UiA University of Agder

# Fault Detection and Diagnostics of Hydraulic Systems in Hydroelectric Power Plants

HANNE KRAKELI

SUPERVISOR
van Khang Huynh

**University of Agder, 2023**
Faculty of Engineering and Science
Department of Engineering and Sciences

## Obligatorisk erklæring

Den enkelte student er selv ansvarlig for å sette seg inn i hva som er lovlige hjelpemidler, retningslinjer for bruk av disse og regler om kildebruk. Erklæringen skal bevisstgjøre studentene på deres ansvar og hvilke konsekvenser fusk kan medføre. Manglende erklæring fritar ikke studentene fra sitt ansvar.

<table>
<tr><td>1.</td><td>Jeg/vi erklærer herved at min/vår besvarelse er mitt/vårt eget arbeid, og at jeg/vi ikke har brukt andre kilder eller har mottatt annen hjelp enn det som er nevnt i besvarelsen.</td><td>Ja</td></tr>
<tr><td>2.</td><td>Jeg/vi erklærer videre at denne besvarelsen:<ul><li>Ikke har vært brukt til annen eksamen ved annen avdeling/universitet/høgskole innenlands eller utenlands.</li><li>Ikke refererer til andres arbeid uten at det er oppgitt.</li><li>Ikke refererer til eget tidligere arbeid uten at det er oppgitt.</li><li>Har alle referansene oppgitt i litteraturlisten.</li><li>Ikke er en kopi, duplikat eller avskrift av andres arbeid eller besvarelse.</li></ul></td><td>Ja</td></tr>
<tr><td>3.</td><td>Jeg/Vi er kjent med at brudd på ovennevnte er å betrakte som fusk og kan medføre annullering av eksamen og utestengelse fra universiteter og høgskoler i Norge, jf. Universitets- og høgskoleloven §§4-7 og 4-8 og Forskrift om eksamen §§ 31.</td><td>Ja</td></tr>
<tr><td>4.</td><td>Jeg/vi er kjent med at alle innleverte oppgaver kan bli plagiatkontrollert.</td><td>Ja</td></tr>
<tr><td>5.</td><td>Jeg/vi er kjent med at Universitetet i Agder vil behandle alle saker hvor det forligger mistanke om fusk etter høgskolens retningslinjer for behandling av saker om fusk.</td><td>Ja</td></tr>
<tr><td>6.</td><td>Jeg/vi har satt meg/oss inn i regler og retningslinjer i bruk av kilder og referanser på biblioteket sine nettsider.</td><td>Ja</td></tr>
</table>

## Publiseringsavtale

Fullmakt til elektronisk publisering av oppgaven Forfatter(ne) har opphavsrett til oppgaven. Det betyr blant annet enerett til å gjøre verket tilgjengelig for allmennheten (Åndsverkloven. §2).
Oppgaver som er unntatt offentlighet eller taushetsbelagt/konfidensiell vil ikke bli publisert.

| | |
|---|---|
| Jeg/vi gir herved Universitetet i Agder en vederlagsfri rett til å gjøre oppgaven tilgjengelig for elektronisk publisering: | Ja |
| Er oppgaven båndlagt (konfidensiell)? | Nei |
| Er oppgaven unntatt offentlighet? | Nei |

# Acknowledgment

This master thesis is written as the final part of my Master's degree in Renewable Energy with the Department of Engineering and Science at the University of Agder. Completed during the spring of 2023. This work has been carried out in collaboration with NORCE, Å Energi, and UiA. Building upon my previous research project [1], the focus of this thesis is to explore the opportunities and challenges associated with fault detection and diagnosis of hydraulic systems used for controlling hydroelectric power plants. While my background primarily encompasses a course in fluid dynamics and hydro power, my understanding of fault detection and diagnosis has been largely self-taught.

Firstly I want to thank my supervisor van Khang Huynh, and the PHM group for support and engaging discussions. Additionally, I extend my appreciation to Å Energi for providing the dataset that serves as the foundation for this thesis. Lastly, I am thankful to my fellow students and friends, who have provided motivation and support during the long hours spent in the master's lab.

Grimstad, May 2023

Hanne Krakeli

Hanne Krakeli

# Abstract

Most hydroelectric power plants experience few faults during their lifetime. However, with the expected increase in volatile energy sources in the energy mix and the use of hydroelectric power plants as balancing mechanisms, the wear and tear on system components may rise. This is where fault detection and diagnosis can significantly improve maintenance regimes. A mechanism within the plant that signals when components are under distress can be both cost-efficient and increase the plant's reliability.

The hydraulic system that regulates the flow of water through the turbine is one of those systems that can be affected if the operation of the plants deviates from established norms. This thesis explores the feasibility of creating a model based on features from the Supervisory Control and Data Acquisition (SCADA) system.

The algorithm selected for this task is neural networks, specifically, a recurrent neural network (RNN) and long short-term memory (LSTM). The two models are used to predict the guide vane position with the use of input features such as oil level, accumulator temperature, and oil pressure. The RNN model proved to be the most accurate of the two, exhibiting low error rates and high $R^2$ score. The LSTM model struggled with accurate prediction, showing poor model fit metrics, even with the introduction of L2-regularization.

Additionally, an investigation was undertaken into the RNN model's performance on synthetic data with anomalous values, and it revealed a significant decrease in accuracy.

# Sammendrag

De fleste vannkraftverk opplever få feil i løpet av levetiden. Imidlertid, med den forventede økningen i volatile energikilder i energimiksen og bruken av vannkraftverk som balanseringsmekanismer, kan slitasje på komponenter i det hydrauliske systemet også øke. Her kan feildeteksjon og diagnose betydelig forbedre vedlikeholdsrutinene. Ved hjelp av overvåkning av komponenter i anlegget som signaliserer når komponenter ikke fungerer optimalt, kan både være kostnadseffektiv og øke anleggets pålitelighet.

Hydraulikksystemet som regulerer vannstrømmen gjennom turbinen, er ett av de systemene som kan påvirkes hvis driften av anleggene avviker fra de etablerte normene. Denne masteroppgaven ønsker å undersøke muligheten for å lage en modell basert på verdier fra Overvåkningskontroll- og Datainnsamling (SCADA) systemet.

Algoritmen som er valgt for denne oppgaven er nevrale nettverk, spesifikt et tilbakevendende nevralt nettverk (RNN) og langt korttidsminne nettverk (LSTM). De to modellene blir brukt til å forutsi ledeappartet sin posisjonen ved hjelp av måleverdier som oljenivå, akkumulatortemperatur og oljetrykk. RNN-modellen viste seg å være den mest nøyaktige av de to, med lave MAE og MSE verdier og høy $R^2$-score. LSTM-modellen slet med nøyaktig prediksjon, og viste dårlige modelltilpasningsmetrikker, selv med innføringen av L2-regularisering.

I tillegg ble det foretatt en undersøkelse av RNN-modellens evne til forutsi på syntetiske data med unormale verdier, og det avslørte en betydelig nedgang i nøyaktighet.

# Contents


**Acknowledgements** **ii**

**Abstract** **iii**

**Sammendrag** **iv**

**List of Figures** **vii**

**List of Tables** **ix**

**1 Introduction** **1**
1.1 Background and Motivation . . . . . 1
1.2 Thesis Definition . . . . . 2
1.3 Limitations . . . . . 2
1.4 Report Structure . . . . . 2

**2 Theory** **5**
2.1 Condtion Monitoring, Fault Detection, and Diagnosis . . . . . 5
2.2 Hydraulic System and Subsystem . . . . . 6
2.2.1 Accumulator . . . . . 6
2.2.2 Oil Reservoir . . . . . 7
2.2.3 Wicket Gate . . . . . 7
2.2.4 Hydraulic Valves . . . . . 8
2.3 Failure Causes in Hydraulic Systems . . . . . 8
2.3.1 Contamination . . . . . 8
2.3.2 Internal and External Leakage . . . . . 9
2.3.3 Other Causes . . . . . 9
2.4 Supervisory Control and Data Acquisition . . . . . 9
2.5 Traditional Fault Detection and Diagnostics of Hydraulic Systems . . . . . 10

**3 Literature Review** **11**
3.1 Fault Diagnosis of a Hydraulic Power System . . . . . 11
3.2 Condition Monitoring and Fault Diagnosis of HEPs . . . . . 12

**4 Utilizing Python for Data Analysis and Implementation** **15**
4.1 Python . . . . . 15
4.1.1 Scikit-learn . . . . . 15
4.1.2 Keras . . . . . 15
4.2 Artificial Neural Network . . . . . 15
4.2.1 Recurrent Neural Network . . . . . 17
4.2.2 Long Short-Term Memory . . . . . 18
4.3 L2-Regularization . . . . . 20
4.4 Metrics . . . . . 20

**5 Data Analyzing and Model Development** **23**
5.1 Supervised Learning . . . 23
5.2 Evenstad Dataset . . . 24
5.3 Preprocessing of the Dataset . . . 25
5.4 Hyperparameterization . . . 26
5.5 Model Development . . . 27

**6 Results and Discussion** **29**
6.1 Hyperparameters . . . 29
6.2 RNN Model . . . 29
6.3 RNN LSTM Model . . . 31
6.4 RNN LSTM Model with L2-regularization . . . 33
6.5 Synthetic Data RNN Model . . . 34
6.6 Comparison of the Models . . . 36
6.7 General Discussion . . . 36

**7 Conclusions** **39**
7.1 Research Findings . . . 39
7.2 Further Work . . . 39

**Bibliography** **41**

# List of Figures


2.1 Steps of fault diagnosis [5] . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
2.2 Example of wicket gate/guide vane [9] . . . . . . . . . . . . . . . . . . . . . . 7
2.3 SCADA system with the three main system components [13] . . . . . . . . . 9

3.1 Selak et al.[17] proposed CMFD model . . . . . . . . . . . . . . . . . . . . . 13

4.1 Architecture of a artificial neural network [20] . . . . . . . . . . . . . . . . . 16
4.2 Architecture of a recurrent neural network [21] . . . . . . . . . . . . . . . . . 17
4.3 Example of a LSTM cell [23] . . . . . . . . . . . . . . . . . . . . . . . . . . . 18
4.4 Bias versus variance tradeoff [24] . . . . . . . . . . . . . . . . . . . . . . . . 20

5.1 Workflow for supervised learning [26] . . . . . . . . . . . . . . . . . . . . . . 23
5.2 Evenstad dataset . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
5.3 Correlation matrix . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25

6.1 Model loss for the RNN model . . . . . . . . . . . . . . . . . . . . . . . . . . 30
6.2 Actual versus predicted values from the RNN model . . . . . . . . . . . . . . 30
6.3 Model loss for the RNN LSTM model . . . . . . . . . . . . . . . . . . . . . . 31
6.4 Actual versus predicted values from the RNN LSTM model . . . . . . . . . . 32
6.5 Model loss for the RNN LSTM model with L2-regularization . . . . . . . . . 33
6.6 Predicted versus actual values for the RNN LSTM model with L2-regularization 34
6.7 Predicted versus actual values for the RNN model with abnormal data . . . 35
6.8 Model loss for the RNN model with abnormal data . . . . . . . . . . . . . . . 35

# List of Tables


5.1 Overview of data gathered from Evenstad hydro power plant . . . . . . . . . . 24
5.2 Hyperparameters . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26
5.3 Alarm Thresholds and Normal Values . . . . . . . . . . . . . . . . . . . . . . . 27

6.1 Result from GridSearchCV . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
6.2 Performance Metrics for RNN Model . . . . . . . . . . . . . . . . . . . . . . . 31
6.3 Performance Metrics for the RNN LSTM Model . . . . . . . . . . . . . . . . . 32
6.4 Performance Metrics for the RNN LSTM Model with L2 regularization . . . 33
6.5 Performance Metrics for the RNN Model with abnormal data . . . . . . . . . 36

# Abbreviations

| | | |
|---|---|---|
| ANN | = | Artificial Neural Network |
| CMFD | = | Condition Monitoring and Fault Diagnostics |
| FD | = | Fault Detection |
| FDI | = | Fault Detection and Isolation |
| FI | = | Fault Isolation |
| HEP | = | Hydroelectrical Power Plant |
| HMI | = | Human Machine Interface |
| IDE | = | Integrated Development Environment |
| IPS2 | = | Industrial Product-service Systems |
| LSTM | = | Long Short-Term Memory |
| ML | = | Machine Learning |
| MAE | = | Mean Absolute Error |
| MSE | = | Mean Squared Error |
| NN | = | Neural Network |
| O&M | = | Operation and Maintenance |
| PLC | = | Programmable Logic Controller |
| RNN | = | Recurrent Neural Network |
| RTU | = | Remote Terminal Unit |
| $R^2$ | = | Coefficient of Determination |
| SCADA | = | Supervisory Control and Data Acquisition |
| SVM | = | Support Vector Machine |
| VDC | = | Virtual Diagnostics Center |

# Chapter 1

# Introduction

The introduction chapter presents the background and motivation for this master thesis. Then the definition of the thesis and objectives. Lastly, this chapter describes the limitations and how the report is structured.

## 1.1 Background and Motivation

The increasing global demand for clean and sustainable energy sources has put renewable energy at the forefront of efforts to reduce greenhouse gas emissions and mitigate the adverse effects of climate change. As one of the most mature and efficient renewable energy technologies, Hydroelectrical power plants (HEP) plays a critical role in meeting these growing energy demands. In Norway they are the most important source of renewable energy and is accounting for approximately 90% of the country's electricity generation according to Statkraft in 2021 [2]. Norway has an ambitious renewable energy goals, and has committed to reducing the greenhouse gas emissions by at least 50% by 2030 compared to the 1990 levels, as part of the global effort to mitigate climate change under the Paris Agreement [3]. To meet these targets it is essential to ensure that HEPs operate efficiently and reliably.

The power sector is currently going through a massive transformation world wide, with a rapid shift towards more renewable technologies. In Norway hydro power plays and will continue to be an important role in providing low emission energy, flexibility and reliability compared to more variable renewable energy systems like wind or solar power [4].

One of the key challenges in maintaining the reliability of HEPs is the detection and diagnosis of faults that can occur, and for this thesis the focus is faults in the hydraulic system of the HEPs. The hydraulic system plays a crucial role in controlling the motion of the power plant, regulating the flow of water to the turbine, and ensuring the stability of the system. The hydraulic components can experience wear and tear over time, leading to malfunctions that result in costly downtime and reduced power generation.

