Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for elkrafteknikk

Vilde Brennmoen

Predictive Maintenance and Analytics in Hydroelectric Power Plants

Masteroppgave i Energi og Miljø
Veileder: Ümit Cali
Medveileder: Ugur Halden
Juni 2023

NTNU
Kunnskap for en bedre verden

Vilde Brennmoen

Predictive Maintenance and Analytics in Hydroelectric Power Plants

Masteroppgave i Energi og Miljø
Veileder: Ümit Cali
Medveileder: Ugur Halden
Juni 2023

Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for elkraftteknikk

NTNU
Kunnskap for en bedre verden

Abstract

In light of the digital green shift, the European energy crisis, and the emergence of a competitive energy market, the exploration of predictive maintenance in the hydropower sector becomes essential. This thesis explores the field of anomaly detection, with a specific emphasis on using artificial intelligence (AI) and machine learning (ML) techniques. An analysis is conducted using real-world data obtained from a power transformer located at Duge, a hydroelectric power plant in western Norway. A test case has been developed based on the key features of winding temperature, hydrogen concentration, and active power. This test case is founded on a comprehensive review of the literature encompassing anomaly detection and predictive maintenance in hydroelectric power systems.

This study presents six models that are based on different algorithms, namely k-nearest neighbors, one-class support vector machines, isolation forests, local outlier factors, artificial neural networks, and autoencoders. Before being tested on a separate test set, each model is trained and tuned with a dedicated training and validation set. The thesis offers a detailed illustration of this procedure in the form of a flowchart. Model evaluation is conducted based on precision, recall, and F1-score, coupled with the receiver operating characteristic curve (ROC). Although none of the models performed exceptionally well, each one showed better predictive abilities than random chance. The area under the curve (AUC) ranged from 0.56 to 0.76. The recall scores for class 1, which is considered anomalous, ranged from 0.15 to 0.54. At the same time, all models were able to maintain a consistent F1-score of 0.99 to 1 for class 0, which is the non-anomalous class.

Although these findings are purely preliminary and the models do not score well enough for reliable use, they provide an effective foundation for future research efforts, which are suggested in the thesis conclusion. The field of predictive maintenance has already been successfully incorporated into other industries, and this thesis hopes to encourage further work to achieve the same for hydroelectric power plants.

Abstrakt

I lys av det digitale grønne skiftet, energikrisen i Europa og fremveksten av et mer konkurransepreget energimarked, er det essensielt å utforske vedlikehold strategien prediktivt vedlikehold innen vannkraft sektoren. Denne masteroppgaven utforsker dette feltet med spesifikk vekt på anomali deteksjon ved bruk av kunstig intelligens og maskinlæring teknikker. En analyse utføres ved hjelp av data hentet fra en krafttransformator tilhørende Duge vannkraftverk som er lokalisert i Vest-Norge. Et testforsøk er konstruert basert på data fra sensorene som registrerer viklingstemperatur, hydrogen konsentrasjon og aktiv effekt. Dette testforsøket er basert på omfattende gjennomgang av litteratur tilhørende prediktivt vedlikehold strategi og anomali deteksjon i vannkraftsystemer.

Denne studien presenterer seks modeller basert på forskjellige algoritmer. Disse algoritmene er k-nearest neighbours, one-class support vector machines, isolation forest, local outlier factor, artificial neural networks og autoencoders. Først bli modellene trent og finjustert for et dedikert trenings- og valideringssett, før de videre blir testet på et separat testsett. Oppgaven tilbyr en detaljert illustrasjon av denne prosedyren i form av et flytskjema. Evalueringen av modellene blir gjort ved hjelp av presisjon, dekning og F1-score, i kombinasjon med en ROC-kurve (receiver operating characteristic). Selv om ingen av modellene presterte eksepsjonelt godt demonstrerte hver enkelt bedre klassifiserings evner enn tilfeldig gjetting. Området under kurven (AUC) varierte fra 0.56 til 0.76. Dekning verdiene for klasse 1, som er klassen tilhørende anomaliene, varierte fra 0.15 til 0.54. Dette uten å ødelegge for verdiene til klasse 0, klassen uten anomalier, som opprettholdt en F1-score på 0.99 til 1.

Selv om disse funnene bare er innledende og modellene ikke presterer bra nok for pålitelig bruk, kan de fungere som et effektivt grunnlag for fremtidig arbeid og testing som er foreslått i slutten av oppgaven. Prediktivt vedlikehold er allerede vellykket innlemmet i andre industrier, og denne avhandlingen håper å bidra og oppmuntre til videre arbeid for å oppnå det samme i vannkraft sektoren.

Preface

This thesis is the final product of my Master's program in Energy and Environmental Engineering, with specialization on Electric Power Systems. I express profound gratitude to those who have provided guidance throughout my journey.

I am deeply grateful to Professor Ümit Cali, my primary supervisor, for his invaluable guidance and for providing me with the opportunity to explore my specific area of interest. I would also like to express my gratitude to my co-supervisor, Ugur Halden, who was always available to discuss and answer questions concerning my thesis work.

I am grateful to the Sira-Kvina power company for their assistance during my summer internship, which provided me with valuable insights into predictive maintenance in hydropower and data for my thesis.

Last but not least, I'd like to extend my heartfelt gratitude to my classmates, especially at Hovedbygget Salen for making every day working on the thesis a little bit more cheerful.

Contents

List of Figures

List of Tables

Abbreviations

- **AE** Autoencoders
- **AI** Artificial Intelligence
- **ANN** Artificial Neural Network
- **AUC** Area Under the Curve
- **BCE** Binary Cross-Entropy
- **BST** Binary Search Tree
- **CBA** Condition-Based Assessment
- **CBM** Condition-Based Monitoring
- **DGA** Dissolve Gas Analysis
- **iForest** Isolation Forest
- **IoT** Internet of Things
- **iTrees** Isolation Trees
- **KNN** K-Nearest Neighbors
- **LOF** Local Outlier Factor
- **LRD** Local Reachability Density
- **ML** Machine Learning
- **MSE** Mean Squared Error
- **OCSVM** One-Class Support Vector Machine
- **PD** Partial Discharge
- **ROC** Receiver Operating Characteristic
- **SCADA** Supervisory Control And Data Acquisition
- **VAR** Vector Autoregression

1 Introduction

1.1 Motivation

The world has faced unprecedented challenges in recent years due to climate change and the rising demand for energy. As a significant contributor to greenhouse gas emissions, the energy industry plays a crucial role in mitigating the effects of climate change. The European Union has set ambitious goals for reducing greenhouse gas emissions and transitioning to renewable energy sources in order to achieve net-zero emissions by 2050[67]. Consequently, there has been a push to expand the production of renewable energy, particularly in Europe. However, sudden and rapid changes in energy distribution in Europe have caused energy crises, highlighting the need for dependable and efficient energy production methods.

Although the expansion of renewable energy sources, such as wind and solar, is essential for addressing these challenges, it is also essential to consider the effects of such expansion on biodiversity. Biodiversity is essential to climate change mitigation because it supports ecosystem functions and services that are vital to human well-being. Therefore, it is essential to strike a balance between the need for renewable energy production and the protection of biodiversity. Streamlining already established facilities could help lower the pressure on biodiversity while also meeting the increasing energy demand.

Hydroelectric power plants are a viable option in this context because they provide a reliable, renewable energy source with relatively low greenhouse gas emissions. However, the efficiency of hydroelectric power plants can be improved to ensure that they meet the increasing demand for energy while minimizing their impact on the environment. The application of artificial intelligence (AI) and machine learning (ML) techniques becomes crucial at this point.

AI and ML techniques have demonstrated their potential in a variety of industries, including the energy production industry, by optimizing processes and enhancing efficiency. It is possible to increase performance and decrease downtime at hydroelectric power plants by employing these techniques. AI and ML can be used, for instance, to develop better maintenance routines, predict equipment failures, and optimize the operation of turbines, thereby maximizing energy production and reducing the need for new infrastructure.

In addition, AI and ML can assist in identifying anomalies in the operation of hydroelectric power plants, enabling early detection of potential faults and preventing timely interventions. This not only ensures the efficient operation of the plants but also reduces their environmental impact by limiting the need for extensive repairs and component replacements and maximizing power production.

Thus, the urgent need to address climate change, rising energy demand, and Europe's ongoing energy crisis served as the inspiration for this master's thesis. The hydropower resources of Norway have enormous potential as a scalable and adaptable energy battery for Europe. By increasing its energy storage capacity and bolstering its interconnections with the European grid, Norway can play a vital role in supplying the European market with a stable and reliable energy supply. The incorporation of advanced data technologies into hydropower production presents an opportunity to reduce downtime and facilitate well-informed decision-making, thereby enhancing productivity. Consequently, this strategy enables increased production without necessitating the expansion of production sites, thereby providing a sustainable solution for meeting energy demands. By investigating the potential of AI and ML techniques for anomaly detection in hydroelectric power plants, this research aims to contribute to the development of more efficient and environmentally friendly methods of energy production. This will ultimately aid in achieving a balance between the need for renewable energy and the preservation of biodiversity, paving the way for a sustainable future.

1.2 Digital Green Shift

This subsection is mainly based on previous work done by the same author[10]

There are five big trends in the energy market right now. They are deregulation, decentralization, decarbonization, digitization, and democraticization, thereby earning the name 5Ds of Energy[12]. These are both technical and political impacts that have emerged over time. Figure 1 gives a view of how they are all pushing towards a digital green shift.

Deregulation and decentralization are at a mature technological level. Deregulation refers to the elimination or decrease of government rules in a particular industry. The objectives are to permit industries to operate more freely, make choices more efficiently, and eliminate corporate limitations. Transferring control and decision-making from a centralized body to a dispersed network is decentralization. Decentralized networks strive to reduce the amount of trust between users and restrict their ability to exercise authority or control over one another in a way that lowers the network's functionality. Deregulation, decentralization, and decarbonization were initially driven by the OPEC crisis in the 1970s[40]. All though deregulation is a well-known and deeply rooted movement, there have been new trends connected to the pandemic, for instance, suggesting deregulation 2.0, which also places it at an early-stage technology level.

Decarbonization and digitalization are emerging technologies. Decarbonization typically refers to the process of reducing carbon intensity, or the quantity of greenhouse gas emissions produced by the combustion of fossil fuels. Typically, this means reducing CO2

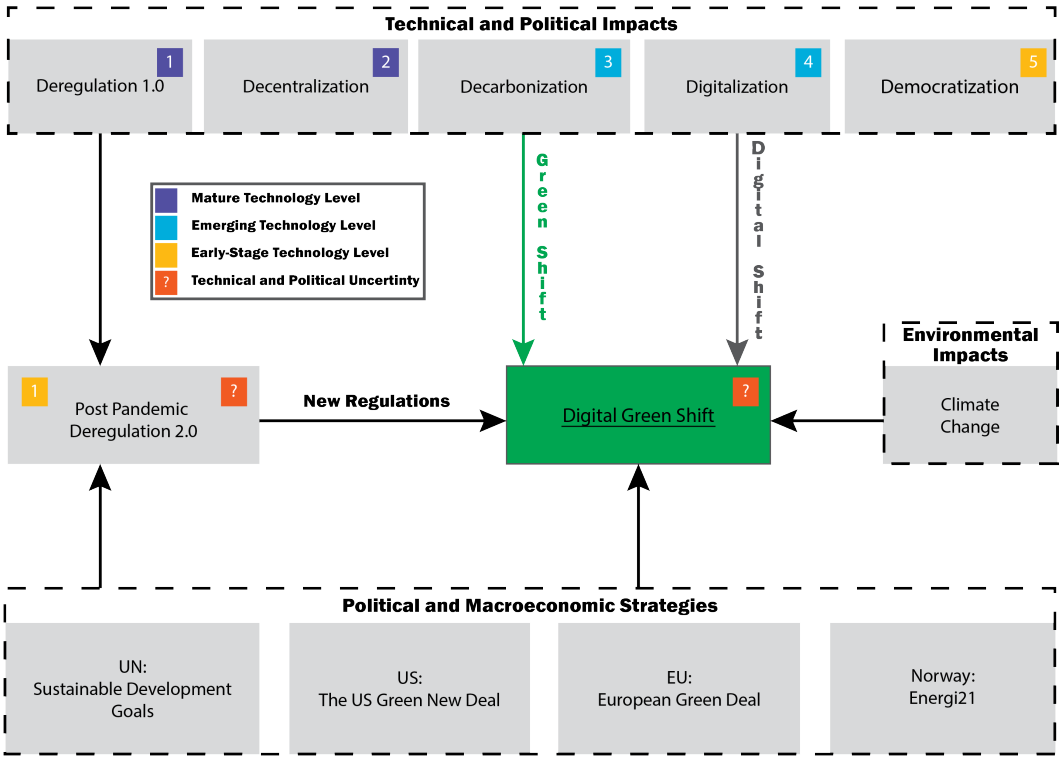


Figure 1: Overview of digital green shift and 5Ds of energy[12].

emissions per unit of generated electricity. Digitization refers to the use of information technology to facilitate the development of digital information as well as its management and utilization. Climate change is a big driver for decarbonization. To maintain a strong social and economic status while reducing greenhouse gas emissions, energy solutions that rely more on renewable energy sources are being favorably considered. However, the fluctuating nature of these sources has led to technical challenges in the energy industry, including supply-and-demand imbalances and a greater need for improved forecasting systems, thus resulting in a digitalization trend. There have been significant advancements in integrated information and communications technology, as well as the capabilities of the Internet of Things (IoT). The digital shift focuses on AI and ML to improve price, supply, and demand forecasts, in addition to technical issues such as fault detection for maintenance.

Democratization is at an early-stage technological level. Energy democracy is a concept established within the environmental justice movement that combines the transition to renewable energy with initiatives to democratize energy resource production and control. As a result of the democratization of the energy market, customers can now participate more actively in generation as prosumers. With increased adoption of blockchain technology, prosumers will be able to participate even more without the need for third-party intermediaries[12]. There are political and macroeconomic impacts as well. Three globally impacting political movements stand out: the United Nations

Sustainable Development Goals, the European Green Deal, and the US Green Deal. These policy packets focus on facing climate change with sustainable development on the social, economic, and environmental levels. Both policies are purposefully designed to increase demand for smart, efficient, and renewable energy solutions by decarbonizing and digitizing the energy sector. Thus, as figure 1 emphasizes, the world is undergoing two significant transitions at the same time: the digital shift and the green shift. Because the two shifts are mutually reinforcing, the transition has been dubbed the Digital Green Shift.

