MAE: Mean Absolute Error

MSE: Mean Squared Error

AEE: Absolute Estimation Error

MAAEE: Moving Average Absolute Estimation Error

ReLU: Rectified Linear Unit

NaN: Not-a-Number

Chapter 1

General Introduction

The field of artificial intelligence (AI) has experienced substantial growth in both academic and industrial contexts, with a wide variety of methodologies and algorithms being utilized. This technology holds the potential to deliver autonomous vehicles, automated language translations, and the substitution of human labor in diverse occupational domains. Although AI has gained widespread recognition among the general public, the operational aspect of this field, known as machine learning, is the primary focus of investigation [4].

Machine learning (ML) has a precise delineation [5]. Broadly, it refers to an algorithm that can forecast the outputs of novel and unfamiliar inputs by leveraging prior experiences in the form of two pairs, input and output, the algorithm can be later trained on these pairs. A rudimentary illustration of this concept is exemplified by the ease with which data can be fitted with linear regression model with a simple mouse click in Excel. Conversely, intricate algorithms like reinforcement learning exhibit the capability to undertake intricate endeavors such as engaging in computer gaming activities [6].

There is a constant development of various machine learning algorithms that aim to solve different types of problems, including data mapping, image analysis, translations, and data generation. In the context of oil and gas well construction operations, a range of machine learning algorithms are utilized, with one promising architecture being Neural Networks. Specifically, for time-dependent

applications, an architecture known as Recurrent Neural Networks (RNN) is employed to capture temporal information by considering not only the current state but also past states. Despite being widely used in commercial applications such as voice recognition, music composition, and speech synthesis, the implementation of RNNs in oil and gas well construction applications is relatively limited [7].

This dissertation examines the application of neural networks in enhancing oil and gas well construction operations, as well as data processing and model evaluation. The research aims to determine the optimal utilization of this technology, assess its potential and limitations, gain a deeper understanding of its functioning, and identify appropriate methods for its implementation in practical contexts.

1.1 Motivation and Objectives

Machine failure in the oil and gas industry significantly affects the financial aspects of well construction operations. In order to mitigate maintenance costs, predictive maintenance is the suggested strategy. To achieve a cost effective maintenance, determining the Remaining Useful Life (RUL) and implementing it as a metric is crucial. The term "Remaining Useful Life" (RUL) pertains to the period remaining until the useful lifespan of a particular tool or machine concludes, denoting the time left before the occurrence of failure in said machine or tool. [8].

From a machine learning standpoint, limited data availability is significant obstacle in predicting and estimating Remaining Useful Life (RUL). This problem is apparent in the oil and gas industry. The accessibility of data for the general public and researchers from service companies is limited. Even when datasets related to oil and gas operations are publicly accessible, they often consist of a limited number of samples. The accuracy of predictions is negatively impacted by this lack of data, therefore Prognosis and Health Management (PHM) systems becomes unreliable and ineffective. [9].

Additionally, accurate prediction of multiple parameters during well construction operations is of utmost importance in order to minimize costs associated

with the operations. Precisely predicting the Rate of Penetration (ROP) and Inclination during drilling activities is a critical aspect of well construction as it presents opportunities for cost reduction and improves operational efficiency, including planning, decision-making, and wellbore stability. Additionally, accurate ROP predictions play a significant role in ensuring the safety of employees and minimizing any negative environmental impacts. Consequently, the success and economic viability of drilling operations heavily rely on precise ROP predictions. Similarly, in the case of horizontally drilled wells, accurate prediction and control of the Inclination are crucial for optimizing well placement, reservoir contact, and productivity. Furthermore, these considerations are essential in guaranteeing drilling operations' financial success while prioritizing safety and environmental concerns [10].

This research has two main objectives. The first objective is to develop an innovative machine learning system in the context of predictive maintenance that can accurately predict the Remaining Useful Life (RUL) of machines . This system aims to overcome challenges related to limited data availability and low accuracy when systems have insufficient data, with the ultimate goal of reducing machine failures in production lines. The second objective is to create a data-driven digital twin that uses multi-target regression to predict and forecast the Rate of Penetration (ROP) and Inclination, this will improve the decision-making of the drillers during a directional drilling operation. This digital twin will incorporate incremental learning to simulate continuous prediction-while-drilling scenarios.

1.2 Methodology

The Cross-industry standard process for data mining (CRISP-DM) is respected in this dissertation research [11], which is widely recognized as the predominant methodology for data science projects. CRISP-DM comprises several key phases, namely:

- Business understanding.

- Data Understanding.
- Data Preparation.
- Modeling.
- Evaluation.
- Deployment.

All stages of the process, except for the deployment phase, were successfully carried out. The business understanding is directly tied to the industry's requirements, as evidenced by the case studies presented in sections 3.2 and 4.2 and references [9, 10]. The case studies address the issue of the prognosis of well completion tools failure as well as the problem of delayed sensor readings in directional drilling, specifically when using the bent sub-assembly, as there is a strong business imperative in positioning the direction sensor in the assembly, as it needs to be in the closest position to the drilling bit. This can be achieved through mechanical means or by utilizing machine learning techniques. Both of these practical challenges are common in oil and gas well construction operations.

The process of understanding data and data preparations, including data collection, data description, uncovering patterns, and evaluating the quality of the data, is conducted in publications [3, 9, 10], which is also mentioned in this dissertation. The data used in this research is real and sourced from two distinct datasets: the '3W dataset' designed for predictive maintenance in the oil and gas industry [1], and the Volve dataset [12], which includes various types of information related to drilling, reservoir modeling, geoscience and production. A part from the research proposed in this dissertation specifically focuses on the drilling data within the Volve dataset.

The modeling and evaluation stages of the CRISP-DM process are the main focus of research publications [9, 10]. Two custom architectures of neural networks are developed: in a predictive maintenance scenario of well-completion tools to mitigate the problem of insufficient data to predict the RUL of machines, the

architecture can be utilized in different applications where only limited data acquisition is possible. Additionally, a Multi-target recurrent neural network (RNN) architecture is developed to predict and forecast multiple drilling parameters simultaneously, the architecture is validated on data acquired during a real directional drilling operation.

1.3 Contributions

The main Contributions of this research are listed below:

- A novel self-adaptive neural network is developed to estimate the RUL of systems with insufficient a priori degradation sequences. The effectiveness of the suggested corrective feedback system is confirmed through the utilization of a dataset that contains a restricted amount of previous real sensor measurements obtained before the malfunction of oil and gas well completion tools, the approach is described in section 3.3 and in [9, 3].
- A data-driven digital twin based on multi-target regression is developed using a customized branched architecture combining Long Short-Term Memory (LSTM) and Multi-Layer Perception (MLP) to predict and forecast the ROP and Inclination during a directional drilling operation while incorporating incremental learning to mimic the continuous prediction-while-drilling scenario, the approach is described in section 4.3 and in [10]

1.4 Scientific Results

The publications [9, 3] explore the proposed approach of overcoming the prognosis challenge of predicting the remaining useful life (RUL) of systems with insufficient degradation sequences accurately. This is achieved through the utilization of a self-adaptive neural network with a corrective feedback mechanism. Health Indicators (HIs) are constructed to offer a more comprehensive understanding of the wellbore

safety closure tool known as the Sub-Surface Safety Valve (SSSV) in emergencies within the oil and gas industry. The successful estimation of RUL with a high level of accuracy showcases the leverage of implementing the proposed prediction approach.

A data-driven digital twin solution is developed in publication [10] to forecast inclination and rate of penetration during directional well-drilling operations. The solution leverages the benefits of multi-target regression to predict multiple well parameters simultaneously by combining branches of LSTM and MLP networks in a single architecture. To simulate real-time data acquisition, an incremental learning scheme is employed where portions of data are continuously fed to the digital twin. The data utilized in this research is obtained from an actual directional drilling operation.

1.5 Manuscript Organization

The thesis manuscript is organized as follows:

Chapter 2 – State of the Art of Fault Prognosis and Parameter Prediction in the Oil and Gas Sector using Neural Networks. This chapter provides a comprehensive overview of the current state of remaining useful life prediction techniques, as well as the utilization of machine learning and neural networks in the oil and gas industry to optimize costs, minimize environmental impacts, and ensure operational safety. The chapter discusses multiple applications of different neural network types within the oil and gas sector.

Chapter 3 - Fault Prognostic Approach with Limited Degradation Sequences. This chapter introduces a methodology that utilizes data-driven techniques to achieve fault prognostics in situations where there is a scarcity of historical degradation sequences. The suggested approach incorporates a feedback mechanism into the architecture of an Artificial Neural Network making it self-adaptive, leading to accurate results due to the updated feature space. to evaluate the prognostic system, real data acquired during oil and gas production

is used, specifically to predict the Time of Failure (ToF) of a Subsurface Safety Valve (SSSV) system. In order to simulate the operational conditions, only a subset of the data is provided to the prediction model, the exact ToF is determined by using a curve-fitting function intersection with the x-axis. Advantageous impact is demonstrated by the proposed methodology results in enhancing prediction accuracy for systems with limited data, as compared to the conventional approaches.

Chapter 4 - Data-driven Digital Twin Based on Multi-target Regression. This chapter introduces a solution for addressing the delay in decision-making caused by the temporal lag in drilling data streams. The proposed approach involves the use of a data-driven digital twin based on multi-target regression, which combines LSTM and MLP branches in a neural network to forecast and predict drilling parameters such as inclination and rate of penetration. An incremental learning framework is implemented to simulate the drilling process, enabling continuous updating of the regression models as drilling data is received. A case study is provided to illustrate the efficacy of the digital twin solution in predicting inclination and rate of penetration in a real-world directional drilling operation.

Chapter 5 - General Conclusion. This chapter provides a summary of the proposed contributions and subsequently discusses the existing challenges and potential future avenues for enhancing the proposed approaches.

1.6 Publications

- Data-Driven Digital Twin Based on Multi-Target Regression for Inclination and ROP Prediction in Oil and Gas Well Drilling (In review).

- Harrouz, A., Salem, H., Toubakh, H. et al. Fault prognosis of subsurface safety valve system with limited real data using self-adaptive neural network. Evolving Systems (2023). https://doi.org/10.1007/s12530-023-09525-w.

- Harrouz, A., Toubakh, H., Kafi, R., Sayed-Mouchaweh, . M., Salem, H. (2022). Self-adaptive learning scheme for Fault prognosis in oil wells and production service lines. Annual Conference of the PHM Society, 14(1). https://doi.org/10.36001/phmconf.2022.v14i1.3227.

Chapter 2

State of the Art of Fault Prognosis and Parameter Prediction in the Oil and Gas Sector using Neural Networks

2.1 Introduction

The profitability and competitiveness of industrial companies are significantly affected by the reliability and availability of their operational systems. As industrial companies upgrade their production equipment and increase their level of automation, they also experience a continuous rise in maintenance and operational expenses. This underscores the necessity for an effective maintenance and operational strategy that can sufficiently safeguard the dependability and accessibility of machinery, procedures, and other productive assets within industrial frameworks, while concurrently reducing periods of inactivity. Production loss costs during operations can be mitigated by implementing performance enhancement strategies.

This chapter provides a comprehensive overview of the integration of Neural

Networks into various machine learning approaches, including the examination of cutting-edge architectures designed for specific applications. Additionally, the chapter emphasizes the application of neural networks in predictive maintenance within the oil and gas industry, addressing various maintenance strategies and challenges. Finally, the chapter explores the utilization of different neural network architectures in the prediction of oil and gas production and drilling parameters, highlighting the limitations that this research aims to address.

2.2 Neural Networks: Definitions, Applications, and Architectures

2.2.1 Definitions

Artificial Neural Networks (ANNs), also known as neural networks or neural nets, are a branch of machine learning models inspired by the principles of neuronal organization discovered in biological neural networks [13]. They are designed to mimic the function and structure of the human brain, using interconnected nodes or neurons to solve complex problems. The key components of ANNs consist of neurons that receive inputs, apply weights, use activation functions, and generate outputs. During training, the weights of connections between neurons are adjusted. Activation functions introduce nonlinearity to enable the network to learn complex patterns. ANNs typically include input, hidden, and output layers. The input layer receives initial data, the hidden layer(s) processes information and identifies patterns, and the output layer produces the final output of the network. The nonlinear model of a neuron is depicted in figure 2.1, while figure 2.2 represents the general structure of the ANN. ANNs can be categorized as deep or shallow. A shallow NN refers to a network with one or no hidden layer, whereas a deep network has multiple hidden layers [14].

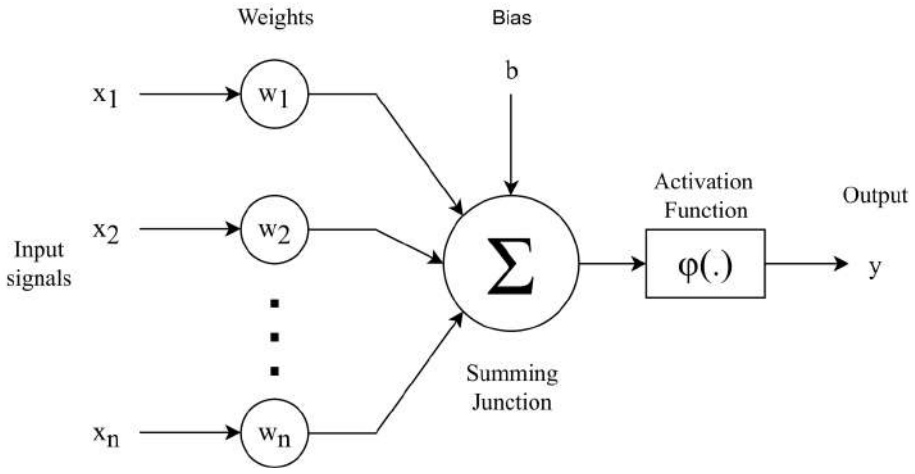


Figure 2.1: Non-Linear Model of a Neuron.

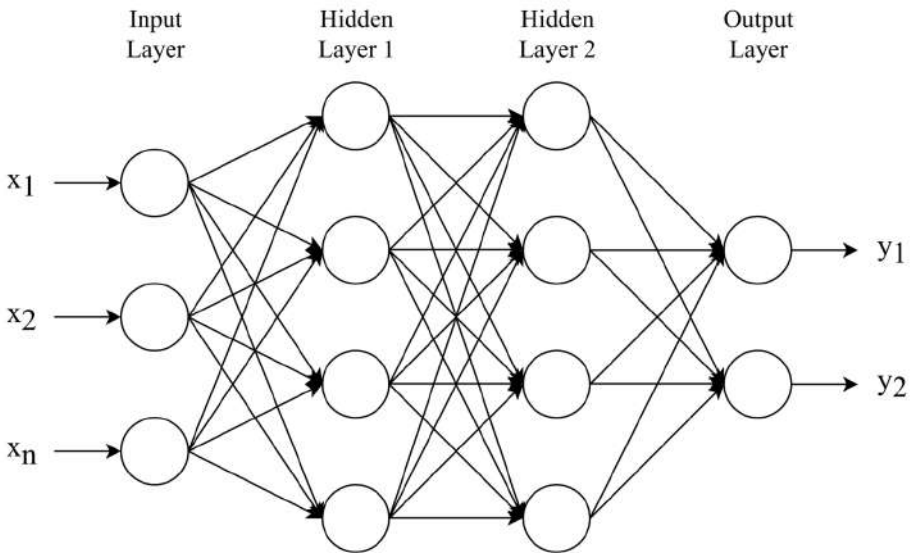


Figure 2.2: Deep Neural Network.

2.2.2 Applications of Neural Networks in different Machine Learning Algorithms

2.2.2.1 Unsupervised Learning

Unsupervised learning refers to the process by which a network can learn to identify certain input patterns in a way that replicates the numerical arrangement of the entire set of input patterns [15]. It is a machine learning technique that

involves inferring a function to determine the underlying structure from unlabeled data. These learning algorithms do not rely on labeled data to guide the training process. In this algorithm, a large amount of data and the characteristics of each observation are provided as inputs, but the desired output is not given. Unsupervised learning is commonly used, such as in clustering, to group images into distinct sets or clusters based on inherent features like color, size, and shape. This algorithm is often referred to as a self-organizing or adaptive learning algorithm because it does not rely on external sources for information. Instead, it utilizes local data and internal processes to organize and categorize training figures and input patterns [16, 17]. A state-of-the-art example of the use of neural networks for unsupervised learning in the oil and gas sector is when Kai Zhang et al. developed an unsupervised learning gas reservoir prediction method using a self-organizing neural network (SOM) to identify and predict tight-sand gas reservoirs in areas with few or no wells [18].

2.2.2.2 Supervised Learning

Supervised Learning is a type of machine learning approach that involves studying a task by mapping input data to output data using a set of example input-output pairs [19]. The objective is to infer a function from the labeled training data, which consists of a collection of training models [20]. In this learning system, an external source provides feedback to the network through a set of stimuli, with the desired output already known. Throughout the execution process, the output results are continually compared to the desired information. After several iterations, the slope descent rule is employed to adjust the connection weights based on the error between the actual output and the target information. This adjustment aims to achieve the closest possible match between the target and the actual output. Therefore, supervised learning relies up on the facts where the true class of the data is known. If the issue at hand pertains to the prediction of categorical classes or labels, it is categorized as a classification problem. Conversely, if the objective is to predict continuous numerical data, it falls under the domain of regression

(figure 2.3). A use case of neural networks in solving a classification problem in the oil and gas industry involves their utilization by Mohammed S. El-Abbasy et al. to help operators assess and predict the condition of existing oil and gas pipelines to prioritize the planning of their inspection and rehabilitation [21]. Additionally, Abdideh, Mohammad implemented neural networks and regression analysis to estimate permeability in Iran oil field [22].

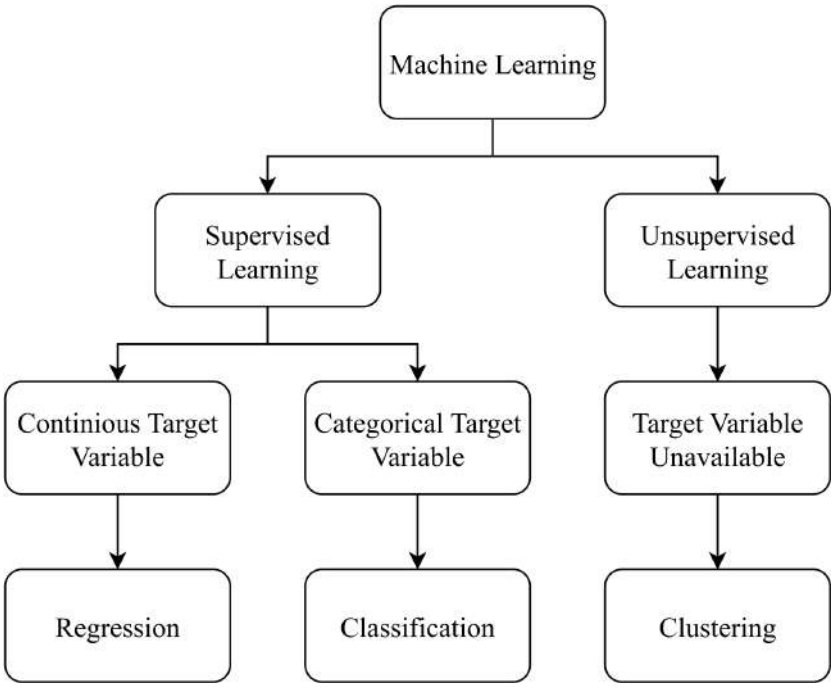


Figure 2.3: Machine Learning Types.

2.2.3 Recurrent Neural Networks Architectures

2.2.3.1 Recurrent Neural Networks

Within the field of neural networks, there exists a wide range of architectures. One of the primary areas of investigation in this dissertation pertains to the application and exploitation of Recurrent Neural Networks (RNNs) [7]. Recurrent Neural Networks (RNNs) have the ability to understand the evolving characteristics of the data by incorporating past values as inputs in order to predict future values, thereby functioning as outputs.

RNNs, specifically their contemporary utilization of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), possess substantial significance within numerous Machine Learning (ML) applications. Despite the fact that RNNs possess the capacity to attain Turing Completeness [23], implying their suitability for solving any computational problem, they demonstrate exceptional proficiency in specific task domains. RNNs are commonly favored for language translation and sequence-to-sequence learning [24].

The Long Short-Term Memory (LSTM) model is extensively utilized in handwriting recognition and speech recognition [25]. Its primary application lies in time series forecasting, which is particularly relevant to real-time drilling forecasting. In this domain, LSTM significantly surpasses non-machine learning methods like ARMIA [26].

The primary advantage of RNN-like architecture lies in its ability to retain the meaning of a sequence of inputs. When using MLP, Decision Tree, or SVM for machine learning, the sequential information of the input is lost. Although this information can be learned during the training process, aligning the architecture of the algorithm with the sequential nature of the data leads to significantly improved performance, similar to how feature engineering enhances machine learning outcomes through basic mathematical operations. Conversely, RNN architectures encounter challenges when handling non-sequential data, as they must overcome their intrinsic nature through learning.

As depicted in figure 2.4, cell A receives input x_t as well as a value v_{t-1}, which represents the previous step recurrent connection. This connection may include the output h_{t-1} obtained from the previous step, however, it can also contain other relevant information depending on the specific architecture of the cell. The result of this process is the generation of output ht. To facilitate practical analysis, the right side of figure 2.4 visualizes the process in its unfolded state.