To address this challenge, the thesis is focused on developing fault detection and diagnostic methods for hydraulic system in HEPs. The goal is to identify and diagnose faults in the hydraulic system before they cause significant damage or downtime, allowing for proactive maintenance and repair. The research is motivated by the need to ensure efficient and reliable operation of HEPs, which are critical for meeting the renewable energy goals and reducing greenhouse gas emissions.

## 1.2 Thesis Definition

This master thesis aims to conduct research of the typical failure mechanisms in hydraulic system used in HEPs, and the negative impact of these failures on efficiency and safety of power generation. The research will conduct an in-depth examination of the causes, implication, and common modes of these failures.

Furthermore, the study intends to investigate alternative methods to conventional maintenance practices in an effort to improve system reliability and efficiency. In an era where innovation and sustainability are integral to energy production, it becomes paramount to discover and implement improved maintenance strategies that are not only more effective but also environment friendly and cost-efficient.

The research will investigate the potential of harnessing real-time data from the Supervisory Control and Data Acquisition (SCADA) system in an ambitious step toward early fault detection and efficient diagnosis. The SCADA system, being the nerve center of modern industrial operations, offers a valuable wealth of data that can be translated into meaningful insights through analysis. The aim here is to establish whether it is possible to develop a data-driven model for early fault detection and diagnosis in hydraulic systems with the use of SCADA data.

Lastly, the thesis will delve into the selection of appropriate model techniques that can be employed for fault detection and diagnosis. The choice of technique is a critical factor that can significantly influence the success and reliability of the proposed model. Therefore, this research will consider some potential techniques, assessing their strengths, weaknesses, and applicability in the context of hydraulic system fault detection and diagnosis in hydroelectric power plants.

## 1.3 Limitations

One of the main limitations of this thesis is that it covers a wide range of research areas that are outside the authors' formal education. This is a challenge because it may have an impact on the quality of the analysis or interpretation of the results. Technical details or nuances that are not immediately apparent may affect decision-making for topics other than the limits of the education. Another limitation is that the literature on fault detection and diagnostics of hydraulic systems in HEPs is limited. While this provides an opportunity to contribute to the field, it also means that there are few resources or studies to compare or analyze. To address this limitation, a literature review was conducted and similar topics and methodologies were investigated.

## 1.4 Report Structure

The report is divided into these chapters:

**Chapter 2** is the theoretical background and concepts relevant to this thesis. It establishes the theoretical framework for the study.

**Chapter 3** is a literature review of similar research and summarizing their key findings.

**Chapter 4** introduction of the software used, and discusses the specific python libraries and tools used in the study.

**Chapter 5** is data analyzing and model development

**Chapter 6** is the result, model evaluation and discussion around the research findings.

**Chapter 7** presents the conclusion of the thesis and suggestions for further work.

# Chapter 2

# Theory

This chapter presents the theory relevant to various aspects of understanding the topic. Firstly, the concept of fault detection and diagnosis is introduced, followed by an explanation of the system on which this research is based. Specifically, the focus is on the hydraulic system and its constituent subparts. The chapter addresses the typical failure causes in hydraulic systems and examines the traditional approaches employed for their maintenance. Furthermore, the chapter goes into how the SCADA system works.

## 2.1 Condtion Monitoring, Fault Detection, and Diagnosis

Fault detection and diagnosis are based on the principles of system dynamics and control theory, signal processing, and statistical analysis. System dynamics and control theory is the field of study that deals with the behavior of systems over time and how they respond to inputs and disturbances. In fault detection and diagnosis, system dynamics and control theory are used to model the system's normal behavior and identify deviations from this behavior that may indicate a fault. Signal processing is a technique used to extract meaningful information from the data collected by sensors. This includes filtering, feature extraction, and pattern recognition. These techniques can be used to detect changes in the signal that may indicate a fault [5].

In addition to these theoretical foundations, practical fault detection and diagnosis methods may also include machine learning algorithms, such as artificial neural networks and decision trees, to perform more advanced data analysis. These algorithms can be trained on data from normal system operation and used to identify abnormal behavior that may indicate a fault. Fault detection and condition monitoring are both techniques used to identify potential problems in a system. The difference between fault detection and condition monitoring is described below.

First, some terminology is used in the concept of fault diagnosis. Fault diagnosis includes fault detection, isolation, estimation, and analysis or identification. The basic steps of fault diagnosis are shown below in figure 2.1. The first step is fault detection. Fault detection is the process of identifying the presence of a fault or defect in a system or component. It involves detecting anomalous behaviors in system performance, such as unexpected variations in output, vibrations, or noise levels, and using these signals to identify the presence of a fault. Fault detection is typically used to identify immediate problems or failures that require immediate attention and corrective action. Fault isolation is to locate where the fault is within the system. Fault estimation is done to reconstruct the time-varying behavior of the fault signals. Lastly, fault analysis is when the fault is characterized by the type of fault, the severity, or the nature of defected faults [5].

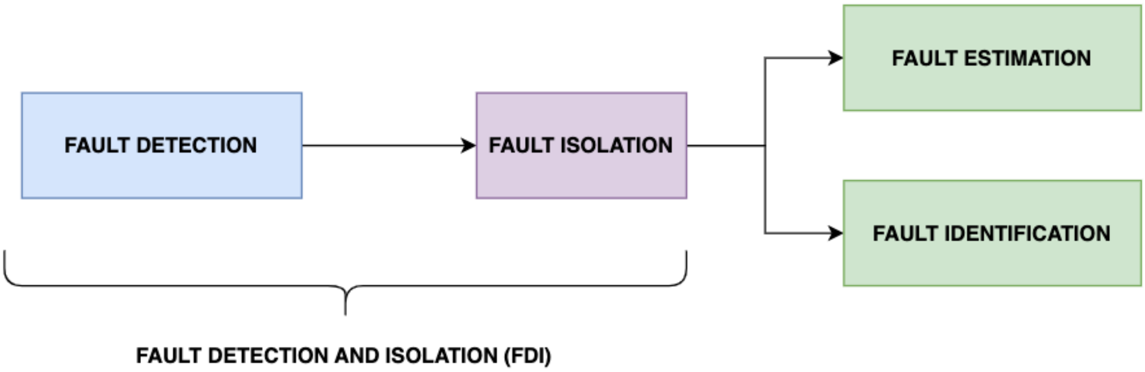


Figure 2.1: Steps of fault diagnosis [5]

Condition monitoring is a proactive approach to maintenance that involves continuously oversight over the condition of the machinery or equipment to identify signs of potential failure before they occur. Condition monitoring involves analyzing data from various sensors and instruments, such as temperature, vibration, pressure, and fluid analysis sensors, to detect changes in equipment conditions that could indicate a developing problem. Condition monitoring aims to design one or more indicators that can tell the state of health of the monitored system [5].

In summary, fault detection and diagnosis is focused on identifying immediate issues or problems that require immediate attention. In contrast, condition monitoring is focused on proactively monitoring equipment to detect signs of potential problems before they become serious issues [5].

## 2.2 Hydraulic System and Subsystem

There are typically three components in machinery: a prime mover, a power transmission, and an implement end-effector. In a hydraulic circuit, the hydraulic system functions as the power transmission. A hydraulic power transmission system is specifically designed to transfer energy from its source to its designated location using pressurized hydraulic fluids. This energy transfer follows a three-step process. Firstly, mechanical energy is converted into potential energy in the form of pressure. Next, this potential energy is delivered to the desired location and converted into useful work. Finally, the potential energy is converted back into kinetic energy. The notable characteristic of fluid power transmission lies in its medium for carrying power. Pressurized fluid can be shaped into various geometrical forms and deliver power in any required direction [6].

In a HEPs a large amount of components is the complex dynamical system that controls for example the motions of the power plant. The role of fault detection and diagnosis of such a large system is not a easy task. This chapter aims to describe the subparts involved in the hydraulic system for this master's thesis.

### 2.2.1 Accumulator

An accumulator is a pressure vessel with multiple purposes in a hydraulic system. The accumulator maintains pressure, stores and recaptures energy, reduces pressure peaks, power chassis suspension, and dampens the circuit's shock, vibration and pulsations. The accumulators task is to store necessary energy to ensure the shutdown of the turbine and perform a quick closing of the guide vanes if their is a sudden load drop. The pressure stored in the accumulator is re-fed into the hydraulic system if needed, and with the accumulator the pressure of the system is maintained. It is a safety precaution to ensure the system pressure is within limits [7].

### 2.2.2 Oil Reservoir

The oil reservoir is where the hydraulic fluid is stored. It is often some kind of tank. The hydraulic pumps are directly connected to the tank with a valve in between. The oil tank can have several monitors, both analog and digital. The analog is often a inspection glass where it possible to see the oil level, and the digital can show the same and temperature of the fluid. It is also normal with a oil guard, that alarms if the oil level is to low, or high [8].

### 2.2.3 Wicket Gate

The wicket gate (guide vanes) is located upstream of the turbine, and it helps manage the water flow through the turbine and modulate its power output. It works with cylinders connected to the hydraulic circuit to adjust the angle of the runner blades. This adjustment influences the angle at which water strikes the turbine, affecting the turbine's output power. The hydraulic system controls the wicket gate. Hydraulic cylinders manipulate the angle of the guide vanes, making them subject to the same wear and tear as other hydraulic components. Kaplan and Francis type turbines utilize wicket gates for water control [9].

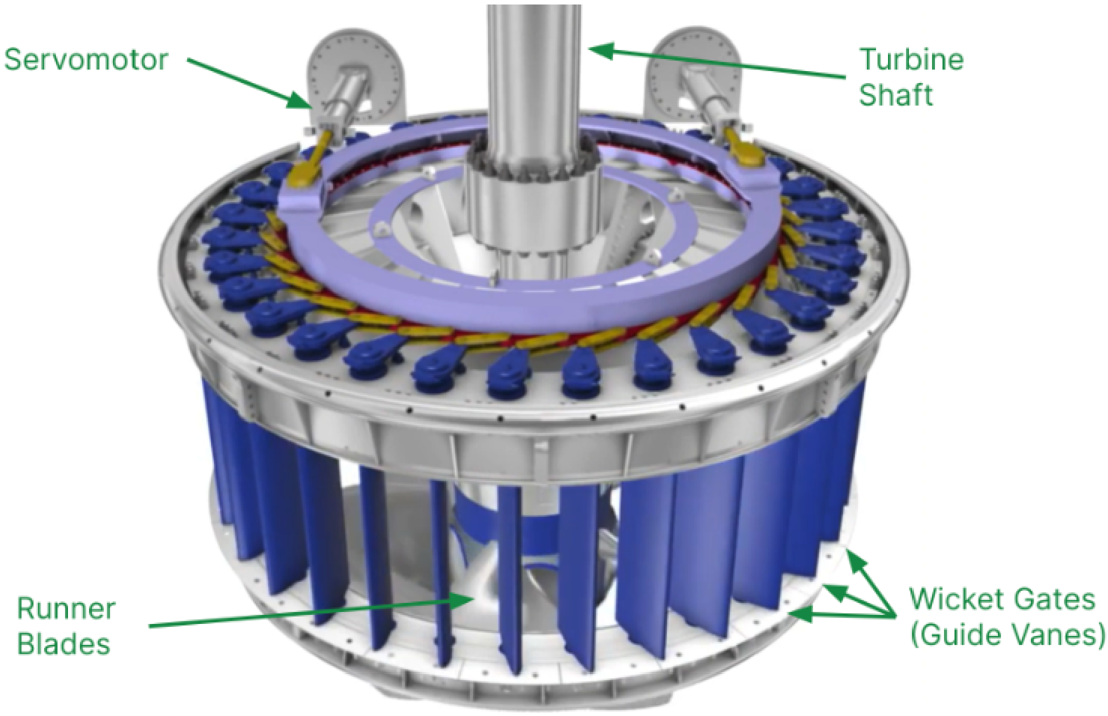


Figure 2.2: Example of wicket gate/guide vane [9]

In Figure 2.2 above is an example of how the turbine, with the wicket gates, runner blades, servomotor and turbine shaft is shown. The hydraulic cylinders are also shown. This part of the system is connected to the hydraulic circuit.

### 2.2.4 Hydraulic Valves

Hydraulic valves are used to control the fluid's direction and adjust the system's pressure, either increasing or decreasing. Hydraulic valves are categorized in three main varieties listed below.

- Hydraulic flow control valves
- Hydraulic pressure control valves
- Hydraulic directional control valves

Hydraulic flow control valves regulate the flow volume of passing fluid in the system. They are compared to movable gates that changes the flow to modify the rate of the fluid in the valve. This types of valves are divided into single or double acting [1, 10].

Hydraulic pressure control valves regulate the fluid pressure passing through a hydraulic component and maintain the pressure at a wanted level. They are placed in the system to ensure that the pressure stays at the wanted level, and not go over the threshold of the component the fluid will pass through. The function as a safety [1, 10].

The last category is hydraulic directional control valves, which track the fluid in the system to the wanted parts. These types of valves shift between discrete positions: extend, retract or neutral. They are used to start, stop and change the direction of flow. They come in various valves with different valve ports that can change in different valve positions [1, 10].

## 2.3 Failure Causes in Hydraulic Systems

In the hydraulic system that controls the HEP the causes of failure are many. In this section the most well known factors will be explained.

### 2.3.1 Contamination

Contamination in a hydraulic system can include dust, air, particles, water, and chemicals. These contaminants can cause wear and tear on the hydraulic components that make up the system. One of the primary sources of contamination is leakage. The same factors that cause oil to leak out of the system can also allow contaminants to enter the system. Dust or particles can lead to premature wear of the components, as they travel with the oil's velocity and grind against the components they encounter. This grinding produces more particles, which are removed from the surface, mix with the fluid, and circulate throughout the system. Contamination can also prevent the mating surfaces of hydraulic valves from achieving a proper seal, creating gaps that can become a source of leakage [1, 11].

Contamination, such as air, can also be harmful. The presence of air in hydraulic systems can have a negative impact on the performance and efficiency. Cavitation is one such issue, which occurs when air bubbles in hydraulic fluid collapse under pressure, resulting in microscopic shock waves. These shock waves can cause hydraulic components to wear and fail prematurely over time. Air trapped in the system can also cause a spongy or sluggish response in hydraulic actuators, reducing their effectiveness and efficiency. This is because air is compressible, unlike hydraulic fluid, leading to inconsistent force transmission.