Energi21 Norway has its own digital green shift strategy. Norway's newest national strategy for climate-friendly energy research and innovation is known as Energi21[20]. Energi21 has been appointed by the Ministry of Oil and Energy. The Energi21 strategy is driven by business and focuses on future energy market opportunities. The current Energi21 strategy places a strong emphasis on digitization, with digital and integrated energy systems being the primary areas of investment. Digital21 presented its digitalization strategy in August 2018, which included 64 recommendations for digitization across the entire business spectrum. The development of digital enabling technologies and the significance these technologies will have in virtually all industries and sectors as a transformative force and a source of new opportunities are central to these recommendations.

In the energy sector, it is critical to prioritize the application of big data processes and digitalization over the development of basic technologies. The energy industry has a lot of potential for using AI effectively. AI can be used to control the entire or a portion of an energy system, predict output from more dynamic energy sources, monitor component condition, and calculate component lifetime. This is especially important for critical energy supply components. Energy system operators, fitters, and maintenance personnel can obtain sufficient real-time data on all critical components of the entire power system by utilizing condition monitoring and AI. Energy and grid suppliers, as well as major energy consumers, will benefit from a more solid foundation for planning, streamlining, and coordinated system optimization.

Energi21 sees digitization as a critical tool for ensuring supply security, achieving cost-effective operations, and developing effective market solutions all at the same time. Digitization also gives companies in the energy sector and at its interfaces new ways to make revenue.

1.2.1 Digitalization

Digitization[20] refers to the utilization of information technology for the purpose of facilitating the creation, management, and utilization of digital information. Digital transformation at the business level refers to the utilization of digital technology to enhance existing services or products, develop novel offerings, optimize operational efficiency, or adopt innovative approaches to task execution.

The implementation of digital transformation can provide significant competitive benefits to companies and facilitate the discovery of novel approaches to value creation. The utilization of data is of the utmost importance. Both private and public businesses collect a great deal of data, but the true benefits result when the data is analyzed and the insights obtained from the analyses are used to enhance optimization and efficiency.

The dynamic nature of digital transformation is characterized by the continuous evolution of the technologies that underpin it, with novel technological innovations being developed on a regular basis. Several technologies are deemed crucial for digital transformation. It is imperative to highlight the significance of cloud services, big data, AI and data analysis, IoT, platforms, and robots[38].

The significance of certain factors for a business is dependent upon components such as the industry of operation, the availability of data, technological advancement, and cost-effectiveness considerations. The significance of digital transformation includes hardware, data centers, and other related infrastructure. Within the energy industry, the process of digitization entails the integration of sensors into a greater number of physical components. These sensors are designed to measure various factors, such as energy consumption and overall condition. The sensors are interconnected within a network that facilitates bidirectional communication. The information is gathered and assessed, and subsequently, regulatory signals are transmitted to enhance elements such as energy consumption and electrical current distribution. The implementation of digital technology will provide us with additional avenues for observation and decision-making.

The implementation of novel digital solutions is expected to facilitate the management and upkeep of the energy system, enhance the safety of electricity supply, and improve the readiness of individuals. The implementation of digitization and enhanced data quality has the potential to improve the precision of investment decisions, while also streamlining various decision-making procedures. The process of digitization facilitates the modification of consumption patterns, enabling the inclusion of unregulated commodities and renewable energy sources. Additionally, it streamlines the integration of distributed energy resources such as solar cells and batteries into the larger energy infrastructure.

Developing novel business models and comprehending consumer behavior will be crucial.

It will be crucial to implement a novel market structure along with new rules and incentives. Indications of this occurrence are already apparent. As the world becomes increasingly digitized and energy systems become more interconnected, the significance of safeguarding data security and privacy will likely escalate[20].

1.2.2 Norwegian Initiatives

Norway is at the forefront of the digital transformation of the energy sector, with several initiatives underway. For example, SmartKraft offers cooperative solutions, while MonitorX is developing a pioneering maintenance framework. These innovative initiatives provide significant perspectives and motivation for utilizing digital technologies to transform the worldwide energy sector.

SmartKraft The Smartkraft[62] initiative in Norway is a collaborative endeavor involving prominent Norwegian companies and research institutions with the objective of devising intelligent and sustainable energy solutions. The initiative is centered around harnessing the potential of digitalization, automation, and electrification in order to establish an energy system that is both more efficient and sustainable.

The primary aim of the Smartkraft program is to create sophisticated energy systems that possess the ability to enhance energy efficiency, minimize inefficiencies, and encourage the utilization of sustainable energy resources. The scope of this work includes the creation of smart grid infrastructure capable of dynamically regulating energy production and consumption, thereby optimizing energy utilization and mitigating energy storage requirements.

The Smartkraft initiative prioritizes the advancement of inventive approaches to energy storage, including the utilization of battery technologies and the incorporation of sustainable energy sources such as wind and solar power. The objective of this initiative is to enhance the energy storage capability in Norway, thereby facilitating the country's function as an eco-friendly energy reservoir for other nations in Europe.

The Smartkraft initiative endeavors to advance the proliferation of electric power in the domains of transportation, industrial operations, and construction. The initiative seeks to mitigate greenhouse gas emissions and foster sustainable development by advocating for the adoption of electric vehicles, intelligent energy management systems, and energy-efficient building solutions.

In general, the Smartkraft initiative constitutes a noteworthy advancement towards establishing a sustainable, effective, and robust energy system in Norway and beyond.

MonitorX The primary aim of the MonitorX project[45] is to devise a theoretical framework and a corresponding software prototype that can facilitate the most efficient deployment of constituent parts in hydroelectric power stations over their operational lifespan. The project is founded upon the notion that the integration of advanced condition monitoring systems and traditional techniques for maintenance and reinvestment analysis culminates in an improved model that facilitates the optimal utilization of component lifetimes in hydroelectric power plants. The MonitorX project focuses on thematic domains that have been given priority in the latest Energi21[20] primary strategy (2014).

The utilization of advanced condition monitoring systems for estimating the remaining lifetime and probability of failure for components and making decisions about maintenance and reinvestments is not yet widespread. These systems employ data from protection and control systems (SCADA) and various types of sensors to detect trends in technical degradation. The proposed combination of the mentioned systems with maintenance and reinvestment planning models and tools, previously developed in prior projects, will culminate in the creation of a novel model (MonitorX). This model will provide a stronger and more refined foundation for prioritizing maintenance and reinvestment measures, in contrast to the utilization of individual models and systems.

To summarize, the objective of the MonitorX project is to devise a comprehensive framework that combines advanced condition monitoring systems with conventional maintenance and reinvestment analysis techniques. The adoption of an integrated approach is expected to enhance the decision-making process related to maintenance and reinvestment in hydroelectric power plants, thereby optimizing the utilization of components over their lifetime. Prioritizing the thematic areas listed in the Energi21 main strategy helps the project contribute to the efficiency and sustainability of the energy sector.

1.2.3 Changes in Production Patterns

The energy sector has experienced notable changes in production patterns due to the emergence of a more competitive energy market and the introduction of Nord Pool[46] over time. Established in 1996, Nord Pool is the largest power market in Europe. Its primary objective is to enable efficient, transparent, and secure electricity trading across the Nordic and Baltic regions. This introduction has resulted in various alterations in the ways in which energy is generated and utilized. A map of Nord Pool Spot bidding areas can be seen in figure 2.

A significant transformation has occurred in the realm of electricity production, whereby the focus has shifted from providing merely for local demands to being influenced by

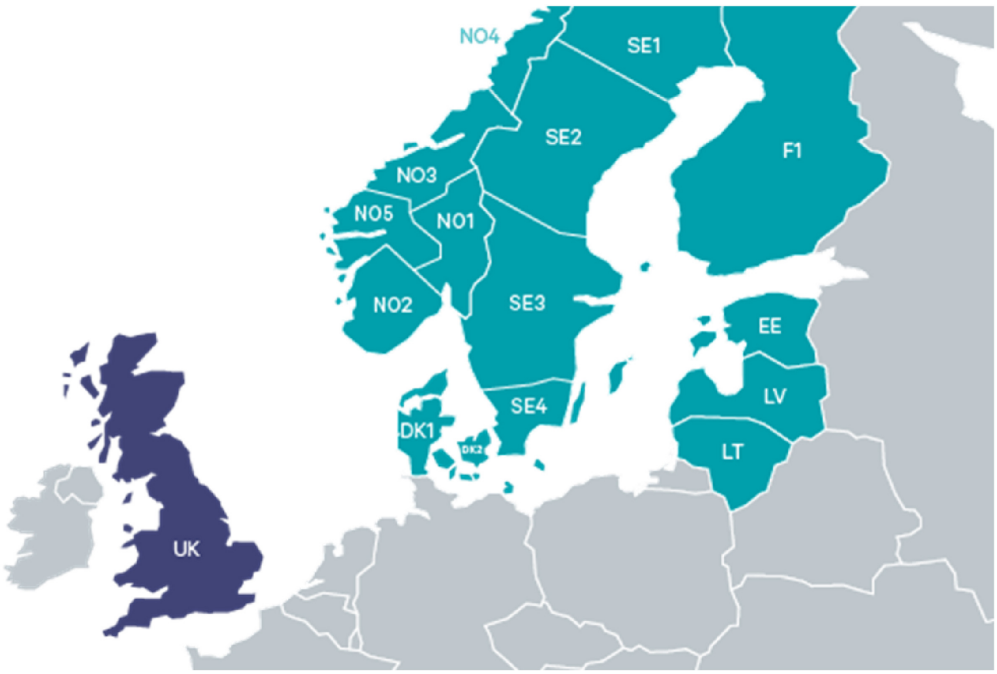


Figure 2: Map over Nord Pool Spot bidding areas[46].

market dynamics[74]. The modification enables power producers to boost their responsiveness to price indications, thereby optimizing their power generation in accordance with current market conditions and maximizing their financial gains. Consequently, power plants are currently encountering a higher frequency of modifications in their production levels, thereby applying more stress on their machinery due to the need to initiate and terminate operations in reaction to the volatile energy market. This contrasts with conventional methods of power generation, which generally comply with more predictable and consistent patterns.

Moreover, the Nord Pool-facilitated competitive energy market encouraged a rise in cross-border electricity trade among the nations involved. The presence of interconnectivity among nations facilitates the optimization of electricity supply and demand, resulting in enhanced energy system efficiency and the encouragement of regional cooperation.

The competitive energy market has supported the integration of renewable energy sources, such as wind and solar power, into the grid. The integration of sources that represent variable and inconsistent production patterns has resulted in increased fluctuations in electricity supply, thereby requiring power systems that are more flexible and responsive. Consequently, the fluctuating production levels created additional stress on the power generation machinery, requiring it to adjust accordingly.

In addition, the competitive nature of the energy market has led to the development of demand-side response mechanisms, which enable consumers to modify their electricity usage in reaction to price indications. The inclusion of flexibility in the energy system

aids in the stabilization of the grid, the minimization of peak demand, and the improved integration of fluctuating renewable energy sources. Nonetheless, this phenomenon leads to increasingly unpredictable production trends, underscoring the necessity for power generation systems that are flexible and capable of adapting to changing circumstances.

In light of the increasingly competitive energy market, power producers have been motivated to enhance the efficiency of their operations and allocate resources towards developing technologies. The pursuit of increased efficiency has resulted in the modernization of power plants, the implementation of advanced monitoring and control systems, and the creation of more effective energy storage alternatives.

1.3 Advancement in Maintenance

In the past seventy years, the maintenance industry has undergone substantial changes. The perception of maintenance within the industry has gone through a transformation, whereby it is no longer viewed as a required work, but rather as a means of generating value[49]. The MonitorX project was launched in 2015 by Sintef and Energi Norge in partnership with multiple major Norwegian power producers[45]. The objective of the project is to use varied information sources and self-learning models to predict maintenance requirements, as mentioned earlier.

Although maintenance advancements have mainly focused on operational and analytical methodologies, the techniques employed are outdated and solidified. The concept of preventive maintenance originated in the 1960s, while predictive maintenance emerged during the 1970s and 1980s. During the 1980s and 1990s, the reliability-based maintenance (RCM) approach and the notion of life cycle costs were introduced[49]. In terms of analytical techniques, FMEA (Failure Modes and Effects Analysis)[18] and HAZOP (Fault Tree Analysis and Hazard Analysis and Prediction)[31] was created in the 1950s and 1960s, respectively. This suggests that the underlying principles of the theory have remained mostly unchanged since the 1990s.

The integration of automated systems and remote monitoring has revolutionized maintenance methodologies in numerous industries. The implementation of automation technology can efficiently reduce the occurrence of human error and optimize operational productivity. Additionally, remote monitoring systems enable operators to observe equipment performance from a centralized control room, thus lowering the need for on-site inspections. Furthermore, the focus on reducing environmental footprints has resulted in the implementation of sustainable maintenance practices such as the choice of eco-friendly lubricants and improved waste disposal methods[55].

In recent years, there has been an increased focus on workforce training and safety measures to ensure that maintenance personnel possess the necessary skills and knowledge

to carry out their tasks in a secure and effective manner. The digital revolution has altered maintenance procedures at hydroelectric power plants as well. New software and tools for advanced data analysis allow for better decision-making based on more precise knowledge of equipment performance, maintenance requirements, and potential risks[52].

Despite these developments, anomaly detection techniques still need to be improved in order to boost the effectiveness and efficiency of maintenance at hydroelectric power plants. Hydroelectric power plants may maximize their operational efficiency by proactively detecting and resolving anomalies, thereby reducing the risk of potential faults, limiting costs, and improving their overall performance. This will support the development of an energy sector that is more dependable and sustainable, thereby supplying the growing demand for clean energy while protecting the environment.

1.4 Problem Definition

This master's thesis investigates and evaluates predictive maintenance and predictive analytics applications, with a particular emphasis on the implementation of anomaly detection techniques driven by AI and ML in the hydroelectric power sector. This study's primary objective is to evaluate the capacity of these proactive technologies to improve the efficacy of maintenance procedures and the operational effectiveness of hydroelectric power systems. The ultimate goal is to increase productivity, decrease downtime, and promote a more reliable and sustainable energy industry.

To achieve these objectives, the study will employ an approach with two phases. The broader aspect involves examining the general use of predictive maintenance and predictive analytics in hydroelectric power plants, gaining an understanding of the current landscape, and identifying gaps where AI and ML could add value. This will provide a broad perspective on the state of these technologies in the industry and lay the groundwork for the next phase of the study.

The narrower aspect will then concentrate on employing diverse ML algorithms for detecting anomalies in power transformers by analyzing recordings of winding temperature and hydrogen concentration. In this section, the chosen algorithms will be applied to a test scenario involving a data set acquired from the Duge power plant. This method will allow for the evaluation of the effectiveness of these algorithms in detecting anomalies and identifying potential faults in the power transformer.

The research will then evaluate the performance of the selected ML algorithms in terms of their success rate in detecting anomalies in the context of power transformers. This performance evaluation will shed light on the potential of these algorithms to increase operational efficiency and decrease downtime.