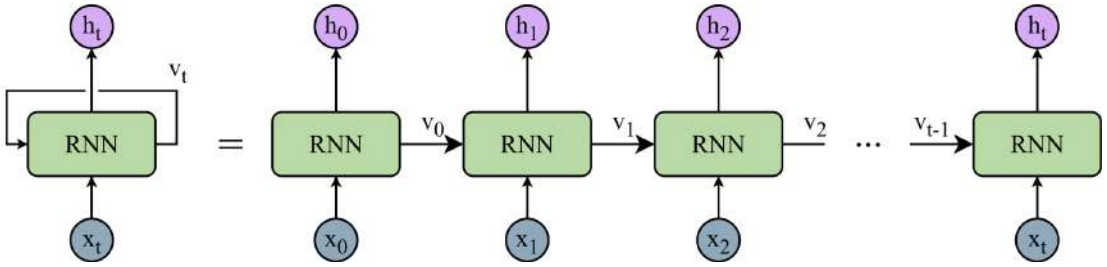


Figure 2.4: Folded and unfolded RNN.

The internal architecture of the basic RNN cell is characterized by its simplicity. This process entails combining the current input vector with the preceding output vector, which is then passed through a *tanh* layer for processing. While figure 2.5 offers a visual representation of the RNN cell structure, it is important to acknowledge that this diagram simplifies the structure and excludes fundamental components of an artificial neuron.

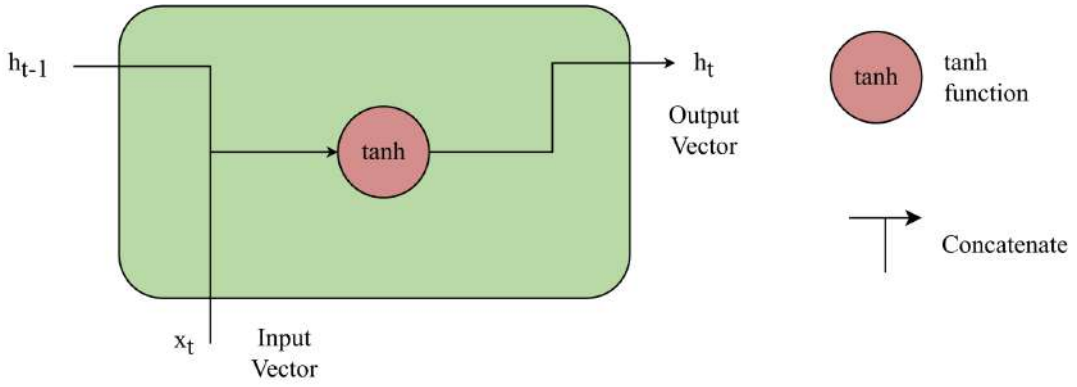


Figure 2.5: Recurrent Neural Network cell.

it is described mathematically as:

$$h_t = \sigma\left(W \cdot x_t + U \cdot h_{t-1} + b\right) \tag{2.1}$$

In figure 2.5, the output vector is denoted as h_t and the activation function is represented by σ. The hyperbolic tangent function is specifically mentioned as the activation function. The input and output vectors are associated with weights W and U, respectively, while the bias is represented by b.

2.2.3.2 Long Short-Term Memory

In order to address various practical challenges in training Recurrent Neural Networks (RNNs), a novel cell structure known as Long Short-Term Memory (LSTM) was introduced [27]. This LSTM cell exhibits a considerably more intricate architecture compared to a conventional RNN cell;figure 2.6 depicts the architecture of an LSTM cell. LSTM cell's design is specifically tailored to enhance its capacity for retaining information over extended durations.

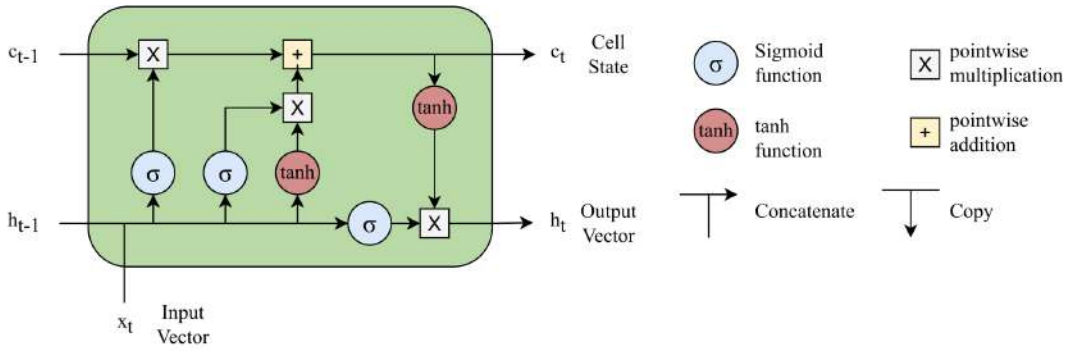


Figure 2.6: LSTM Cell.

To express the structure of the cell, the following mathematical equation is provided:

$$f_t = \sigma_g \left(W_f x_t + U_f h_{t-1} + b_f \right) \tag{2.2}$$

The activation vector of the forget gate is represented by f_t. Matrix W denote the weights of the input, U denotes the recurrent connections, with subscripts indicating the specific vector. The activation function σ, also identified by subscripts, varies depending on the relevant vector. In the case of LSTM,the sigmoid function is represented by σ_g, σ_c and σ_h represents the hyperbolic tangent function. The subscripts t and $t-1$ indicate the current and previous time-steps, respectively. x_t represents the input vector, and b represents the bias, which is unique to each relevant vector.

$$i_t = \sigma_g \left(W_i x_t + U_i h_{t-1} + b_i \right) \tag{2.3}$$

the input gate's activation vector is represented by i_t

$$o_t = \sigma_g \left(W_o x_t + U_o h_{t-1} + b_o \right) \tag{2.4}$$

the output gate's activation vector is represented by o_t

$$\tilde{c}_t = \sigma_c \left(W_c x_t + U_c h_{t-1} + b_c \right) \tag{2.5}$$

the cell input activation vector is represented by $\tilde{c}_t$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \tag{2.6}$$

the cell state vector is represented by c_t

$$h_t = o_t \circ \sigma_h \left(c_t \right) \tag{2.7}$$

The h_t represents the resultant vector obtained from the LSTM unit. The $\circ$ symbol denotes a point-wise multiplication, also known as an element-wise or Hadamard product. In this operation, two matrices A and B with identical dimensions act as inputs, resulting in a matrix C with the same dimensions. The elements of matrix C are computed as $c_{ij} = a_{ij} b_{ij}$.

2.3 Applications of Neural Networks in Predictive Maintenance

2.3.1 Maintenance Strategies

To facilitate maintenance planning, four fundamental maintenance strategies can be delineated, namely, corrective maintenance, preventive maintenance, condition-based maintenance, and predictive maintenance. Over time, the implementation of these strategies has undergone advancements aimed at minimizing the overall costs associated with the lifespan of systems [28].

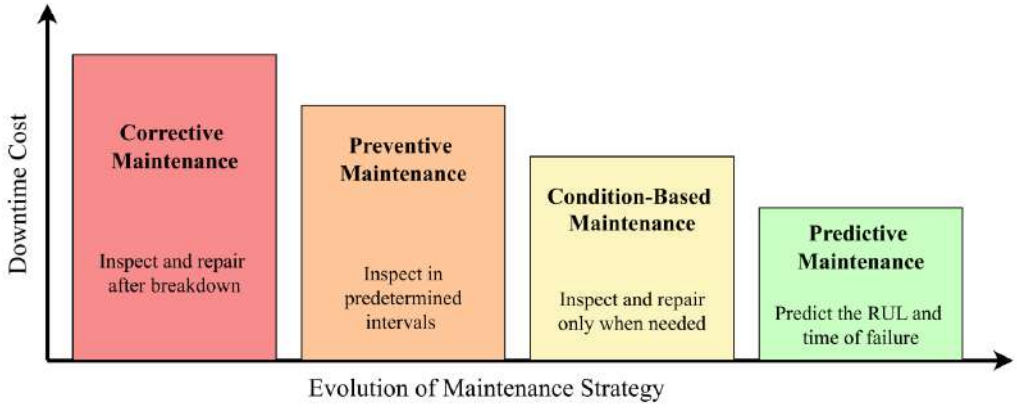


Figure 2.7: Maintenance Strategies.

Figure 2.7 illustrates the progression of maintenance strategy aimed at minimizing downtime costs. Initially, the strategy involved relying on corrective maintenance and preventive maintenance methods, which were later replaced by the adoption of condition-based maintenance and predictive maintenance. Corrective maintenance entails operating the system until it malfunctions, whereas preventive maintenance involves scheduled inspections at predetermined intervals. Conversely, condition-based maintenance allows for real-time monitoring of the system and suggests maintenance actions solely when necessary. Predictive maintenance utilizes the same fundamental principle as condition-based maintenance (CBM), however, it incorporates prediction tools, specifically machine learning, to foresee the requirement for maintenance before a specific threshold of deterioration is attained. This threshold of deterioration has the potential to ultimately result in system failure. The subsequent section presents a thorough examination of these strategies.

2.3.1.1 Corrective Maintenance

Referred to as unplanned, reactive, or breakdown maintenance, this maintenance strategy holds the distinction of being the most ancient. It does not involve regular maintenance tasks and is focused on repairing a system only after it has failed. This maintenance strategy is appropriate when product quality and revenue are

not impacted with equipment shutdowns, and the repair cost is in the acceptable range. However, This approach is correlated with an extended period of time to respond, specifically in situations where spare parts are not easily accessible, resulting in the inability to proactively plan for maintenance actions prior to a breakdown. Consequently, this strategy results in higher costs due to the frequent and unplanned nature of downtime events, which are often of longer duration.

2.3.1.2 Preventive Maintenance

Referred to as planned maintenance, involves scheduling regular inspections of a system to prevent failures and their associated consequences. This approach involves replacing key components at predetermined intervals, regardless of their current condition. Although this approach may be financially advantageous when all components are projected to malfunction at the same time, this scenario is frequently implausible in practical systems. As a result, maintenance costs increase as unnecessary replacements are made. Additionally, scheduled inspections require equipment downtime, leading to increased costs. This approach has the potential to mitigate system failures and prolong the system's longevity, it is a labor-intensive process that is dependent on time rather than the actual condition of the system, further adding to maintenance costs.

2.3.1.3 Condition based maintenance

The complexity of modern industrial systems has increased with the advancement of technology, the complexity of modern industrial systems has increased, leading to a greater potential for failure. As a result, the costs associated with maintaining these systems, particularly through preventive maintenance, have become quite expensive. To mitigate these costs while ensuring the reliability and safety of the systems, condition-based maintenance has emerged as a promising solution. This approach, also known as condition-directed maintenance, aims to address the limitations of preventive maintenance. Unlike preventive maintenance, which is solely time-based and disregards the health of the system, condition-based main-

tenance relies on the current health conditions of the system. By identifying and addressing degraded parts before product quality is compromised or failures occur, condition-based maintenance can prevent costly issues. While implementing, operating, and maintaining condition-based maintenance does require investment costs, these costs are still lower than the losses incurred from production downtime [29].

2.3.1.4 Predictive Maintenance

This approach is maximally efficient in cases where the system's deterioration pattern is well understood and undergoes changes over time, and when there are quantifiable factors regarding the system's state that can be obtained via sensor data gathering. In addition, predictive maintenance does not encroach upon the system's constituents and does not necessitate system cessation for examination, as inspections are only triggered when degradation is detected. This strategy leverages the use of prediction and forecasting techniques (like Machine learning) to predict the system failure or the life expectancy of a system, usually referred to as the Remaining Useful Life (RUL).

2.3.2 Prognostics and Health Management

In scholarly literature, an increasing number of studies are exploring the Prognostics and Health Management (PHM) strategy, which shares similarities with CBM maintenance. While CBM and PHM are often defined in the same manner, there exists a subtle distinction between these two maintenance approaches. CBM primarily focuses on diagnosing the current conditions and identifying appropriate maintenance actions to detect a fault before it transforms into a failure. On the other hand, PHM adopts a predictive approach, aiming to determine the expected time until a fault occurs in a system based on its present operating conditions which makes it a predictive maintenance strategy [30].

PHM is a strategy that primarily focuses on detecting faults at an early stage, assessing the current health of a system, and predicting its remaining useful life

[31]. The application of the PHM strategy serves several objectives, such as anticipating failures in advance, minimizing unscheduled maintenance, increasing system availability, Reducing maintenance expenses through the reduction of inspection costs, the minimization of costs associated with downtime, and the optimization of maintenance operations. Jardine and colleagues delineate three primary phases in a Prognostics and Health Management (PHM) initiative. These encompass data acquisition, wherein pertinent data is gathered to monitor the well-being of the system; data processing, wherein advanced techniques are employed to analyze the amassed data for fault diagnosis and prognosis; and maintenance decision making, wherein optimal maintenance actions are recommended [32]. Callan et al. propose a PHM architecture consisting of five steps: Data Manipulation, Condition Monitoring, Health Assessment, Prognostics, and automatic decision reasoning [33]. In a similar vein, Kim and colleagues outline four primary stages within the Prognostics and Health Management (PHM) strategy. These stages encompass the acquisition of data, encompassing the collection of condition monitoring data and the extraction of pertinent features. Additionally, the diagnostic stage involves identifying faults and evaluating their severity. The prognostic stage entails predicting the remaining useful life, while the health management stage focuses on implementing optimal maintenance and logistics management. [33]. Overall, figure 2.8 illustrates the primary steps involved in implementing the PHM strategy.

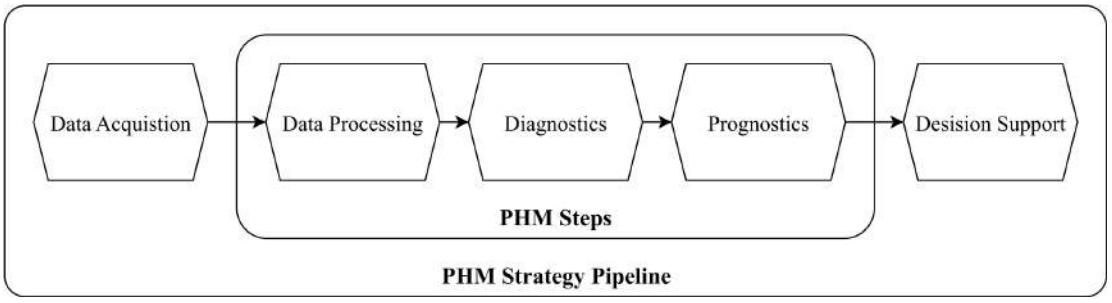


Figure 2.8: PHM Strategy.

The data acquisition stage is responsible for gathering relevant data from sen-

sors placed on critical components. This data is then used in the prognostics stage to estimate the Remaining Useful Life (RUL) of these components, potentially leading to significant cost reductions in maintenance. During the data processing stage, the collected data is analyzed to extract important indicators or features that reflect the health evolution of the system, which will be useful for subsequent steps. The diagnostics stage enables the detection, isolation, or localization of incipient faults. The initiation of degradation detection signals the commencement of the prognostics phase, during which the RUL is estimated. Subsequently, the decision support phase utilizes the information acquired from preceding states, encompassing the root cause, health state,and RUL, in conjunction with other pertinent factors like logistics and priority, to propose the most advantageous maintenance measures. Data processing and prognostics phases of the PHM strategy are discussed in Chapter 3 of this thesis, with special attention devoted to prognostics (figure 2.8).

2.3.3 Neural Networks and RUL Estimation

The growing use of sensors has resulted in an exponential growth of Big Data, thereby increasing the appeal and relevance of machine learning approaches. Among the various machine learning techniques, ANNs are widely recognized as highly effective solutions for data-driven predictive maintenance. Trend direction of bearings was forecasted by Wei Teng et al. using a developed data based model by incorporating ANN to wind turbine gearboxes. This was achieved by integrating predicted and training features and fitting a polynomial curve to capture the long-term degradation process [34]. Additionally, Mohamed Elforjani and Suliman Shanbr compared three supervised machine learning approaches (Support Vector Machine Regression, Artificial Neural Network, and Gaussian Process Regression) to estimate the Remaining Useful Life (RUL) of machine components through the utilization of acoustic emission technique [35]. Numerous studies utilize Neural networks to predict the life condition (or RUL) of oil and gas pipelines; in a study conducted by Nagoor Basha Shaik et al., a neural network was created that uti-

lized various pipe parameters (including pressure, corrosion, wall thinning, age, nominal thickness, outer radius, and product type) to forecast the remaining useful life of piping [36]. In another study, the researchers developed a Feed-Forward Back Propagation Network (FFBPN) which relied on historical inspection data from oil and gas fields to predict the condition of crude oil pipelines. This prediction was specifically focused on factors such as metal loss anomalies (including length, width, and depth), wall thickness, weld anomalies, and pressure flow [37].

2.3.4 Neural Networks and Insufficiency of Training Data

The training process for machine learning models can be particularly challenging when there is a limited amount of degradation data available. Nevertheless, this problem can be resolved by enhancing the prediction approach or employing an augmentation techniques to the data to improve the quality of the training data. Although there is a lack of studies that specifically tackle the problem of low Remaining Useful Life (RUL) prediction accuracy caused by insufficient sample data, researchers such as Wenbai Chen et al. have proposed a solution using a deep convolutional neural network-bidirectional long short-term memory network (DCNN-Bi-LSTM) and domain adaptation [38]. Additionally, Emmanuel Ramasso and Rafael Gouriveau have suggested an approach that involves classifying predictions based on a neuro-fuzzy system and the theory of belief functions to overcome data scarcity [39]. Another proposed method is the Sequence Adaptation Adversarial Network (SAAN) developed by Haixin Lv et al., where data from a similar system is incorporated to expend the training dataset. This approach expended the data with multiple degradation modes to helps improve the model performance [40]. Chapter 3 of this study dives into the issue of prognostics for systems with insufficient data, a predicament that is particularly evident within the oil and gas industry owing to the confidential and critical nature of the data amassed by service providers.

2.4 Neural Networks Applications in Parameter Prediction in Oil and Gas Operations

Well construction is the term used to describe the process of drilling and completing a subsurface well for various purposes, such as extracting hydrocarbons, geothermal energy extraction, waste storage, or rock sample collection. The geological environment deep beneath the earth's surface is extremely complex and presents numerous challenges, which make drilling a costly and unpredictable endeavor. These uncertainties and variations also have an impact on well-construction activities. As a result, it is crucial to closely monitor equipment and process variables during drilling operations in order to prevent failures, ensure safety, minimize non-productive time, reduce troubleshooting expenses, and avoid potential human casualties [41].

Continuous measurements of direction and inclination are of utmost importance in the decision-making process during directional drilling. In a study conducted by William G. et al., they successfully incorporated continuous inclination measurements, drilling parameters from both the surface and downhole, and a distinct model to accurately calibrate and predict the directional inclination of a Bottom Hole Assembly (BHA) in real-time. As a result, the reliance on regional factors in directional drilling operations was significantly reduced. The research effectively addressed the time delay in sensor measurements that are far from the actual drill bit [42].

Andrzej T. Tunkiel and colleagues proposed a machine learning methodology that utilizes Recurrent Neural Networks (RNN), specifically a Gated Recurrent Unit (GRU), for the purpose of predicting the inclination of a directional drilling operation that employs a bent motor. This approach yielded a Mean Absolute Error (MAE) of 0.6° when forecasting the inclination of the drilling operation up to 23 meters ahead of the sensors located in the BHA [43, 44].

There is a considerable body of literature containing numerous models for predicting ROP based on data analysis [45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56,

57, 58, 59]. However, there has been limited research conducted on the development of a unified ROP prediction model that can effectively adapt to changes in geology and subsurface layers. Cesar Soares and Kenneth Gray attempted to address this challenge by applying the concept of continuous learning and an expanding dataset in their prediction model, utilizing a random forest algorithm [60]. However, reproducing or improving upon their findings is not feasible without access to the necessary data, as previous studies have highlighted a significant lack of data availability in the oil and gas industry. Fortunately, the issue has been tackled by Andrzej T. Tunkiel et al. through the provision of a database sourced from Equinor's Volve dataset that is accessible to the public. This database establishes a benchmark for assessing the effectiveness of ROP prediction techniques and explores the efficacy of different machine learning algorithms (such as Gradient Boosting, XGBoost, Random Forest, AdaBoost, KNeighbors, and Bingham) under a continuous learning scenario [12].

Despite their significance and practicality, multi-target regression models are rarely applied in the oil and gas industry and have not been utilized in the context of drilling optimization. Research conducted by Helmi Helmiriawan and Zaid Al-Ars introduced a multi-target regression method that integrates deep learning (RNNs) and the Cumulative sum (CUSUM) technique for predictive maintenance in oil refineries. The efficacy of this approach was verified using a real-world dataset from the industry [61]. Additionally, Liang Xue et al. devised a data-driven model based on a multi-objective random forest technique to forecast shale gas production performance [62].