Moreover, air in hydraulic fluid can reduce its ability to lubricate and cool system components, increasing the risk of wear and overheating. Another issue associated with air in hydraulic systems is increased noise levels due to the vibration and cavitation it causes. Additionally, air in the hydraulic fluid can reduce the system's overall efficiency by introducing compressibility, leading to energy losses and decreased performance [12].

### 2.3.2 Internal and External Leakage

In a hydraulic system, leakage is classified into two types. External and internal leakage. External leakage occurs when hydraulic fluid escapes the system, often visible as oil spills. External leakage poses hazards to both workers and the environment. Large leaks can also affect the system's efficiency by lowering it. When external leakage occurs, the system's internal pressure decreases, affecting the system's response time and slowing it down. It can also cause vibrations that put stress on the components, leading to failure [11]. Internal leakage occurs within the system and can prevent it from responding effectively to load manipulation. For example internal leakage can happen inside of a hydraulic cylinder. Various problems can arise from internal leakage, depending on its location within the system. It can impact temperature, pressure, flow, and velocity, causing the system to operate inefficiently. Both external and internal leakages can potentially cause the system to malfunction, fail, and harm the environment as hydraulic fluid could leak into surrounding nature [1].

### 2.3.3 Other Causes

There are other factors that can contribute to the failure of a hydraulic system. Overheating can cause fluid degradation, reducing the system's effectiveness and leading to component failure. If the system pressure increases beyond the maximum workload, it can potentially cause components in the system, such as seals, hoses, and fittings, to fail. The increase in pressure can directly result from overheating, but it may also stem from other causes. Another cause of failure could be power loss, which might lead to the hydraulic pump's malfunction [12].

## 2.4 Supervisory Control and Data Acquisition

SCADA systems are utilized in the control, monitoring, and analysis of industrial devices and processes. A SCADA system enables the company that owns the site to remotely manage the industrial plants without physically being present. A SCADA system is made up of three major components. Figure 2.3 depicts the three main components.

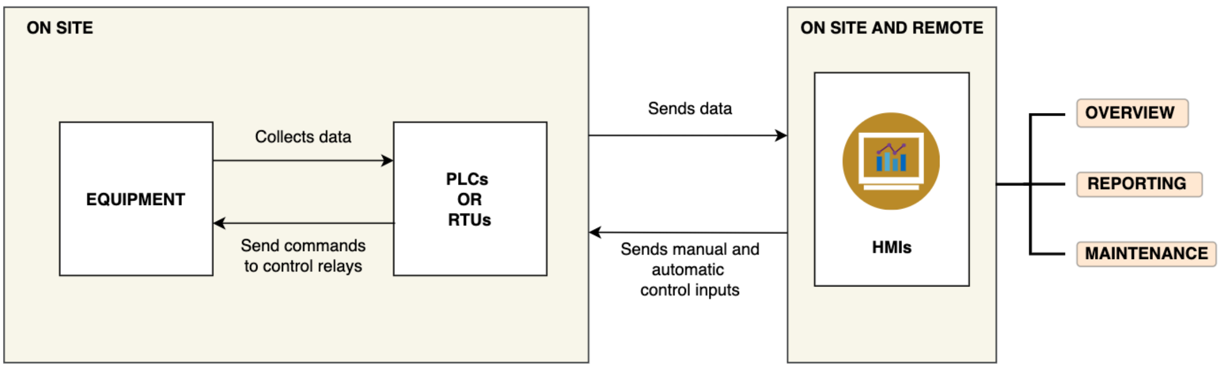


Figure 2.3: SCADA system with the three main system components [13]

The first of the three main system components are the equipment for which the SCADA is used, which could be a solar plant, a hydro plant, or other equipment. PLCs or RTUs and HMI are also part of the SCADA system. PLCs and RTUs are local data collection

points that send and translate information to a remote or on-site HMI. The HMI reacts to control commands from the equipment. The HMI allows the operators to access data, and the software interprets and displays the data in an easy-to-understand manner, allowing the operators to analyze and respond to alarms [13].
With terms of fault detection and diagnosis and the use of already implemented SCADA system their are some advantages. When using already installed software there are no need for more investment in hardware [14].

## 2.5 Traditional Fault Detection and Diagnostics of Hydraulic Systems

The traditional detection and diagnosis methods can be categorized into three types of maintenance: corrective, preventative, and predictive. Corrective maintenance is performed after an unexpected breakdown of machinery. This type of maintenance is unscheduled and can be the most expensive. When a fault occurs unexpectedly, it can trigger additional faults, causing further damage to other parts of the system [15].

Preventative maintenance is scheduled based on a predetermined time schedule and is typically provided by the manufacturer or based on past failure data. During preventative maintenance, critical components of the system are replaced or serviced before failure occurs due to fatigue. While this incurs costs for the plant owner, it is considered a budgeted expense. The replaced or serviced parts may still be functional, and since the replacement is based on a schedule, the replaced parts may have worked for an unknown period. Although predicting the exact timing of failure is challenging, replacing parts during preventative maintenance is generally considered cheaper than replacing them after a failure has occurred [15].

Predictive maintenance involves monitoring the health of the equipment and performing service or part replacement according to an optimized schedule. This method aims to detect potential faults or failures before they happen, based on data analysis and condition monitoring. By adopting predictive maintenance strategies, the maintenance activities can be better planned, reducing unexpected breakdowns and optimizing the utilization of resources [15].

# Chapter 3

# Literature Review

This chapter reviews similar literature on the topic, and their findings.

## 3.1 Fault Diagnosis of a Hydraulic Power System

In [16], the authors propose the use of an artificial neural network for fault diagnosis in a hydraulic system. The paper begins by discussing the significance of fault diagnosis in modern process automation, asserting that it provides a reliable and fundamental design feature for complex engineering systems. The authors identify two specific faults of interest: internal leakage inside the cylinder and spool blockage of the directional control valve, stating that these are the most common faults in a hydraulic power system. Subsequently, they propose a mathematical model for these faults.

The mathematical model of the oil leakage fault is based on the continuity equation, with some assumptions. Equations (3.1) and (3.2) yield the continuity equation for each of the chambers in the actuator during extension motion:

$$Q_1 - Q_{IL} = A_1\dot{x}_p + \dot{P}_1(V_{o1} + A_1 x_p)/\beta_e \tag{3.1}$$

$$Q_{IL} - Q_2 = -A_2\dot{x}_p + \dot{P}_2(V_{o2} + A_2 x_p)/\beta_e \tag{3.2}$$

The internal leakage inside of the actuator are derived from Equation (3.1) and (3.2) as follows:

$$Q_{IL} = \frac{1}{2}\left[Q_2 - Q_1 - (A_1 + A_2)\dot{x}_p - (\dot{P}_1(V_{o1} + A_1 x_p) - \dot{P}_2(V_{o2} + A_2 x_p))/\beta_e\right] \tag{3.3}$$

$Q_1$ and $Q_2$ are the flow rates at the high pressure side and the low pressure side. $Q_{IL}$ is the internal oil leakage flow. $V_{o1}$ and $V_{o2}$ is the initial volumes of the oil of the actuator sides. $A_1$ and $A_2$ is the areas of the piston side. $x_p$ and $\dot{x}_p$ is the displacement and the velocity of the piston rod.

The article assumes that leakage through the cylinder ports are negligible, and therefor the relationship for $Q_1$ and $Q_2$ is,

$$Q_1 = C_d A_p(x_s)\sqrt{2(P_s - P_1)/\rho} \tag{3.4}$$

$$Q_2 = C_d A_p(x_s)\sqrt{2(P_2 - P_e)/\rho} \tag{3.5}$$

where $C_d$ is the valve flow coefficient, $A_p(x_s)$ is the valve port area in regards to the valve spool travel $x_s$. When substituting $Q_1$ and $Q_2$ from Equation (3.4) and (3.5) in Equation (3.3), the internal leakage flow rate across the actuator sides can be expressed as follows,

$$Q_{IL} = \frac{1}{2}\left[\frac{2}{\rho}C_d A_p(\sqrt{P_2 - P_e} - \sqrt{P_s - P_1} - (A_1 + A_2)\dot{x}_p \right. \\ \left. - (\dot{P}_1(V_{o1} + A_1 x_p) - \dot{P}_2(V_{o2} + A_2 x_p))/\beta_e\right] \tag{3.6}$$

Then they continue to make a mathematical model of the valve spool-blockage fault as the travel of the valve spool $x_s$ as follows,

$$x_s = \left[Q_{IL} + A_1\dot{x}_p + \dot{P}_1(V_{o1} + A_1 x_p)/\beta_e\right] / C_d w\sqrt{2(P_s - P_1)/\rho} \tag{3.7}$$

where $w$ is the width of the valve port.
They present parameters for fault indices that they uses in the ANN modeling. The fault indices they use for internal leakage across the actuators chambers,

$$a_1 = \frac{Q_{IL}}{Q_1} \tag{3.8}$$

For valve spool-blockage fault:

$$a_2 = \frac{x_s}{s_s} \tag{3.9}$$

Equation (3.8) and (3.9) are dimensionless. $s_s$ is the total stroke of the valve spool. The values of the fault range between 0 to 1 according to the fault severity. With a healthy system the fault index is zero.

To implement these mathematical models in an ANN, the addition of three more sensors is needed in the hydraulic system: two pressure sensors and one velocity sensor. The paper later explains how these mathematical models can be used to train the ANN model, proposing a fault diagnosis scheme for a hydraulic power system. The focus is solely on two faults, internal leakage in the cylinder and valve spool-blockage fault. Other faults that could occur in a hydraulic system are not accounted for. The authors use an experimental setup to validate the model, concluding that a trained ANN-based model can isolate, detect, and identify the faults in focus.

While this article does not directly apply to this thesis, which uses data from a SCADA system for a HEP, it provides valuable insights into the training of neural networks and understanding hydraulic systems.

## 3.2 Condition Monitoring and Fault Diagnosis of HEPs

Selak et al. [17] presents a condition monitoring and fault diagnostics (CMFD) system for hydropower plants. The authors base their article on the concept of Industrial Product-Service Systems (IPS2) and propose using Support Vector Machines (SVM) for fault diagnostics. The SVM classifier is implemented on a HEP with three Kaplan turbines.

SVM is a supervised learning method that can be used for classification or regression. The authors state that SVM is superior to neural networks trained with the backpropagation algorithm and linear discriminant analysis.

The proposed IPS2 framework for CMFD consists of the following components:

- Work system: This involves the energy transformation process in the HEP, machinery and subsystems, and an operator responsible for system supervision and control.

- CMFD architecture: It performs CMFD and provides necessary information to service providers.
- Operation and maintenance (O&M): O&M carries out activities essential for successful work system operation.
- Service providers: They deliver O&M services.

According to Selak et al. [17], the condition monitoring system measures selected input and output process parameters and processes the data through signal processing, data acquisition, and feature extraction. The data is then transferred to the Virtual Diagnostics Center (VDC) via an information communication infrastructure and data replication. The VDC serves as a data and knowledge base, hosts machine learning applications, web reports, and a collaboration platform. The proposed machine learning method is SVM, which utilizes a knowledge base with a causality table of faults and indicators consistent with faulty components.

In the study, the CMFD system is divided into six steps: signal acquisition, signal analysis and storage, data transfer and storage, data selection, SVM training, and SVM testing. These steps are depicted in Fig. 3.1. The goal of CMFD is to identify if a fault is occurring and determine the cause. The system classifies data into two classes: Class A without faults and Class B with faults. SVM training determines the classes. The data from the pilot project is filtered into four operation regimes: steady state operations, stops and low-power operations, operations during high water flow, and transitions between two operation regimes. These regimes are transformed into four datasets and classified using SVM training. They propose separating the data in four dataset to reduce the amount of data stored.

The classification achieves high training accuracy (99.68%) and reasonable accuracy for test data (97.02%). The study identifies appropriate fault identification models for the four datasets. Dataset 1 is suitable for identifying faults during stops, Dataset 2 for detecting significant changes during operation, and Datasets 3 and 4 for identifying small changes at early stages. Additionally, the research shows that the number of failures increases during operations close to system margins, such as high water flow or below 7 MW power. The HEP in the pilot project has a nominal power of 13MW.

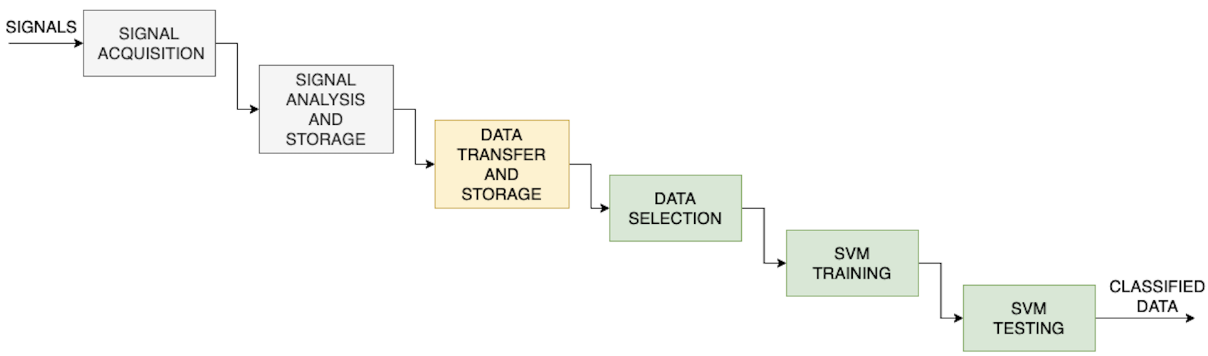


Figure 3.1: Selak et al.[17] proposed CMFD model

They presented system's advantages and contributions include enabling condition monitoring and system assessment based on the IPS2 concept, integrating customers, original equipment manufacturer, and service providers into HEP operations and maintenance. It combines low-frequency data acquisition from the control system and high-frequency data acquisition for monitoring slowly evolving and high-frequency phenomena. The SVM method used for fault diagnostics exhibits high accuracy and enables sensor failure identification. The VDC facilitates access to operational information for service providers, supports condition-based maintenance, and allows collaboration among service providers for better maintenance scheduling.