The thesis will conclude with a discussion of the practical challenges and benefits associated with the deployment of AI and ML-driven anomaly detection techniques within a predictive maintenance and analytics framework. This evaluation will take into account the hydroelectric power plant's unique operational characteristics and standards.

This thesis aims to provide a comprehensive overview of the potential benefits and obstacles associated with the integration of AI and ML within the predictive maintenance and analytics of the hydroelectric power sector through a detailed analysis of these integral aspects. The ultimate objective is to contribute to the development of a more sustainable and reliable energy future by advancing maintenance procedures and operational efficiency in the renewable energy sector.

1.5 Limitations

This project faces several challenges. First, establishing a thorough literature review and gaining in-depth knowledge on such a broad topic might require more time than a single semester can provide. Therefore, it may be challenging to reach a conclusion at the level of an expert within this time frame.

Although the author is knowledgeable about hydroelectric power systems, AI and ML are relatively new fields of study. This project also requires high-quality data, which can be challenging to access due to power companies' strict security measures. This data acquisition process could take longer time than anticipated.

In addition, errors in the production of hydroelectric power are less frequent than in other industries, sometimes occurring years apart. This rarity makes it challenging to choose a suitable test case for the study. Lastly, the sensitive nature of the involved data necessitates careful handling and high privacy measures. All of these factors pose notable challenges to the depth and scope of this study.

1.6 Report Outline

The report follows an organized outline, beginning with an introduction that provides the motivation and context for the study, discussing the broader trends of digitalization and their influence on production patterns, and introducing the advancements in maintenance.

In the second section of the report, background information related to the study is discussed, including aspects of hydroelectric power plants, maintenance strategies, anomaly detection methods, and a literature review. This section also discusses the data

recording process, data preprocessing techniques, and transformer health assessment methods.

The third section describes the study's methodology, including the flowchart, Python use, data collection, data analysis and test case development, data preprocessing, and model creation. A comprehensive description of this procedure is provided.

The results of the study, including performance scores and receiver operating characteristic (ROC) curves, are presented in the fourth section. The report concludes with a discussion and conclusion section in which the study's findings are considered and their implications are discussed, ultimately leading to a conclusion and further suggested work.

2 Background Information

This chapter offers a comprehensive introduction and contextual background for the key themes and concepts that are explored in this thesis. The purpose of this text is to provide readers with essential knowledge and understanding of the relevant fields. This will facilitate a better comprehension of the analysis and conclusions presented in the study that follows.

2.1 Hydroelectric Power Plants

This subsection is mainly based on previous work done by the same author[10]

Hydropower is a significant form of renewable energy that has been utilized for centuries to generate electricity and operate machinery. Hydropower was responsible for supplying more than 17% [23] of the global electricity demand in the year 2020, thereby establishing itself as the foremost renewable energy source for electricity generation. Hydropower is a sustainable source of clean energy that emits neither air pollution nor greenhouse gases. Additionally, its impact on biodiversity can be effectively managed through regulation. Hydropower is a reliable and consistent energy source that can be harnessed to meet base-load electricity demand at a reasonable expense. This characteristic makes it a desirable option for numerous countries aiming to reduce their dependence on non-renewable energy sources and have the geographical topology necessary.

A power system[17] is an intricate system of connected components that facilitate the generation, transmission, and distribution of electrical energy to consumers. The process consists of the conversion of some sort of energy source into electrical energy. The power system includes a range of interconnected components, such as the synchronous generator, motor, transformer, circuit breaker, and conductor, among other elements. The power system involves six fundamental components, namely the power plant, transformer, transmission line, substations, distribution line, and distribution transformer. The power plant is responsible for producing the energy, which is then elevated by the transformer prior to being transmitted. The transmission line serves the purpose of transferring electrical power to the substations. Electric power passes through substations and is then reduced by the distribution transformer to a level that is appropriate for usage by the end user.

The process of generating electricity involves the conversion of a particular form of energy source into electrical energy through the utilization of a mechanical energy harvester and a generator at a power plant. The sources of energy can be classified into two categories: nonrenewable sources, which include oil, gas, and coal, and renewable sources, which include wind, solar, and hydro. A power system has the potential to contain a

combination of various energy sources. The illustration depicted in Figure 3 illustrates the different components of a hydroelectric power infrastructure. Hydroelectric power is produced by harnessing the energy of precipitation, including rainfall and snowfall. Norway possesses ideal conditions for the generation of hydropower. Apart from the presence of mountains, waterfalls, and lakes, the area has consistent precipitation for most of the year. The water from the lakes is used as reservoirs and then pumped to power plants when necessary.

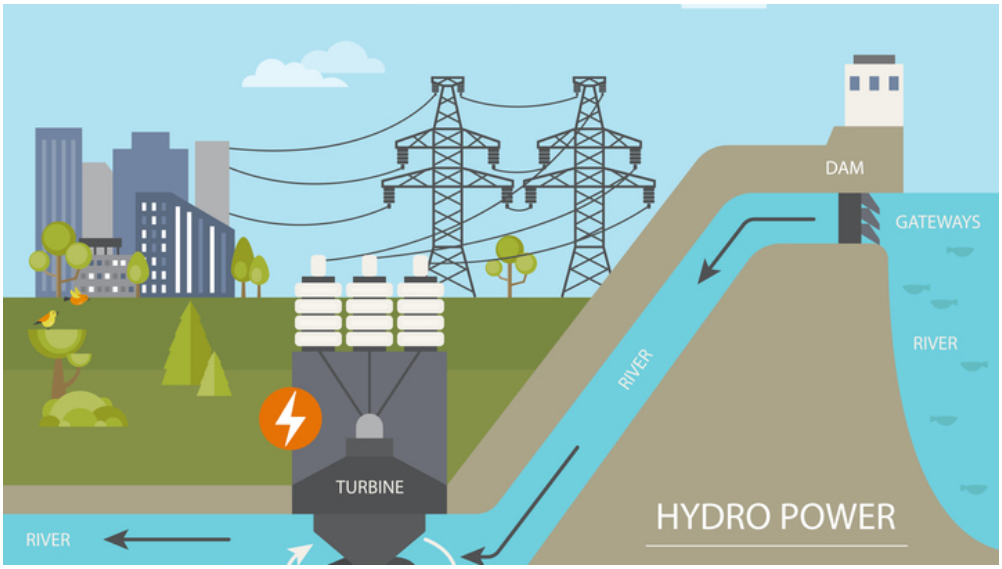


Figure 3: Illustration of a hydroelectric power system[70].

Water flows through inclined pipes within the power plant. Thus, the water achieves a high velocity prior to contacting the turbine wheel. The turbine drives the generator, converting the kinetic energy of the moving water into electrical energy. Once the electrical current is produced, it travels through a transformer, which then distributes it across high-voltage power lines until it finally arrives at the electrical outlet within one's residence. Upon completion of its cycle through the power plant, the water is released back into the surrounding waters. In this cycle, water undergoes evaporation and eventually returns as precipitation. This cycle is what makes hydropower a renewable and sustainable source of energy. In comparison to reservoir power plants or high-pressure plants, a river power plant is a hydroelectric facility that lacks extended tunnels or notable water access regulations and typically operates under low pressure. Hydropower plants are typically classified into two primary categories based on various factors, including size. These categories are large-scale power plants, which have an installed power capacity exceeding 10 MW, and small-scale power plants, which have an installed power capacity of less than 10 MW[76].

The high-voltage distribution lines in Norway can be classified into three categories, namely, transmission network, regional network, and distribution network[66]. At the

national level, the transmission network serves as an infrastructure for connecting major producers and consumers. The transmission network includes intercontinental connections as well. The transmission grid typically functions at a high voltage range of 300 to 420 kV in the majority of the country, although specific regions may also incorporate 132 kV lines. The connection between the transmission grid and distribution grid is a common feature of the regional grid, which might also include high-voltage production and consumption radials. The regional grid operates within a voltage range of 33 kV to 132 kV. Distribution networks refer to the local power networks that are responsible for distributing electricity to relatively smaller end users. The maximum voltage of the distribution network is 22 kV, with a clear distinction between high-voltage and low-voltage distribution networks. The separator operates at 1 kV, whereas the low-voltage distribution system utilized for common consumption purposes is typically set at 400 or 230 volts. The regional or distribution network is linked to smaller production facilities. Production facilities of greater scale have connections with the transmission or regional network.

2.1.1 Turbine

The turbine is an essential part of a hydropower plant. The turbine converts the potential energy of water into mechanical energy. The hydro turbines can be classified into two distinct categories, namely impulse turbines and reaction turbines[32]. The impulse turbine operates by directing a continuous flow of water through a nozzle, which subsequently impacts the rotating buckets at a constant pressure. A reaction turbine operates by directing water through a series of guide vanes and then through the rotating blades, resulting in lower pressure. The Francis, Pelton, and Kaplan turbines are often used in hydropower facilities located in Norway[77]. Both the Francis turbine and the Kaplan turbine are classified as reaction turbines. In contrast, the Pelton turbine is classified as an impulse turbine. The Francis turbine is deemed optimal for drop heights that are in the small to medium range, whereas the Pelton turbine is considered to be the most suitable option for drop heights that are relatively larger. The optimal performance of the Kaplan turbine is achieved under conditions where the head is low and the flow rate of water is high or fluctuating.

The two most commonly observed defects in hydroturbines are cavitation and erosion[32]. Cavitation is a phenomenon that occurs when the pressure of a liquid drops below its corresponding vapor pressure. Subsequently, the liquid will undergo a local phase transition from its liquid state to a gaseous state, resulting in the formation of gas bubbles. This phenomenon frequently occurs when a fluid flows past an object at a high velocity, potentially resulting in physical damage to the object. Erosive wear is caused by the impact of particles, such as sand, on the surface of the turbine. The occurrence

of cavitation and erosion has been observed to result in an elevation of temperature and vibration levels in bearings. The issue of erosion and cavitation is more pronounced in reaction turbines due to their higher susceptibility to erosion, which is directly related to their greater operational speed. The degradation of the machine's performance is attributed to cavitation and erosion.

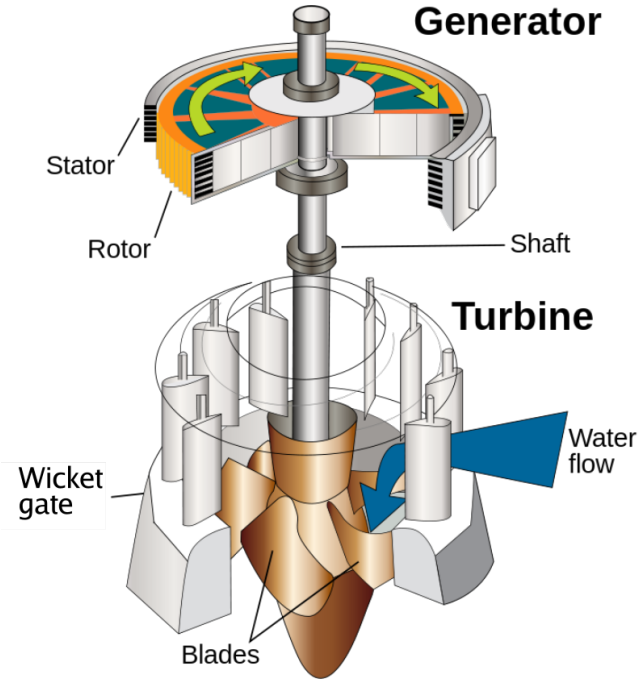


Figure 4: Simple illustration of a hydropower generator and turbine[21].

2.1.2 Generator

Figure 4 illustrates the connection between a hydropower generator and turbine. The hydroelectric power plant uses a generator to transform the mechanical rotational energy generated by the turbine into electrical energy. Synchronous generators are commonly used in hydropower[32]. The illustration presented in Figure 4 shows that the generator consists of two fundamental components, namely a rotor and a stator. The rotor is linked to the turbine through the shaft, which facilitates its rotation around the stator. The rotor consists of windings that carry direct current and possess magnetic features. At rest, the stator consists of numerous copper wire windings. The magnetization of the rotor is achieved through the utilization of slip rings from a distinct rectifier. The magnetic field produced by a rotating, magnetized rotor causes voltage to be induced in the stator coils. The windings are connected to the electrical circuit and provide energy to the grid. The stator is connected to multiple circuits. The circuit's inability to generate a variable magnetic field using a single circuit results in the absence of any induced voltage. The rotational movement of the rotor will result in a change in the

magnetic flux within the coil.

The primary function of the shaft line is to transmit the rotational energy produced by the turbine to the rotor of the generator[8]. In relation to the upper and lower cross, the bearings play a crucial role in maintaining the position of the axle string, rotor, and turbine. Ensuring that the rotating components are securely fixed in position to enable rotation around the central axis is critical. Minor deviations from the central axis may result in notable vibrations, thereby causing increased wear and tear on the various components. Significant deviations can result in serious consequences. The primary objective of the brakes and braking system is to decelerate the rotational speed of the turbine's rotor subsequent to the reduction of water flow. Brush extraction installed in the generator's top is responsible for collecting the brush dust generated by the brushes' operation. On each brush contact with the slip ring, the system acts as a vacuum cleaner. The standard frequency of the alternating current on the Norwegian power grid is 50 hertz. This must be taken into consideration when determining the type and speed of the generator.

The cooling system[9] is another important component of the generator. The release of heat generated by alternators in synchronous generators is an essential concern, as natural cooling is insufficient. There are two distinct methods for cooling the system: air and water. The forced air cooling system facilitates the transfer of heat from the surface of the alternator by inducing air flow into it. Ducts are incorporated into the stator and rotor cores as well as the field coils of generators and machines to enhance the surface area exposed to cooling air. The orientation of the airflows determines whether the channels are radial or axial. Water cooling systems use hydrogen to directly cool the rotors and demineralized water to directly cool the stator windings. The water is circulated by means of an AC-powered centrifugal pump. Cartridge filters are often used in the process of water filtration. These filters are designed to prevent metallic particles that are corrosive from entering the winding hole conductors. A generator with an air-based cooling system is typically larger than one with a water-based cooling system, as water can cool the generator in a more compact manner.

2.1.3 Transformer

The main function of the power transformer within a hydroelectric power system is to elevate the voltage produced by the synchronous generator to an elevated regional transmission level[26]. The transformation is of utmost importance in mitigating the amount of energy loss that occurs while transmitting electricity. Power transformers are fundamentally linked to the synchronous generator, which is the source of the electrical energy in the system. The generator produces electrical energy at a designated voltage level, commonly within the range of a few kilovolts (kV). Yet, the transfer of electricity

at such a magnitude would lead to a notable loss of energy owing to the resistance encountered in the transmission lines. In order to reduce these losses, it is necessary to convert the voltage to a significantly higher level, typically in the range of tens to hundreds of kilovolts (kV). Herein lies the significance of the power transformer.