Various types of recurrent neural networks (RNNs) have been employed in previous studies to predict different parameters in oil and gas operations. For instance, Wanxing Zhang et al. employed a combination of Generative Adversarial Network (GAN) and LSTM called GAN-LSTM to simultaneously predict wellhead pressure and circulation pressure, as well as predict the ROP and total weight in a coiled tubing drilling operation using LSTM [63]. Similarly, Shaowei Pan et al. utilized a combination of Convolutional Neural Networks (CNN), LSTM, and

Self-Attention (SA) referred to as CNN-LSTM-SA to predict oil well production, achieving higher accuracy compared to traditional machine learning and deep learning methods [64]. B. Sirisha et al. developed a Deep Stacked Bidirectional LSTM (SBiLSTM) model to forecast petroleum production, which resulted in a 17% and 14% increase in accuracy compared to vanilla RNN and GRU [65]. Indrajeet Kumar et al. forecasted oil production using an attention-based LSTM network [66]. Empirical evaluation conducted by Junyoung Chung et al. on the Ubisoft dataset demonstrated that both GRU and LSTM outperformed vanilla RNN, with LSTM performing better on Ubisoft dataset A and GRU performing better on dataset B [67]. Despite the prevalent use of LSTM models in the oil and gas domain, these models have not been utilized for predicting inclination in directional drilling operations.

2.5 Conclusion

This chapter provides an overview of various machine learning algorithms and their application in the maintenance and oil and gas industries. Machine learning techniques can be implemented using neural networks, which use different approaches to address specific tasks. Unsupervised learning, for example, is used when the target data is unknown or unlabeled, and it helps organize datasets. On the other hand, supervised learning is employed to predict outputs or targets based on input data. If the predicted data falls into discrete categories, it is considered a classification problem, while continuous targets are referred to as regression problems. The chapter also explores different neural network architectures, with a particular focus on recurrent neural networks, which leverage the time dependency of sequential data. Various maintenance strategies, including Prognosis and Health Management, are discussed, highlighting the integration of machine learning to estimate the remaining useful life of different systems in the industry. The chapter showcases multiple applications of various neural network architectures, specifically in the oil and gas operations and production field.

The upcoming chapter introduces a data-focused methodology for fault prognostics, specifically designed for systems with a scarcity of past data, aiming to forecast the RUL by employing a tailored neural network model. The efficacy of this approach is substantiated through its application to genuine operational data obtained from an oil and gas extraction site.

Chapter 3

Fault Prognostic Approach of Systems with Limited Historical Data

3.1 Introduction

In the context of industrial systems, there is typically an abundance of historical data pertaining to the typical functioning of these systems, whereas data concerning degraded or faulty conditions is often lacking. This scarcity of data may be attributed to the high expenses involved in generating degradation data under controlled laboratory conditions, concerns related to safety, or the absence of such data for newly installed machinery. Additionally, the degradation behavior of a system in real operating conditions frequently differs from that observed in laboratory settings, primarily due to variations in environmental and load conditions.

This chapter suggests a data-driven prognostic method to address the issue of inadequate degradation sequences. It conducts prognostics in situations where there is a lack of run-to-failure data or a limited amount that is not enough for a dependable and accurate prediction of RUL. The methodology consists of three primary stages: the development of a health indicator, the training of multiple machine learning algorithms, and the implementation of a feedback technique

on the most successful model to improve its efficacy and address the issue of predicting the RUL of systems with restricted data availability.

The structure of this chapter is as follows: section 3.2 introduces the case study that was utilized to validate the proposed approach. In section 3.3, the main components of the approach are described. Section 3.4 outlines the experimental procedure and presents the results obtained. Section 3.5 initiates a discussion regarding the obtained outcomes and their applicability in comparable real-life situations. Section 3.6 concludes the chapter.

3.2 Case Study

The process of oil and gas production involves multiple stages, starting from extraction and ending with delivery. Following the drilling and cementing phases, the well undergoes completion. In this stage, a tool positioned in the production string called the Sub-Surface Safety Valve (SSSV), typically at a depth of 60-100 meters below the surface (see figure 3.1). The implementation of this tool as a standard in oil and gas production wells can be attributed to the Gulf War oil spill, which caused significant environmental damage due to millions of barrels of oil being released onto the surface. The SSSV serves as a safety mechanism, ensuring the closure of the production tubing during emergencies or the hydraulic control line physical disconnection connected to the tool. Unfortunately, there are instances where the closure function of the SSSV fails unexpectedly, resulting in loss of production, without any indication being provided at the surface.

Marvin Rausand and Jorn Vatn examined The issue of the SSSV's reliability, they explored the occurrence of blowout events using a Weibull life distribution instead of an exponential distribution [68]. In a recent study conducted by Danilo Colombo and colleagues, finite element machines employing regression techniques were employed to model the dependability of the SSSV system. [68]. While the failure prognosis of the SSSV was not addressed by these studies, it is of utmost importance to anticipate potential failures to avert production losses that may

arise from the unintentional shutdown of the SSSV.

The depiction in figure 3.2 elucidates the operational mechanism of a Subsurface Safety Valve (SSSV). The SSSV functions by having the flapper transition between closed and open states in tandem with the vertical movement of the control sleeve. The pressure within the pressure chamber dictates the position of the control sleeve, thereby regulating the valve's operation. The position of the control sleeve is determined by the pressure within the pressure chamber. When there is no pressure in the chamber ($Pressure = 0psi$), the control sleeve moves to the highest position and the flapper closes. Conversely, when the pressure within the chamber reaches a certain threshold ($1100psi < Pressure < 2200psi$), dependent on the type of SSSV, the flapper opens due to the movement of the control sleeve to the lowest position. The pressure within the chamber is maintained from the surface using the hydraulic control line. In the event of a disconnection or leakage in either the control line or the pressure chamber, the flapper is automatically shut. Furthermore, the SSSV can be deliberately closed in emergency situations or during surface maintenance.. However, it is worth noting that this mechanism can sometimes fail, resulting in the SSSV remaining in the closed position and causing a loss in production. To address this failure, an exercising tool can be employed to forcibly open the flapper, and if necessary, the tool can be replaced through the production tubing.

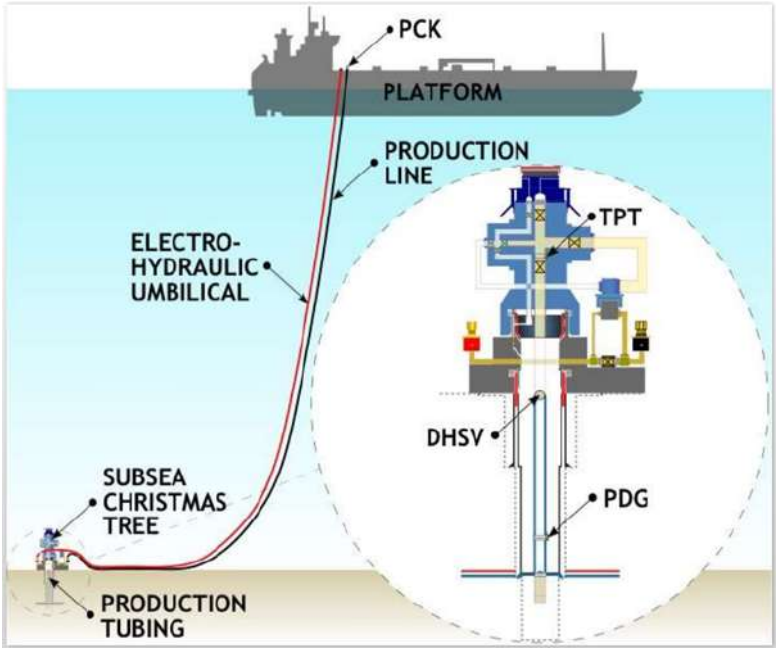


Figure 3.1: Simplified schematic of a typical offshore well [1].

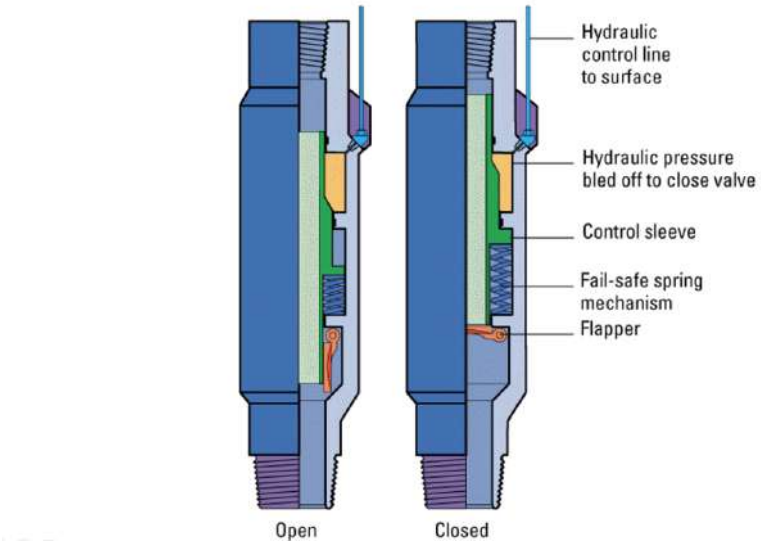


Figure 3.2: Cross-sectional view and internal elements of a Subsurface Safety Valve.[2].

3.3 The Proposed Approach

3.3.1 Prognosis System

In this segment, diverse approaches for forecasting Remaining Useful Life (RUL) are examined. Initially, Health Indicators (HIs) are formulated to assess the operational condition of individual SSSVs, offering a dynamic depiction of system deterioration. The utilization of HIs as a predictive feature for estimating RUL necessitates the incorporation of sensor measurements during the construction of these Health Indicators. HI starts from a healthy state and continues to decrease until HI reaches zero ($HI = 0$), this instance is called Time of Failure (ToF). Calculating the RUL for a specific instance requires the following equation:

$$RUL = ToF - Real\ Actual\ Time \tag{3.1}$$

Forecasting accurately with limited data poses a challenge due to insufficient features within the feature space, leading to decreased predictive precision. To address this issue, the presented methodology suggests a remedy by integrating the anticipated response from the test back into the feature space, acting as a corrective feedback mechanism for predictions. This approach was originally introduced in a prior study [3].

The flow chart in figure 3.3 illustrates the proposed prediction system. In the pre-processing stage, HIs are constructed through the fusion of sensor measurements using regression techniques. These HIs serve as features for training a machine learning model based on regression. The model undergoes testing with unseen data, the resultant Predicted Remaining Useful Life (PRUL) is stored in a matrix along with its corresponding mean squared error (MSE) value. This PRUL serves as input for the model's retraining in subsequent iterations of the system, which comprise regression model iterations and corrective feedback system iterations. System optimization involves identifying the global minimum of the MSE function across all predicted RULs, leading to the determination of the Optimal Predicted RUL (OPRUL). A curve-fitting function is then employed to

estimate the Remaining Useful Life (RUL) based on OPRUL, and the accuracy of the prediction and estimation outcomes is assessed.

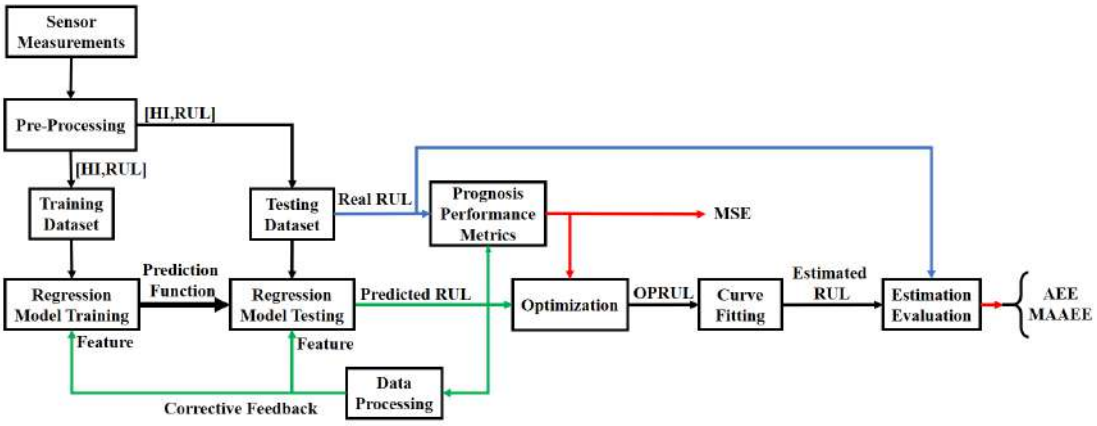


Figure 3.3: Flowchart of the proposed RUL prediction approach.

3.3.2 Data Processing

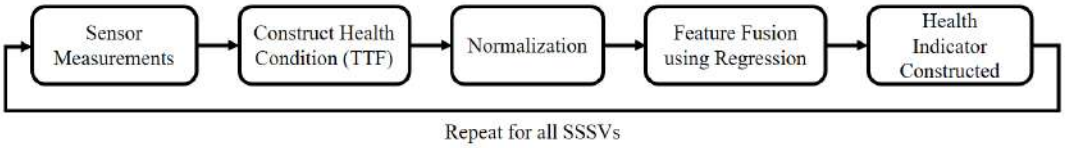


Figure 3.4: The proposed preprocessing methodology.

Data preparation is conducted during the preprocessing phase, subsequently, the data is used for training the regression model. The primary aim is to obtain a Health indicator (HI) that accurately characterizes the tool's health state dynamics throughout the degradation phase.

The information obtained from the temperature and pressure transducer (TPT) positioned at the surface level comprises pressure and temperature measurements. These measurements serve the purpose of assessing the well's health status or predicting time-to-failure (TTF). In the event of a failure, referred to as spurious closure, there is a gradual decline in both pressure and temperature. The rate of decline may fluctuate based on well conditions. Other sensor readings within the

dataset remain constant, leading to a variance of zero. Such measurements with zero variance are omitted during the training phase.

The process of identifying the most suitable point for the optimization algorithm is simplified when variations in pressure and temperature amplitudes among different wells are overlooked. Consequently, it is necessary to normalize the training and test datasets. The normalization process is carried out using the min-max scale method, as outlined in the following equation:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.2}$$

The HI construction is a crucial component in the pursuit of prognostics. It serves to portray the performance and condition of a system, enabling the monitoring of the health state of a particular component or system over a period of time. The presence of failure is signaled by a deviation from the normal or healthy state, as manifested by a decrease or rise in the HI trajectory [69]. Figure 3.6 displays the sensor measurements after normalization for a sample SSSV during the degradation phase, along with the TTF function.

The TTF or degradation is determined based on the data collected by the sensor. The healthy phase is indicated by a value of 1, while the faulty phase is represented by zero (figure 3.9). The transition (degradation) follows a linear decline from 1 to 0, depending on the rate of degradation.

The construction of the Health Indicator (HI) involves merging sensor measurements into a unified HI. This fusion technique relies on linear regression to uphold the linear relationship between the sensor measurements and the degradation function. The degradation function acts as the dependent variable in the regression, with sensor measurements serving as independent variables or features. The regression model can be expressed through equation 3.3. The HI for a sample SSSV is established and depicted in figure 3.6.

$$TTF = a + b_1 Pressure + b_2 Temperature \tag{3.3}$$

The process is iterated for all specific sets of variables for each of the twelve

wells under varying circumstances. The HIs depicted in figure 3.7 are subjected to smoothing using a moving average filter and are regarded as the primary dataset for the feedback system aimed at rectification; this dataset is further divided into training and testing subsets.

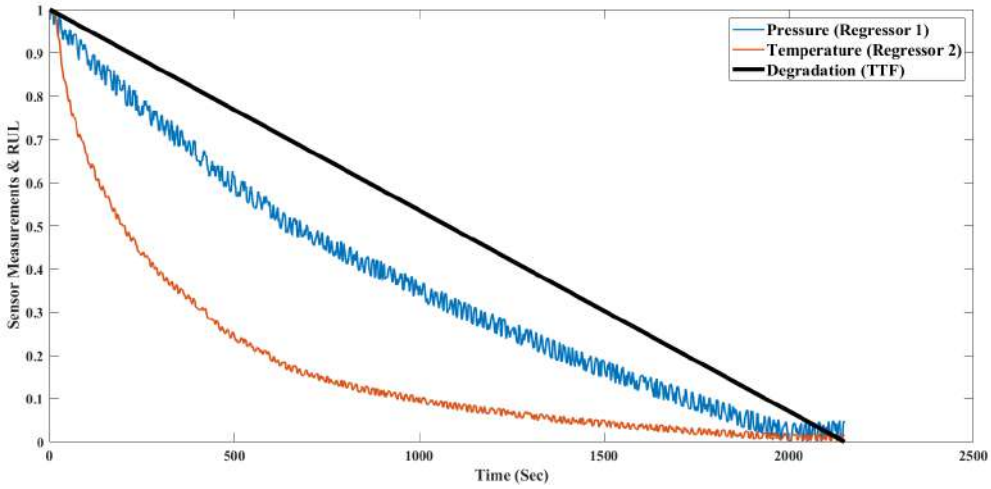


Figure 3.5: Pressure, Temperature Measurements, & and degradation of an SSSV sample.

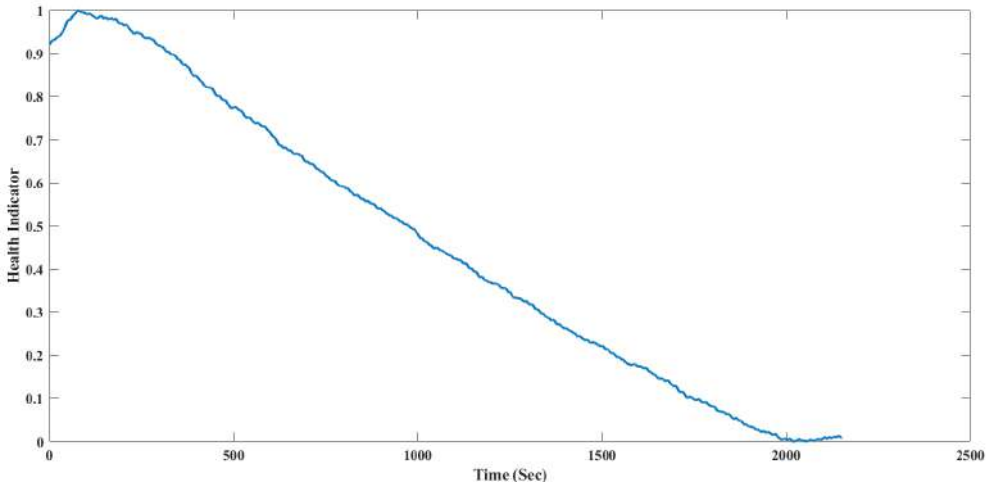


Figure 3.6: The Constructed HI for an SSSV sample.

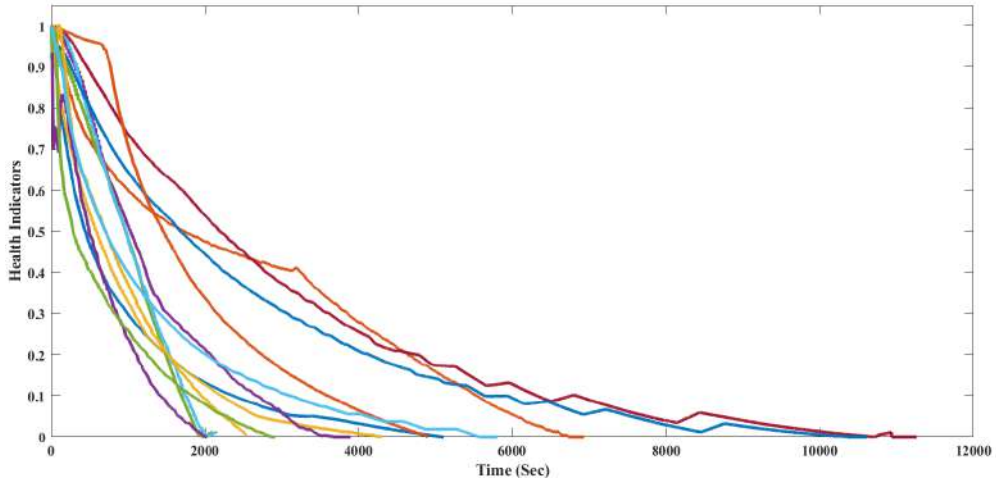


Figure 3.7: Health Indicators used as Training, Validation and testing sets.

3.3.3 Regression Algorithm Selection

Based on prior studies, the predominant models employed in predictive maintenance within the research domain include Artificial Neural Networks (ANN), Decision Trees (DT), and Support Vector Machines (SVM) [70]. The inclusion of the proposed Corrective Feedback approach in these models has been found to enhance prediction accuracy. However, the selection of the optimal model relies on two crucial factors, namely the consistency of results and the speed of training. SVMR models are associated with lengthy training periods, which can hinder the effectiveness of the proposed approach due to the need for frequent retraining. In contrast, Bagged and Boosted Decision Trees exhibit comparatively faster training durations; however, the resulting signal is discretized instead of being continuous. This can result in inaccurate outcomes when used in combination with a curve-fitting algorithm, as it lacks the continuous information flow necessary for accurate predictions over a given period. ANN, on the other hand, allows for flexible training times based on the designed network complexity. The level of complexity of the neural networks investigated includes One Layer NN (OLNN), Bi-layered (BLNN), and Tri-layered (TLNN), in addition to Quadratic SVM (QSVM) and Linear SVM (LSVM). The Mean Squared Error (MSE) and

training time values of the examined regression models are presented in table 3.1.

Table 3.1: Training results of multiple models.