One drawback mentioned is the long computational time required for SVM analysis. The SVM detector is also only trained for two classes, normal and abnormal data. They claim that two-dimensional models are enough to catch all known failure modes. The paper suggests optimizing the calculation time by reducing the number of training data and using parallel calculation of classification models. The proposed system has reduced condition assessment and diagnostics time, but further enhancements to the VDC's functionalities, such as virtual collaborative environments, are being researched to improve system condition reporting and recording of service activities [17].

# Chapter 4

# Utilizing Python for Data Analysis and Implementation

This chapter describes the programming language chosen for this thesis, and theory about the libraries used.

## 4.1 Python

The code that used for this master thesis is developed in python 3.6 through Anaconda and Spyder. Python 3.6 was released in December 2016. Spyder is an integrated development environment (IDE) that comes pre-installed with Anaconda. The SCADA data that is the variables for this thesis is preprocessed and analyzed with python. For the machine learning process python packages will be discussed below.

### 4.1.1 Scikit-learn

Scikit-learn is a open source toolbox available for Python, and are used to preprocess and analyze the data. It is also used to split the data into training and testing sets that will be the input in the machine learning processes.

### 4.1.2 Keras

To implement a neural network system in the python environment the Keras libraries are implemented. There are other libraries that could have been chosen to preform the NN analysis, but Keras was chosen for this thesis. Keras is an open-source library written in python, and is used for building and training deep learning models. It was devoloped by François Chollet, and was first released in March 2015. Keras is built on top of other deep learning frameworks such as Tenserflow and Theano, and has a user-friendly interface to create and train deep learning models. [18]

## 4.2 Artificial Neural Network

A neural network is a parallel-distributed processor made up of simple processing units, these are called neurons. Artificial Neural Networks (ANN) are a subset of machine learning and are the heart of deep learning algorithms. ANN are inspired by the human brain mimicking the way biological neurons signal to one another. It has the capability to perform pattern recognition and diagnosis that are difficult to describe with analytical diagnosis algorithms since it learns the input patterns by itself [16]. ANNs are comprised of node layers, containing a input layer, one or more hidden layers and an output layer. Each node connects to another and has an associated weight and threshold. If the output of the individual node is above the specified threshold value, the node is activated and the data is send to the next later of the network. Number of nodes, and layers is determined in the code and increasing/decreasing the number, or layer can improve the accuracy of the neural network system. The neural networks rely on training data to learn and improve their accuracy over time. One of the most well-known neural network is Google´s search algorithm [19]. In Figure 4.1 below is a visual of how a neural network operates as described.

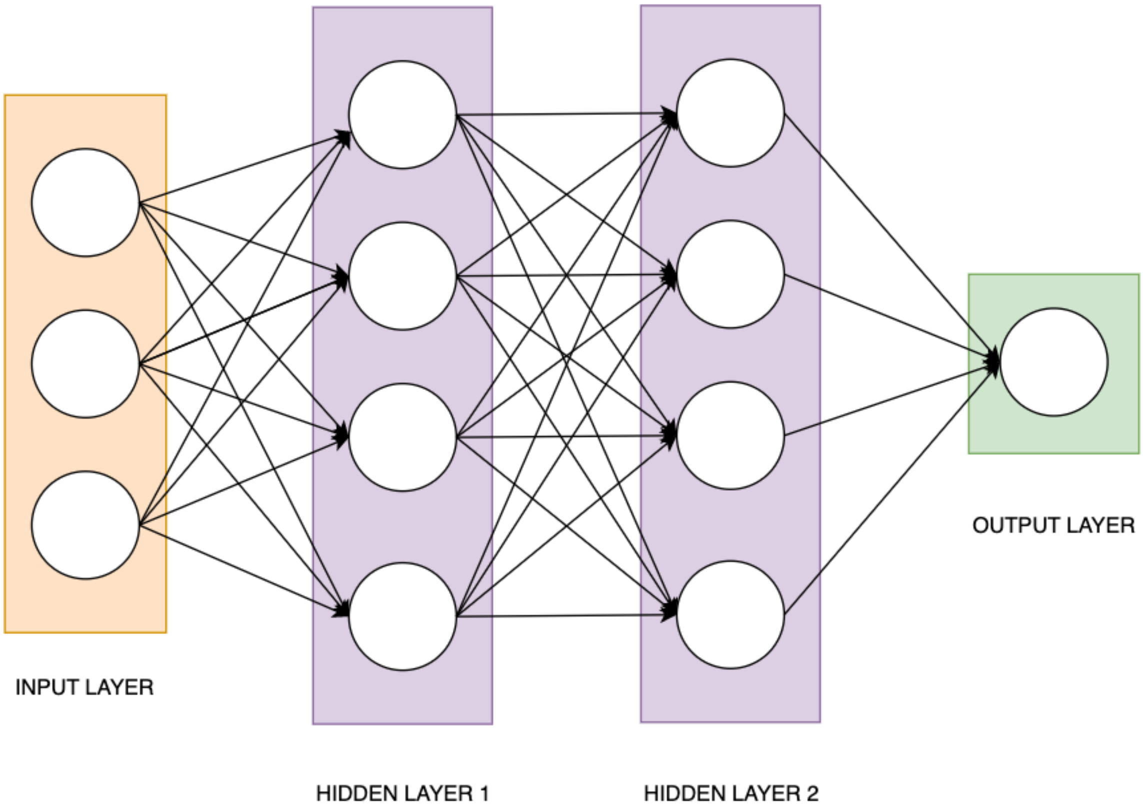


Figure 4.1: Architecture of a artificial neural network [20]

### 4.2.1 Recurrent Neural Network

Recurrent Neural Networks (RNN) are a deep learning technique with an architecture distinct from previously mentioned feedforward architectures. RNNs process data by incorporating a loop in the middle hidden layer, which allows data to be fed back into the hidden layer. This design enables RNNs to have a memory concept, allowing them to store the states of information from previous inputs to generate subsequent outputs. This architecture is particularly suited for handling time series data or sequences.

In this case, since the data depends on previous data to recognize faults in the system, the memory of the RNN treats the data points as interconnected rather than independent. This is achieved by saving the output of a particular layer and feeding it back to the input to predict the next output. Figure 4.2 illustrates the RNN architecture. It takes input X and passes it into the middle layer h, which may consist of multiple hidden layers. The data passes through the hidden layers and loops back as many times as necessary. The primary purpose of this process is to learn sequential or time-varying patterns [21, 22].

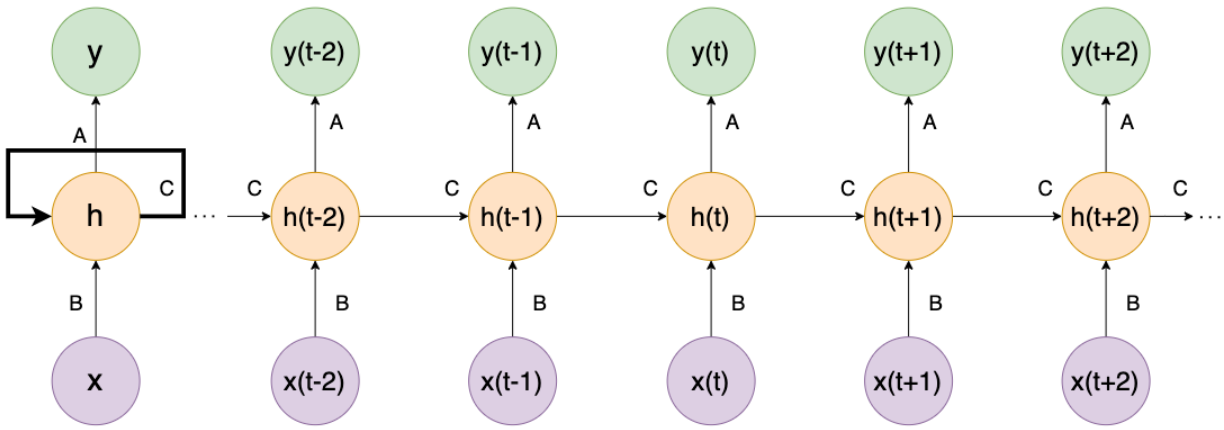


Figure 4.2: Architecture of a recurrent neural network [21]

RNNs use activation functions to introduce non-linearity to the model. The most common activation functions are:

- **Sigmoid function** - The sigmoid function has a range between 0 and 1, and is useful for binary classification task.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

- **Hyperbolic tangent (tanh) function** - The Tanh function has a range between -1 and 1, and is useful for non-linear classification tasks.

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

- **Rectified Linear Unit (ReLU) Function** - The ReLU function has a range between 0 and infinity, and are used in deep neural networks where it is required postive outputs.

$$ReLU(x) = max(0, x)$$

### 4.2.2 Long Short-Term Memory

Long short-term memory (LSTM) networks are a type of deep learning sequential neural network that allows information to persist over time. They are based on RNNs but specifically designed to address the vanishing gradient problem common in standard RNNs. An LSTM unit consists of three gates and an LSTM cell. The LSTM cell can be thought of as a layer of neurons in a traditional feed-forward neural network. The three gates control the flow of information in and out of the LSTM cell [23]. Figure 4.3 provides a visualization of the LSTM cell.

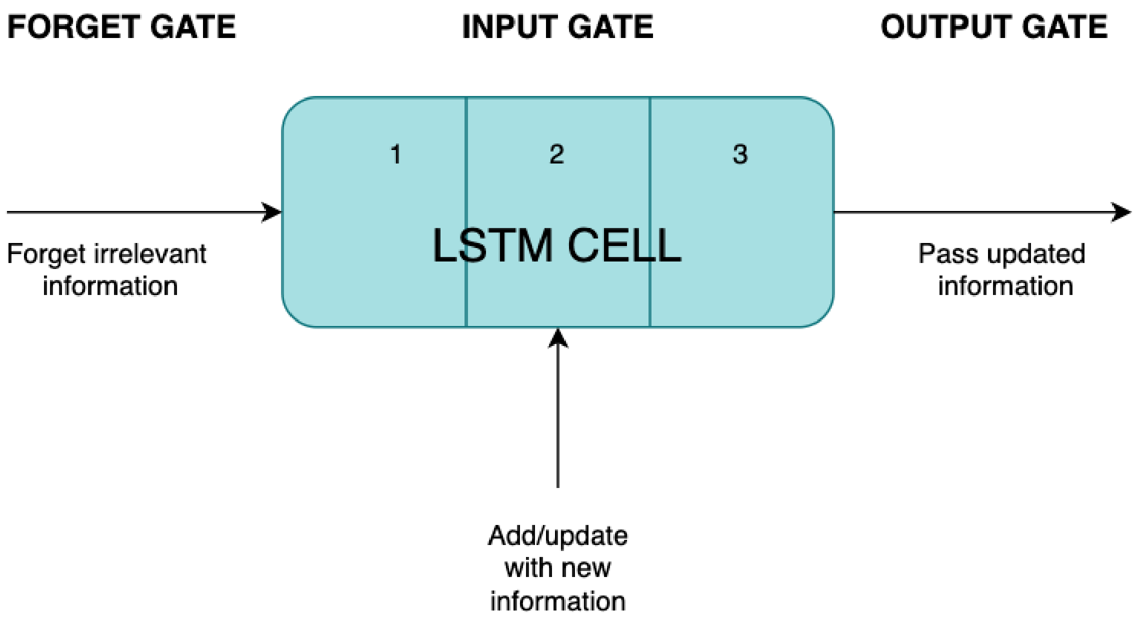


Figure 4.3: Example of a LSTM cell [23]

The first gate is called the forget gate, which determines whether information from the previous time step $Ct-1$ should be retained or discarded. It uses a sigmoid activation function to produce a value between 0 and 1 for each element in the cell state. A value close to 0 means the information should be forgotten, while a value close to 1 means it should be retained [23].

Equation for the forget gate:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{4.1}$$

where $f_t$ is the forget gate's output at time step t. $\sigma$ is the sigmoid activation function, $h_{t-1}$ represent the hidden state from the previous time step and carries information from the past time steps and used as input to the forget gate. $[h_{t-1}, x_t]$ is a concatenation of the previous hidden state and current input. $b_f$ is the bias term for the forget gate [23].

The second gate, known as the input gate, quantifies the importance of new information carried by the input. This also uses a sigmoid activation function to produce a value between 0 and 1. The input gate have a cell state update, this is called $C_t$ [23].

Equation for the input gate:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4.2}$$

where $i_t$ represents the output of the input gate at time step t. $W_i$ is the weight matrix for the input gate and is a set of learned parameters that determine the importance of the input and previous hidden state. $x_t$ is the input vector at the current time step t and represent new information processed by the LSTM cell. The cell state update uses a tanh activation function to generate a candidate cell state, this value ranges from -1 to 1 [23]. The equation for updating the cell state is:

$$C \sim_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4.3}$$

The cell state is updated by combining the results of the forget gate, input gate and cell state update.

The third gate, called the output gate decides which information from the updated cell state should be output as the current hidden state. This gate also have a sigmoid activation function [23].

Equation for the output gate:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{4.4}$$

$o_t$ represents the output gate at time step t.

These three gates and equations enables the LSTM to selectively remember or forget information from the input sequence and effectively learn long-term dependencies [23].

## 4.3 L2-Regularization

In machine learning, the objective is to ensure the selected algorithm performs optimally on both the training data and unseen test data. Should difficulties arise, there are certain strategies that can be implemented to address them. Regularization is one such technique. Regularization is any adjustment or modification to the learning algorithm that aids in reducing errors on the test data set. It's important to note that regularization does not affect the training data set [24].