The electricity generated is fed into the power transformer, where the voltage is stepped up. Taking advantage of a higher voltage allows the transmission of electricity over longer distances while minimizing losses. At the final destination (such as a town or city), additional transformers reduce the voltage to a level suitable for use in homes, businesses, and other consumers.

Two wire coils (windings) and a magnetic field are used to increase the voltage in a power transformer according to the electromagnetic induction principle. A transformer consists of two distinct windings, namely the primary winding, connected to the energy input, and the secondary winding, connected to the energy output. The windings are coiled around a shared iron core, thereby enabling efficient magnetic coupling between them. The synchronous generator feeds alternating current (AC) into the primary winding. This alternating current induces a voltage in the secondary winding by producing a variable magnetic field around the primary winding.

Along with their significance, power transformers are complicated and expensive devices. The consequences of malfunctions can carry substantial weight, and the time required to obtain replacement components can be long[15]. Moreover, the maintenance and operation of transformers pose a challenging task for their owners due to their closed-unit design and inaccessible critical components. The process of determining the appropriate timing for reinvestments and significant maintenance activities requires a careful and strategic approach that takes into account the requirement for reliability while also considering the expenses associated with maintenance and replacement.

2.1.4 Duge Power Plant

The Sira-Kvina power company operates multiple power plants across the Sira and Kvina river systems, providing the region with clean, renewable energy. The Duge power plant[22] is one of the relatively small hydroelectric facilities in the Sira-Kvina network. It is equipped with two reversible Francis turbines, and the average annual net production of the power plant is 248 GWh. Between the Svartevatn and Gravatn reservoirs, the plant utilizes a 240-meter drop in altitude. Due to the significant regulation height in Svartevatn and the use of pump turbines, the reservoir has two water intakes at different levels.

During times of low electricity costs and high storage capacity in the Svartevatn reservoir, water is pumped through the 12-kilometer-long drainage tunnel from Gravatn back

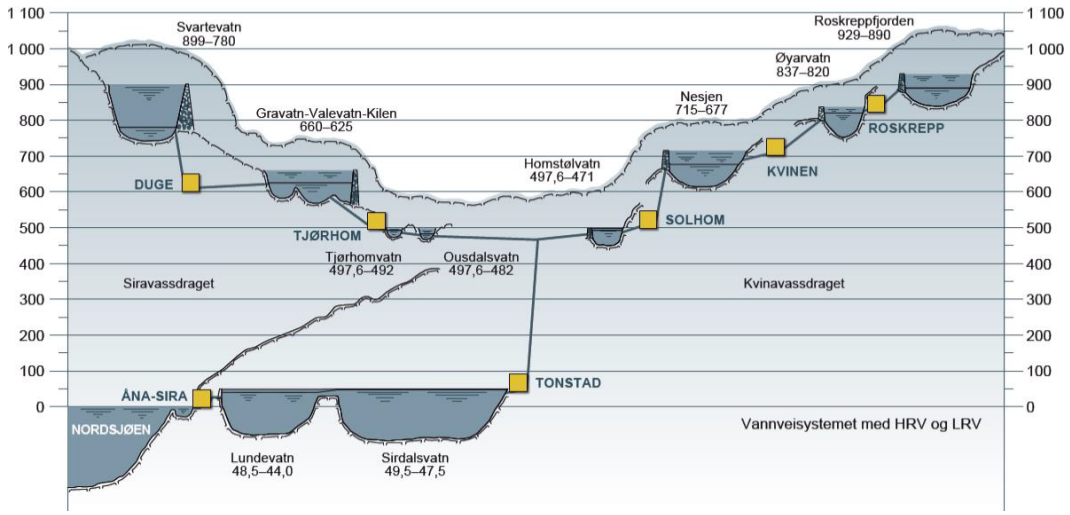


Figure 5: Illustration of the waterway system for Sira-Kvinas hydropower production[28].

into Svartevatn. Figure 5 illustrates the waterway system of Sira-Kvinas hydropower production. Duge power plant is found in the upper left of the illustration between Svartevatn and Gravatn, at approximately 600m altitude.

The majority of pumping occurs in the spring and fall, as well as on summer weekends. Market prices and reservoir conditions can fluctuate significantly from year to year, resulting in highly variable pumping demand. In 1974, construction of the Duge power plant began. In 1979, after a few months of delay caused by a landslide in the drainage tunnel, the facility went into operation. An overview of the technical data is found in table 1. Generator and transformer performance, rotational speed, and flow capacity of the power plant are listed here.

Completed	1979
Turbine Type	Francis turbines, reversible for pump operation
Capacity	2 x 100 MW
Rotational Speed	375 rpm
Flow Capacity - Turbine Operation	2 x 53 m^3/s
Flow Capacity - Pump Operation	2 x 42 m^3/s
Generator Performance - Nominal	2 x 120 MVA
Generator Performance - Pump Operation	2 x 106 MVA
Generator Voltage	13 kV
Transformer Performance	2 x 120 MVA
Transformer Voltage	320/12 kV

Table 1: Technical data from Duge power plant[22].

2.2 Maintenance Strategies

This subsection is partly based on previous work done by the same author[10]

The fundamental aim of a maintenance management strategy is to optimize the accessibility, quality, consistency, and flexibility of a system while ensuring cost-effectiveness[49]. In the context of hydroelectric power systems, maintenance is essential for ensuring the long-term viability, efficiency, and dependability of the power generation process. The classifications of maintenance[32] in a hydropower plant are categorized in Figure 6. Hydroelectric power systems are typically subjected to three primary maintenance strategies: corrective, preventive, and predictive maintenance.

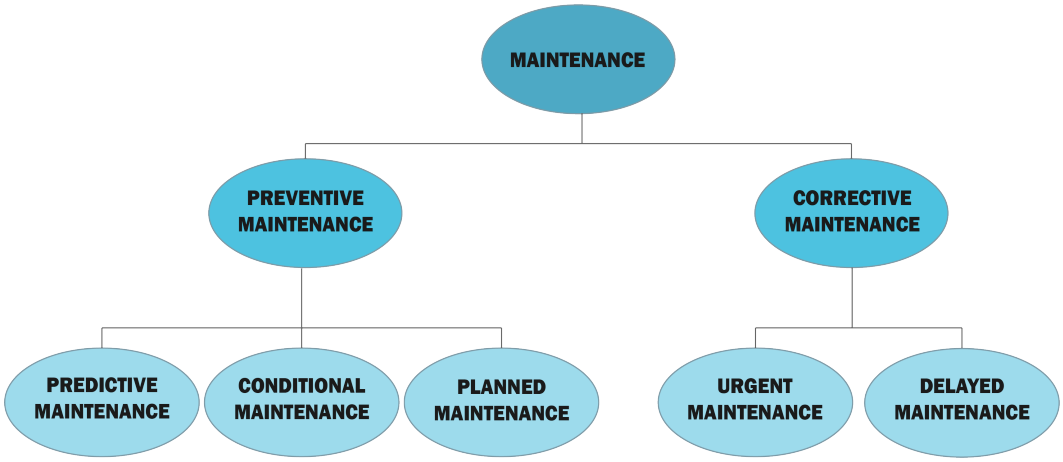


Figure 6: Classification of maintenance in a hydropower plant[32].

2.2.1 Corrective Maintenance

Corrective maintenance, which is also referred to as reactive maintenance, is a maintenance strategy that implies the repair or replacement of equipment components solely after their failure[25]. The condition of the equipment is not monitored, and no maintenance tasks are planned in advance of a failure with this maintenance strategy. Maintenance tasks are executed reactively in response to unforeseen failures and malfunctions.

When equipment failure has minimal consequences and repair costs, such as when changing light bulbs, corrective maintenance can be financially advantageous. However, corrective maintenance is frequently associated with increased downtime, higher repair costs, and decreased equipment availability, making it the most expensive approach for many types of equipment in hydroelectric power systems[43]. In addition, the absence of preventive measures and the uncertain nature of equipment malfunctions can lead to compromised safety as a consequence of corrective maintenance.

Corrective maintenance can be categorized into urgent and delayed corrective maintenance[32]. Urgent corrective maintenance is when the consequences of malfunction cannot be ignored and production must seize immediately. On the other hand, delayed corrective maintenance does not involve critical components, and maintenance can be scheduled in the near future without having to stop production at once.

2.2.2 Preventive Maintenance

Preventive maintenance is a strategy that adopts a proactive stance by carrying out maintenance tasks periodically or according to the equipment's utilization patterns[43]. The aim is to minimize the probability of equipment failure and prolong its operational lifespan. The maintenance approach in question includes activities such as oil modifications, lubrication, filter replacements, and bearing replacements[9].

The practice of preventive maintenance can be categorized into three types: conditional maintenance, planned maintenance, and predictive maintenance[32]. Planned maintenance is a maintenance strategy that involves carrying out maintenance activities at predetermined time intervals. In contrast, conditional maintenance is a maintenance approach that is dependent on the equipment's operational time and activity. The challenge in both of these cases is to determine the optimal maintenance interval to minimize costs and maximize equipment reliability.

Preventive maintenance can reduce the probability of equipment failure and lengthen its useful life, resulting in long-term cost savings. The primary obstacle associated with preventive maintenance relates to calculating the most suitable maintenance interval to prevent unnecessary maintenance actions or unexpected equipment malfunctions. Additionally, because it is predicated on the idea that equipment failure rates are predictable and constant over time, preventive maintenance may not be appropriate for all equipment types or operating circumstances.

2.2.3 Predictive Maintenance

Predictive maintenance is a recently developed approach that falls under the category of preventive maintenance. In order to detect potential failures before they happen, this advanced maintenance strategy depends on continuously monitoring the equipment's performance and condition. The approach involves the application of diverse technologies and techniques, including vibration analysis, thermography, and oil analysis, to evaluate the state of the equipment and strategize maintenance operations based on the equipment's current condition[44]. AI and ML techniques are applied to discover hidden patterns and anomalies that would not look like a fault or error to the human

eye.

The implementation of predictive maintenance provides various benefits, such as improved equipment reliability, reduced maintenance expenses, and enhanced operational efficacy[41]. Predictive maintenance allows maintenance personnel to schedule maintenance tasks more efficiently by predicting potential failures, resulting in reduced downtime and repair expenses. Moreover, the implementation of predictive maintenance can result in improved resource allocation and increased safety measures.

However, predictive maintenance is not without its challenges and disadvantages. A significant initial investment in specialized hardware, software, and training is required to implement a predictive maintenance program. A predictive maintenance plan's effectiveness depends on the consistency and accuracy of the data collected, which can vary depending on factors like the caliber of the sensors, data transmission, and human error. Additionally, because predictive maintenance demands a certain degree of predictability and consistency in equipment performance, it might not be appropriate for all equipment types or operating circumstances[42].

2.2.4 Selecting a Strategy

The choice of the optimal maintenance approach for a hydroelectric power system is dependent upon many factors, such as the equipment's rating and age, resource availability, and the company's primary objectives and risk tolerance. To maximize the performance and dependability of a hydroelectric power system, a combination of maintenance strategies can be used.

In certain cases, corrective maintenance may be deemed appropriate for equipment that is non-critical, has low failure rates, and has low repair costs. On the other hand, preventive maintenance may be implemented for equipment that has predictable failure patterns and limited repair costs. The implementation of predictive maintenance can prove to be a viable approach for important machinery that experiences serious repair expenses and has a noticeable impact on the overall system's performance.

2.3 Predictive Analytics

The terms "predictive maintenance" and "predictive analytics" are frequently used interchangeably, and it would be advantageous to establish a precise definition for each term. Predictive maintenance constitutes an important part of predictive analytics [51]. Throughout this thesis, the literature reviewed may refer to predictive maintenance in terms of the benefits derived from predictive analytics. This is not incorrect, as the terms are used interchangeably, but predictive maintenance is, in essence, a stepping stone toward predictive analytics, increasing proactiveness. Figure 7 illustrates a hierarchical depiction of all maintenance strategies discussed across this thesis.

As previously explained, predictive maintenance is intertwined with condition-based maintenance and revolves around the continuous monitoring of vital components in a machine or system in real-time. By carefully monitoring the condition of these components, it becomes possible to anticipate potential malfunctions and take preventative measures to avert unexpected downtime and costly repairs. Fundamentally, it emphasizes immediate action based on current or near-future conditions or observations.

Predictive analytics, on the other hand, refers to a more advanced level of maintenance approach in which condition-based data is collected over time, combined with expert knowledge, and then processed by ML or AI techniques to predict future events or failures. The goal here is not merely to reduce the likelihood of failure but rather

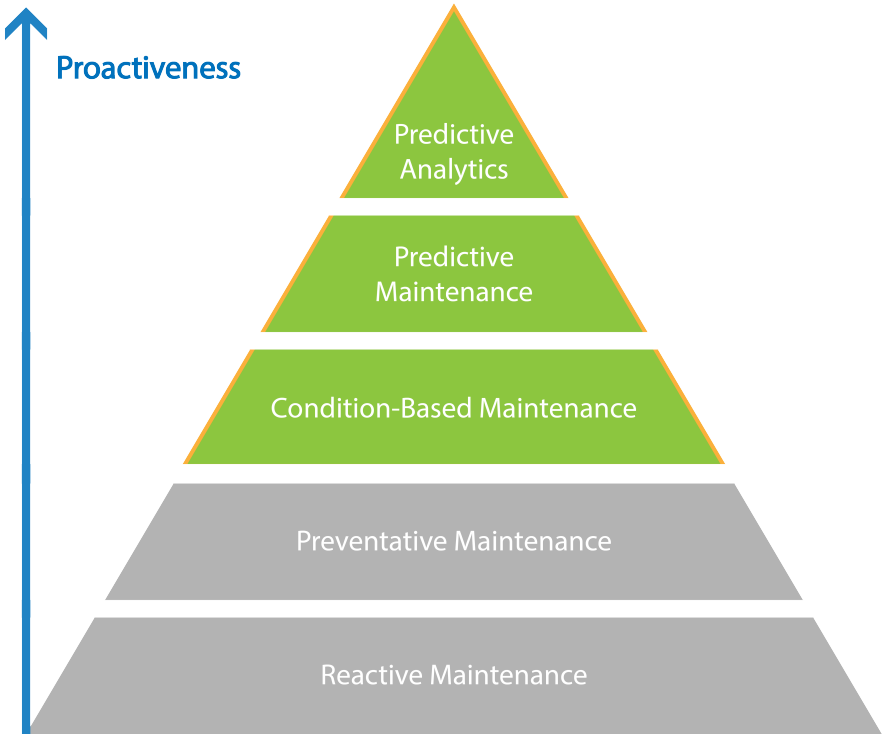


Figure 7: A hierarchical representation of maintenance strategies[14].

to eliminate it entirely. This strategy is highly dependent on the quality of the data tracked, the duration of data collection, and the system's expert knowledge.