Model	LSVM	QSVM	TLNN	BLNN	**OLNN**	Boosted Trees	Bagged Trees
MSE	0.266	1.039	0.084	0.084	**0.084**	0.087	0.185
Training time	5702 s	5360 s	40 s	25 s	**15 s**	12 s	11 s

The parameters of the neural network are adjusted through a process of trial and error, with a focus on two crucial factors: performance and training time. The objective is to achieve optimal performance while minimizing the duration of the training process. The neural network is comprised of one fully connected layer containing ten neurons. Although the default iteration count for the neural network is set at 1000, it was noted that convergence of the training loss occurred as early as the 200^{th} iteration, prompting a decrease in the number of iterations. Increasing the number of layers, neurons, and iterations results in a longer training time without any performance improvement. The Rectified Linear Unit (ReLU) activation function is utilized due to its versatility in working with various types of neural networks [71]. The function is defined in the following equation.

$$\mathrm{ReLU}(x) = \begin{cases} 0, & x \leq 0 \\ x, & x > 0 \end{cases} \tag{3.4}$$

The number of iterations in the regression model is distinct from the number of iterations in the corrective feedback system. Specifically, the neural network regression model undergoes 200 iterations during both the training and testing phases. On the contrary, the corrective feedback system functions at an elevated system level, necessitating the regression model to undergo 200 iterations to facilitate a single advancement in the corrective feedback system. To illustrate,

when the corrective system undergoes 100 iterations, the regression model must complete 20,000 iterations.

The inclusion of an adaptive variable in the form of an adaptive input neuron affects the structure of regression models at a lower level. In the initial iteration of the system, the HIs are the features of the training dataset alongside an empty feature vector, which remains empty during this iteration.

The regression model undergoes training to establish a prediction function, encompassing the neural network's weights and bias terms. Subsequently, this prediction function is employed to assess the model's performance on a novel dataset not previously encountered. The model's effectiveness is evaluated on unseen data, yielding a new prediction for Remaining Useful Life (RUL). The predicted RUL is then integrated as a feature, along with the nominal system HI, into the vacant adaptive feature vector for retraining the regression model. The process involves recalculating weights and bias terms, using them to test the newly trained model on additional unseen data. This iterative cycle persists similarly. At a more granular level, the adaptive feature is manifested as an adaptive neuron updated in each system iteration by incorporating the latest predicted RUL from the testing phase on unseen data. The corrective feedback mechanism empowers the neural network to leverage predicted responses from testing as novel features for learning, thus supplying supplementary data for model enhancement within the feature space matrix. This facilitates the neural network's adaptation and augmentation of prediction accuracy by assimilating optimal predictions.(See figure 3.8 for a visual representation).

The following set of equations describes The proposed NN structure:

$$N_{L1} = ReLU\left(HI \cdot W_1 + \widehat{RUL} \cdot W_2 + B_1\right) \tag{3.5}$$

$$RUL = ReLU\left(N_{L1} \cdot W_3 + b_{2,1}\right) \tag{3.6}$$

From equation 3.5 and 3.6

$$RUL = ReLU\left[ReLU\left(HI \cdot W_1 + \widehat{RUL} \cdot W_2 + B_1\right) \cdot W_3 + b_{2,1}\right] \tag{3.7}$$

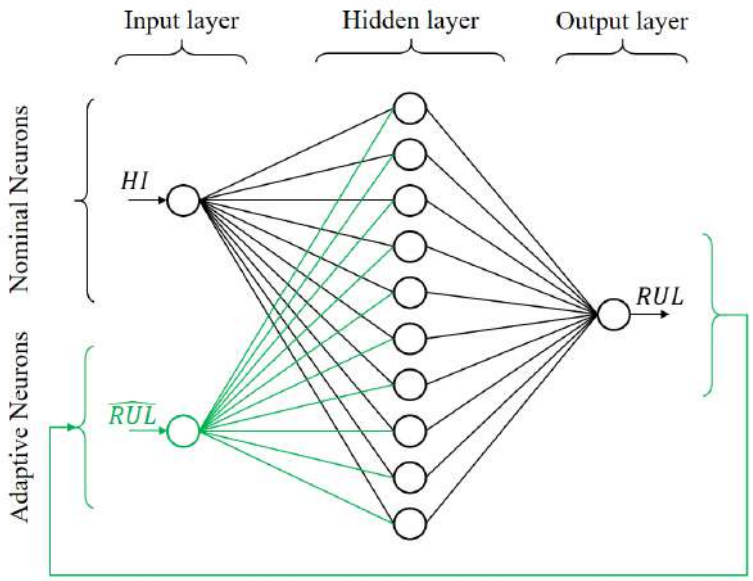


Figure 3.8: The Proposed NN structure with adaptive neurons.

Where:

$$
\begin{aligned}
W_1 &= \begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} & \cdots & w_{1,10} \end{bmatrix} \\
W_2 &= \begin{bmatrix} w_{2,1} & w_{2,2} & w_{2,3} & \cdots & w_{2,10} \end{bmatrix} \\
W_3 &= \begin{bmatrix} w_{3,1} & w_{3,2} & w_{3,3} & \cdots & w_{3,10} \end{bmatrix} \\
B_1 &= \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} & \cdots & b_{1,10} \end{bmatrix}
\end{aligned} \tag{3.8}
$$

Where:

N_{L1}: hidden layer activations vector.

HI: HIs are the input layer neuron.

$\widehat{RUL}$: PRUL ; $\widehat{RUL_0} = 0$.

W_1, W_2, and W_3: are the weight vectors.

B_1 and $b_{1,2}$: are the bias vector of the first layer and bias coefficient of the output.

$ReLU$: is the used activation function.

To analyze the impact of iterations of the corrective feedback system on neural networks, the following set of equations is employed:

$$
\begin{aligned}
\widehat{RUL}_1 &= ReLU\left[ReLU\left(HI \cdot W_1 + 0 \cdot W_2 + B_1\right) \cdot W_3 + b_{2,1}\right] \\
\widehat{RUL}_2 &= ReLU\left[ReLU\left(HI \cdot W_1 + \widehat{RUL}_1 \cdot W_2 + B_1\right) \cdot W_3 + b_{2,1}\right] \\
\widehat{RUL}_3 &= ReLU\left[ReLU\left(HI \cdot W_1 + \widehat{RUL}_2 \cdot W_2 + B_1\right) \cdot W_3 + b_{2,1}\right] \\
&\vdots \\
\widehat{RUL}_n &= ReLU\left[ReLU\left(HI \cdot W_1 + \widehat{RUL}_{n-1} \cdot W_2 + B_1\right) \cdot W_3 + b_{2,1}\right]
\end{aligned}
\tag{3.9}
$$

In the initial iteration of the system, the predicted remaining useful life ($\widehat{RUL}_n$) for unseen data is not yet calculated, so it is assumed to be zero ($\widehat{RUL}_0 = 0$). After the initial iteration, the first predicted Remaining Useful Life (RUL) serves as input for subsequent iteration. In this investigation, the number of iterations is arbitrarily established at one hundred (100). The predicted RULs are systematically arranged in a matrix alongside their respective Mean Squared Error (MSE) values.

The proposed approach expands the input dataset of the neural network model by increasing the feature space dimensions. Initially, the feature space dimensions were $(k \times 1)$, where "1" represents the health indicator (HI) and "k" is the number of samples in the HI. After implementing the proposed approach, the feature space size is increased to $(k \times 2)$, with the addition of the previous RUL ($\widehat{RUL}_{n-1}$) as an extra row in the feature space. Therefore, it can be said that the proposed approach has a quasi-data augmentation effect.

3.3.4 Optimal Model Selection

A refinement process is utilized to ascertain the most accurate prediction within a collection of 100 Predicted Remaining Useful Life (PRUL). These PRULs are arranged in a matrix alongside their corresponding Mean Squared Error (MSE) values. In each iteration, a novel Remaining Useful Life (RUL) is forecasted and evaluated, and the optimization algorithm concurrently identifies the PRUL associated with the lowest MSE. Ultimately, the achieved minimum MSE corre-

sponds to the Optimal Predicted Remaining Useful Life (OPRUL) is obtained. The OPRUL is determined using the following equations:

$$MSE_{opt} = min\,(MSE_i)^T = min\,(MSE_1, MSE_2, MSE_3, \cdots, MSE_{100})^T \quad (3.10)$$

$$OPRUL = PRUL\{MSE_{opt}\} \quad (3.11)$$

3.3.5 Curve fitting function used for RUL Estimation

Estimating the RUL is a common practice in the industry when assessing the lifespan of a system or component. It involves subjecting the system to real-life scenarios and testing its limitations. The prediction system is fed with data in incremental portions, starting with 10% of the known data. As new data becomes available, the RUL estimation process is repeated, gradually increasing the amount of data used for estimation until reaching 90% of the known data.

Estimating the Remaining Useful Life using a subset of available data necessitates the utilization of a curve-fitting algorithm. This algorithm is designed to maintain the consistency of vector length with the provided data. If only 10% of the data is used for prediction, only 10% of the RUL result can be determined. Therefore, employing a curve-fitting algorithm becomes crucial in estimating the intersection between the x-axis and the prediction.

Curve fitting algorithms are specifically designed to learn and estimate the dynamics of the RUL. Consequently, the selection of an appropriate curve-fitting function is of utmost importance. Polynomial and exponential functions are commonly used in curve-fitting techniques. As the RUL is represented by a linear function, opting for a first-degree polynomial function proves to be the most appropriate selection, given its similarity to the signal shape and its capacity to minimize the MSE between the predicted and fitted RUL values. Once the predicted signal is fitted, the point of intersection between the curve and the time axis represents the Time of Failure (ToF). The RUL is then calculated using equation

3.1. The fitting function used is described in equation 3.12.

$$f(x) = Ax + b \tag{3.12}$$

3.4 Experimentation

3.4.1 Dataset description

The study utilizes the public 3W Dataset provided by Petrobras [1] to address abnormal events, identify their underlying causes, and make appropriate control decisions to restore the systems to a safe operational state. The overarching strategy employed is denoted as Abnormal Event Management (AEM). The dataset is composed of a blend of authentic, simulated, and manually generated data. Real instances were extracted from the plant information system utilized by Petrobras Operational Unit located in Espírito Santo, Brazil (PI System [72]) without any preprocessing to maintain their authenticity. Simulated instances were generated using OLGA [73], a dynamic multiphase flow simulator widely adopted by various oil companies worldwide [74]. Hand-crafted examples refer to templates generated by a knowledgeable professional, which underwent image processing through a script after detailing all chart attributes such as variables, event types, and states (normal, transient, faulty).

The dataset comprises a variety of abnormal and infrequent occurrences that take place in the process of oil and gas operation and production (Table 3.2):

- **Abrupt Increase of BSW:** The term Basic Sediment and Water (BSW) refers to the proportion of water and sediment flow rate to the liquid flow rate. When there is a sudden rise in BSW, it gives rise to various issues. Recognizing such occurrences automatically enables proactive measures to prevent production complications.

- **Spurious Closure of the SSSV:** The SSSV, or surface-controlled subsurface safety valve, serves as a protective measure in oil and gas wells.

Any disruption or malfunction in the system will lead to the closure of the SSSV. However, there are instances where the closure mechanism may fail. Detecting and addressing these instances promptly can enable the valve to be reopened using appropriate operational procedures, thereby preventing losses in production and additional expenses (section 3.2).

- **Severe Slugging:** This occurrence exhibits two discernible attributes, namely periodicity and intensity. Consequently, it has the potential to inflict harm upon the machinery employed throughout the production line. However, measures can be implemented to rectify the situation.

- **Flow Instability:** This phenomenon also is marked by intermittent and false increases in the flow of liquid and gas. However, these increases are less severe and do not encompass the entire sequence of liquid blockage followed by a surge of gas. Failure to address flow instability may lead to the development of severe slugging.

- **Rapid Productivity Loss:** The productivity of a well that flows naturally is influenced by various characteristics, and alterations to these characteristics can result in energy losses. If the energy losses exceed the system's energy, the flow rate may decrease or cease entirely. It is crucial to detect this occurrence early to take appropriate measures and prevent any decrease in production.

- **Quick Restriction in the Production Choke (PCK):** The production choke valve installed at the production unit plays a crucial role in maintaining control over the flow of fluids from the surface. However, if this valve is manually operated, it can potentially result in unintended sudden restrictions, which can have a direct impact on oil production. The automatic identification of such occurrences can facilitate faster remedial action.

- **Scaling in PCK:** The presence of inorganic deposits at the PCK site leads to a decline in oil production, making it necessary to identify this occurrence

at an early stage. Swift detection is preferred as it enables the implementation of remedial measures to prevent further losses in production.

- **Hydrate in Production Line:** The occurrence of hydrate formation in wells and production/injection lines poses a significant challenge in the industry, as promptly identifying this undesirable phenomenon enables the prevention of extended periods of production losses.

The primary objective of this investigation pertains to the occurrence of the second failure event, which involves the spurious closure of the SSSV. This event is particularly relevant as it incorporates sensor measurements associated with a tangible component (i.e., SSSV) that can be effectively maintained. In contrast, other failure events in this context are predominantly of a chemical nature or are linked to reservoir instability, rendering them unsuitable for the purposes of a predictive maintenance (PdM) and remaining useful life (RUL) estimation study.

Table 3.2: 3W database per event, a quantitative Description [3]

Class	Description	Real	Simulated	Sketched	Total
0	Normal	597	0	0	597
1	Abrupt BSW Increase	5	114	10	129
2	**Spurious SSSV Closure**	**22**	**16**	**0**	**36**
3	Severe Slugging	32	74	0	106
4	Flow Instability	344	0	0	344
5	Rapid Productivity Loss	12	439	0	451
6	Quick PCK Restriction	6	215	0	221
7	Scaling in PCK	4	0	10	14
8	Hydrate in Prod. Line	3	81	0	84
	Total	1025	939	20	1984

The sensor readings encompassed information pertaining to both faulty and health states, along with observations during the transitional phase of deterioration. The 3W dataset used in Petrobras offshore naturally flowing wells contains the most commonly monitored variables, which are listed in table 3.3. Table 3.3 displays different tags representing measurements from various sensors, such as Pressure and Temperature. In the context of the examined event in this research, the analysis focused solely on the Pressure at the Temperature and Pressure Transducer (P-TPT) and the Temperature at TPT (T-TPT). These sensors, situated at the surface, were the primary variables of interest. Other measurements were either constant or absent, rendering them inconsequential for integration into the envisioned machine learning system.

Out of the recorded instances of failure in the SSSV tool, there were twenty-two cases in total. However, only twelve instances have complete data of the signal's entire range, known as run-to-failure data. These twelve instances are utilized to estimate the RUL of the SSSV tool. The measurements of the SSSV tool's degradation vary depending on the conditions of the well. Some well conditions result in rapid degradation, while others exhibit a gradual and consistent degradation process. The degradation process allows for the detection of the SSSV tool's behavior.

The initial dataset for the second failure event consists of three segments recorded throughout the lifespan of each SSSV. The first segment represents the normal behavior of the SSSV, where there is no observed deviation in the sensor measurements. This segment is labeled as class zero (0) in the dataset. The second segment is the transitional phase, where the SSSV gradually shifts from normal to faulty behavior. This is indicated by a sudden decrease in pressure and temperature measurements at the surface sensors, signifying a gradual closure of the SSSV. This segment is labeled as class one hundred and two (102) in the dataset. The third and final segment represents the faulty behavior of the SSSV, where the pressure and temperature measurements at the surface sensors are non-existent, indicating a complete closure of the SSSV and a cessation of oil

or gas production. This segment is labeled as class two (2) in the dataset. Figure 3.9 illustrates the normalized sensor measurements for one of the SSSVs during these three segments before preprocessing. The dataset also includes the Time to Failure (TTF) parameter, which describes the health condition of the SSSV. During the normal and transitional segments, TTF is fixed at "1" and linearly decreases until it reaches "0" during the faulty segment. The stable normal behavior and fully closed faulty behavior are disregarded, as the focus is on investigating the dynamics of spurious closure during the transitional degradation phase.

Table 3.3: Tags in the 3W dataset [3]

Name	Description	Unit
P-PDG	Pressure at the permanent downhole gauge (PDG)	Pa
P-TPT	**Pressure at temperature/pressure transducer**	**Pa**
T-TPT	**Temperature at temperature/pressure transducer**	**°C**
P-MON-CKP	Pressure upstream of production choke (CKP)	Pa
T-JUS-CKP	Temperature downstream of production choke (CKP)	°C
P-JUS-CKGL	Pressure downstream of gas lift choke (CKGL)	Pa
T-JUS-CKGL	Temperature downstream of gas lift choke (CKGL)	°C
QGL	Gas lift flow rate	m^3/s

3.4.2 Experiment Settings

The MATLAB R2021a software was used to simulate the proposed approach on a computer system running 64-bit Windows 10 and equipped with an Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz 2.40 GHz processor and 12.00 GB of RAM. To prevent overfitting of the dataset, a fivefold cross-validation technique was employed.

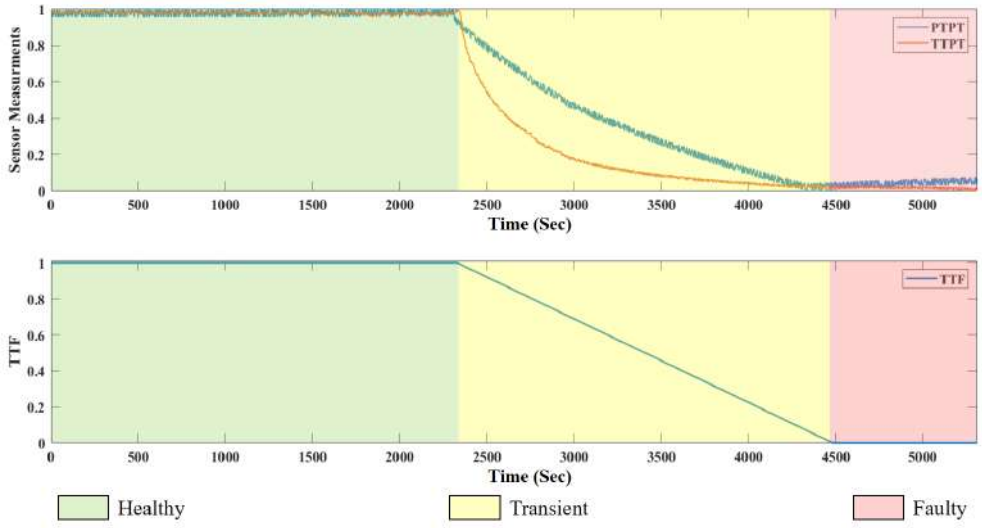


Figure 3.9: Representation in three phases, Healthy, Transient (Degradation), and Faulty. Plot 1: temperature & Pressure, Plot 2: Time to Failure

3.4.3 Evaluation

3.4.3.1 Metrics

The methodology in this study undergoes two evaluations. The initial evaluation takes place during the regression model training. It assesses the performance of the PRULs in comparison to the actual RUL using MSE as a metric, as shown in equation 3.13.

The second evaluation occurs when estimating the Time of Failure or Remaining Useful Life (RUL), in which the Absolute Estimation Error (AEE) is determined using equation 3.14. This error measurement quantifies the difference between the estimated value and the actual value. Relying solely on one estimation moment does not provide a comprehensive perspective, as maintenance decisions are based on multiple, if not all, previous RUL estimates. Thus, a Moving Average Absolute Estimation Error (MAAEE) is proposed and utilized to assess the RUL estimates. Early-stage estimations are excluded from the MAAEE calculation due to their lack of accuracy, and only estimations performed after 20 units of known data are considered.

$$MSE = \overline{(\widehat{PRUL} - RealRUL)^2} \tag{3.13}$$

$$AEE = |Estimated\ ToF - Real\ ToF| \tag{3.14}$$

$$MAAEE = \left| \frac{1}{n} \sum_{n}^{i=1} [Estimated\ ToF(i)] - Real\ ToF \right| \tag{3.15}$$

$$MAAEE = |Moving\ Average\ Estimated\ ToF - Real\ ToF| \tag{3.16}$$

3.4.3.2 Training and prediction

The proposed methodology is validated using actual measurements obtained from twelve wells equipped with an installed SSSV and TPT at the surface. Health indicators (HIs) are created and utilized to train and test the prediction system. The mean squared error (MSE) is employed as a performance metric for both the training and testing of the regression model, as well as the corrective feedback system. Figure 3.10 illustrates the loss function during the training and testing of the regression model over 200 iterations during the initial phase of the corrective feedback system.

The regression model undergoes 200 iterations for each iteration of the corrective feedback system. The corrective feedback system itself goes through 100 iterations, resulting in a total of 20,000 training and testing iterations for the neural network model. Figure 3.11 and figure 3.12 display the final MSE value achieved by the regression model for each of the 100 iterations of the correction system. These figures represent the training and testing loss of the corrective feedback system. When the cost function (MSE) is plotted on a linear scale, a significant decrease in value is observed between the first and second iteration. To provide better clarity, a logarithmically scaled MSE plot is used.

Figure 3.13 illustrates the 100 predicted RULs of the corrective feedback system in comparison to the actual RUL. Initially, the initial projected RUL is the least precise estimation. Nevertheless, when this estimation is incorporated as an additional feature for retraining the regression model, the precision of the PRULs improves.