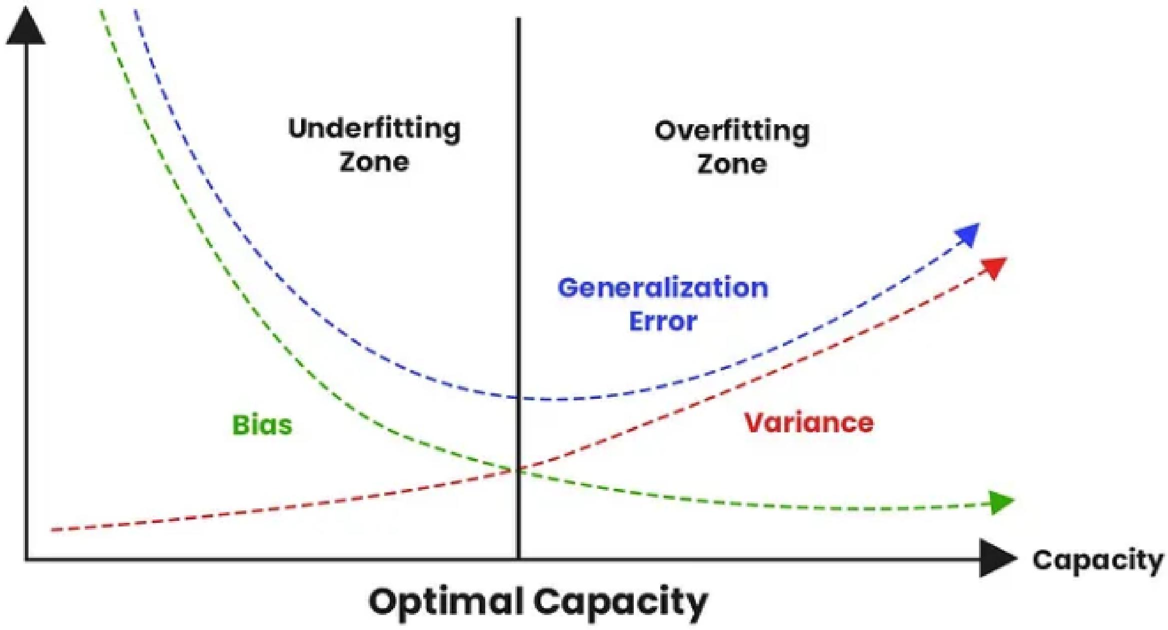


Figure 4.4: Bias versus variance tradeoff [24]

Figure 4.4 illustrates the optimal capacity between the underfitting and overfitting zones. Effective regularization achieves a balance between bias and variance, the outcome of which should ideally be a reduction in variance at the optimum bias level [24].

L2-regularization, also known as Ridge Regression, is represented by Equation 4.5 [24].

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 + \lambda\sum_{j=i}^{n}\theta_j^2\right] \tag{4.5}$$

L2-regularization imposes a penalty as the model's complexity increases. The regularization parameter, lambda, penalizes all parameters except for the intercept, aiding the model in generalizing to the data rather than overfitting. L2-regularization promotes smaller weights in the model, reducing the likelihood of overfitting on the training data and enhancing the model's ability to generalize on the test data [24].

## 4.4 Metrics

To evaluate the models accuracy some statistical metrics will be used. The mean absolute error (MAE) represents the average magnitude of errors between actual and predicted values in the dataset, without considering their direction. It measures the average of the residuals in the dataset [25].

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \tag{4.6}$$

where $y_i$ is the actial value, and $\hat{y}_i$ the predicted value.

Mean squared error (MSE) represents the average the squared difference between the predicted and true values in the dataset. When squaring the difference, it gives more weight to larger error, making it more sensitive to outliers compared to MAE [25].

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y})^2 \tag{4.7}$$

Coefficient of determination or R-squared ($R^2$) represent the proportion of the variance in the dependent variable that is predictable from the independent variable(s). $R^2$ is a measure that indicates how well the predicted values fit the actual values. It ranges from 0 to 1, with higher value indicating a better fit [25].

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} \tag{4.8}$$

where $\bar{y}$ is the mean of the actual values.

# Chapter 5

# Data Analyzing and Model Development

This chapter describes the method used to make the models.

## 5.1 Supervised Learning

Supervised learning combines the data from various sensors in the system to determine if a component has failed. With the data together with machine learning algorithms it is possible to design a model of the system.
The model decides when is the ideal time for a component to having maintenance, reducing the risk of the equipment breaking down [26].

Figure 5.1 is a workflow diagram with the steps used in supervised learning. The first is to prepare the data that is going to be used. Then choosing an algorithm to use to train the model. Supervised learning typically needs multiple iterations of building and evaluating different models. The process begins with a data set. Choose a type of model, then train it with the data. Once the model is trained its performance needs to be evaluated. This is typically done by comparing the model's predictions with known responses. If the model is not satisfying the goal of the model, it is possible to try other models either by making adjustments to the various options for the particular algorithm or by changing the algorithm entirely. Once the model satisfies the needs it is ready to make predictions on new data [26].

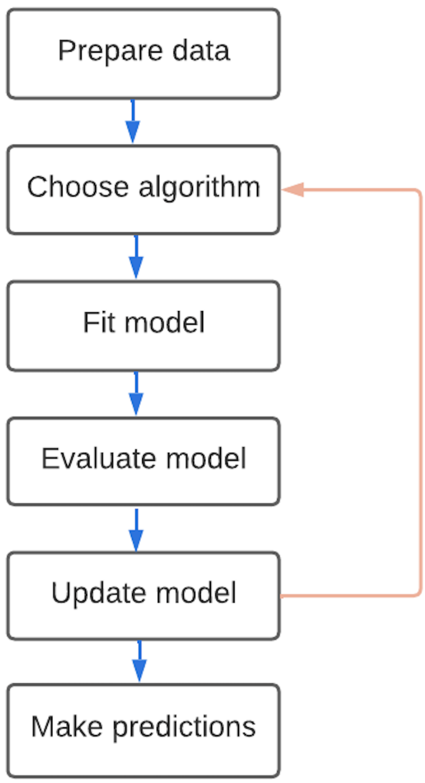


Figure 5.1: Workflow for supervised learning [26]

## 5.2 Evenstad Dataset

The measuring data is from Evenstad hydro power plant, located in Fevik, Grimstad. The data is from various sensors. In Table 5.1 is an overview of the sensors in Evenstad, and some of the measured data is used to calculate other data. The sensor on the guide vanes is used to calculate how extended the cylinder used to open or close the guide vanes in mm. The data from the oil level in percentage is used to calculate how the oil level is in mm.

Table 5.1: Overview of data gathered from Evenstad hydro power plant

| **Sensors** | **Calculated** |
|---|---|
| Guide vane [%][1] | Cylinder position [mm] |
| Oil level [%][2] | Oil level [mm] |
| Oil pressure [bar] | |
| Oil temperature [°C] | |
| Accumulator temperature [°C] | |

The original dataset from Evenstad is shown in Fig. 5.2.

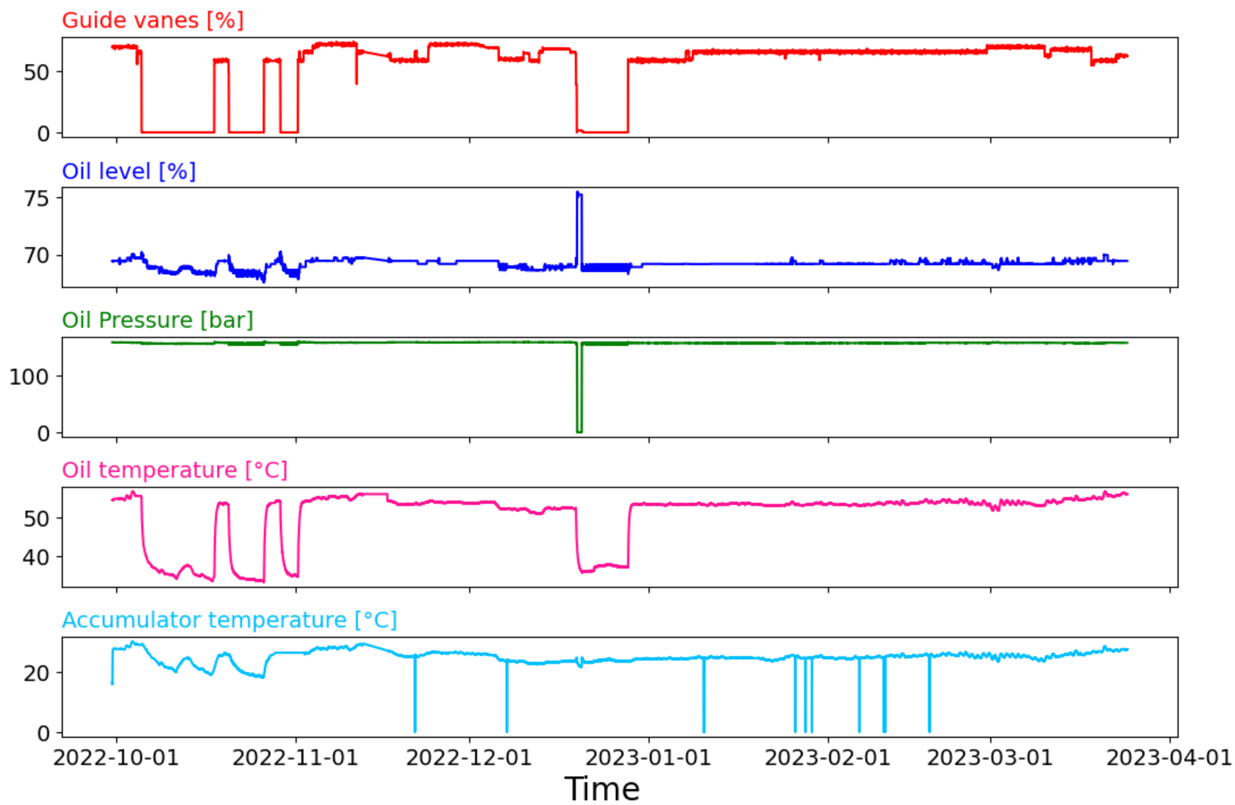


Figure 5.2: Evenstad dataset

The figure shows the five features. Two features are excluded as they are the same as the guide vane position %, and oil level % only shown in mm. The red graph shows the guide vane position in %. The range is from 0% to 100%. 0% means the guide vane is closed, and 100% means that the guide vane is completely open. The normal range is 0% to around 75%. The dark blue graph is the oil level, also shown in percentage. The oil level has a sudden increase and this is assumed a sensor error. From the graph it is possible to see that the normal range for the oil level is around 65% to 74%. The The green graph is the oil

[1]Percentage of accumulator tank volume
[2]Percentage of oil tank volume

pressure, shown in bar. This graph has little to no change and the pressure is around 150 bar. The dip in the pressure is likely a sensor error or trouble sending data to the SCADA system. The pink graph is the oil temperature shown in °C. The sensor is measured from the inside of the oil reservoir. The increase in temperature correlates with when the guide vane is opened. Lastly is the light blue graph, which is the accumulator temperature. This is a new sensor, and is placed outside of the accumulator nitrogen tank. The sensor has multiple sudden decreases, and in the dataset this dips has no value. The reason could be sensor issues.

To see how the different features in the dataset correlates to each other a correlation matrix is produced. The result from this is shown in Fig. 5.3

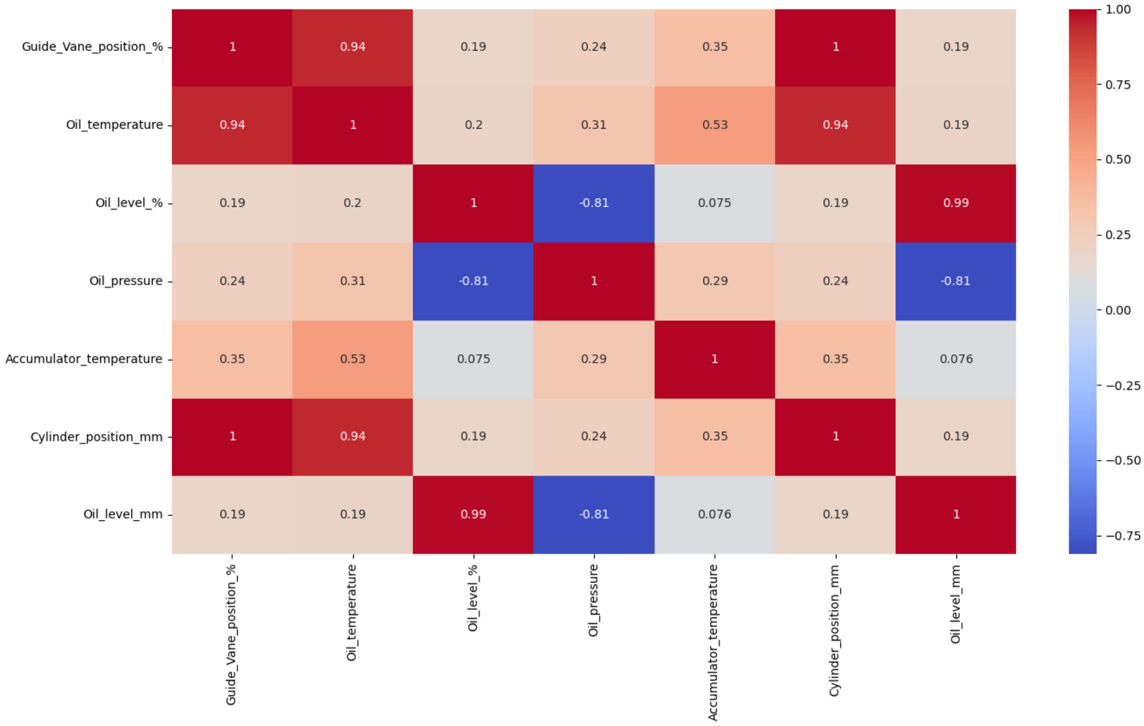


Figure 5.3: Correlation matrix

The correlation matrix includes all of the dataset's features. The reason for the correlation equal to 1 and 0.99 between the guide vane position and oil level shown in percentage to the guide vane position and oil level shown in mm is that they are calculated from each other and are the same variables, only shown in different units. The correlation matrix shows that the correlation between the guide vane position and oil temperature is strong, with a correlation of 0.94. This is also visible to see in Figure 5.2.

## 5.3 Preprocessing of the Dataset

In order to analyze the dataset from Evenstad, it is necessary to preprocess it. As a first step, the initial 363 rows of the dataset were removed because the sensor that samples the temperature of the accumulator was not yet mounted in the time period for the received data. The decision to remove these rows was made because the number of rows affected was relatively small and the impact was considered to be negligible. The next step was to analyze whether the dataset had NaN values. The datset did not have many NaN values, most of them where removed with the removal of the first 363 rows. It is important to analyze if the dataset has NaN values as they can affect the accuracy and reliability of the data analysis results.