The primary distinction between predictive maintenance and predictive analytics is the amount of time between the identification of a potential problem and the implementation of corrective measures. Predictive maintenance is primarily concerned with the swift and efficient resolution of immediate or imminent issues. In contrast, predictive analytics adopts a more comprehensive and strategic outlook, providing valuable perspectives for enhancing system performance over the long haul and preventing any potential downtime.

Moreover, the main objective of predictive maintenance is to mitigate unscheduled downtime. However, with its advanced capabilities, predictive analytics promises to not only prevent unplanned downtime but also reduce planned downtime, resulting in increased operational efficiency and cost savings. The utilization of predictive analytics software enables the timely detection of subtle variations in equipment performance that may indicate impending failure. This affords personnel sufficient time to take proactive measures to address the equipment issues.

2.4 Anomaly Detection

Detecting anomalies is an important part of predictive maintenance and a crucial aspect of data analysis, particularly when it comes to monitoring systems and minimizing faults. Anomaly detection refers to the process of identifying events, data points, or patterns that behave differently from what to expect within the data set[3]. Deviations, which are also referred to as anomalies or outliers, may suggest the existence of underlying faults, inaccuracies, or unfavorable behaviors that place a need for further examination.

There are various types of anomalies, such as point anomalies, contextual anomalies, and collective anomalies[79]. Point anomalies are individual data points that demonstrate significant deviations from the remaining data. Contextual anomalies are data points that show unusual behavior within a specific context or situation. However, they may not be considered outliers when analyzed individually. Collective anomalies are a set of interrelated data points that have uncommon patterns when examined together, even though each individual data point may not appear to be anomalous.

Numerous methods and strategies exist for identifying anomalies[71]. For example, detecting anomalies in statistical analysis relies on statistical models and assumptions about the data distribution, or anomalies can be detected by observing data points that fall outside of a predetermined range or threshold, which could be a specific number of standard deviations away from the mean.

Supervised learning methods include building a model using labeled data, where examples are categorized as either regular or anomalous. After being trained, the model can be used to classify new and unlabeled data points and detect anomalies. Unsupervised learning methods differ from supervised learning methods in that they do not rely on classified data. Instead, they identify anomalies by examining the underlying structure and distribution of the data. In general, unsupervised anomaly detection methods rely on techniques like clustering, density estimation, and dimensional reduction.

Semi-supervised learning is a technique that blends supervised and unsupervised learning[60]. It involves using a limited amount of labeled data to guide learning and enhance the accuracy of the anomaly detection model. Advanced AI and ML techniques, such as neural networks, deep learning, and reinforcement learning, have demonstrated the potential for detecting anomalies. By applying these techniques, it becomes possible to create complex data structures and patterns that aid in identifying intricate anomalies that may be challenging to detect using traditional methods.

In the context of anomaly detection in hydroelectric power plants, it is common to use multiple ML algorithms due to their unique strengths and limitations. It is rare for a single algorithm to possess all the necessary qualities. Using a combination of supervised and unsupervised techniques in algorithmic testing can be advantageous for assessing their individual performances. These six algorithms chosen possess specific traits that are deemed advantageous for the given test case. Based on the literature review, there are indications that they are capable of effectively detecting anomalies in the test case.

2.4.1 K-Nearest Neighbors

This section is based on Hastie T. et al. "The Elements of Statistical Learning" [27].

The K-nearest neighbors (KNN) algorithm is a type of instance-based learning approach that is categorized under the family of lazy learning algorithms. Lazy learning algorithms differ from eager learning algorithms in that they do not construct a global model based on the entire data set. Instead, they store the training data and use it to make local approximations for new instances. KNN utilizes the concept of similarity between instances to generate predictions. Specifically, it selects the K most similar instances, also known as neighbors, to the instance under consideration for the given task. A conceptual illustration of how KNN separates anomalies is given in Figure 8. Two classes of data are clustered in purple and yellow, while the data point that is furthest from the two classes is identified as an anomaly and colored red.

Given a data set $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ with n instances, where $x_i \epsilon X$ is

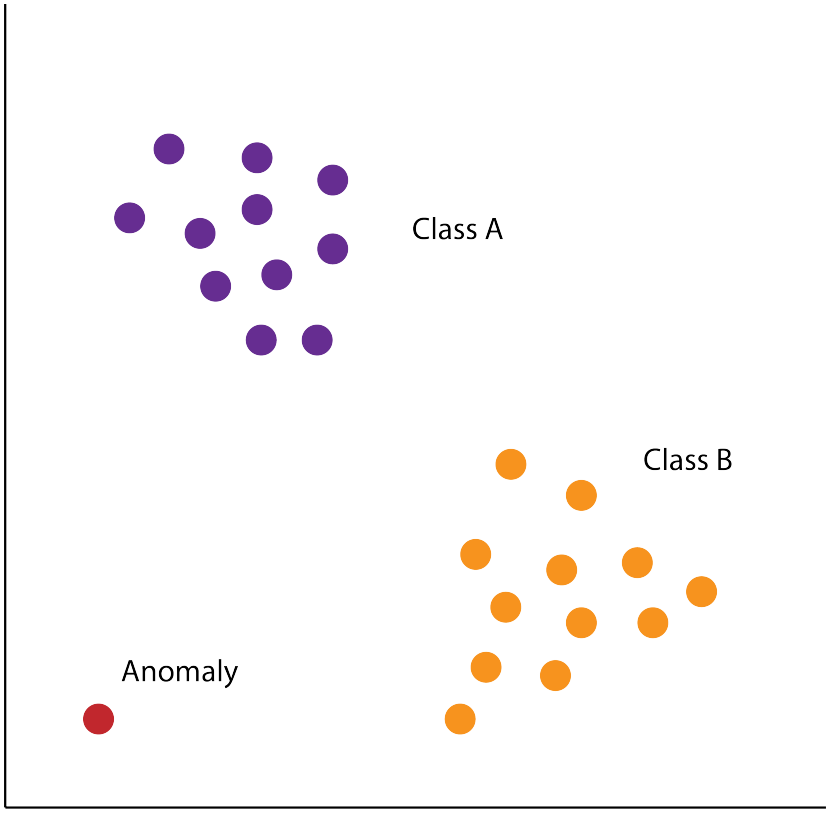


Figure 8: Conceptual illustration of how the K-NN algorithm distinguishes anomalies[56].

an input vector with d features and $y_i \epsilon Y$ is the corresponding output or class label, the KNN algorithm estimates the output for a new instance $x \epsilon X$ by the following steps:

- First, the distance between the newly introduced instance x and all instances present in the data set D is calculated. One may utilize a distance metric such as Euclidean, Manhattan, Minkowski, Chebychev distance, or correlation for this purpose.
- Then, the K instances that have the shortest distances to x is chosen. These instances are commonly denoted as the K-nearest neighbors of x.
- Lastly, based on the K-nearest neighbor outputs (class labels or continuous values), a prediction is made.

In classification tasks, the prediction is determined by selecting the predominant class from the K-nearest neighbors. For regression tasks, predictions are derived by averaging the outputs (continuous values) of the K-nearest neighbors. In order to detect anomalies, the KNN algorithm measures the dissimilarity between instances. Anomalies are instances that have a large average distance to their K-nearest neighbors.

When presented with the data set D and a new instance x, one can calculate the anomaly score, $A(x)$, by determining the average distance to its K-nearest neighbors. Mathematically, this is expressed as:

$$A(x) = (\frac{1}{K})\Sigma[dist(x, x_i)] \quad \text{for} \quad i \quad \epsilon \quad \text{K-nearest neighbours.} \tag{1}$$

A high $A(x)$ value indicates that the instance x is an outlier.

The choice of the distance metric and the K value has a significant impact on the performance of the KNN algorithm. Selecting a suitable distance metric that accurately represents the fundamental structure of the data is crucial. Choosing the appropriate value for K is essential because selecting a small K value can cause overfitting, whereas a large K value can lead to underfitting.

2.4.2 One-Class Support Vector Machines

This section is based on Hastie T. et al. "The Elements of Statistical Learning" [27].

The One-Class Support Vector Machine (OCSVM) is a variant of the Support Vector Machine (SVM) algorithm that has been specifically developed to address the task of anomaly detection. The One-Class Support Vector Machine (OCSVM) is distinct from the conventional Two-Class SVM in that it endeavors to identify the optimal decision boundary or hyperplane that effectively distinguishes the normal data points from the origin within the feature space. This is in contrast to the Two-Class SVM, which aims to separate data points belonging to two distinct classes. Figure 9 depicts a conceptual representation of how the OCSVM algorithm discerns anomalies from the remaining data. The decision boundary, which is formed by support vectors, serves to separate the anomalous data points, depicted in red, from the regular data points, depicted in purple.

The OCSVM algorithm endeavors to optimize the margin between the origin and the data points in the feature space, while permitting a portion of the data points to exist outside the margin. This is accomplished by considering a data set D comprising n instances, where each instance X_i is an input vector with d features. The One-Class Support Vector Machine (OCSVM) is expressed as a quadratic optimization problem in the following manner:

$$\begin{aligned} \min_{w,\xi,\rho} \quad & \frac{1}{2}\|w\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & \langle w, \phi(x_i)\rangle \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n, \end{aligned} \tag{2}$$

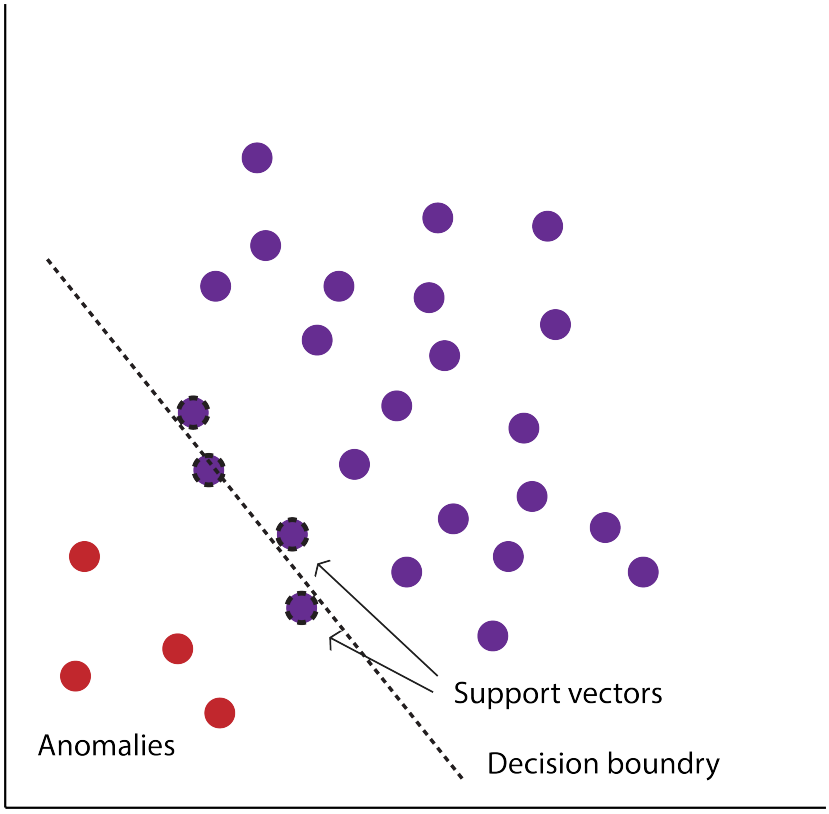


Figure 9: Conceptual illustration of how the OCSVM algorithm distinguish anomalies[2].

The equation for a hyperplane is defined by the weight vector, denoted as w, and the offset from the origin, denoted as ρ. The slack variables, represented by ξ_i, permit certain data points to exist outside the margin. Additionally, a hyperparameter, denoted as ν, is utilized to balance the objective of maximizing the margin while minimizing the number of outliers. The function denoted by $\phi(x)$ is responsible for mapping the data points from the input space X to a feature space of higher dimensionality.

The optimization problem mentioned above has a corresponding dual problem that can be expressed as follows:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \frac{1}{\nu n}, \quad \sum_{i=1}^{n} \alpha_i = 1, \quad i = 1, \ldots, n, \end{aligned} \tag{3}$$

The Lagrange multipliers, denoted by α_i, are utilized in conjunction with the kernel function $K(x_i, x_j)$. This function calculates the inner product of the mapped feature vectors $\phi(x_i)$ and $\phi(x_j)$ in the feature space. The OCSVM benefits from a dual formulation that leverages the kernel trick. This technique empowers the algorithm to identify non-linear decision boundaries in the input space, without the need to explicitly

compute the mapping function $\phi(x)$.

The performance of the OCSVM is significantly influenced by the selection of the kernel function and the hyperparameters. These two factors are of utmost importance in ensuring optimal performance. Kernel functions that are frequently used include the linear kernel, polynomial kernel, and Gaussian Radial Basis Function (RBF) kernel. The linear kernel can be defined as follows:

$$K(x_i, x_j) = \langle x_i, x_j \rangle \tag{4}$$

The polynomial kernel is defined as:

$$K(x_i, x_j) = (\gamma \langle x_i, x_j \rangle + r)^d \tag{5}$$

where $\gamma > 0$ is the kernel coefficient, r is the kernel offset, and d is the degree of the polynomial.

The Gaussian RBF kernel is defined as:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{6}$$

where $\gamma > 0$ is the kernel coefficient.

The selection of the kernel function is dependent upon the characteristics of the data and the fundamental issue at hand. The linear kernel is appropriate for data that can be separated linearly. Conversely, the polynomial and Gaussian RBF kernels are capable of representing intricate, non-linear decision boundaries. It is noteworthy that the utilization of excessively intricate kernels can result in overfitting, particularly in cases where the number of instances is limited.

2.4.3 Isolation Forest

This section is based on Liu F.T. et al. "Isolation Forest" [72]

The Isolation Forest (iForest) algorithm is a technique for detecting anomalies in an unsupervised learning situation. It operates on the premise of isolating anomalies rather than characterizing normal instances. The iForest algorithm constructs a collection of Isolation Trees (iTrees) and leverages the mean path length of observations within the iTrees to determine their level of abnormality. This is illustrated in Figure 10. Thanks to its distinctive procedure, iForest exhibits remarkable efficiency, particularly in data sets with high dimensions, and provides a resilient technique for identifying anomalies.

The Isolation Forest algorithm comprises a dual-step process, namely, forest construction and anomaly score estimation. The forest is built through a recursive process of partitioning the feature space. Splitting values and randomly selecting features enable this. Every step of partitioning generates an iTree that comprises a root, internal nodes, and external nodes (leaves). The procedure continues until an end criterion is met, which could be getting to the maximum tree depth or the separation of an instance in a leaf node.