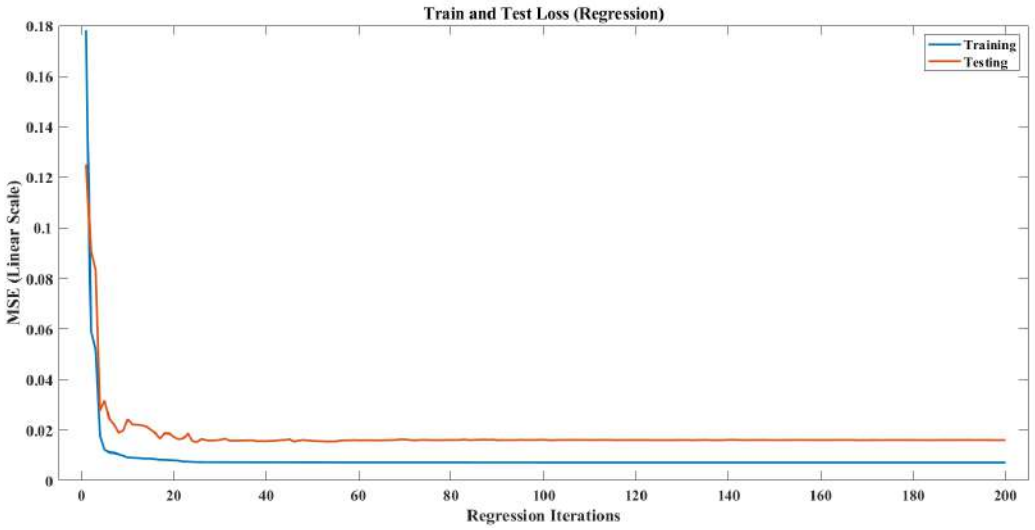


Figure 3.10: Loss functions of the proposed model.

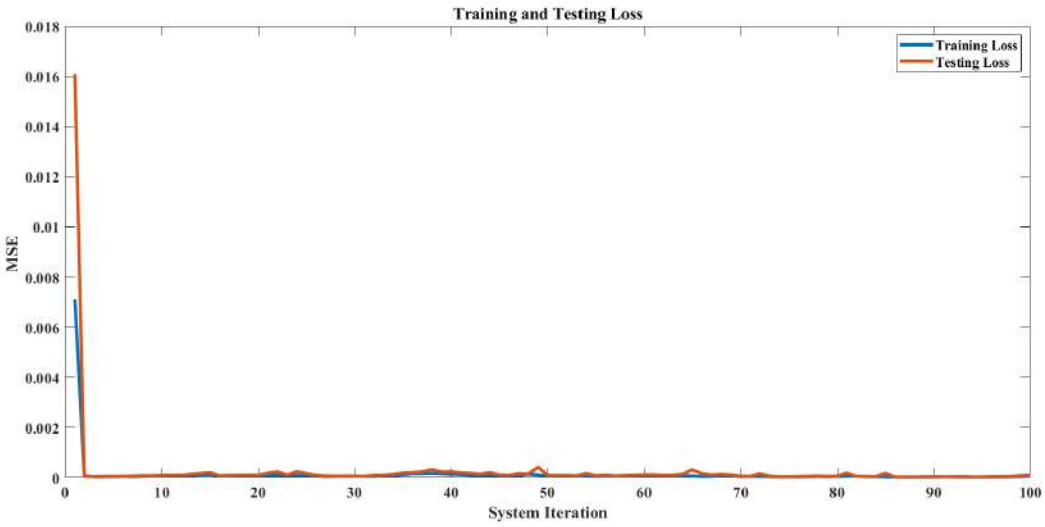


Figure 3.11: Linearly scaled loss functions of the Corrective feedback system.

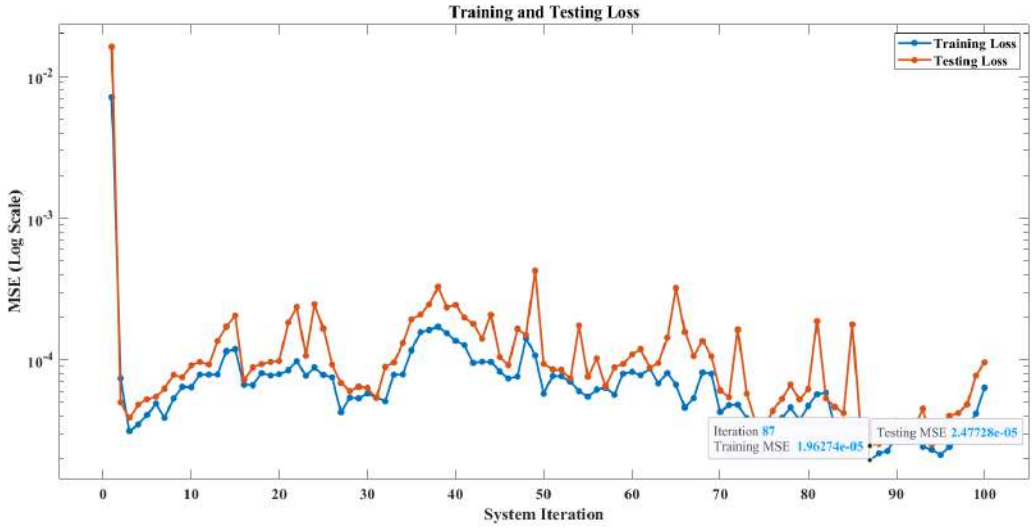


Figure 3.12: Logarithmically scaled loss functions of the Corrective feedback system.

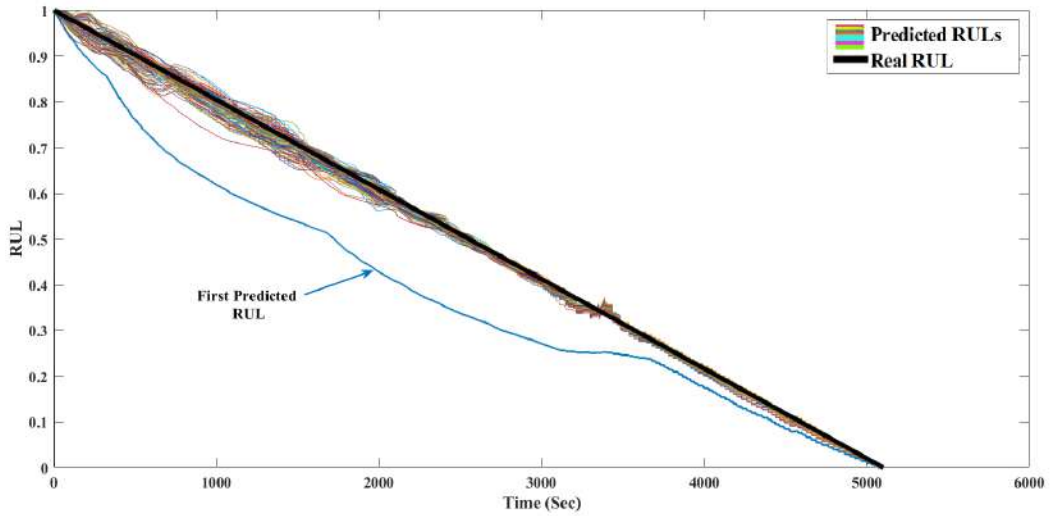


Figure 3.13: Predicted RULs VS real RUL.

An optimization algorithm is utilized to choose the most accurate prediction from a set of 100 system predictions. The optimum prediction, with a minimum MSE of 1.9627×10^{-5}, is found at the 87^{th} iteration of the system. During testing on new data, the global minimum MSE achieved is 2.4772×10^{-5}. Hence, the 87^{th} predicted remaining useful life (PRUL) is considered to be the Optimal Predicted RUL (OPRUL). Figure 3.14 illustrates the similarity between the true

RUL and the OPRUL. In order to further analyze the optimization results, the RUL estimation is performed by fitting a curve to the OPRUL.

Table 3.4: Training and testing MSE of predicted RULs

Iteration	1	2	...	**87**	...	99	100
PRUL	PRUL 1	PRUL 2	...	**PRUL 87**	...	PRUL 99	PRUL 100
Training MSE	0.0071	0.000074	...	**0.000019**	...	0.000041	0.000063
Testing MSE	0.0161	0.000051	...	**0.000024**	...	0.000077	0.000095

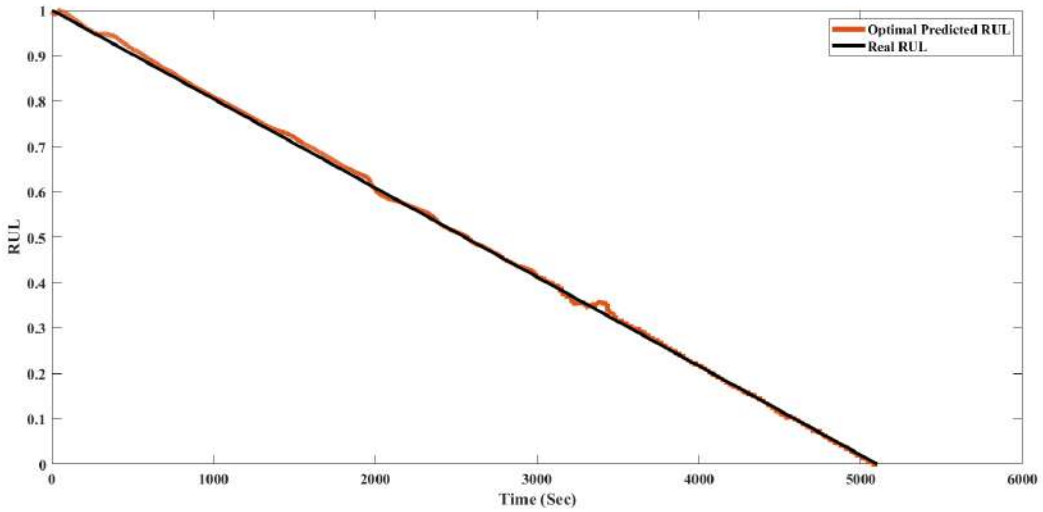


Figure 3.14: Final Predicted RUL VS Real RUL.

3.4.3.3 RUL and Time of Failure Estimation

The RUL is a time deteriorating function and is directly related to both the Time of Failure (ToF) and Real Actual Time as equation 3.1 depicted. Presenting the estimation results in terms of ToF helps provide a clearer understanding, allowing for conversion to estimated RUL values. As mentioned in section 3.3.5, segments of data are fed into the regression model. The ToF for the monitored SSSV is known in advance ($ToF = 5103s$). Estimations of the ToF are performed at

intervals of 10% of the full degradation time, approximately every 510 seconds. The accuracy of the estimation results is assessed using the AEE and MAAEE metrics.

Figures 3.15 and 3.16 present graphical representations of the ToF estimates for both long and short distances. The forecasting model employs different portions of the data, utilizing 20% for long-range estimates conducted at 1020 seconds, and 80% for close-range estimates conducted at 4082 seconds. The graph plots the PRUL against a linear fit, and the estimated ToF is determined by the point of intersection between the time axis and the Fitted Remaining Useful Life (FRUL) ($FRUL = 0$).

In a long-range estimation using only 20% of the available data, the ToF is estimated to be at the 5134^{th} second ($ToF_{20\%} = 5134s$), which corresponds to an absolute error of 40 seconds ($AEE_{20\%} = 40s$). On the other hand, in a short-range estimation using 80% of the available data, the ToF is estimated to be at the 5103^{rd} second ($ToF_{80\%} = 5103s$), resulting in no absolute error ($AEE_{80\%} = 0s$).

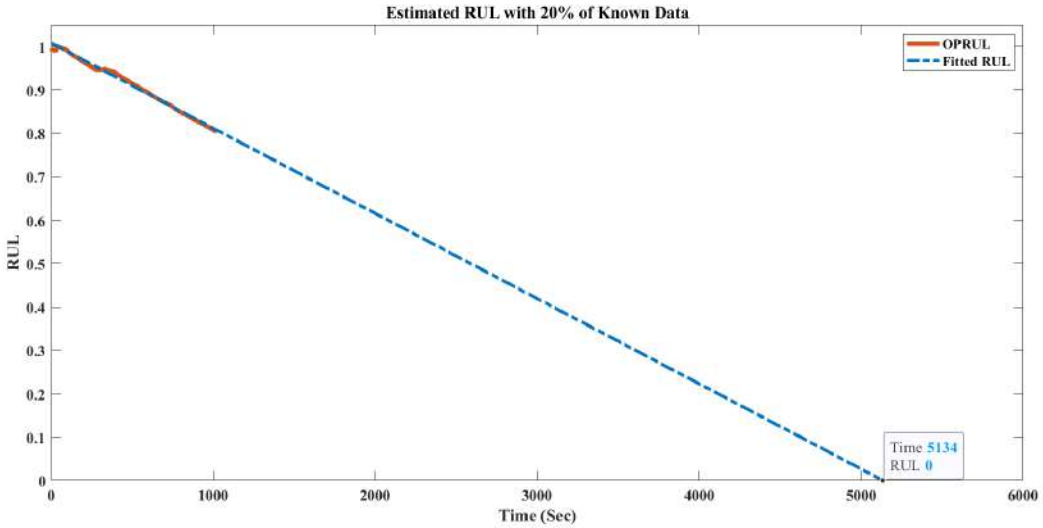


Figure 3.15: Fitting of OPRUL at an ealy stage.

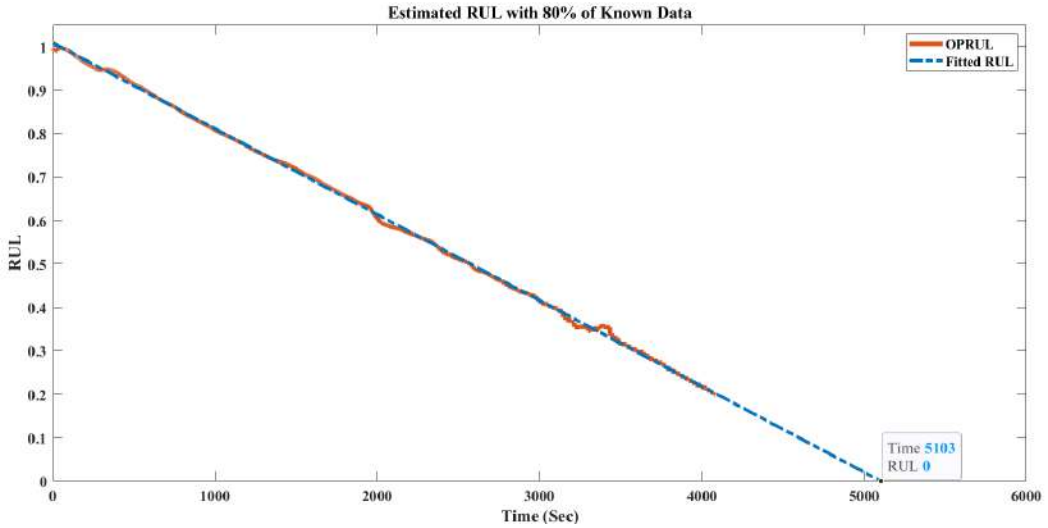


Figure 3.16: Fitting of OPRUL at a late stage.

AEE values are computed and graphed in figure 3.17 for each estimation instance. Initially, higher AEE values are observed due to limited data availability and the development of system dynamics. The lowest AEE value is achieved when 80% of the data is fed to the prediction model ($AEE_{80\%} = 0s$). However, during the last few estimates, a parabolic function behavior is observed where the AEE is higher than the previously recorded minimum, despite the model being provided with more data. This discrepancy occurs because the AEE calculation only considers instantaneous estimates and does not accumulate previous estimates. To account for previous estimates, a moving average type of error (MAAEE) is introduced (Figure 3.18).Nevertheless, not all prior estimates are considered, given that initial estimates exhibit greater inaccuracies owing to restricted data inputs in the prediction system. In this investigation, estimations are deemed reliable only when a minimum of 20% of the data is provided to the prediction system. The lowest MAAEE value is found when 90% of the data is fed into the model, with $MAAEE_{90\%} = 19s$ (Table 3.5).

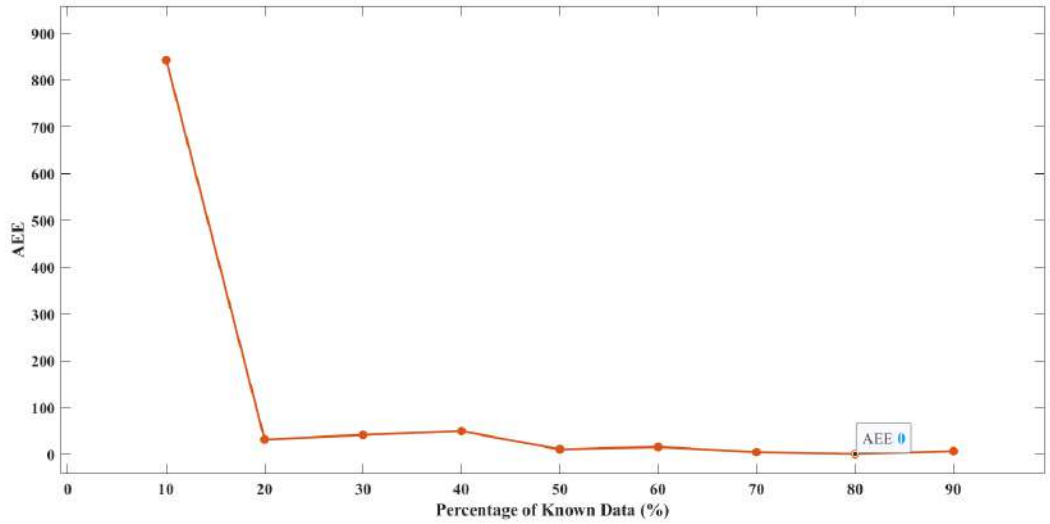


Figure 3.17: The proposed approach AEE.

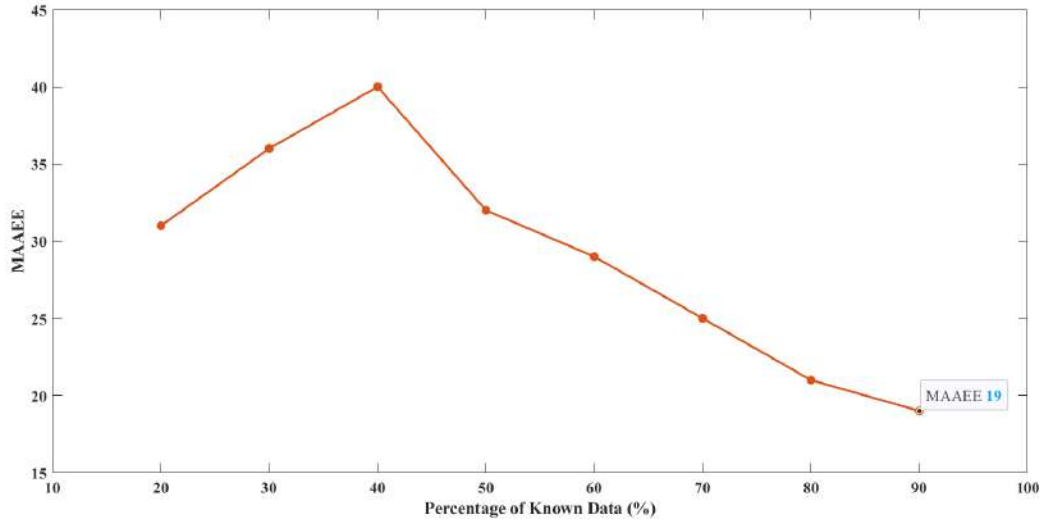


Figure 3.18: The proposed approach MAAEE.

3.5 Discussion

The results underscore the importance of incorporating health indicators (HIs) in the pre-processing phase for predicting the Remaining Useful Life of proposed validation system. The current study outperforms a previous study [3] which did

Table 3.5: Long, medium, and short-range Estimation results

Prediction Type	Long Range			Medium Range			Short Range		
Known Data	10%	20%	30%	40%	50%	60%	70%	**80%**	**90%**
Real-Time	510	1021	1531	2041	2552	3062	3572	4082	4593
Real ToF	5103	5103	5103	5103	5103	5103	5103	5103	5103
Estimated ToF	5945	5134	5144	5152	5093	5088	5099	**5103**	5097
Real RUL	4593	4082	3572	3062	2552	2041	1531	1021	510
Estimated RUL	5435	4113	3613	3111	2542	2026	1527	1021	504
AEE	842	31	41	49	10	15	4	**0**	6
MAAEE	N/A	31	36	40	32	29	25	21	**19**

not incorporate HIs, demonstrating the favorable influence of HIs, particularly when utilized to mitigate noise and enhance the accuracy of prediction outcomes.

The most commonly utilized machine learning models for estimating RUL and predicting maintenance needs are ANN, Decision Trees (DT), and Support Vector Machines (SVM) (Section 3.3.3). Variations of multiple regression models were trained on the dataset, with the Mean Squared Error (MSE) results depicted in table 3.1. The LSVM regression model yielded an MSE of 0.266, DTs achieved an MSE of 0.087, and the NN family achieved an MSE of 0.084. Comparing these values with the final training results of the proposed approach (table 3.4), a significant improvement is evident. The proposed approach achieved a training MSE of 0.000019 at the 87^{th} iteration of the corrective feedback system. These results highlight the superiority of the proposed approach over traditional methods commonly employed for RUL estimation.