In the original dataset the features where sampled in 20-minute intervals. This was resampled in python to be a interval of 10-minutes. The resampling was done by interpolating the column above and below and the result was submitted in the new column in the middle. The code is shown below.

```
#Load excel file
df = pd.read_excel('EVENSTAD.xlsx', index_col='TimeUTC', parse_dates=True)

#Remove the first 363 rows
df1 = df.iloc[363:]

#Resample from 20-minutes to 10-minute intervals
df_resampled = df1.resample('10T').mean()

#Interpolate
df_interpolated = df1_resampled.interpolate(method='linear', ...
    limit_direction='both')

#Store in a new csv file
df_interpolated.to_csv('evenstad_10min.csv')
```

Before the data was used in RNN, and RNN LSTM it was scaled in the range 0-1 with MinMaxScaler. MinMaxScaler is a preprocessing class from scikit-learn library. It works by computing the minimum and maximum value of each feature in the dataset, and then scaling each value to the new range, in this case from 0 to 1.

## 5.4 Hyperparameterization

In machine learning the chosen model has a set of parameters that are learned during training to optimize the performance of the model. There are also other parameters, these are called hyperparameters. Hyperparameters are set prior to the training and determines how the model learns. Hyperparameters can be set by the user, rather than learned from the data. The goal of hyperparameter tuning is to find the best combination of hyperparameters for a given model and task to maximize its performance [27].

There are several methods to hyperparameter tuning, such as: Grid Search, Random Search, Bayesian Optimization etc. For this dataset Grid Search is chosen and it is done with GridSearchCV from Scikit-learn. Grid search is automating the process of trying all possible values to know the optimal values instead of doing it manually. It is done by defining a dictionary in python with a particular hyperparameter with the values it can take [27]. For the RNN LSTM model the parameters for the grid search is mentioned in Tab. 5.2.

Table 5.2: Hyperparameters

| | |
|---|---|
| Neurons | 50, 100, 150 |
| Dropout rate | 0.2, 0.3, 0.4 |
| Activation | relu, tanh |
| Optimizer | adam, rmsprop |
| Layers | 1,2,3 |
| Epochs | 50, 100, 200 |

## 5.5 Model Development

Multiple neural network model are used to make a model based of the dataset from Å Energi. The first model was a simple recurrent neural network. The goal of the model is that the model is going to notice when the system is in distress. The biggest issue with working with HEPs is that they run for years without problems, which can be seen in the data. For the data provided by the company all of the chosen features was inside of normal operation. The options are training the model so that if there will be a change in the operation the model will rule this out as abnormal behavior. To test the model for this a new dataframe with synthetic data is made.

The synthetic data has samples that are outside of the normal operation values. The threshold values are chosen based on the limitation the system has. Such as when the alarms for the system is set to go off. Values that do not have alarms connected are chosen based on the features' normal range, and then the threshold is set outside of this. The threshold values and alarms/normal values are shown in Tab. 5.3

Table 5.3: Alarm Thresholds and Normal Values

| **Feature** | **Alarm/ Normal values** | **Threshold** |
|---|---|---|
| Oil level | 87% - high level<br>50% - low level<br>47.5% - low low level | (46,88) |
| Guide vane | 0 - Closed<br>0-74 - Normal<br>100 - Open | 75 |
| Oil pressure | 160 bar - Maximum workload | 160 |
| Accumulator temperature | 15-30 | (14,32) |

# Chapter 6

# Results and Discussion

This chapter presents the findings for this thesis, and discussion around the findings. Two types of neural network are used, RNN and LSTM.

## 6.1 Hyperparameters

To find the best hyperparameters for the model it was chosen to perform a grid search. The hyperparameters are used as a guideline for improving the model performance. The result from the grid search is presented in Table 6.1 below.

Table 6.1: Result from GridSearchCV

| **Parameter** | **Value** |
|---|---|
| Activation | ReLU |
| Batch Size | 64 |
| Dropout Rate | 0.3 |
| Epochs | 200 |
| Neurons | 50 |
| Optimizer | Adam |
| Number of Layers | 1 |

## 6.2 RNN Model

This RNN model, designed to predict the guide vane position, utilizes input features such as oil level, accumulator temperature, and oil pressure. Its performance has been evaluated through a plot showcasing the model's loss over several epochs.

Python's matplotlib library was employed to generate a plot with two distinctive lines: the green line represents training loss, while the orange line indicates validation loss. The green line illustrates how well the model learns and fits the training data. The loss here is computed by applying the model's current state to the training data and comparing the model's predictions to the actual values using a 'mean_squared_error' loss function. The objective of training is to minimize this loss. As depicted in Fig. 6.1, the training loss commences at a value of 0.05 and decreases to approximately 0 during the initial 20 epochs, indicating effective learning and improved predictions on the training data.

Conversely, the orange line portrays the validation loss, or error for the validation data, providing insights into the model's generalization capability for new, unseen data. If the orange line had increased over the epochs, it would have indicated poor performance on unseen data. However, for this model, the validation loss remained relatively consistent throughout the training process.

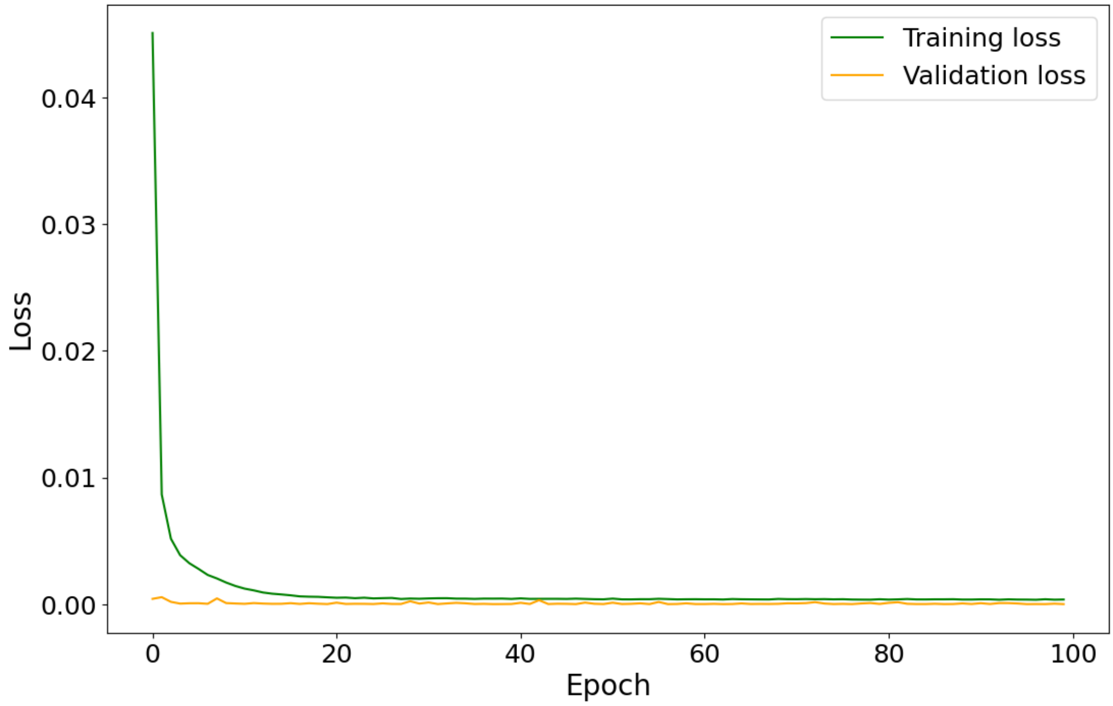


Figure 6.1: Model loss for the RNN model

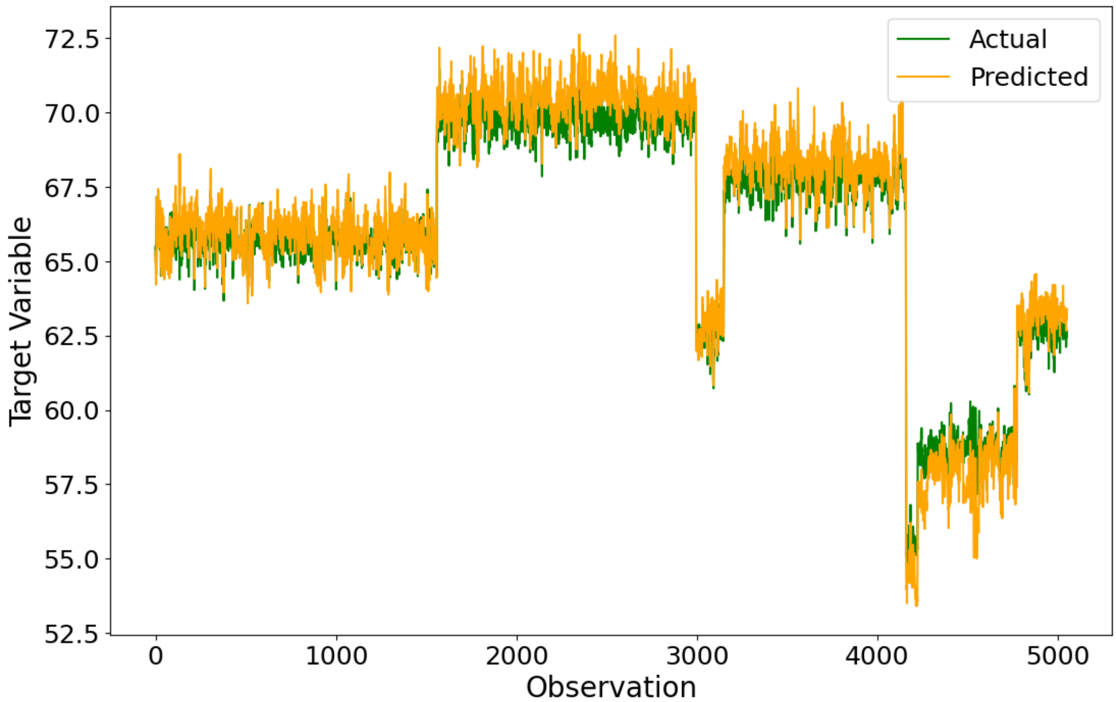


Figure 6.2: Actual versus predicted values from the RNN model

In Fig. 6.2, the green line shows actual values, while the orange line displays predicted values. The close alignment between these two suggests the model's predictions are in good accord with the actual data, as further evidenced by the metrics in Tab. 6.2.

Analysis of Tab. 6.2 reveals the model's Mean Squared Error (MSE) is 0.2959, indicating a low average squared difference between the actual and predicted values, thus implying a good model fit. The Mean Absolute Error (MAE) is 0.4358, meaning the model's predictions are, on average, 0.4358 units away from the actual values, demonstrating a respectable level of prediction accuracy. The $R^2$ score of 0.9775 suggests the model explains roughly 97.75% of the variability in the guide vane position, further attesting to a good model fit.

The results suggest that the RNN model has performed well in predicting the guide vane position, showing strong prediction accuracy and robustness as illustrated in Fig. 6.2. This model is only based on normal values from the dataset.

Table 6.2: Performance Metrics for RNN Model

| **Metric** | **Value** |
|---|---|
| Mean Squared Error (MSE) | 0.2959 |
| Mean Absolute Error (MAE) | 0.4358 |
| $R^2$ Score | 0.9775 |

## 6.3 RNN LSTM Model

The RNN LSTM model has the same input as the RNN model from section 6.2. Fig. 6.3 illustrates the model loss for the RNN LSTM model. It employs the same calculations mentioned in the RNN model results. The training loss, depicted in green, starts at approximately 0.16 and decreases to 0.04 at around 70 epochs. Conversely, the validation loss shown in orange starts at approximately 0.04 and exhibits oscillations between 0.04 and 0.01 throughout the epochs duration. Although the model achieves a lower loss during training, the oscillating validation loss suggests it struggles to generalize well to unseen data.

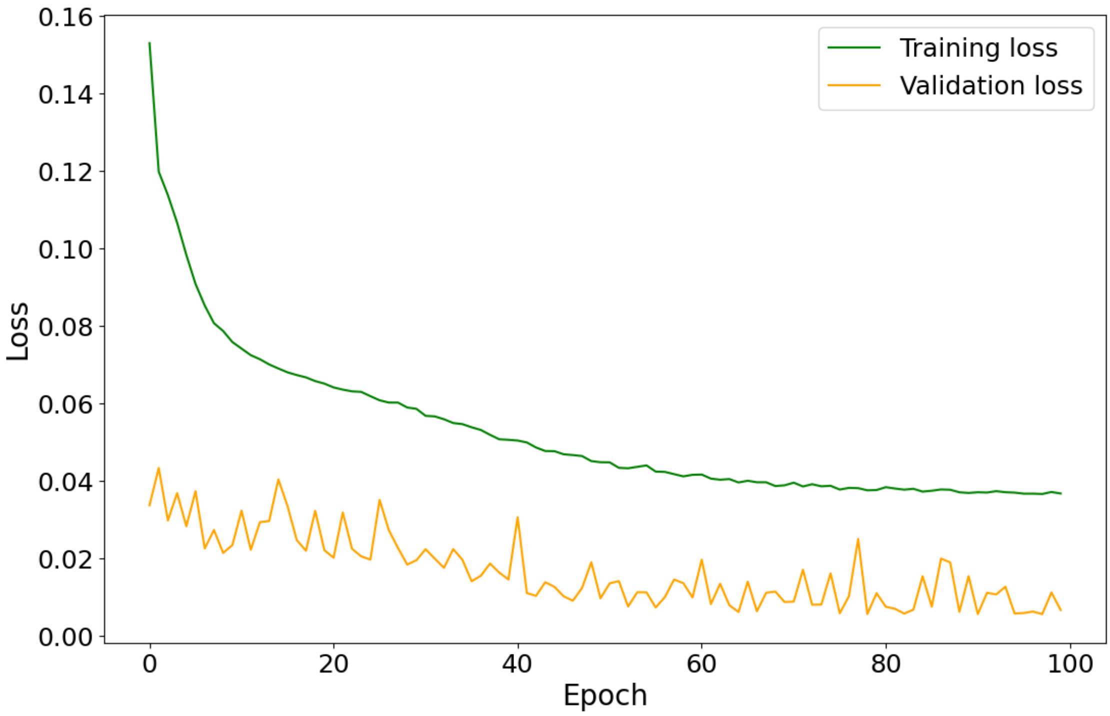


Figure 6.3: Model loss for the RNN LSTM model

Figure 6.4 displays the actual values (green plot) versus the predicted values (orange plot) from the RNN LSTM model. It is evident that the predicted values do not closely follow the actual values. While the initial predictions may appear reasonable, the model's performance deteriorates as the observations progress. This indicates that the RNN LSTM model struggles to capture the underlying patterns and dependencies in the data.