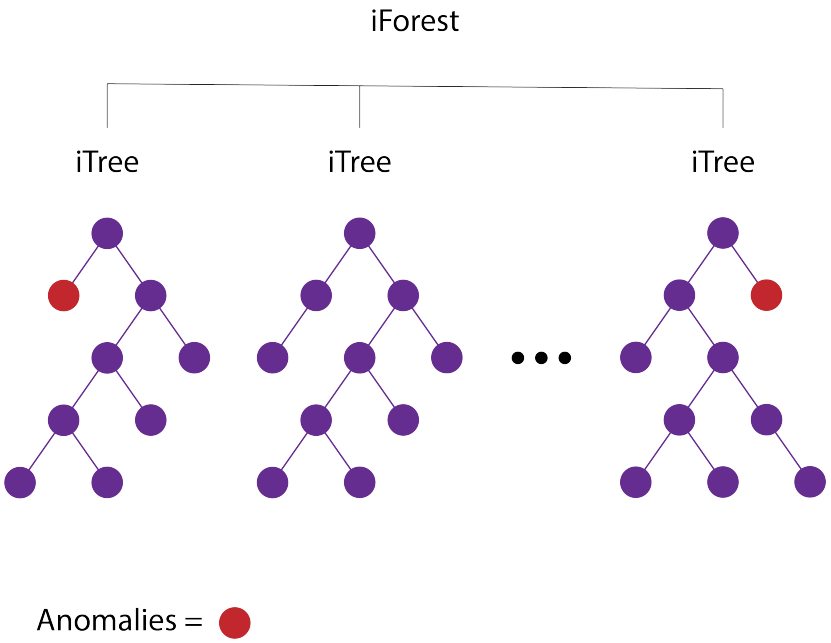


Figure 10: Conceptual illustration of how the iForest algorithm separate anomalies[48].

The Isolation Forest algorithm operates on a data set $D = \{x_1, x_2, ..., x_n\}$ comprising n instances, where each instance x_i is an input vector with d features belonging to the set X. The algorithm follows the subsequent steps:

- To generate a forest, it is necessary to construct iTrees, where each iTree is created by utilizing a random sub sample of the data set D.
- To determine the average path length $h(x)$ for each instance x in the data set D, one must traverse the iTrees in the forest.
- To determine the anomaly scores $A(x)$ for each instance x, one can utilize the average path length $h(x)$ and the expected path length for a specific tree size. This approach allows for the estimation of anomaly scores.

In an iTree, the distance traveled from the root node to the leaf node that holds a given instance x is referred to as the path length, denoted as $h(x)$. Instances with shorter

path lengths are more likely to be considered anomalies, as they can be readily isolated from normal instances.

To calculate the anomaly score $A(x)$ for a given instance x, one can use the following formula:

$$A(x) = \frac{2^{\frac{-h(x)}{c(n)}}}{\psi(n)} \tag{7}$$

In the context of iTrees, the function $h(x)$ denotes the mean path length of a given instance x. Meanwhile, $c(n)$ represents the average path length of an unsuccessful search in a Binary Search Tree (BST) that has n external nodes. Additionally, $\psi(n)$ serves as a normalization factor. The role of the normalization factor $\psi(n)$ is to ensure that the anomaly score is confined within the range of 0 to 1. This score is indicative of the probability of an instance being an anomaly, with a score closer to 1 implying a higher likelihood of an anomaly.

2.4.4 Local Outlier Factor

This section is based on Hastie T. et al. "The Elements of Statistical Learning" [27], and Kuna H.D. et al. "Outlier detection in audit logs for application systems"[33].

The Local Outlier Factor (LOF) algorithm is a widely used method for identifying anomalies in data sets that exhibit non-uniform instance distributions. The LOF algorithm operates under the assumption that anomalies are data points that exhibit a lower density in the nearest area compared to their neighboring data points. The LOF algorithm is capable of detecting instances that exhibit significant deviation from the majority of the data by comparing their local density with that of their nearest neighbors, as illustrated in Figure 11.

Given a data set $D = \{x_1, x_2, ..., x_n\}$ with n instances, where $x_i \epsilon X$ is an input vector with d features. The algorithm estimates the local density for each instance and subsequently computes the LOF score by comparing the ratio of local densities. The LOF algorithm proceeds through a series of steps as outlined below:

- To obtain the k-nearest neighbors for every instance x in the data set D, perform a computational process that involves identifying the k instances in D that are closest to x based on a chosen distance metric.
- Using their k-nearest neighbors, calculate the reachability distance between each pair of instances x and y.

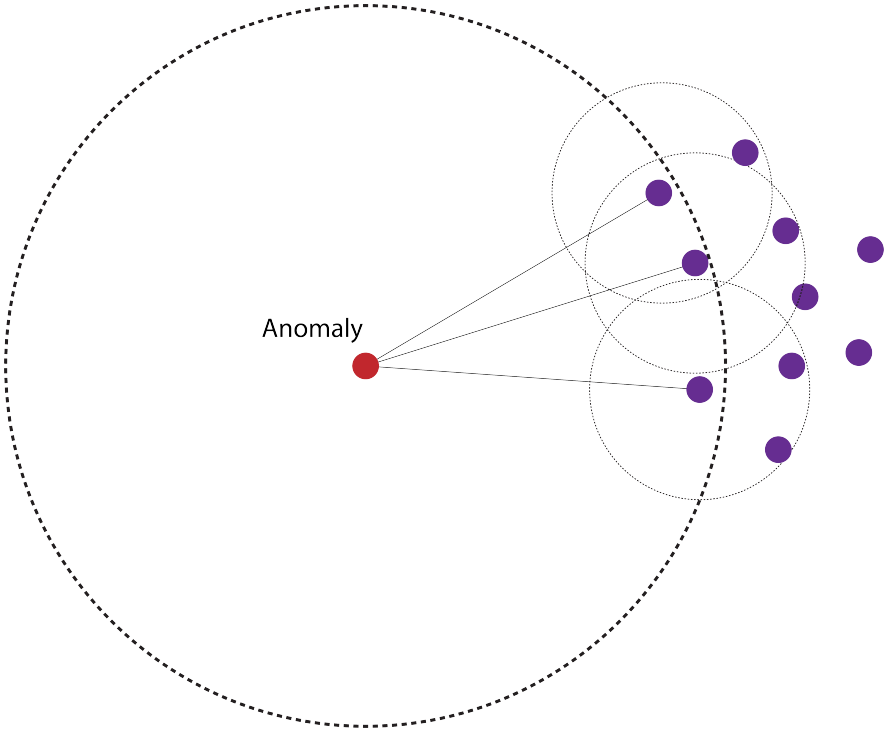


Figure 11: Conceptual illustration of how the LOF algorithm separate anomalies[39].

- Based on the reachability distance, compute the local reachability density (LRD) for each instance x.
- To evaluate the Local Outlier Factor (LOF) score for each instance x, it is necessary to compare its Local Reachability Density (LRD) with the LRDs of its k-nearest neighbors.

The distance of reachability between two instances, x and y, is formally defined as:

$$r_k(x, y) = \max\{k\text{-distance}(y), d(x, y)\} \tag{8}$$

Where k-distance(y) refers to the distance between a point y and its k-nearest neighbor. $d(x, y)$ denotes the Euclidean distance between two points x and y.

The LRD pertains to an instance x and is mathematically expressed as the inverse of the mean reachability distance linking x to its k-nearest neighbors:

$$LRD_k(x) = \frac{1}{\frac{\sum_{y \in N_k(x)} r_k(x,y)}{|N_k(x)|}} \tag{9}$$

where $N_k(x)$ denotes the set of k-nearest neighbors of x.

To determine the LOF score of a given instance x, one must compute the ratio of the LRD of its k-nearest neighbors to its own LRD:

$$LOF_k(x) = \frac{\frac{\sum_{y \in N_k(x)} LRD_k(y)}{|N_k(x)|}}{LRD_k(x)} \tag{10}$$

When the LOF score is higher, it implies that the instance has a lower local density in comparison to its neighboring data points. This observation leads to the conclusion that the instance is more likely to be an anomaly.

2.4.5 Artificial Neural Network

This section is based on Hastie T. et al. "The Elements of Statistical Learning" [27].

Artificial Neural Networks, commonly referred to as ANNs, are computational models that are specifically designed to imitate the intricate processes of biological neurons in terms of information processing and transmission. ANNs have demonstrated remarkable efficacy in diverse domains, including but not limited to image recognition, natural language processing, and game play. Artificial neural networks are composed of interconnected neurons that are arranged in layers. Each neuron receives input from other neurons and produces an output based on a predetermined activation function.

An artificial neuron is capable of receiving input from various sources, performing a weighted summation of these inputs, and subsequently applying an activation function to generate an output. The artificial neuron can be mathematically expressed in the way that follows:

$$y_i = f\left(\sum_{j=1}^{n} w_{ij} x_j + b_i\right) \tag{11}$$

In this context, the variable x_j represents the input originating from the j-th neuron. The weight that connects the i-th and j-th neurons is denoted by w_{ij}, while b_i refers to the bias term. Also, the function $f(\cdot)$ is the activation function.

ANNs are commonly composed of three fundamental layers: an input layer, one or more hidden layers, and an output layer. In a neural network, each layer is comprised of a distinct number of neurons, and these neurons are interconnected via weighted edges. A conceptual illustration of a feedforward ANN is given in Figure 12

The process of feedforward in an ANN involves the transmission of input signals through the network's layers to generate an output. The computation of the output of each

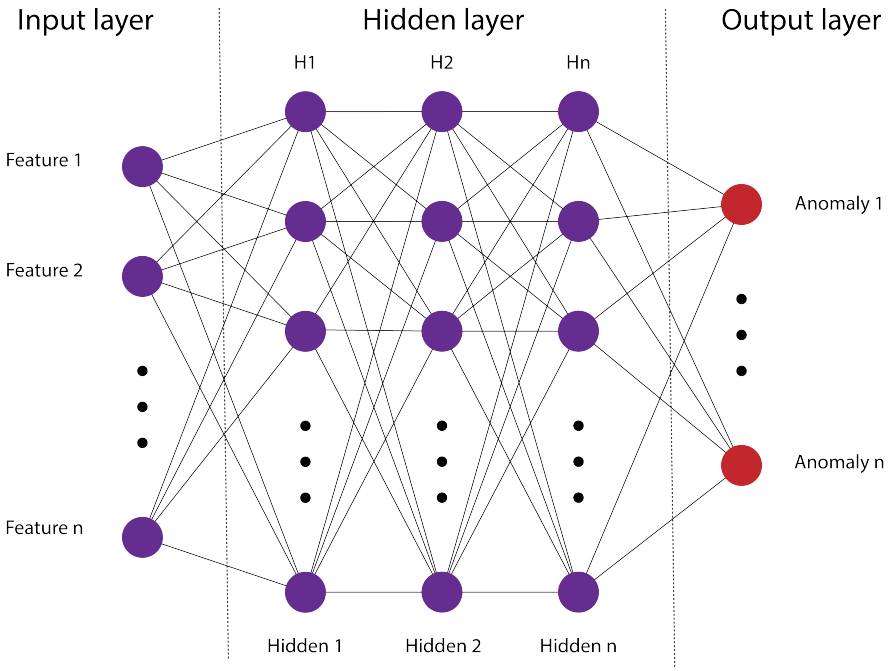


Figure 12: Conceptual illustration of how the feedforward ANN algorithm separate anomalies[57].

neuron in the network involves the weighted summation of its inputs and the activation function, as stated in the previous section.

The backpropagation algorithm is a technique for training artificial neural networks by minimizing the difference between the predicted output and the ground truth output. The algorithm can be broken down into two primary stages: first, the computation of the error gradient concerning the weights and biases; and second, the adjustment of the weights and biases based on the computed gradients.

To compute the gradient of the error E in relation to the weights and biases, one can apply the chain rule of calculus.

$$\frac{\partial E}{\partial w_{ij}^{(l)}} = \frac{\partial E}{\partial y_i^{(l)}} \cdot \frac{\partial y_i^{(l)}}{\partial w_{ij}^{(l)}} \tag{12}$$

The output of the i-th neuron in the l-th layer is denoted by $y_i^{(l)}$. The weight connecting the i-th neuron in the l-th layer to the j-th neuron in the (l-1)-th layer is represented by $w_{ij}^{(l)}$. After computing the gradients, the weights and biases undergo an update process utilizing gradient descent or alternative optimization techniques.

2.4.6 Autoencoders

This section is based on Li P. et al. "A comprehensive survey on design and application of autoencoder in deep learning" [34].

Autoencoders (AE) are a type of artificial neural network that is capable of learning how to compress input data into a lower-dimensional representation. This compressed representation is then used to reconstruct the original data. It consists of two primary constituents: an encoder, which facilitates the mapping of input data to a latent space of lower dimensions, and a decoder, which enables the mapping of the latent space back to the original data space. This is demonstrated in Figure 13. AE prove to be particularly advantageous in scenarios involving unsupervised learning tasks, where the availability of labeled data is limited or nonexistent.

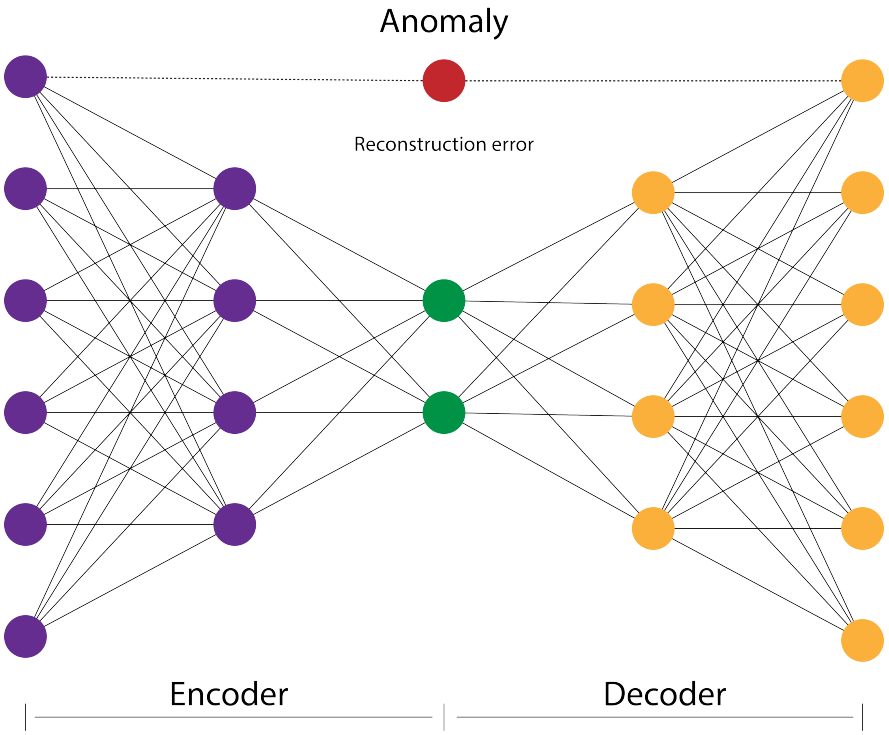


Figure 13: Conceptual illustration of how the Autoencoders algorithm distinguish anomalies[78].