The selected number system's iterations is, in this research, 100 iterations are conducted, the most accurate prediction is found at the 87^{th} iteration. A clear convergence of the MSE during both the training and testing phases of the corrective feedback system is demonstrated, particularly when graphed on a linear scale, there is a noticeable improvement in accuracy during the second iteration.

From the second iteration onwards the MSE values appear to be the same, but when plotted on a logarithmic scale, the discrepancy in their values becomes apparent.

An inquiry may arise regarding the optimal number of iterations necessary to reach the OPRUL. It is recommended to implement the proposed corrective feedback system for a minimum of one iteration to assess a substantial enhancement in prediction accuracy. Nevertheless, achieving the ultimate OPRUL depends on the extent of computational resources allocated to the system.

The interpretation of estimation results initially involves the use of AEE to focus solely on the discrepancy between the estimated and actual ToF. This metric serves as a suitable measure for immediate evaluation. However, it is limited in scope as it does not consider previous estimates. On the other hand, the proposed evaluation metric, known as MAAEE, takes into account previous estimates, thus offering a more comprehensive understanding of the estimation approach.

As outlined in section 3.4.1, the scope of this study was limited to the transient (degradation) aspect of the SSSV tool. Consequently, the RUL of the tool is relatively short compared to other systems. Although the SSSV is known for its durability and ability to function without issues for extended periods, any failure that does occur necessitates costly maintenance and a comprehensive well WorkOver. The findings of the study demonstrated a promising level of accuracy in estimating the RUL using a prediction system based on neural network, which offered a satisfactory maintenance timeframe.

The proposed approach demonstrates effective performance when utilized in conjunction with Neural networks. However, if applied to other machine learning models like linear regression, the PRUL convergence is observed to occur in the second iteration without any significant enhancements in outcomes. Consequently, it is advisable to employ Neural Networks for optimal results. Additionally, incorporating an imprecise prediction as a feature can result in unsatisfactory outcomes.

This research presents empirical results derived from a singular case study,

focusing on the SSSV system that exhibits distinct degradation dynamics, and Remaining Useful Life. The suggested methodology has shown significant effectiveness in forecasting the RUL of the examined system. As a result, it is reasonable to assume that with minor adaptations and modifications to the system, the methodology can effectively be applied to other applications. Consequently, future research is recommended to implement the proposed approach in diverse real-life applications.

3.6 Conclusion

This chapter introduces a machine-learning approach for predicting maintenance requirements in production lines. The research investigates the feasibility of accurately forecasting the Remaining Useful Life (RUL) of systems with limited data by employing a corrective feedback mechanism. The utilization of regression and preprocessing techniques facilitates the creation of Health Indicators (HIs) that enhance the understanding of the condition of subsurface safety valves. The proposed model incorporates a feedback loop wherein predicted RULs from testing are used to retrain the model for making new predictions. An optimization method is applied to determine the optimal prediction, and a curve-fitting function is employed to estimate the time of failure during the tool's service. Accurate predictions are achieved in the experimentation of the approach allowing the operator to minimize downtime.

Chapter 4

Data-Driven Digital Twin Based on Multi-Target Regression

4.1 Introduction

Accurate prediction and forecasting of multiple parameters is of utmost importance in the construction of oil and gas wells, particularly during the drilling phase. The development of a digital replica of the drilling operation, encompassing various activities such as drilling, completion, transport, and fleet management, is essential for minimizing failures, safety risks, non-productive time (NPT), troubleshooting costs, and potential human casualties. In smart drilling systems, a Measurement While Drilling (MWD) system is typically situated a considerable distance away from the drill bit, resulting in delayed measurement transmission to the drilling process and ultimately leading to suboptimal decision-making.

This chapter introduces an approach for developing a data-driven digital twin for predictive analysis and forecasting of various drilling parameters in an actual directional drilling scenario. The proposed prediction model utilizes a branched deep neural network, which integrates Dense and LSTM layers. To validate the approach, the Volve dataset, comprising real data collected during directional drilling operations, is employed. The experimentation also incorporates the concept of incremental learning, wherein the model is trained using small subsets of

data to monitor its performance over time.

The chapter is structured in the following manner: section 4.2 outlines the case study and the dataset used to validate the approach. Section 4.3 delves into the proposed approach. Section 4.4 provides a comprehensive overview of the procedures employed during experimentation. Section 4.5 presents the results obtained from the implementation of the proposed approach. Section 4.6 critically analyzes these findings and compares them to existing approaches in the field. Finally, section 4.7 concludes the chapter by emphasizing the significant discoveries made.

4.2 Case Study

4.2.1 Dataset

Equinor has released the Volve dataset, which pertains to the Volve oil field situated in the southern section of the North Sea near the Norwegian coastline. The field operated between 2008 and 2016. This dataset contains a range of information encompassing geoscience, production, reservoir modeling, and drilling. Despite the dataset's reliability and accessibility, it has yet to gain popularity within the research community due to the absence of any preprocessing. The file formatting of the drilling logs necessitates separate processing to be converted from WITSML files. Efforts have been made to convert real-time drilling logs into CSV files [75], thereby facilitating the handling of the data for data science purposes.

The drilling data contains numerous files of measurements from different wells, the data is logged in two methods, depth-based real-time logs, and time-based real-time logs; in this study, depth-based real-time logs are used due to long data gaps and the small file size compared to the time-based real-time logs. The well *F9A* is selected for this case study, this well is chosen as it has a relatively long section of the well that does not contain data issues in the depth-based logs. To put in perspective, the difference between the time-based and the depth-based logs, the

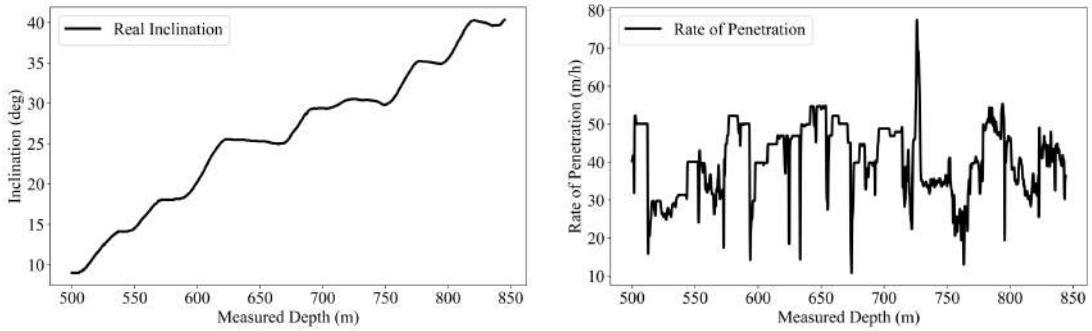
selected well *(F9A)* file size is: 417,882KB for the time-based, and 5,466KB for the depth-based, this elucidates that the time-based file contains long strides of data gaps and/or repeated data points.

4.2.2 Case Study

The *F9A* well has a total measured depth (MD) of 1206 meters; the well is drilled in three runs (a run in the oil and gas drilling field is referred to as the number of times the BHA is changed or replaced), the first run starts from 0 meters of MD until 273 meters, this is the vertically drilled portion of the well, the BHA used in this portion is a drilling system specialized in drilling vertical wells, the first BHA is tripped-out of the well and replaced with a bent sub motor BHA, specialized in inclined drilling, the second run starts from 273 meters of MD until 847 meters, in this portion the measured inclination is between [0 - 41] degrees, the BHA of inclined drilling is tripped out, and replaced with a BHA specialized in horizontal drilling, the third run is between [880 - 1206] meters, the change in inclination is between [48 - 58] degrees.

The discrepancies in measured depth and inclination between the second and third runs can be attributed to a change in drilling equipment. This change can result from variations in calibration and measurement techniques or simply different lengths of bottom hole assemblies (BHAs). The system responsible for collecting real-time measurements during drilling operations, known as Measurement While Drilling (MWD), is offered by companies such as Schlumberger, Baker Hughes, and Halliburton. These companies compete to secure contracts from clients who own oil or gas reservoirs, leading to situations where different technologies from various companies are utilized, resulting in limited synergy in oil and gas operations. From a data science perspective, this issue creates significant gaps in the data, which can impact the performance of machine learning algorithms. To prevent biased and misleading results, this study avoids analyzing portions of the well with large data gaps that are difficult to recover. However, smaller data gaps can be addressed using traditional data imputation techniques (Section 4.3.1.3).

Therefore, the focus of this study is on the section between 500 and 845 meters of measured depth, ensuring that both the targets (ROP and inclination) and the features do not have substantial data gaps (Figure 4.1).



(a) (b)

Figure 4.1: Case study well inclination profile (a), and the Rate of Penetration (b)

4.3 The Proposed Approach

The digital twinning process generally consists of three main parts [29, 30, 31]:

- A real physical system with its associated sensors for data collection.
- A virtual representation of the integrated models to mimic the system behavior for performance analysis and optimization.
- The final part is a data bus to allow communication and information exchange between the real system and the Digital Twin.

This study focuses mainly on the second part, which includes numerous steps:

a) The first step is to identify the outputs of the Digital Twin, for instance, in our case, the outputs are the predicted values of the ROP and Inclinations.

b) The next step is to identify the Digital Twin performance metrics; the objective of the twin is to predict the future values of the ROP and inclination

ahead of the drill bit, since we have measurement of physical parameters, the Mean Absolute Error (MAE) is used as a performance indicator.

c) Once Digital Twin outputs and performance metrics are established, the subsequent phase involves the implementation of data-driven techniques (If the application is data-driven) namely preprocessing and machine learning (Regression model design).

d) Subsequently, based on the outputs and inputs of various models, the necessary data streams are identified (assessing the influence of different inputs on targets)

e) The twinning system is used for prediction, performance enhancement, and scenario analysis.

A detailed, step-by-step exploration of the process outlined in Part C, where data preprocessing, machine learning, and model design are explored in this section.

The process of developing a data-driven digital twin used for directional drilling parameters prediction is described, the data used in this study is imported from a real dataset (Section 4.2); data analysis, imputation, resampling, and scaling are conducted; data splitting for machine learning purposed is performed; multi-target regression models are developed. The work carried out in this section is done using Python, including multiple libraries: Numpy, Pandas, Matplotlib, Scikit Learn, and Keras from TensorFlow; each library with its designated purpose (vector manipulation, data manipulation, data visualization, machine learning, and predictive data analysis, deep learning).

4.3.1 Data Preparation

4.3.1.1 Raw Dataset Import

The data of the *'F9A'* well are imported from the CSV file *'Norway-NA-15_47_9-F-9 A depth.csv'*, all the data contained in this file are real-time data acquired

from the MWD system during a directional drilling operation, the total measured parameters during this operation is 115, evidently, not all the measured parameters are useful for machine learning studies due to existing gaps and low variance; a total of 16670 instances (measurements) are available in the depth-based file. As stated in section 4.2, only the section from 500 to 845 meters of MD is considered; resulting in approximately 8104 usable instances.

4.3.1.2 Data gaps and variance analysis

In machine learning applications, large data gaps and low variance variables either harm or do not contribute to the learning process; data gaps (missing data) occur in real datasets due to various reasons; short data gaps occur generally in the presence of 'uneven or out of sync logging' [76]; while drilling equipment change, well radius change or operation change may lead to long data gaps occurrence. Short data gaps can be recovered using data imputation techniques (Section 4.3.1.3), whereas long data gaps are not easily recoverable.

A data gap analysis is performed using equation 4.1, where the percentage of Not-a-Number (NaN) values in each measured parameter (column) is calculated against the total number of instances (measurements), if the percentage is above 95% the column is removed (dropped) from the imported data. Additionally, columns with low variance are identified using equation 4.2; parameters with variance below 0.1 are equally removed. After the performance analysis the number of useful parameters is decreased to 53, among these parameters: are the two targets (*'Rate of Penetration m/h'* and *'MWD Continuous Inclination dega'*), and the *'Measured Depth m'* (MD) that is used as an index for the measurements instances, leaving 50 drilling parameters to be used as features for the prediction model.

$$\frac{(\sum NaN) \times 100}{\text{Total Number of Samples}} > 95\% \tag{4.1}$$

$$\sigma^2 = \frac{\sum \left(\text{ Paramater }_i - \overline{\text{ Paramater }} \right)}{N} < 0.1 \tag{4.2}$$

4.3.1.3 Data Imputation

Several data imputation techniques exist in the literature (forward/backward filling, interpolation, regression, mean filling... etc.). Nevertheless, certain methodologies may not be appropriate for the process of filling missing values in data logs of drilling, or may be unfeasible for datasets that possess missing values across the entirety of the dataset, such as regression.[76]. A combination of linear interpolation, forward filling, and backward filling is used in this study; the first imputation technique implemented is linear interpolation, as it establishes a linear relation between two existing values and uses it to fill the missing data between the known values. If the imputation phase starts with backward filling, this means data is fed backward even if it is not the ground-truth data at that instant (one cannot assume that an event happened before its occurrence); if the imputation starts with forward filling on the other hand, the data will contain a lot of singular transitions. For example, a sequence of data has the following values considered (NaN, 56, NaN, 58, NaN, 60, NaN, 62, Nan), if the backward filling is applied first, the sequence will become (56, 56, 58, 60, 60, 62, 62, NaN), feeding real-time data backward is considered, in some cases, as data corruption; on the contrary, if interpolation is applied first, the sequence become (NaN, 56, 57, 58, 59, 60, 61, 62, NaN), filling the first and last NaN data is done with forward and backward filling, the final sequence is (56, 56, 57, 58, 59, 60, 61, 62, 62). In this study, a similar approach is applied, linear interpolation is used as the first phase of data imputation, followed by forward and backward filling to achieve fully imputed data.

4.3.1.4 Data Resampling and Interpolation

Resampling drilling logs is highly advised before using temporal machine learning models with Recurrent Neural Networks (RNN). This is required due to the variance in the data's sampling rate throughout the drilling operation, resulting in differences in the signal's duration. As for the data's resolution, i.e. resampling, the interval between successive samples, either in terms of time or distance

(depth), can be regulated.

The assumption made when using an RNN (LSTM in this study) is that the time intervals between consecutive data points are uniformly distributed, however, this is not the case with the imported data. When investigating the distance (depth) between each measurement, a fluctuation in values between 0.002 meters and 0.15 meters is found, it is very common to have asynchronous readings in real-time drilling longs; additionally, when calculating the mean distance between each measurement the mean value found is 0.048 meters, Henceforth, 1 meter contains 20.83 measurements. The primary goal of the resampling phase is to modify the dataset in order to recalculate the subsequent sensor readings based on predetermined index values. Several approaches exist to perform data resampling: K-nearest Neighbor (KNN) and Fixed Radius Neighbor (FRN) regression are explored by Andrzej T. Tunkiel [76], cubic splines, and windowed resampling are also viable; however, the investigation of the optimal approach for data resampling is out of the scope of this research. A simple yet effective approach is linear interpolation; it is an effective technique implemented when data points are sampled at irregular intervals and the goal is to regulate the sampling interval. The desired interval after conducting the resampling is 0.05 meters between each measurement, Henceforth, 1 meter contains 20 measurements. Regulated sampling intervals grant control over the desired training, validation, and test distances and depths.

4.3.1.5 Features and Targets

As mentioned in section 4.3, a crucial part of the development of a digital twin is to appropriately identify the inputs and outputs of the system; the two outputs of the twin are the predicted values of the ROP and the forecasted values of the Inclination. The inputs on the other hand are quite intricate, given that two different types of machine learning models are applied simultaneously (Section 4.3); a Multi-Layer Perception (MLP) network branch is applied mainly to predict the ROP using 50 different drilling parameters as features; concurrently, an LSTM

network branch is applied to forecast the Inclination values using previous inclination measurements. Therefore, two types of inputs are fed to the digital twin, 50 drilling parameters scaled to (0,1), and previous instances of inclination values.

Scaled inputs are typically optimal for machine learning algorithms, this is the case with the 50 drilling parameters fed to the MLP network, utilized mainly for ROP prediction, a simple min-max normalizer is applied to keep the input data scaled to (0, 1); the scaling process is conducted using the following equation:

$$\text{Paramater}_{Norm} = \frac{\text{Paramater} - \text{Paramater}_{min}}{\text{Paramater}_{max} - \text{Paramater}_{min}} \tag{4.3}$$

Only the 50 drilling parameters that are subsequently used as features go through this scaling process. The Inclination instances are used without scaling.

Preparing the Inclination input data is conducted by converting the forecasting problem at hand into a supervised learning problem; this approach is common in machine learning for time series forecasting. A reconstruction of the time series data to a set of targets and features is performed; this is undertaken by creating lags from the Inclination sequence where each column represents an instance ($t - N$, ..., $t-3$, $t-2$, $t-1$), the data measured in these instances is used to predict the instance at (t). Assuming the following inclination sequence (in degrees) is acquired from the MWD system (9, 10, 11, 12, 13, 14), supposing we have a learning window of 4 lags ($N = 4$), the input instances (features) are (9, 10, 11, 12), and the target is (13); sliding the learning window by one step updates the input instances to (10, 11, 12, 13), and the new target is (14). This process is repeated until the entire Inclination data is converted to lags used as features, and measurements at time (t) used as targets.

4.3.2 Machine Learning Model

4.3.2.1 Training, Validation, and Testing Data Splitting

A crucial aspect in the development of a machine learning model, is the created structure of the training and testing datasets, especially when it comes to drilling,

due to the time series nature of the drilling logs. The common train/test split is the Random Split, where random samples (measurements) from the dataset are used for training and the rest for testing; this type of data splitting cannot be used for prediction models in drilling, due to the inflation of the testing results caused by spurious correlations [43].

Three scenarios of model validations are proposed in the literature [12]:

- All for One: this scenario implicates the use of data acquired from multiple wells for training and validation, and leaving one well for testing; the issue associated with this scenario is the potential bias of the training dataset, where the acquired data is from a certain region and a certain geology, the prediction model will perform accurately if the well used for testing is near the region of the wells used in the training; on the contrary, the prediction model is ineffective in different regions.

- One for All: this scenario suggests the use of one well for training and multiple wells for testing; this is anticipated to result in under-fitting, and high prediction error, due to the insufficient training data.

- Continuous learning: This scenario contributes to the real-time prediction approach, where each well is evaluated separately using incremental sequential data splitting, eliminating the bias problem that the All for One scenario has, and the under-fitting issue of the One for All approach. An argument can be made concerning the small size of the training dataset during the initial drilling; it is a natural phenomenon for a machine learning algorithm to learn over time, additionally, it is plausible to start the training and prediction after acquiring an adequate amount of data.

The correct data splitting approach related to drilling is an incremental sequential data split (Continuous Learning); where the first meters of the drilled well are used for training and the rest for testing, applying this split strategy simulates the process of training and predicting while drilling the well in real-time.

4.3.2.2 Model Design

The model consists of two branches, an RNN branch using the LSTM layer, and an MLP branch. Several dropout layers are implemented to prevent over-fitting, and the final dense layer of each branch is concatenated such that the model has two input layers and two target variables. All used layers are imported from the Keras library [77]. The transformed input Inclination data is fed into the LSTM branch. ROP as a target alongside the 50 well drilling parameters are fed into the MLP branch (Figure 4.2). Training and validation are performed for 40 epochs, with a callback for the model with the best validation (best model during the training session), saved at a checkpoint during the training; the best model is then used for testing. The Mean Squared Error (MSE) is chosen as the loss function. The loss functions for both training and validation data throughout a training session are visible in Figure 4.3.

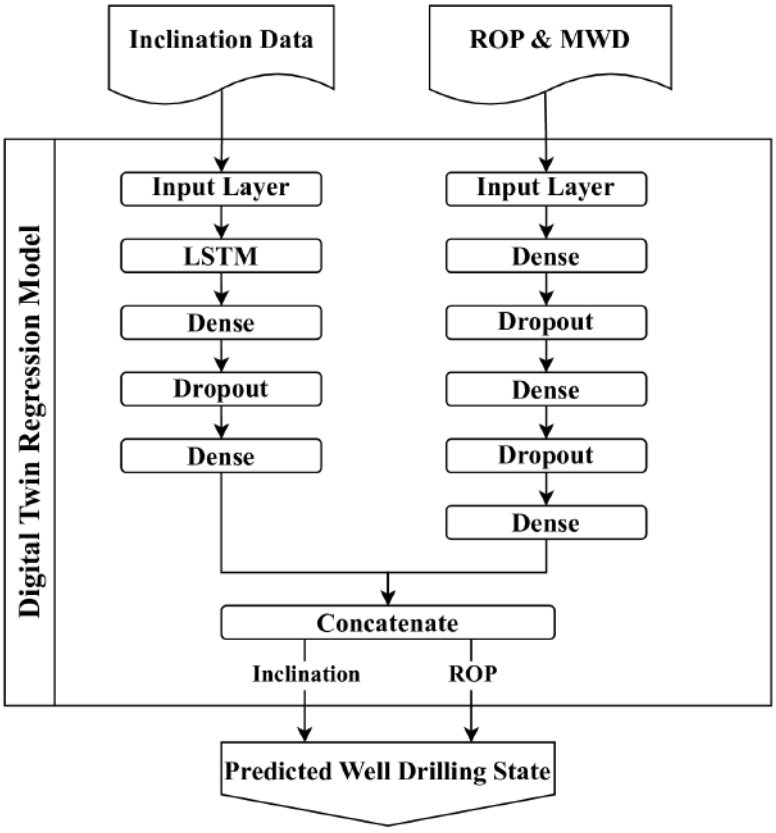


Figure 4.2: Neural Network Architecture.