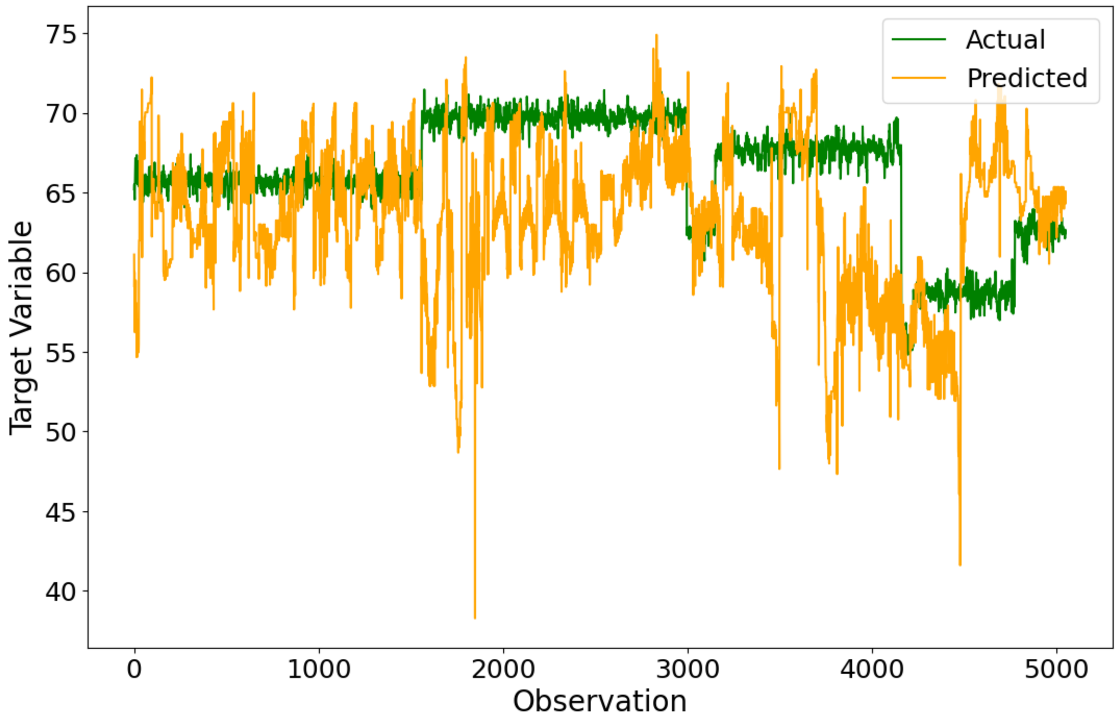


Figure 6.4: Actual versus predicted values from the RNN LSTM model

Tab. 6.3 presents the performance metrics for the RNN LSTM model. The Mean Absolute Error (MAE) is calculated to be 4.6973, implying that the model's predictions deviate from the actual values by approximately 4.70 units on average. The Mean Squared Error (MSE) is computed as 36.8465, indicating an average deviation of 36.85 units squared between the model's predictions and the actual values. The R-squared ($R^2$) score, -1.8011, confirms the observations made in Fig. 6.4. The negative $R^2$ score signifies that the model's predictions do not fit the data well and perform worse than a model that simply predicts the mean of the target variable.

The RNN LSTM model struggles to accurately predict the target variable. The model's predictions do not closely align with the actual values, and the loss values show instability without significant improvement during training. The relatively high MAE and MSE, along with the negative $R^2$ score, further indicate that the model's performance is subpar and does not effectively capture the underlying patterns in the data.

Table 6.3: Performance Metrics for the RNN LSTM Model

| **Metric** | **Value** |
|---|---|
| Mean Squared Error (MSE) | 36.8465 |
| Mean Absolute Error (MAE) | 4.6973 |
| $R^2$ Score | -1.8011 |

## 6.4 RNN LSTM Model with L2-regularization

The RNN LSTM model struggled to generalize well to unseen data, it could be beneficial to implement L2 regularization. This regularization technique can help reduce overfitting and improve the model's generalization ability to new observations.

The model loss for the RNN LSTM model with L2 regularization is shown in Figure 6.5. By comparing Figure 6.3 and Figure 6.5, it can be observed that the loss improves with the addition of L2 regularization.

However, there is no improvement when comparing the actual and predicted values for the RNN LSTM model without and with L2 regularization. In some cases, the addition of L2 regularization may even make the predictions appear worse. While L2 regularization can help improve the model's generalization ability, it may not necessarily improve the accuracy or closeness of the predicted values to the actual values. Other factors, such as the model architecture and hyperparameter tuning, could improve the performance of the RNN LSTM model in predicting the target variable accurately.

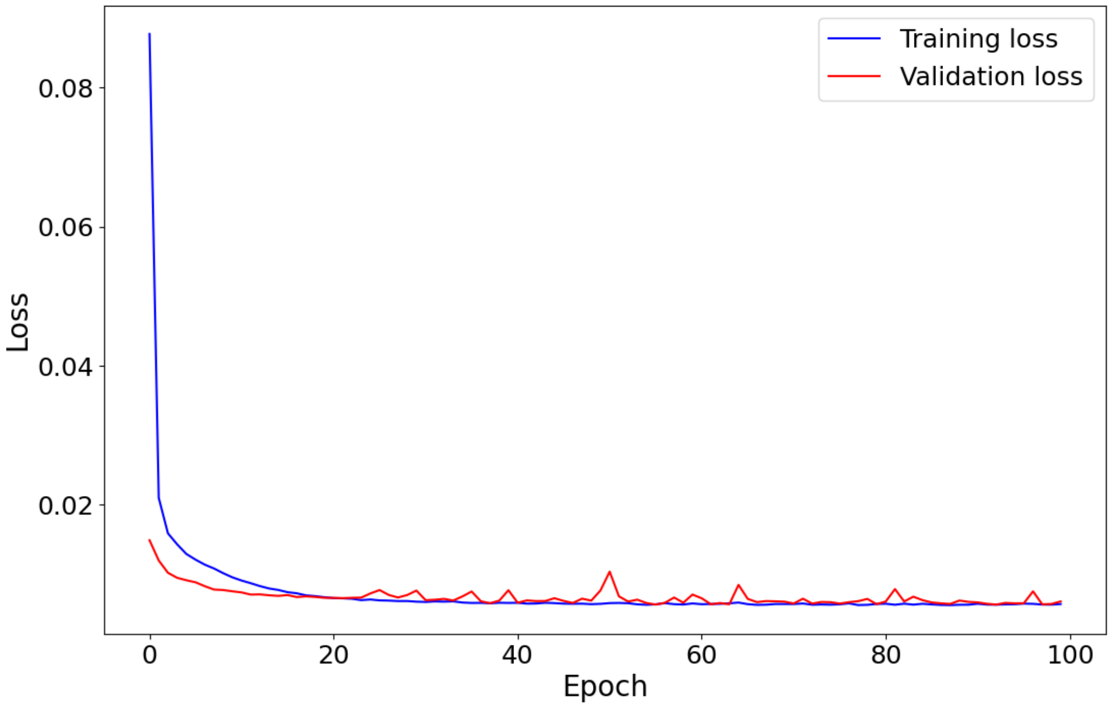


Figure 6.5: Model loss for the RNN LSTM model with L2-regularization

Table 6.4 presents the metrics for the RNN LSTM model with L2-regularization. When comparing these metrics to those of the RNN LSTM model without L2-regularization, there are some improvements, although they are insignificant.

Table 6.4: Performance Metrics for the RNN LSTM Model with L2 regularization

| **Metric** | **Value** |
|---|---|
| Mean Squared Error (MSE) | 29.6410 |
| Mean Absolute Error (MAE) | 4.0649 |
| $R^2$ Score | -1.2534 |

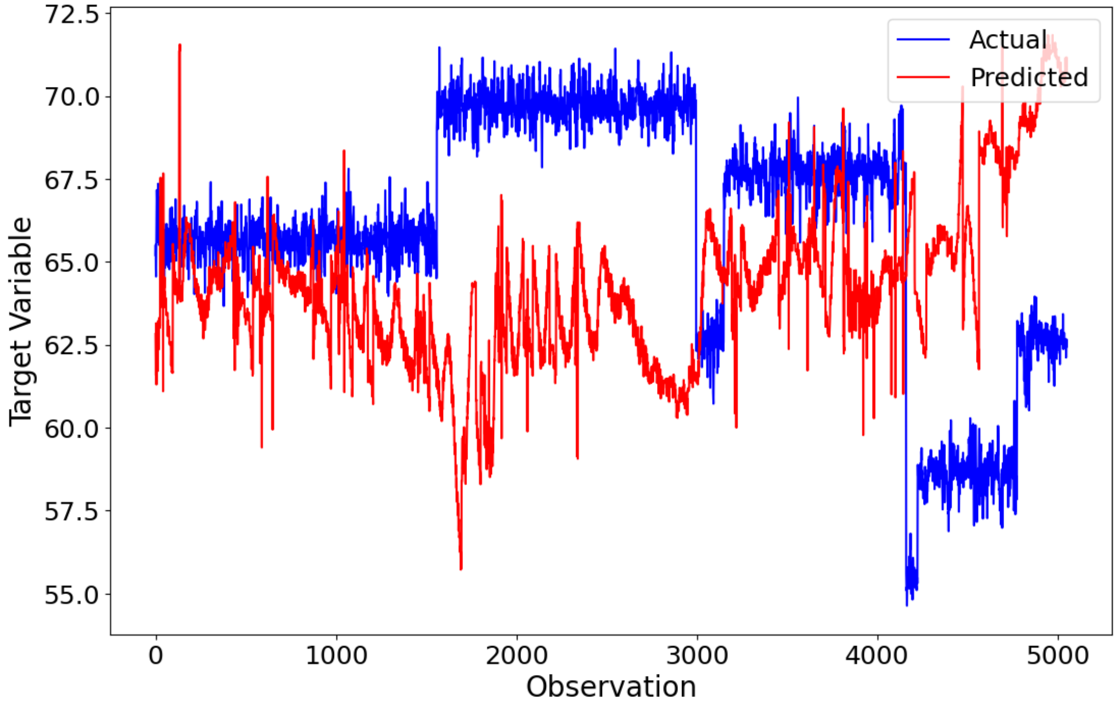


Figure 6.6: Predicted versus actual values for the RNN LSTM model with L2-regularization

## 6.5 Synthetic Data RNN Model

The original dataset contained no values outside the scope of normal operations, therefor a new dataset that included synthetic abnormalities was created to assess how the model responds to faulty data. These abnormal data points were generated based on the system's alarm settings, which trigger an alarm in the SCADA system if a set threshold is exceeded. Some features, however, did not have any associated alarm settings. For these features, the faults implemented were chosen based on an analysis of the values during steady-state and start/stop operations. The values in the original dataset reflected the system's behavior during these states, so the threshold values for abnormal conditions were chosen to be outside these operational ranges. The threshold values are detailed in Table 5.3 in Chapter 5.

The dataset incorporating abnormal values was implemented in the RNN model, as it demonstrated superior performance compared to the LSTM variant. Fig.6.7 illustrates the losses experienced by the RNN model when processing the abnormal data. The model losses are more pronounced than those experienced by the RNN model using the original dataset, now resembling the loss patterns of the LSTM model shown earlier. While the training loss decreases over the epochs, the validation losses oscillate between 0.02 and 0.04, occasionally spike up to 0.06.

Fig.6.7 also presents the predicted values against the actual values for the RNN model handling anomalous data. The predicted values are marked in green, while the actual values are represented in red. Black dots indicate anomalies. The model's performance appears to decline when the actual values incorporate anomalies. For instance, around the 1500-observation mark, the anomalies first appear in the actual values. The predicted values during this period oscillate significantly. This oscillation pattern repeats in subsequent observations when the actual values contain anomalies. This may suggest that the RNN model struggles to comprehend the anomalies, and as a result, cannot accurately predict their occurrence.