An autoencoder consists of two essential components: an encoder network and a decoder network. The encoder network is commonly a feedforward neural network that receives input data x and transforms it into a latent representation z:

$$z = f_{\text{enc}}(x) \tag{13}$$

The decoder network is another feedforward neural network that operates by taking

the latent representation z and using it to reconstruct the original data x':

$$x' = f_{\text{dec}}(z) \tag{14}$$

The primary objective of the autoencoder is to acquire the optimal encoder and decoder functions, which can minimize the reconstruction error between the input data x and the reconstructed data x'.

Autoencoders are trained through the process of minimizing the reconstruction error that exists between the input data x and the reconstructed data x'. In general, the evaluation of the accuracy of a reconstructed output is done by employing a loss function, which can be either the mean squared error (MSE) or the binary cross-entropy (BCE):

$$L(x, x') = \frac{1}{N} \sum_{i=1}^{N} l(x_i, x'_i) \tag{15}$$

where N represents the number of input instances and $l(x_i, x'_i)$ represents the loss for the i-th instance.

Similar to the back propagation algorithm used to train artificial neural networks, the training process entails updating the weights and biases of the encoder and decoder networks by means of gradient descent or other optimization techniques.

AEs purpose is to learn how to compress input data into representations that are both efficient and compact. Additionally, they are able to reconstruct the original data from these compressed representations. AE have proven to be valuable in a multitude of applications, including but not limited to dimensionality reduction, denoising, and anomaly detection. The training process involves minimizing the reconstruction error by updating the weights and biases of both the encoder and decoder networks. This is achieved by minimizing the difference between the input data and the reconstructed data.

2.5 Time Series Data

This subsection is mainly based on previous work done by the same author[10]

The field of anomaly detection is vast and encompasses a wide range of techniques. AI and other ML algorithms represent just a small part of all strategies. Statistics and mathematical optimization are complementary technologies that are better suited for tackling certain challanges.

Time series data refers to a collection of data points that are recorded at regular intervals

of time, and their order is of utmost significance. Every individual data point within the series denotes the numerical value of a particular variable of interest at a distinct moment in time. The defining characteristic of these intervals is their consistency, which may vary from yearly to quarterly, monthly, daily, or even in microseconds, depending on the nature of the data. Time series analysis is a powerful tool that utilizes data of this nature to uncover fundamental patterns, trends, and seasonal fluctuations.

Time series analysis[36] is a statistical technique used to examine a collection of data points gathered over time. It is used to comprehend how a variable or combination of variables evolves over time and to recognize patterns and trends. Time series analysis may be utilized to determine the reasons for data changes and forecast future values. Time series analysis may be utilized for a variety of reasons, including forecasting, seasonal trend analysis, and data set comparison. There were several possibly relevant statistics discovered for data preparation and processing:

- Feature engineering: This is the process of making new features from existing data, like the difference between two points or the change in percentage over time.
- Decomposition: This is the process of breaking a time series into sequences like trend, seasonality, and residuals.
- Autocorrelation: Which measures the correlation between lagged versions of the same time series. Autocorrelation can be used to identify trends, seasonality, and other patterns in the data.
- Vector Autoregression (VAR): This is a statistical model for analyzing the interdependent relationships between various time series. It is used to discover how one time series affects another.
- Moving Average: Which smooths and removes short-term variations from a data collection while highlighting long-term trends and cycles.

2.6 Literature Review

This subsection is mainly based on previous work done by the same author[10]

In 2004, C. FU et al. published *Predictive Maintenance in Intellegient-Control-Maintenance-Management System for Hydroelectric Generating Unit*[24]. Predictive maintenance within the framework of an intelligent control-maintenance-management system (ICMMS) is presented. An artificial neural network (ANN)-based identification and diagnosis model is set up to implement the predictive maintenance of the electrohydraulic servomechanism in the hydroelectric generating unit. An ANN-based

identification and diagnosis model is put together with existing computer supervisory and control systems and management information systems to form an ICMMS in Geheyan Hydropower.

In 2017, I. Buaphan et al. published *Development of Expert System for Fault Diagnosis of an 8-MW Bulb Turbine Downstream Irrigation Hydro Power Plant*[11]. The author uses an expert system that allows inexpert maintenance crews to solve the problem as fast as the expert system can. Fault tree analysis (FTA) is one technique of failure analysis that can explicitly find out the failure causes that will affect the power plant's operation. The expert system can speed up the experience of the maintenance crews using this technique. Artificial Neural Networks (ANN) and Artificial Intelligence (AI) are used to help develop expert systems for fault diagnosis of small hydropower plant failures.

In 2014, L. Selak et al. published *Condition Monitoring and Fault Diagnostics for Hydropower Plants*[59]. A condition monitoring and fault diagnostics system for hydropower plants has been developed. The proposed system consists of signal acquisition, data transfer to the virtual diagnostics center, and fault diagnostics. A support vector machine (SVM) classifier has been used as part of the system. The data are stored in a database and transmitted through a secure VPN to the virtual diagnostics center, where fault diagnostics are performed. Data classification of the data was performed using the support vector machine method, which was on average 99.68% accurate.

In 2022, Y. Chen et al. published *Fault Anomaly Detection of Synchronous Machine Winding Based on Isolation Forest and Impulse Frequency Response Analysis*[16]. Isolation forest, which is a form of random forest, and impulse frequency response analysis (IFRA) are used for anomaly detection of synchronous machine winding faults. The basic principle of the anomaly detection method is introduced, and mathematical-statistical indicators of IFRA signatures are then explained. The experimental results show that the proposed method is superior to the existing conventional supervised learning method. The method does not need to predict the machine fault type in advance but instead analyzes a large amount of health data as normal and abnormal. It has a strong ability to distinguish faults and a fast calculation speed for periodic fault detection in synchronous machine winds. This is for synchronous machine windings. It considers a small amount of failure data abnormal.

In 2022, J. Pöppelbaum et al. published *Contrastive Learning Based on Self-Supervised Time-Series Analysis*[50]. This work aims to introduce a unique method for self-supervised learning-based time-series analysis based on SimCLR contrastive learning. They compare multiple fault classification algorithms on the benchmark Tennessee Eastman data set and report an accuracy improvement of 81.43 %, surpassing other multi-classification approaches. The author suggests MLP could instead be done with a

SVM.

In 2021, A. Betti et al. published *Condition Monitoring and Predictive Maintenance Methodologies for Hydropower Plant Equipment*[6]. The paper included a case study on two hydropower plants located in Italy. They propose a novel Key Performance Indicator (KPI) and argue that it outperforms conventional multivariable process control charts, like Hotelling t^2 index. The KPI is based on an appropriately trained Self-Organizing Map (SOM). The authors also claim that "Asymetric distribution of correct and faulty patterns advise against the usage of other classic supervised learning methodologies". Their study suggests SOMs are better suited than SVMs since unsupervised learning methodologies are usually more efficient to adopt to represent the structure and distribution of nominal data.

In 2020, P. Calvo-Bascones et al. published *Anomaly Detection Method Basen on the Deep Knowledge Behind Behavior Patterns in Industrial Components. Application to a Hydropower Plant*[13] The study describes a new technique that intends to fill a gap in the identification of abnormalities and diagnostics of industrial component behaviors. They propose an approach based on the generation of behavior patterns using unsupervised ML algorithms such as K-means and Self-Organizing maps (SOMs). The behavior patterns generated are compared with indicators of similarity and deviation combined. The case study is carried out at the Nygard hydropower plant, located in Norway.

In 2017, A. Dhieep et al. published *Thermal Stress Monitoring and Pre-Fault Detection Dystem in Power Transformers Using Fibre Optic Technology*[5]. The paper discusses the importance of monitoring and protection systems for power transformers, as faults in transformers significantly affect power systems' reliability. Traditional protective devices are described, but they often lead to system disconnections. The paper suggests the use of optical fiber sensing technology for more advanced fault detection and prevention. This method uses the phosphorescence of the material for temperature sensing. It allows for continuous monitoring of select hotspots within the transformers, feeding this information back to the protection system and cooling systems. Early detection can activate protective equipment before a circuit breaker operates, improving system reliability. The paper also discusses the effects of thermal stress on transformers and the proposed pre-fault detection system's field tests. The data transmission techniques and the specifics of the fiber optic sensing system are detailed as well.

In 2018, A. Sasidharan et al. published *Monitoring of Winding Temperature in Power Transformers*[4]. The paper discusses the concept of 'hotspots' within transformers where temperature varies with loading conditions, the identification and monitoring of which are crucial for performance evaluation and lifecycle monitoring. Different methods for hotspot temperature rise prediction are examined, including thermal modeling and soft

computing techniques such as fuzzy logic, neural networks, and ANFIS. The paper covers thermal stress monitoring techniques, discussing the development of a numerical model, the importance of sensor allocation in transformer winding temperature monitoring, and the Fibre Bragg Grating (FBG) principle of fiber optic sensing systems.

In 2022, A. Abbasi published *Fault Detection and Diagnosis in Power Transformers: a Comprehensive Review and Classification of Publications and Methods*[1]. This paper reviewed several methods, including thermography analysis, quantitative AI methods, signal processing, and a comparison of knowledge-driven, data-driven, and value-driven methods. The comparative analysis showed that each method has its own strengths and weaknesses, depending on specific conditions. For instance, knowledge-driven methods offer simplicity, rapid diagnosis, and strong explanatory power, making them effective with small data volumes. However, as data volume and complexity increase, extracting relevant knowledge becomes more challenging, making data-driven methods a more suitable choice due to their ability to exploit fault information from large data sets effectively. Value-driven methods, while having similarities with data-driven methods, are computationally expensive and time-consuming, making them less feasible for large data volumes. Moreover, the paper touched upon the evolution of FDD studies from offline to online methods and from focusing on systems' physical mechanisms and structures to accuracy and computational efficiency.

In 2021, R. Soni et al. published *Review on Asset Management of Power Transformer by Diagnosing Incipient Faults and Fault Identification Using Various Testing Methodologies*[63]. This paper focuses on various tests and methodologies used to identify potential faults in power transformers and estimate the residual life of these critical electrical assets. The paper reviews various tests for chemical, electrical, and other faults, such as dissolved gas analysis, oil testing methods, sweep frequency response analysis, recovery voltage method, partial discharge, infrared thermograph test, turns ratio test, dielectric dissipation factor, transformer winding resistance, core to ground test, and insulation resistance calculation. The document discusses the importance of condition-based assessment (CBA) and condition-based monitoring (CBM) in understanding the health and lifecycle of transformers, as well as predicting possible failures. The authors also describe the financial feasibility and benefits of these practices for both utilities and corporations. They highlight the importance of understanding transformer insulation decay, noting that it can increase the likelihood of failure and decrease the transformer's residual life.

In 2009, M.A. Taghikhani et al. published *Prediction of Hottest Spot Temperature in Power Transformer Windings with Non-Direct and Direct Oil-Forced Cooling*[68]. This technical paper discusses a numerical method for determining the hottest spot temperature (HST) in power transformers. The hottest spot in a transformer is an area of interest because exceeding the prescribed HST can cause insulation failures. Thus,

accurate predictions of the HST are crucial for the transformer's operation and life expectancy. The authors argue that existing models, such as the IEEE loading guide, often overlook the differences in winding structure and heat loss/unit volume between two transformers of identical ratings.

In 2022, D. Zou et al. published *Outlier Detection and Data Filling Based on KNN and LOF for Power Transformer Operation Data Classification*[83]. The paper proposes a data preprocessing method for power transformer operation data. The method uses the K-nearest neighbor (KNN) and local outlier factor (LOF) algorithms to classify and detect abnormalities in the data. The researchers calculate the local reachable density of the input data using the LOF algorithm, determining a local outlier factor score for the data. Data with abnormal scores is outputted as abnormal. The KNN algorithm is then applied to classify the data around these abnormal values and any missing values. The method was found to effectively detect and correct abnormal and missing data, providing accurate data samples for fault diagnosis. It was concluded that the LOF algorithm is effective in identifying outliers in the transformer data and that the KNN algorithm can correctly classify data and fill in missing values. The method was deemed suitable for handling time series data for transformers.

In 2023, N. Islam et al. published *Power Transformer Health Condition Evaluation: A Deep Generative Model Aided Intelligent Framework*[30]. The paper introduces a deep generative model-aided intelligent framework for effectively evaluating the health conditions of power grid transformers. The proposed model employs a multi-layer perception generative model paired with a logistic regression classifier. The developed model utilizes twelve input layers, which allows the model to effectively compress the data set, and eight categories in the output classification layers. The model was tested on real-world testing data from 608 transformers across 31 categories. The performance of the proposed framework was confirmed to be effective in precisely evaluating the health condition of transformers, with the model outperforming existing machine-learning models by achieving an accuracy of 99%. The model was created by integrating a multi-layer perception auto-encoder model.

In 2023, Z. Xing et al. published *Health Evaluation of Power Transformers Using Deep Learning Neural Network*[81]. The paper presents a deep learning neural network (DLNN) approach for detecting the health condition of power transformers. This method seeks to address issues such as redundancy, complexity, and the small sample size of the data set, which can affect the performance of health evaluation methods. The proposed DLNN is composed of three networks: generation, extension, and extraction. First, the Leaky Echo State Network (Leaky-ESN) generates data related to the original data set, solving the issue of the small sample size. Then, significant features are extracted by the DLNN, which incorporates improved Deep Residual Shrinkage Networks (IDRSN) and a one-dimensional convolutional neural network (1DCNN). Finally, the health

status of the power transformer is obtained by the Concat layer and Softmax layer. The experiments show that the DLNN outperforms other existing neural networks and health assessment methods for power transformers, achieving higher accuracy. However, the CPU time is still high, suggesting a need for future work to reduce computation time while maintaining high accuracy. The paper concludes by indicating that the proposed DLNN has a strong ability to detect the health condition of power transformers.

In 2023, Z. Xing et al. published *Multi-Modal Information Analysis for Fault Diagnosis with Time Series Data from Power Transformer*[80]. The paper discusses a novel approach to diagnosing faults in power transformers using a multi-modal information analysis method that includes a Selective Kernel Network, a bidirectional gated recurrent unit, and a cross-attention mechanism. This method is designed to work with time sequences and multi-modal data, addressing limitations such as multi-modal heterogeneity of data and missing samples that have hampered the effective use of ML in this context. The proposed method is tested with data sets of dissolved gas and infrared image modes collected from real power transformers and historical data. The results demonstrate that the proposed method has higher diagnostic accuracy and a quicker diagnostic time than comparison methods, outperforming other neural networks.