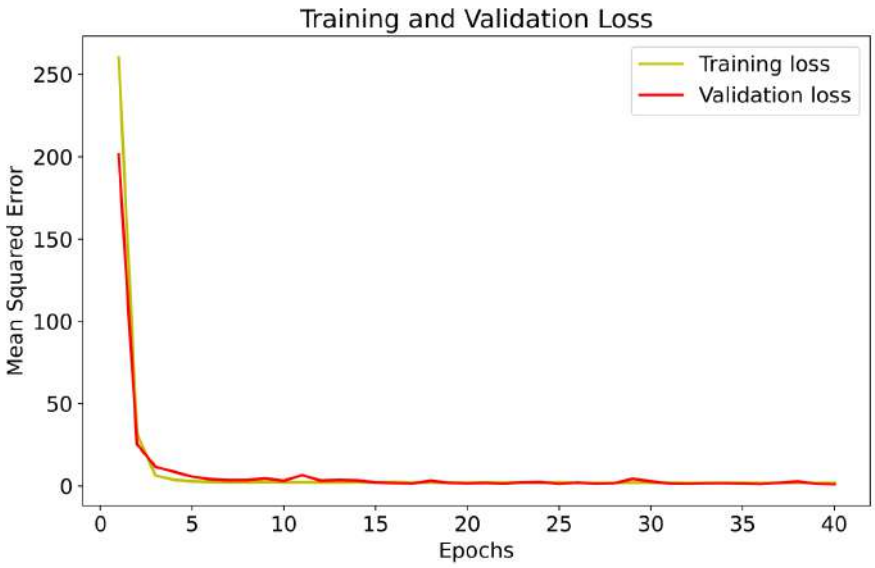


Figure 4.3: Training and validation Loss for a single run.

4.4 Experimentation

4.4.1 Incremental Learning

The proposed experimental setup is illustrated in Figure 4.4; the first iteration starts when the drilling reaches 600 meters of MD (100 meters after the measurement started), the multi-target regression model is trained with data from 500 to 600 meters, the validation dataset is exactly after the training dataset to achieve best results [43], the validation dataset has a fixed length of 40 meters after the final measurement of the training dataset. Once the training of the regression model is conducted for 40 epochs, the model with the best validation (minimum validation MSE) is saved and tested on new unseen data. The testing dataset has a fixed length of 25 meters after the validation dataset. ROP and Inclination are predicted and compared to true values; the Mean Absolute Error (MAE) is used for evaluation. The MD is increased by 20 meters and the process is repeated, this increase in MD simulates the continuous training and prediction each 20 meters while the well is being drilled. The training interval of 20 meters can be decreased or increased as desired, however, decreasing the interval requires longer training

time; for the arbitrarily chosen interval (20 meters), the depth of 845 meters is reached in 10 iterations, in the final iteration the model is trained from 500 to 780 meter, validated from 780 to 820 meters, and tested on unseen data from 820 to 845 meters.

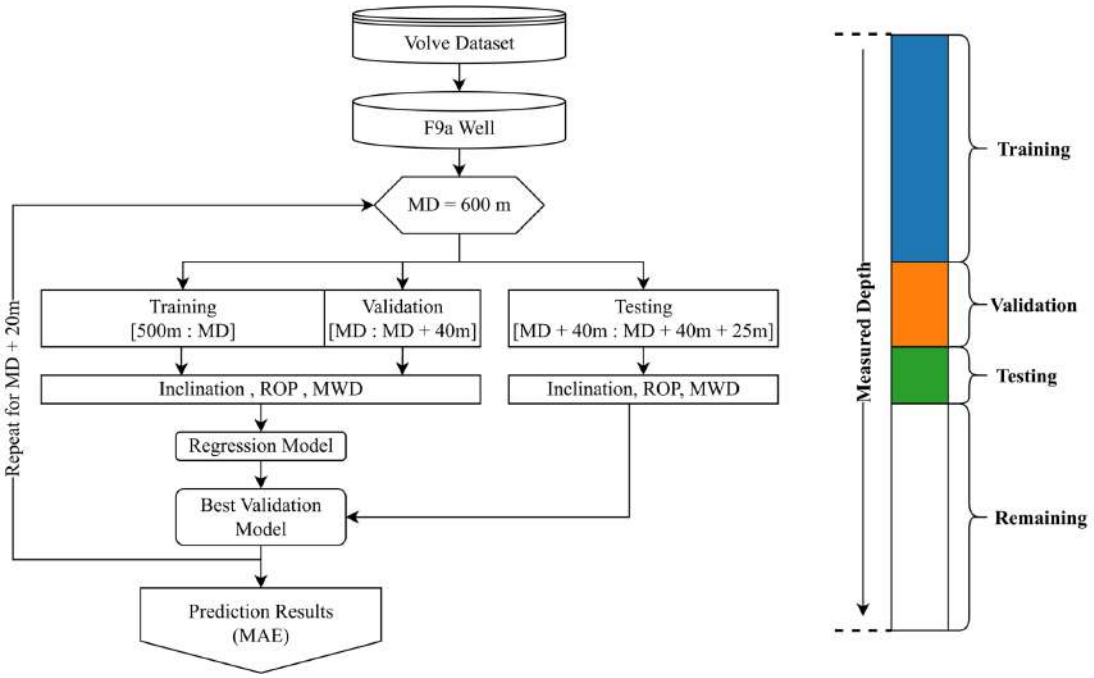


Figure 4.4: Incremental learning process workflow.

4.4.2 Evaluation Metrics and Training Environment

During the training of the multi-target regression model, the Mean Squared Error (MSE) is used as the loss function (Equation 4.4); however, the prediction results are evaluated with the Mean Absolute Error (MAE) metric, due to its ease of interpretability (Equation 4.5).

$$MSE = \overline{(\ \widehat{Multi\ Target}\ -\ Multi\ Target\ )} \tag{4.4}$$

$$MAE = \overline{(\ \widehat{Target}\ -\ Target\ )} \tag{4.5}$$

The experiments are simulated on both local and hosted environments; the local simulation uses TensorFlow 2.10.0 with integrated Keras library and Python

3.9.18; Model training was performed on 11th Gen Intel(R) Core(TM) i7-1185G7 @ 3.00GHz 1.80 GHz, 32 GB of RAM, the training was CPU based and it required 6-8 hours per experiment; however, the hosted environment leverages the power of the T4 GPU provided by Google Colaboratory allowing for faster training time; the experiment is conducted only in 25-30 min.

4.5 Results

4.5.1 Single Run Results

Results from a single run can visualized by plotting the true ROP and Inclination values alongside the predicted results during the training, validations, and most importantly the testing phase. This provides an interpretable illustration of the practical results, Figure 4.5 displays the results of an early experiment conducted at 640 meters of MD, where the data from 500 to 640 meters is used for training and validation to provide predicted inclination and ROP parameter values from 640 to 665 meters of MD. Observe how both the predicted inclination and ROP adhere to the same pattern, and have close values to the real parameter's values. Figure 4.5 highlights the predicted results (test results) allowing for clear insight. To visualize the working principle of the proposed data-driven digital twin, Figure 4.6 is created. The digital twin takes the previous inclination values and MWD parameters as inputs to provide the predicted ROP and Inclination values. It is essential to distinguish between the intervals of the provided inclination and MWD parameters data as inputs, the LSTM branch needs previous sequential data to predict the upcoming inclination values; the MLP branch on the other hand needs to be trained on previous MWD parameters and link each instance to the corresponding ROP values to find the relation between the independent variables and the dependent variable, additionally, this branch needs the present MWD parameters to predict the future ROP values. Sequential inclination values from 500 to 640 meters of MD are provided to the twin, in addition to the MWD parameters from 500 to 665 meters of MDs, this yields the prediction of Inclination

and ROP from 640 to 665 meters of MD.

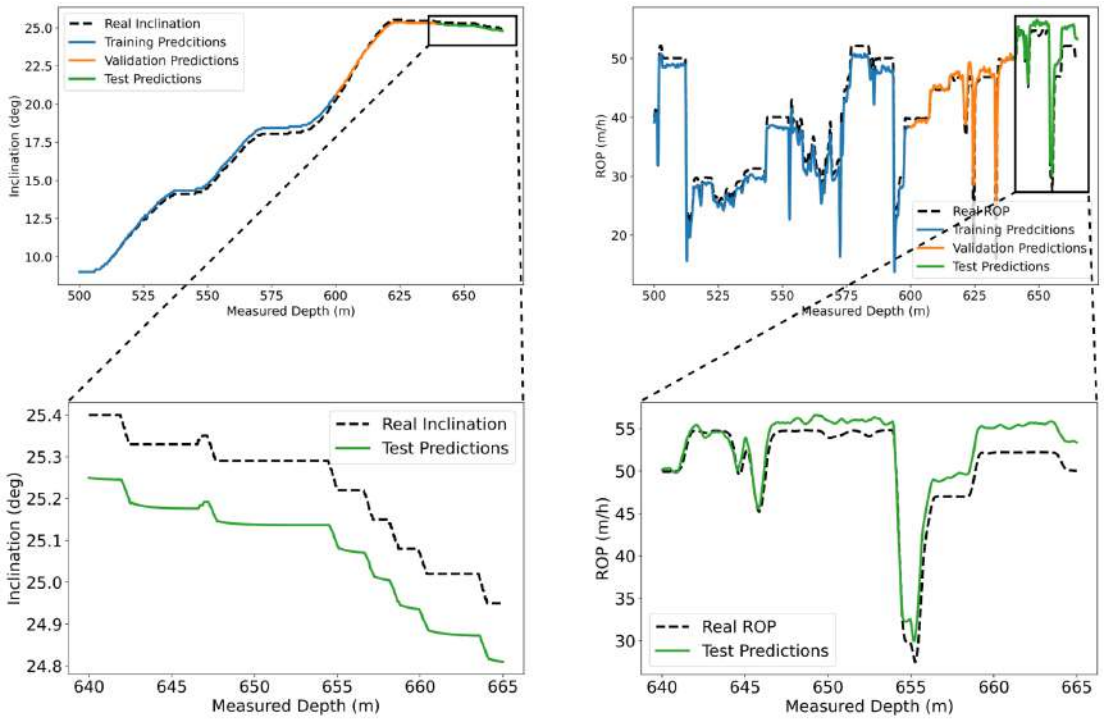


Figure 4.5: Real Inclinations and ROP versus performed predictions with emphasis on the testing results at an early stage.

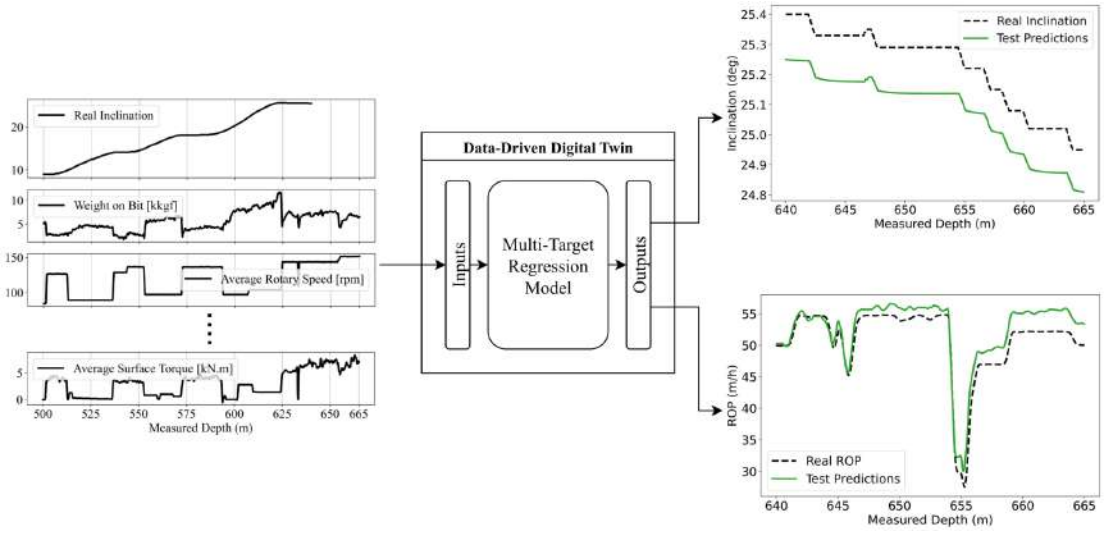


Figure 4.6: Digital twin workflow process: received inputs, high-level system's block diagram, and predicted outputs.

4.5.2 Experimental Results

While training the developed model, it is noticeable that the results fluctuate between good and bad, even if the training was conducted on the same data. This arises from the stochastic characteristics of the training process of neural networks (deep neural networks as well); for this reason, the experimentation is conducted 10 times aiming to compensate for the stochastic aspect of the results.

A scatter plot is constructed from the 10 conducted experiments showing the MAE values of each experiment throughout the drilling of the well; as stated in section 4.4, predictions take place every 20 meters of MD. For every prediction step or iteration, 10 scattered dots are plotted representing the 10 conducted experiments on a particular depth. The mean MAE values of each step are calculated for all 10 experiments and plotted with a bold black line. The mean of all mean MAE values throughout the drilling of the well is calculated at the end of the experiments and plotted in dashed lines; this mean of mean values provides an overall performance indicator for all the conducted experiments during the whole drilling operation.

Figure 4.7 presents the MAE values for the multi-target predictions, individual MAE values alternate between 3.6 and 0.33. Mean MAE values are between 2.63 and 0.52. Relatively high MAE values are witnessed at the beginning of the drilling operation, which is very logical due to the small training dataset; an apparent spike is noticed around the depth of 700 meters for all 10 experiments indicating a hard-to-predict zone, followed by an easy to predict zone at 740 meters. Fairly small and stable MAE values are recorded at the last portion of the well due to the substantial amount of the provided data during the training of the model. The total recorded mean in this study for the multi-target prediction is 1.45.

Although presenting multi-target prediction results provides insight into the general performance of the model; however, diving into each one of the targets is more profound; the MAE values discussed earlier remain devoid of a specific measurement unit due to their association with multiple parameters. Figure 4.8

exhibits MAE values for the inclination target predictions, individual MAE values demonstrate a stable performance for all experiments throughout the well drilling aside from three instances around the 700 meters of depth that the model struggles to predict with relative accuracy. Mean MAE values oscillate between 0.9 degrees and 0.05 degrees, which is very accurate. The total observed mean MAE for the inclination target throughout this study is 0.5 degrees. MAE values for the ROP target are depicted in figure 4.9, showing a similar trend to the multi-target results, with relatively high MAE values at the beginning and low values when enough data is supplied; the MAE values vary from 6.1 to 0.62 m/h. The calculated mean for MAE values is between 4.38 and 1 m/h. the total recorded mean MAE for the ROP target in this study is 2.39 m/h.

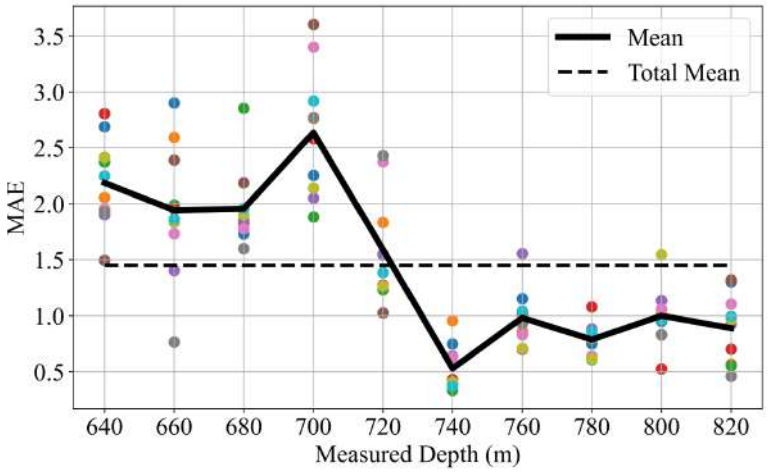


Figure 4.7: Multi-targets MAE results, the mean, and total mean of ten experiments.

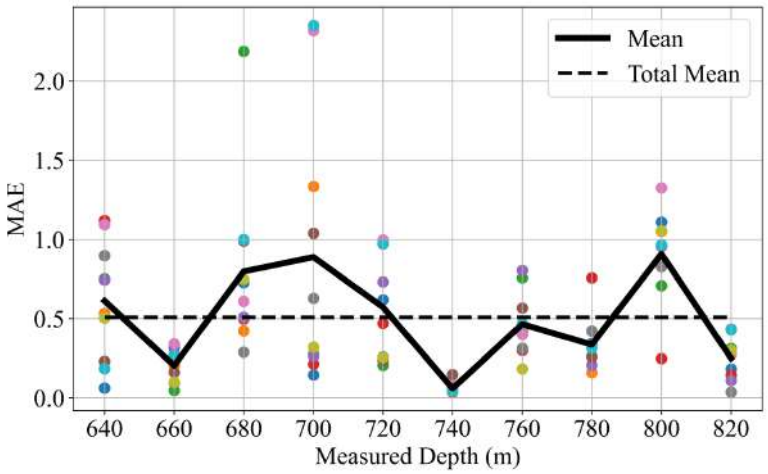


Figure 4.8: Inclination MAE results, the mean, and total mean of ten experiments.

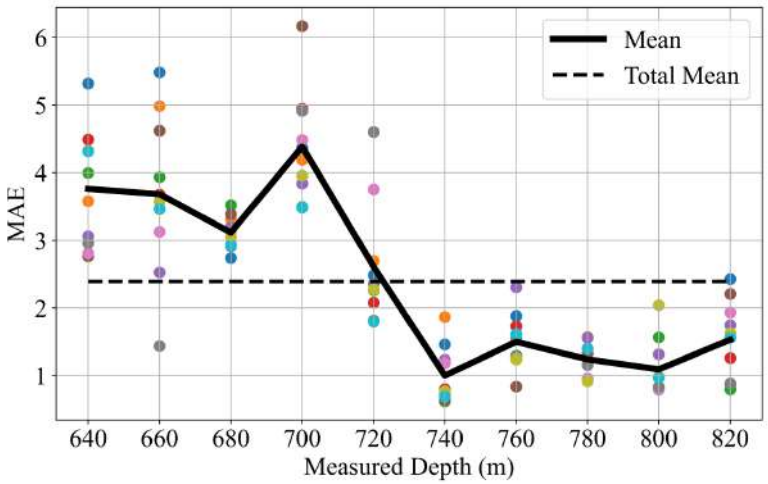


Figure 4.9: ROP MAE results, the mean, and total mean of ten experiments.

4.6 Discussion

In the context of the case study presented in this chapter, it was possible to obtain inclination prediction results with a mean absolute error of around 0.05 degrees in some parts of the well, and ROP prediction results of around 1 m/h. It is worth noting that these results are achieved for a prediction horizon of 25 meters after only 140 meters of drilling; other studies experimented on 23 meters [43, 44] and 10 meters prediction horizons, our study extended the horizon to provide a safe

and large confidence interval; drillers can use BHAs that contains sensors (MWD systems) up to 25 meters behind the drill bit. Despite the experimentation with an extended prediction horizon, the proposed approach achieved better results than the literature. The starting dataset of 140 meters can be lowered further by conducting additional experiments. A brief comparison between the findings of this study and other state-of-the-art implementations is validated in table 4.1, other studies use 30 meters prediction horizon for ROP prediction, while our study uses 25 meters, to maintain a synchronous prediction horizon between inclination and ROP.

Table 4.1: State-of-the-art results vs. the proposed approach.

		Training Data (Meters)	Prediction Horizon (Meters)	Average MAE of 10 Experiments	Best MAE of 10 Experiments
Inclination	Our Study	140	25	0.5 degrees	0.38 degrees
	State-of-art	180	23	0.65 degrees	0.61 degrees
ROP	Our Study	140	25	2.39 m/h	2.19 m/h
	State-of-art	N/A	30	7.33 m/h	N/A

Difficulties can be faced when deploying the digital twin in the field. Considerations have to be made regarding the BHA replacement during the initial vertical phase and the horizontal phase of the drilling, different MWD systems may have distinct calibration methodologies and referencing in addition to various BHA lengths, in such cases, adjusting the prediction horizon is the first consideration to make, another consideration is to decide whether the data from the previous BHA is usable, according to the used calibration methodologies and external environment a decision can be made by the drilling data analyst.

Improving the approach and its outcome can be achieved by implementing the following concepts:

- A combination of the current approach and the 'inclination change model' suggested by Tunkiel, A. et al. in [43, 44] can improve the prediction of inclination values in terms of precision. The proposed 'inclination change model' in the literature [43] contributed to a 40% performance increase compared to the 'Nominal Inclination model'; the approach proposed in this study contributed up to a 50% performance increase to the 'Nominal Inclination model' and a 10% performance increase than the 'inclination change model', therefore, combining the two approaches has a high chance for achieving a better performance.

- Geological data acquired during the exploration of the reservoir can be used as an additional feature to improve both the ROP and Inclination predictions, although it may provide imprecise or approximated values indicating the depth of different layers, it is always helpful to know the nature of the current and upcoming sub-surface layers.