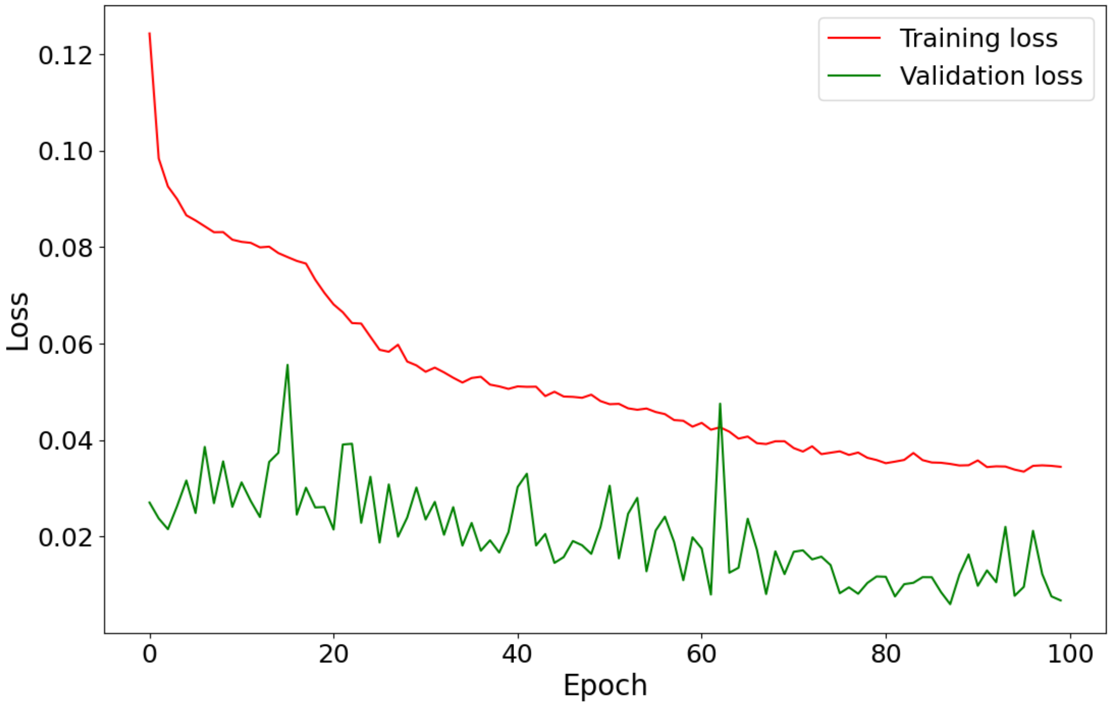


Figure 6.7: Predicted versus actual values for the RNN model with abnormal data

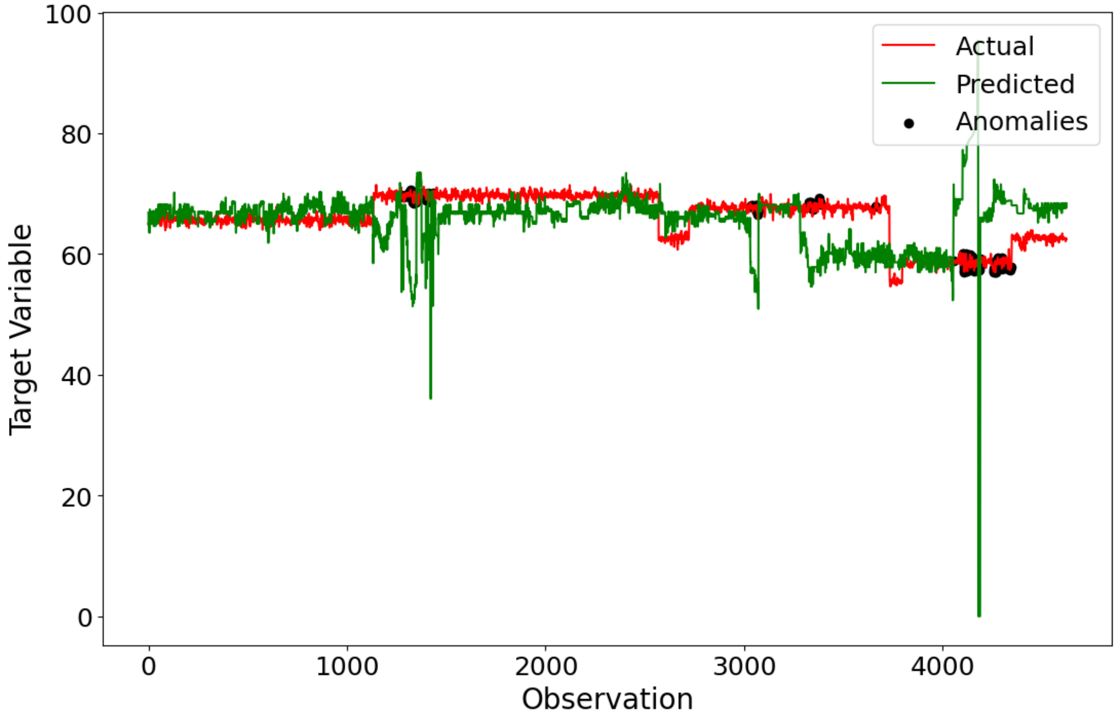


Figure 6.8: Model loss for the RNN model with abnormal data

The performance metric for the RNN model with abnormal data is shown in Tab.6.5. The MSE was recorded at 36.9305. This relatively high value indicates substantial disparity between the model's prediction and the actual data. The model's performance in terms of prediction precision was measured with the MAE value was 3.9285. This value again highlights a significant error level in the model's predictions. The $R^2$ score was -1.5785. Negative $R^2$ score suggest that the model fails to accurately capture the variation in the data, thus indication poor performance in it predictive ability.

Table 6.5: Performance Metrics for the RNN Model with abnormal data

| **Metric** | **Value** |
|---|---|
| Mean Squared Error (MSE) | 36.9305 |
| Mean Absolute Error (MAE) | 3.9285 |
| $R^2$ Score | -1.5785 |

## 6.6 Comparison of the Models

When comparing the performance of the RNN model and RNN LSTM model, several factors can contribute to the RNN model outperforming the LSTM model. One possible explanation is the difference in model architecture. RNNs and LSTMs have distinct architectural designs, with LSTMs specifically designed to capture long-term dependencies in sequential data. If the data does not exhibit significant long-term dependencies or the LSTM model is not properly configured, it may not perform as well as the simpler RNN model. The architectural design choices play a crucial role in determining model performance, and an inadequate configuration can be the result in inferiors metrics for the RNN model.

Another factor could be the hyperparameter tuning. Eventhough the code has been run with multiple configurations, the results can indicate that it needs more tuning to predict more efficiently.

It can also be because of the amount of training data available. Deep learning like LSTM generally requires much training data to effectively learn complex patterns. The training data is limited because the company provides the data. RNN can be more robust to smaller datasets, allowing it to outperform the LSTMs in these scenarios.

Overfitting is an important consideration in evaluating the models. Overfitting occurs when a model becomes too specialized in the training data and fails to generalize well to unseen data. LSTMs are more susceptible to overfitting, particularly when many parameters and the training data are limited.

## 6.7 General Discussion

It is evident that implementing a fault detection and diagnosis model in the power plant system could provide significant value. The growing reliance on volatile energy sources in the grid requires a more reliable alarm system. This, in turn, leads to increased start and stop operations of the hydro power plant to maintain grid balance. Monitoring the health of the system components responsible for controlling these start and stop operations is a logical step towards effective monitoring.

However, implementing fault detection and diagnosis solely based on SCADA data has proven challenging in this thesis. The SCADA measurements have challenges in that case that many of the measurement can be effected by the surroundings, and what the measured value means for the system. For example an increase in oil temperature can mean that the system's pressure is increasing but what makes the system pressure increase. Several faults can make changes in the values, so to differentiate what this means is difficult. Another example is that a decrease in oil level can indicate a leak, but the system is large, so finding where the leak is not easy. Especially if you have internal leakage explained in section 2.3.2.

The most viable approach may be the need for more precise measurements and the adoption of mathematical models, as suggested in [16]. The unique aspect of hydro power is its prolonged fault-free operation over many years, complemented by time-based maintenance practices to ensure reliability. Consequently, analyzing faults without faulty data has proven to be difficult. The synthetic data generated did not effectively train the model to comprehend the impact of values exceeding the alarm threshold on the system and predict faults.

A model trained solely on normal or healthy data could effectively understand and predict outcomes under these conditions. However, this narrow focus might be challenging when the model encounters data points outside the normal operational values. This limitation becomes particularly significant if the model is expected to detect faults and abnormalities for efficient maintenance. Should a fault occur, the model might fail to detect it due to the drastic differences between the healthy data it was trained on and the faulty data. The model might mistake or overlook the fault data for noise, missing the crucial early warning signs of potential system failure. This oversight could increase maintenance costs, equipment damage, and unplanned downtime.

# Chapter 7

# Conclusions

This chapter concludes the research findings of the thesis and suggestions for further work.

## 7.1 Research Findings

In the course of this thesis, two neural network models where created and assessed to predict the guide vane position based on various input features such as oil level, accumulator temperature and oil pressure. The models developed and evaluated were an RNN model and an RNN LSTM model.

Among these, the RNN model exhibited superior performance in term of prediction accuracy and robustness, as evidenced by the low MSE and MAE values, and high $R^2$ score. Its performance remained consistent with the validation data, demonstrating good generalization capabilities. Moreover, the model's predicted values closely aligned with the actual values.

Contrastingly, the RNN LSTM model, despite the potential of capturing long-term dependencies, struggled to predict the target variable accurately. The model had a high MSE and MAE value, and negative $R^2$ score. The introduction of L2-regularization to the RNN LSTM model showed some improvement in loss, but not significantly in the accuracy of predictions. Additionally, a synthetic data RNN model was developed, incorporating abnormal values to assess how the model handles faulty data. The RNN model's performance declined with the introduction of anomalies, with validation losses oscillating significantly, and the predicted values deviating from the actual values.

## 7.2 Further Work

The results of the current models did not meet the expected outcomes, pointing towards a need for further investigation and refinement in several key areas.

- **Hyperparameter Optimization:** Both neural network models, namely the Recurrent Neural Network (RNN) and the Long Short-Term Memory (LSTM) model, showed room for improvement. Fine-tuning the hyperparameters might lead to better model performance.

- **Investigation of Different Algorithms:** While this thesis focuses on using neural networks, other machine learning algorithms might be effective for this task. Algorithms such as decision trees, support vector machines, or ensemble methods could be explored, depending on the nature of the SCADA data. Additionally, unsupervised learning algorithms or anomaly detection techniques could be useful in identifying abnormal patterns in the data, potentially leading to more effective fault detection.

- **Experimental Setup for Known Failures:** As the current model struggled to identify and respond effectively to faults, a valuable next step would be to create an experimental setup that allows for the generation of known failures. Such a setup could provide valuable data for training and validating the model, allowing it to effectively learn and respond to specific fault conditions. This would involve manipulating the system under controlled conditions to induce known faults and then accurately collecting and labelling this data. This enriched training set could then help improve the model's capability in real-world fault detection and diagnosis.

# Bibliography


[1] H. Krakeli. *Fault Detection and Diagnostics of Hydraulic Systems in Hydroelectric Power Plants [unpublished manuscript]. Faculty of Engineering and Science, University of Agder.* 2022.

[2] Statkraft. *Hydropower*. https://www.statkraft.com/what-we-do/hydropower/. Accessed: 2023-04-10.

[3] Norwegian Ministry of Climate and Environment. *Norway's Climate Action Plan for 2021-2030*. 2021. URL: https://tinyurl.com/25mepn9y.

[4] Statkraft. *Statkraft's Low Emissions Scenario: The Energy Transition in a conflicted world.* 2022. URL: https://www.statkraft.com/lowemissions/.

[5] M. Mazzoleni, G. D. Rito, and F. Previdi. *Electro-Mechancical Actuators for the More Electric Aircraft*. Springer, 2021. ISBN: 978-3-030-61798-1. DOI: https://doi.org/10.1007/978-3-030-61799-8.

[6] Q. Zhang. *Basics of Hydraulic Systems*. CRC Press, 2009.

[7] Freudenberg. *Accumulators*. 1.04.2023. URL: https://www.fst.com/sealing/products/accumulators/hydraulic-accumulators/.

[8] Power & Motion. *Fundamentals of Hydraulic Reservoirs*. https://www.powermotiontech.com/hydraulics/reservoirs-accessories/article/21882642/fundamentals-of-hydraulic-reservoirs. Powermotiontech.com. Jan. 2012. (Visited on 03/08/2023).

[9] Ö. F. Eker. *Hydropower part I: application and equipment*. 1.04.2023. URL: https://www.kavaken.com/blog/hydropower-part-i-application-and-equipment.

[10] Linquip Team. *4 Types of Hydraulic Valves and Their Working Principle*. 5.08.2021. URL: https://www.linquip.com/blog/types-of-hydraulic-valves/.

[11] M. Heestand. "Manage leakage in hydraulic-valve design." In: *Plant Engineering* (2016). URL: https://www.plantengineering.com/articles/manage-leakage-in-hydraulic-valve-design/.

[12] V.G. Magorien. *Effects of Air on Hydraulic Systems*. Accessed: 2023-02-23. Power Motion Technology. 1967. URL: https://www.powermotiontech.com/hydraulics/hydraulic-fluids/article/21883212/effects-of-air-on-hydraulic-systems.

[13] SCADA International. *What is SCADA?* [Accessed: 19-02-2023]. URL: https://scada-international.com/what-is-scada/.

[14] W. Yang, R. Court, and J. Jiang. "Wind turbine condition monitoring by the approach of SCADA Data Analysis." In: *Renewable Energy* 53 (2013), pp. 365–376. DOI: 10.1016/j.renene.2012.11.030.

[15] K. Khan et al. "Recent trends and challenges in predictive maintenance of aircraft's engine and hydraulic system." In: *Journal of the Brazilian Society of Mechanical Sciences and Engineering* 43 (Aug. 2021). DOI: 10.1007/s40430-021-03121-2.

[16] A. El-Betar et al. "Fault Diagnosis of a Hydraulic Power System Using an Artificial Neural Network." In: *Journal of King Abdulaziz University-Engineering Sciences* 17 (Jan. 2006). DOI: 10.4197/Eng.17-1.7.

[17] L. Selak, P. Butala, and A. Sluga. "Condition monitoring and fault diagnostics for hydropower plants." In: *Computers in Industry* 65 (Mar. 2013). DOI: https://doi.org/10.1016/j.compind.2014.02.006.

[18] Wikipedia contributors. *Keras — Wikipedia, The Free Encyclopedia*. [Accessed 21-04-2023]. 2023. URL: https://en.wikipedia.org/wiki/Keras.

[19] IBM. *What are neural networks?* 10.04.2023. URL: https://www.ibm.com/topics/neural-networks.

[20] L. Dormehl. *What is an artificial neural network? Here's everything you need to know*. 10.04.2023. URL: https://www.digitaltrends.com/computing/what-is-an-artificial-neural-network/.

[21] Simplilearn. *Recurrent Neural Networks (RNN): A Beginner's Guide*. https://www.simplilearn.com/tutorials/deep-learning-tutorial/rnn. 2021.

[22] F. Chollet. *Deep Learning with Python*. Manning Publications, 2017.

[23] S. Saxena. *Introduction to Long Short-Term Memory (LSTM)*. Analytics Vidhya. Mar. 2021. URL: https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/.

[24] U. Tewari. *Regularization - Understanding L1 and L2 regularization for Deep Learning*. https://medium.com/analytics-vidhya/regularization-understanding-l1-and-l2-regularization-for-deep-learning-a7b9e4a409bf. Accessed: 2023-03-16. Nov. 2021.

[25] A. Chugh. *MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared: Which Metric is Better?* https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e. 8.12.2020.

[26] A. Bayas. *Machine Learning with MATLAB*. Online. URL: https://matlabacademy.mathworks.com/details/machine-learning-with-matlab/mlml.

[27] Great Learning Team. *Hyperparameter Tuning with GridSearchCV*. [Accessed: 29-04-2023]. 2022. URL: https://www.mygreatlearning.com/blog/gridsearchcv/.