4.7 Conclusion

This chapter presents a data-driven digital twin solution to forecast inclination and rate of penetration during well-drilling operations. The solution leverages the benefits of multi-target regression to predict multiple well parameters simultaneously by combining branches of LSTM and MLP networks in a single architecture. Real-time data acquisition is simulated using an incremental learning scheme, by feeding portions of data continuously to the digital twin. The data used in this study is acquired during a real directional drilling operation from the MWD system. Utilizing a prediction horizon of 25 meters and 140 meters of drilling data, the proposed approach successfully forecasted both inclination and ROP with precision, leading to a total MAE of 0.5 degrees and 2.39 m/h for Inclination and ROP. The methodology's performance can be improved by combining state-of-the-art models with the proposed approach coupled with smart integration of geological data (Logging data).

Chapter 5

General Conclusion

5.1 Thesis Summary

Industrial companies can suffer significant economic losses due to system failures or underperformance, which can result in costs for repairs and downtime. Therefore, The implementation of an efficient maintenance strategy is crucial in improving the dependability and accessibility of industrial systems, all while minimizing the costs associated with maintenance efforts. The maintenance strategy has undergone a transformation from the traditional methods of corrective and preventive approaches to a more advanced and sophisticated approach known as predictive maintenance strategy, which is also referred to as Prognostics and Health Management (PHM) strategy. The implementation of a predictive maintenance approach proves to be highly efficient as it addresses the drawbacks of alternative strategies by utilizing the health status of the system to prompt maintenance activities. Typically, the PHM strategy encompasses five distinct phases, namely data collection, data manipulation, fault analysis, fault prediction, and health management decisions facilitation. Fault prognostics, a key stage in implementing the PHM strategy, aim to estimate the Remaining Useful Life (RUL) of the system before failure occurs. This estimation helps in planning maintenance actions in advance, thereby preventing system downtime and reducing revenue losses.

In addition to estimating the Remaining Useful Life (RUL), it is crucial to forecast critical operational parameters in the oil and gas industry. The development of a tool that optimizes operational procedures is highly necessary, particularly if it leads to cost reduction and increased productivity. The current cutting-edge tool involves creating a data-driven replica of the specific operation, known as a data-driven digital twin. Digital twins come in various forms depending on their objective. This study focuses on a variation of a digital twin for drilling optimization, which predicts multiple essential drilling parameters throughout the operation. This enables more informed decision-making and ultimately leads to an optimized drilling operation.

This manuscript addresses the following challenges:

- Estimating the remaining useful life (RUL) becomes a challenging task when there are limited a priori sequences available that only partially capture the degradation dynamics and conditions. How can these degradation sequences be used to improve the accuracy of estimating the RUL?

- The ability to accurately predict parameters is of utmost importance in oil and gas well drilling operations. Despite the abundance of big data available during the drilling process, it remains a challenge to develop a tool that can effectively forecast specific parameters in order to optimize operations and facilitate decision-making.

- Is it possible for a tool developed to predict one parameter in drilling operations to also predict other parameters, even if they are different in nature (sequential and non-sequential parameters), as there is no single crucial parameter but rather multiple important parameters?

This manuscript presents two novel approaches that utilize customized neural network architectures. The first approach addresses the challenge of predicting the Remaining Useful Life (RUL) of systems or tools with limited data. To overcome this issue, a corrective feedback mechanism is incorporated into the neural

network architecture, enabling the model to learn from previous predictions. This is achieved by feeding the predictions made by the output layer neurons back to the input layer neurons. From a broader system perspective, this approach combines the predicted values with the nominal system features, allowing the model to incorporate knowledge from previous predictions. Compared to conventional models that lack this corrective feedback mechanism, this approach yields significantly more accurate predictions. Furthermore, it can be applied to various machine learning applications and regression models.

The second approach effectively integrated two branches of deep neural networks to create a customized architecture for implementation in a data-driven digital twin system. This digital twin leverages multi-target regression to accurately forecast multiple parameters concurrently. Moreover, the proposed approach not only offers valuable insights for optimizing drilling operations but also outperforms existing methods by incorporating LSTM layers in the architecture. This approach specifically addresses the aforementioned challenges.

5.2 Open issues and Future work

- The methodology outlined in chapter 3, while specifically tailored for the validation system, has the potential to be adapted for use in other systems. Although it was utilized to forecast the Remaining Useful Life (RUL) of the SSSV system, it can also be employed to predict any continuous variable or parameter using a regression-based approach. Furthermore, this methodology is not limited to the use of neural networks; the corrective feedback can be applied to any regression algorithm to significantly enhance prediction accuracy. As a result, readers are encouraged to incorporate the corrective feedback into their customized regression algorithms and evaluate the effectiveness of their architectures in various applications.

- An important consideration arises when optimizing the corrective feedback approach: the convergence of the system mean squared error (MSE) or mean

absolute error (MAE). It is crucial to recognize that when the predictions are used as a feature for the regression algorithm, they can either enhance or impair the accuracy of the subsequent prediction. If the initial prediction is highly accurate, it will undoubtedly improve the subsequent prediction. However, if the first prediction is highly inaccurate, incorporating it as a feature will decrease the accuracy of the next prediction rather than improve it. Thus, it is advisable for the accuracy of the first prediction to exceed a certain threshold in order for the corrective feedback approach to yield good results. Investigating this issue further is suggested as future work, and readers are encouraged to explore this matter.

- The methodology described in chapter 4, which entails the utilization of data-driven digital twin technology and the implementation of multi-target regression to predict multiple parameters simultaneously, has broad applicability beyond the particular case study presented. It is important to emphasize that this tailored framework can be employed across diverse industries and applications that require accurate prediction of crucial parameters. Industries such as aviation and motorsports, such as Formula One, are examples of sectors that can gain advantages from this approach.

Bibliography

[1] Ricardo Emanuel Vaz Vargas, Celso José Munaro, Patrick Marques Ciarelli, André Gonçalves Medeiros, Bruno Guberfain do Amaral, Daniel Centurion Barrionuevo, Jean Carlos Dias de Araújo, Jorge Lins Ribeiro, and Lucas Pierezan Magalhães. A realistic and public dataset with rare undesirable real events in oil wells. *Journal of Petroleum Science and Engineering*, 181:106223, 2019.

[2] PAA Garcia, Carlos MC Jacinto, BP Lima, and EAL Droguett. Optimizing downhole safety valve test scheduling using a multiobjective genetic algorithm. In *International Conference on Probabilistic Safety Assessment and Management*, 2006.

[3] Aymen Harrouz, Houari Toubakh, Redouane Kafi, Moamar Sayed-Mouchaweh, and Hajer Salem. Self adaptive learning scheme for fault prognosis in oil wells and production & service lines. In *Annual Conference of the PHM Society*, volume 14, 2022.

[4] Andrzej Tunkiel. Prediction, interpolation and extrapolation of drilling data with deep learning. 2022.

[5] Jake VanderPlas. *Python data science handbook: Essential tools for working with data*. " O'Reilly Media, Inc.", 2016.

[6] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq

Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

[7] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[8] Xiao-Sheng Si, Wenbin Wang, Chang-Hua Hu, and Dong-Hua Zhou. Remaining useful life estimation–a review on the statistical data driven approaches. *European journal of operational research*, 213(1):1–14, 2011.

[9] Aymen Harrouz, Hajer Salem, Houari Toubakh, Redouane Mohamed Kafi, and Moamar Sayed-Mouchaweh. Fault prognosis of subsurface safety valve system with limited real data using self-adaptive neural network. *Evolving Systems*, pages 1–19, 2023.

[10] Harrouz Aymen. Data-driven digital twin based on multi-target regression for inclinatio and rop prediction in oil and gas well drilling.

[11] Colin Shearer. The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22, 2000.

[12] Andrzej T Tunkiel, Dan Sui, and Tomasz Wiktorski. Reference dataset for rate of penetration benchmarking. *Journal of Petroleum Science and Engineering*, 196:108069, 2021.

[13] Erkam Guresen and Gulgun Kayakutlu. Definition of artificial neural networks with comparison to other networks. *Procedia Computer Science*, 3:426–433, 2011.

[14] Balázs Csanád Csáji et al. Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, 24(48):7, 2001.

[15] Yusman Yusof, HM Asri H Mansor, and Adizul Ahmad. Utilizing unsupervised weightless neural network as autonomous states classifier in reinforcement learning algorithm. In *2017 IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA)*, pages 264–269. IEEE, 2017.

[16] Happiness Ugochi Dike, Yimin Zhou, Kranthi Kumar Deveerasetty, and Qingtian Wu. Unsupervised learning based on artificial neural network: A review. In *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, pages 322–327. IEEE, 2018.

[17] Jean-Paul Haton. Introduction à l'intelligence artificielle et à la reconnaissance des formes.

[18] Kai Zhang, Niantian Lin, Gaopeng Tian, Jiuqiang Yang, Deying Wang, and Zhiwei Jin. Unsupervised-learning based self-organizing neural network using multi-component seismic data: Application to xujiahe tight-sand gas reservoir in china. *Journal of Petroleum Science and Engineering*, 209:109964, 2022.

[19] MAJ Van Gerven and Sander M Bohte. Artificial neural networks as models of neural information processing. 2017.

[20] Stuart J Russell and Peter Norvig. *Artificial intelligence a modern approach*. London, 2010.

[21] Mohammed S El-Abbasy, Ahmed Senouci, Tarek Zayed, Farid Mirahadi, and Laya Parvizsedghy. Artificial neural network models for predicting condition of offshore oil and gas pipelines. *Automation in Construction*, 45:50–65, 2014.

[22] Mohammad Abdideh. Estimation of permeability using artificial neural networks and regression analysis in an iran oil field. *International Journal of the Physical Sciences*, 7(34):5308–5313, 2012.

[23] Mirac Suzgun, Yonatan Belinkov, and Stuart M Shieber. On evaluating the generalization of lstm models in formal languages. *arXiv preprint arXiv:1811.01001*, 2018.

[24] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[25] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016.

[26] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. A comparison of arima and lstm in forecasting time series. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 1394–1401. IEEE, 2018.

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[28] Nam-Ho Kim, Dawn An, and Joo-Ho Choi. Prognostics and health management of engineering systems. *Switzerland: Springer International Publishing*, 2017.

[29] Gang Niu et al. Data-driven technology for engineering systems health management. *Springer Singapore*, 10:978–981, 2017.

[30] Sreerupa Das, Richard Hall, Amar Patel, Steve McNamara, and Jonathan Todd. An open architecture for enabling cbm/phm capabilities in ground vehicles. In *2012 IEEE Conference on Prognostics and Health Management*, pages 1–8. IEEE, 2012.

[31] Jay Lee, Fangji Wu, Wenyu Zhao, Masoud Ghaffari, Linxia Liao, and David Siegel. Prognostics and health management design for rotary machinery systems—reviews, methodology and applications. *Mechanical systems and signal processing*, 42(1-2):314–334, 2014.

[32] Andrew KS Jardine, Daming Lin, and Dragan Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, 20(7):1483–1510, 2006.

[33] Rob Callan, Brian Larder, and John Sandiford. An integrated approach to the development of an intelligent prognostic health management system. In *2006 IEEE Aerospace Conference*, pages 12–pp. IEEE, 2006.

[34] Wei Teng, Xiaolong Zhang, Yibing Liu, Andrew Kusiak, and Zhiyong Ma. Prognosis of the remaining useful life of bearings in a wind turbine gearbox. *Energies*, 10(1):32, 2016.

[35] Mohamed Elforjani and Suliman Shanbr. Prognosis of bearing acoustic emission signals using supervised machine learning. *IEEE Transactions on industrial electronics*, 65(7):5864–5871, 2017.

[36] Nagoor Basha Shaik, Srinivasa Rao Pedapati, and Faizul Azly BA Dzubir. Remaining useful life prediction of a piping system using artificial neural networks: A case study. *Ain Shams Engineering Journal*, 13(2):101535, 2022.

[37] Nagoor Basha Shaik, Srinivasa Rao Pedapati, Syed Ali Ammar Taqvi, AR Othman, and Faizul Azly Abd Dzubir. A feed-forward back propagation neural network approach to predict the life condition of crude oil pipeline. *Processes*, 8(6):661, 2020.

[38] Wenbai Chen, Weizhao Chen, Huixiang Liu, Yiqun Wang, Chunli Bi, and Yu Gu. A rul prediction method of small sample equipment based on dcnn-bilstm and domain adaptation. *Mathematics*, 10(7):1022, 2022.

[39] Linxia Liao and Felix Köttig. Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction. *IEEE Transactions on Reliability*, 63(1):191–207, 2014.

[40] Haixin Lv, Jinglong Chen, and Tongyang Pan. Sequence adaptation adversarial network for remaining useful life prediction using small data set. In *2020 IEEE 18th International Conference on Industrial Informatics (INDIN)*, volume 1, pages 115–118. IEEE, 2020.

[41] Gurtej Singh Saini, AmirHossein Fallah, Pradeepkumar Ashok, and Eric van Oort. Digital twins for real-time scenario analysis during well construction operations. *Energies*, 15(18):6584, 2022.

[42] William G Lesso Jr, Iain M Rezmer-Cooper, and Minh Chau. Continuous direction and inclination measurements revolutionize real-time directional drilling decision-making. In *SPE/IADC Drilling Conference and Exhibition*, pages SPE–67752. SPE, 2001.

[43] Andrzej T Tunkiel, Dan Sui, and Tomasz Wiktorski. Training-while-drilling approach to inclination prediction in directional drilling utilizing recurrent neural networks. *Journal of Petroleum Science and Engineering*, 196:108128, 2021.

[44] Andrzej T Tunkiel, Tomasz Wiktorski, and Dan Sui. Continuous drilling sensor data reconstruction and prediction via recurrent neural networks. In *International Conference on Offshore Mechanics and Arctic Engineering*, volume 84317, page V001T01A002. American Society of Mechanical Engineers, 2020.

[45] Abdulmalek Ahmed, Abdulwahab Ali, Salaheldin Elkatatny, and Abdulazeez Abdulraheem. New artificial neural networks model for predicting rate of penetration in deep shale formation. *Sustainability*, 11(22):6527, 2019.

[46] Omogbolahan S Ahmed, Ahmed A Adeniran, and Ariffin Samsuri. Computational intelligence based prediction of drilling rate of penetration: A comparative study. *Journal of Petroleum Science and Engineering*, 172:1–12, 2019.

[47] Chiranth Hegde and KE Gray. Use of machine learning and data analytics to increase drilling efficiency for nearby wells. *Journal of Natural Gas Science and Engineering*, 40:327–335, 2017.

[48] Chiranth Hegde, Scott Wallace, and Ken Gray. Using trees, bagging, and random forests to predict rate of penetration during drilling. In *SPE Middle East Intelligent Oil and Gas Symposium*, page D011S001R003. SPE, 2015.

[49] Chiranth Hegde, Scott Wallace, and Ken Gray. Using trees, bagging, and random forests to predict rate of penetration during drilling. In *SPE Middle East Intelligent Oil and Gas Symposium*, page D011S001R003. SPE, 2015.

[50] Cesar Soares and Kenneth Gray. Real-time predictive capabilities of analytical and machine learning rate of penetration (rop) models. *Journal of Petroleum Science and Engineering*, 172:934–959, 2019.

[51] Mohammad Sabah, Mohsen Talebkeikhah, David A Wood, Rasool Khosravanian, Mohammad Anemangely, and Alireza Younesi. A machine learning approach to predict drilling rate using petrophysical and mud logging data. *Earth Science Informatics*, 12:319–339, 2019.

[52] Jiahang Han, Yanji Sun, and Shaoning Zhang. A data driven approach of rop prediction and drilling performance estimation. In *International Petroleum Technology Conference*, page D011S010R006. IPTC, 2019.

[53] Xian Shi, Gang Liu, Xiaoling Gong, Jialin Zhang, Jian Wang, Hongning Zhang, et al. An efficient approach for real-time prediction of rate of penetration in offshore drilling. *Mathematical Problems in Engineering*, 2016, 2016.

[54] Bharat Mantha and Robello Samuel. Rop optimization using artificial intelligence techniques with statistical regression coupling. In *SPE Annual Technical Conference and Exhibition?*, page D031S041R007. SPE, 2016.

[55] Tuna Eren and M Evren Ozbayoglu. Real time optimization of drilling parameters during drilling operations. In *SPE Oil and Gas India Conference and Exhibition?*, pages SPE–129126. SPE, 2010.

[56] Cesar Soares, Hugh Daigle, and Ken Gray. Evaluation of pdc bit rop models and the effect of rock strength on model coefficients. *Journal of Natural Gas Science and Engineering*, 34:1225–1236, 2016.

[57] Khoukhi Amar and Alarfaj Ibrahim. Rate of penetration prediction and optimization using advances in artificial neural networks, a comparative study. In *Proceedings of the 4th International Joint Conference on Computational Intelligence, Barcelona, Spain*, pages 5–7, 2012.

[58] Ping Yi, Aniket Kumar, and Robello Samuel. Real-time rate of penetration optimization using the shuffled frog leaping algorithm (sfla). In *SPE Intelligent Energy Conference & Exhibition*. OnePetro, 2014.

[59] Wanyi Jiang and Robello Samuel. Optimization of rate of penetration in a convoluted drilling framework using ant colony optimization. In *IADC/SPE Drilling Conference and Exhibition*. OnePetro, 2016.

[60] Cesar Soares and Kenneth Gray. Real-time predictive capabilities of analytical and machine learning rate of penetration (rop) models. *Journal of Petroleum Science and Engineering*, 172:934–959, 2019.

[61] Helmi Helmiriawan and Zaid Al-Ars. Multi-target regression approach for predictive maintenance in oil refineries using deep learning. *International Journal of Neural Networks and Advanced Applications*, 6:18–24, 2019.

[62] Liang Xue, Yuetian Liu, Yifei Xiong, Yanli Liu, Xuehui Cui, and Gang Lei. A data-driven shale gas production forecasting method based on the multi-objective random forest regression. *Journal of Petroleum Science and Engineering*, 196:107801, 2021.

[63] Wanxing Zhang, Kai Bai, et al. Parameter prediction of coiled tubing drilling based on gan-lstm. 2023.

[64] Shaowei Pan, Bo Yang, Shukai Wang, Zhi Guo, Lin Wang, Jinhua Liu, and Siyu Wu. Oil well production prediction based on cnn-lstm model with self-attention mechanism. *Energy*, 284:128701, 2023.

[65] B Sirisha, Katamouni Kranthi Chaitanya Goud, and BTV Saketh Rohit. A deep stacked bidirectional lstm (sbilstm) model for petroleum production forecasting. *Procedia Computer Science*, 218:2767–2775, 2023.

[66] Indrajeet Kumar, Bineet Kumar Tripathi, and Anugrah Singh. Attention-based lstm network-assisted time series forecasting models for petroleum production. *Engineering Applications of Artificial Intelligence*, 123:106440, 2023.

[67] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[68] Marvin Rausand and Jørn Vatn. Reliability modeling of surface controlled subsurface safety valves. *Reliability Engineering & System Safety*, 61(1-2):159–166, 1998.

[69] Koceila Abid. *Data-driven Approach for Fault Prognostics of Industrial Systems-From Using No, Insufficient, to Multiple Historical Degradation Sequences*. PhD thesis, Ecole nationale supérieure Mines-Télécom Lille Douai, 2020.

[70] Ziqiu Kang, Cagatay Catal, and Bedir Tekinerdogan. Machine learning applications in production lines: A systematic literature review. *Computers & Industrial Engineering*, 149:106773, 2020.

[71] Bekhzod Olimov, Sanjar Karshiev, Eungyeong Jang, Sadia Din, Anand Paul, and Jeonghong Kim. Weight initialization based-rectified linear unit activation function to improve the performance of a convolutional neural network model. *Concurrency and Computation: Practice and Experience*, 33(22):e6143, 2021.

[72] OSIsoft. OSIsoft website. https://www.osisoft.com/pi-system/.

[73] Schlumberger. Schlumberger website. https://www.software.slb.com/products/olga.

[74] Ivanilto Andreolli. Introdução à elevação e escoamento monofásico e multifásico de petróleo. *Rio de Janeiro: Interciência*, 2016.

[75] Andrzej T Tunkiel, Tomasz Wiktorski, and Dan Sui. Drilling dataset exploration, processing and interpretation using volve field data. In *International Conference on Offshore Mechanics and Arctic Engineering*, volume 84430, page V011T11A076. American Society of Mechanical Engineers, 2020.

[76] Andrzej T Tunkiel, Dan Sui, and Tomasz Wiktorski. Impact of data preprocessing techniques on recurrent neural network performance in context of real-time drilling logs in an automated prediction framework. *Journal of Petroleum Science and Engineering*, 208:109760, 2022.

[77] François Chollet et al. Keras documentation. *keras. io*, 33, 2015.

[78] Nam-Ho Kim, Dawn An, and Joo-Ho Choi. Prognostics and health management of engineering systems. *Switzerland: Springer International Publishing*, 2017.

[79] Danilo Colombo, Gilson Brito Alves Lima, Danillo Roberto Pereira, and João P Papa. Regression-based finite element machines for reliability modeling of downhole safety valves. *Reliability Engineering & System Safety*, 198:106894, 2020.

[80] Michael Grieves. Digital twin: manufacturing excellence through virtual factory replication. *White paper*, 1(2014):1–7, 2014.

[81] Michael Grieves and John Vickers. Mitigating unpredictable, undesirable emergent behavior in complex systems (excerpt).

[82] A Boxall. Inside mclaren's secretive f1 mission control room| digital trends, 2016.