2.7 Health of Transformers

According to the literature review, it is crucial to monitor the health and condition of power transformers to ensure the dependable functioning of electrical power systems. The typical techniques used to evaluate the condition of power transformers involve the analysis of gas and oil, such as dissolved gas analysis (DGA). This method provides significant insights into the insulating materials, oil state, and likely faults of the transformer. In the case of large transformers installed in power plants that have generating capacities above 100 MW, it is typical to include additional sensors, such as winding temperature and hydrogen concentration sensors [7]. The use of these sensors allows the CBM in real-time, leading to the early detection of faults.

The literature review indicates that none of the studies have exclusively used winding temperature and hydrogen concentration for anomaly detection. However, several studies have acknowledged the potential of these parameters for detecting faults. A rise in hydrogen concentration and winding temperature may be an early indicator of undesirable thermal behavior. Determining how well AI and ML algorithms can perform anomaly detection with such data is thus an interesting approach.

2.7.1 Winding Temperature Sensors

Winding temperature sensors are typically installed to monitor the temperature of either the hotspots or the average temperature of the windings. Continuous temperature data is provided by these devices, which can be implemented to monitor the thermal performance of the transformer, prevent overloading, and identify abnormal conditions. High temperatures in the winding can suggest problems such as overloading, degradation of insulation, or the existence of hotspots[73].

2.7.2 Hydrogen Sensors

Hydrogen sensors are utilized for the purpose of monitoring the levels of hydrogen gas present in the insulating oil of transformers on a continuous basis. Particularly when subjected to intense electrical and thermal stress, hydrogen is one of the main gases produced during the breakdown of insulating oil and paper insulation. Elevated levels of hydrogen concentration may signify potential problems such as overheating, arcing, or PD. The implementation of real-time hydrogen monitoring promotes rapid detection of potential faults, thereby preventing timely corrective measures and maintenance[29].

2.7.3 Anomaly Detection Using Winding Temperature and Hydrogen

Continuous monitoring of winding temperature and hydrogen concentration permits early detection of anomalies, which may be indicative of numerous transformer faults. Comparing the current temperature with the anticipated or typical temperature range for operation, for example, can help detect anomalies in winding temperature. Similar to this, a sharp rise or a sustained upward trend in hydrogen concentration may call for additional research to determine the real cause of the problem. AI and ML techniques may also reveal hidden trends and patterns indicating abnormal behavior.

By utilizing winding temperature and hydrogen sensors, various types of faults can be identified in the power transformers. These faults include overloading, hotspots, insulation degradation, PD, and overheating. Detecting these faults at an early stage enables rapid action, thereby preventing fatal breakdowns and improving the operational life of the transformer.

Although winding temperature and hydrogen sensors are useful in detecting anomalies and potential faults in power transformers, they do have some limitations. Depending solely on these sensors may not offer a comprehensive evaluation of the transformer's state, as they may not detect all categories of defects. Combining monitoring and diagnostic techniques like DGA, electrical testing, vibration analysis, and PD analysis is advised to gain a more thorough understanding of the transformer's health. Still, this is a new field, and using this approach may reveal new discoveries.

3 Methodology

This chapter presents the methodology adapted for this study, offering a thorough explanation of the procedures, methods, and approaches employed. The process encompasses the gathering and organization of data, developing models, and evaluating metrics.

3.1 Flowchart

This section contains a flowchart that serves as a visual guide to the research procedure. It provides a clear overview of the methodology, from data collection to execution and evaluation. The flowchart is presented in Figure 14.

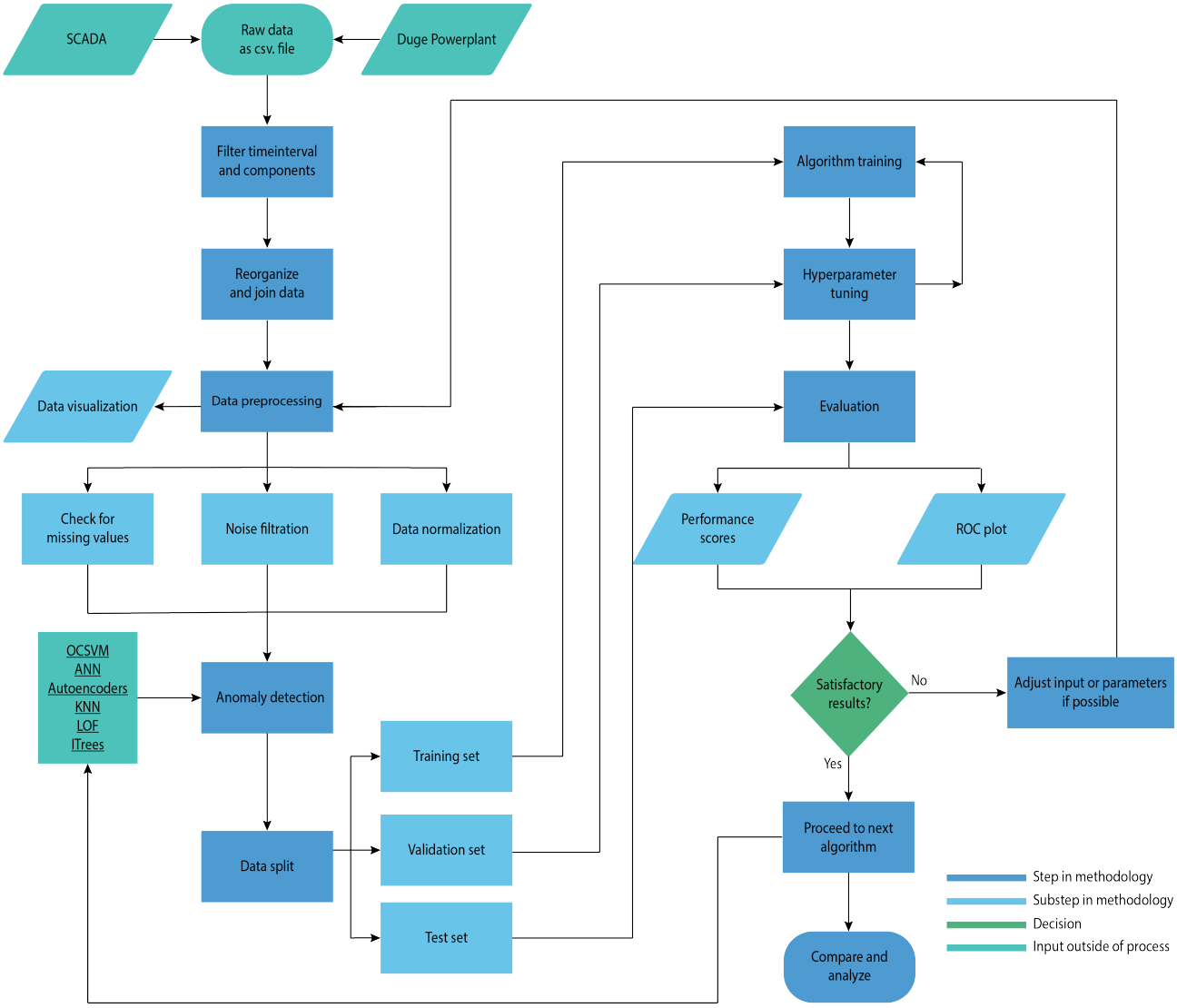


Figure 14: Flowchart over the methodology utilized.

At first, raw data is gathered from SCADA and the power plant, and it is stored in CSV files. The data is transferred into a suitable working environment, which in this case is PyCharm and Python, as utilized for this thesis. The data is effectively filtered to ensure that it is applicable to the chosen intervals and relevant components for the selected test case. Subsequently, the information is reorganized and merged. To prepare the data for anomaly detection, a series of data preprocessing measures and data visualization steps are required. When developing the model, the focus is on a single algorithm at a time. To assist in the tuning of the hyperparameters, the data is divided into three sets: the training set, the validation set, and the test set. Next, the algorithm is trained with the optimal parameters that were discovered. After that, the performance is assessed, and if deemed satisfactory, it proceeds to the next algorithm. If the evaluation scores are not satisfactory, it is necessary to pay extra attention to the parameters and input data. If possible, adjustments should be made accordingly. Once all the algorithms have been tested, they are compared, and the results are analyzed.

3.2 Implementation

The following section presents the implementation of the system and the software applications utilized throughout the thesis.

3.2.1 Python

Python 3.9[54] is the primary programming language used for this thesis, and PyCharm is the preferred integrated development environment (IDE). PyCharm, a software developed by JetBrains[53], offers a range of features that enhance the coding process. The features encompass smart code assistance, debugging and testing support, and seamless integration with diverse version control systems.

PyCharm, the code editor, is designed to improve the coding process by offering features like code completion, error detection, and quick navigation. These abilities enhance the effectiveness and precision of coding. The entire code is written during the master's thesis.

3.2.2 Standard Libraries

Python has a set of commonly used packages that are frequently used in the fields of data analysis and model development. These packages are well-known for their robust data manipulation and visualization capabilities.

Pandas[75] is a well-known Python package that is renowned for its speed, adaptability,

and ability to manage labeled and relational data structures. The goal of Pandas is to function as a crucial top-level element for conducting practical, real-world data analysis in Python, streamlining operations for enhancing and structuring data.

NumPy[47] is a powerful Python package that is recognized for its speed and versatility. The principles of vectorization, indexing, and broadcasting in NumPy are widely recognized as the best practices in array computation. In addition, NumPy offers a wide variety of mathematical functions, generators of random numbers, routines for linear algebra, Fourier transforms, and other functionalities.

Matplotlib[37] is a flexible visualization library for Python that can produce a wide range of graphics, including static, animated, and interactive ones. The software provides a wide variety of plot options, such as histograms, scatter plots, and violin plots, which come in various styles and colors.

3.2.3 Scikit-learn

The sklearn library, also known as the Scikit-learn package[58], is a thorough machine learning (ML) library that supports a variety of algorithms and methodologies. This resource is essential for ML tasks as it provides a variety of tools and functionalities for data preprocessing, tuning, and evaluation.

Scikit-learn provides extensive support for both supervised and unsupervised learning algorithms. In the context of this thesis, this library plays a crucial role in implementing several ML techniques. These include the Local Outlier Factor (LOF), the One-Class Support Vector Machine (OCSVM), the K-Nearest Neighbors (KNN), and the Isolation Forest(iForest).

Apart from implementing these models, Scikit-learn is also heavily used for preprocessing tasks, which include cleaning data, handling missing values, and feature scaling, among others. In addition, Scikit-learn offers reliable resources for optimizing hyperparameters.

3.2.4 Tensorflow

Scikit-learn does not support neural networks, so the TensorFlow[69] package developed by Google Brain was installed additionally. TensorFlow is a powerful and popular open-source library that includes models, layers, and optimizers necessary to construct artificial neural networks (ANN) and autoencoders (AE). TensorFlow's flexible and intuitive interface simplifies the process of building these networks, facilitating various architectures ranging from simple feed-forward networks to more complex structures like convolutional neural networks (CNN) and recurrent neural networks (RNN).

TensorFlow also provides robust tools for model evaluation and optimization. It allows for easy computation of various performance metrics and offers powerful optimizers for gradient descent, a fundamental algorithm used to minimize the loss function in neural networks.

3.3 Data Collection

It is often challenging to carry out tests using power system data due to the sensitive nature of information related to power production and producers. This data is typically considered confidential and not easily available to the public. This is also relevant to the data collected for this thesis.

The Sira-Kvina power company provided the data for this study. They exchanged information about one of their power plants named Duge, as described in the introduction. The company provided access to nearly three years of recorded data, containing both hourly and minutely recorded recordings on different parts of the generators and transformers.

Apart from the recorded data, Sira-Kvina also provided an indication log that was obtained from their Supervisory Control and Data Acquisition (SCADA) system. The log consists of data related to the different indicators of the transformers, and each record is designated as either '1' or '2'. The value '1' denotes the activation of both the single-point and dual-point indicators, whereas the value '2' signifies their deactivation. The log provides valuable information regarding the functioning of transformers and can aid in detecting recurring trends or relationships with other recorded variables. Upon receiving the data from Sira-Kvina, numerous steps were taken to organize and prepare the data for analysis.

3.4 Analysing the Data and Building a Test Case

The research is applicable to the detection of anomalies in both generators and transformers, but since the indication log was recorded on the transformer, that was where the test case was constructed. The review carried out within the literature showed potential for detecting anomalies in the transformer through the use of recordings of winding temperature and hydrogen. As a result, these two components were selected, along with active power recordings, to gain an understanding of the behavior patterns at various production levels. Based on the available data and indication log, a time interval spanning from October one year to December the next year was selected as the period with the most intriguing indications. The decision was made to concentrate solely on transformer 1, as the signals of interest were specific to this particular transformer. The

features selected and their respected units of measure are listed in Table 2.

Feature	**Unit**
Active power	Megawatt $[MW]$
Hydrogen	Parts per million $[ppm]$
Winding temperature	Degrees Celsius [℃]
Indication	1 = on, 2 = off

Table 2: Features selected and their respected units of measure.

The frequency of logging ranges for different components and indications, with some being recorded every minute, some every second minute, and some every five to ten minutes, for example. After choosing the desired time interval and combining the components and indication data, the resulting reconstructed data will have a structure that resembles the example shown in Table 3.

Timestamp	**Hydrogen**	**Winding temp**	**Active power**	**Indication 1**	...	**Indication n**
Time 1	$H_2[ppm]$	Temp[℃]	Power$[MW]$	2	...	1
Time 2	NaN	NaN	Power$[MW]$	2	...	2
Time 3	NaN	Temp[℃]	NaN	2	...	NaN
⋮	⋮	⋮	⋮	⋮		⋮
Time n	$H_2[ppm]$	NaN	Power$[MW]$	NaN	...	1

Table 3: Example of how the data is structured after reconstruction for the test case, before data preprocessing.

3.4.1 Indications

From the indication log, 10 different incidents are recurring during the time interval chosen. Some of them are triggered several times over longer periods, some of them only a few times for a short period. The incidents and the number of times their sensor was triggered are listed in Table 4. In total, this results in 63 anomalies in the data set.

Incident	Occurrences
Gas	14
Gas critical	5
Oil temperature	6
Winding temperature	4
Temperature critical	4
Water circulation	11
CO_2 triggered	7
Control current	1
Cable minimum oil pressure	8
Cable minimum oil pressure critical	3

Table 4: Name of incident and how many times they are triggered.

3.4.2 Visualizing data

The data has been visualized using line plots, as shown in Figure 15. The data has been normalized to enable better representation. The recordings of winding temperature are colored blue, the recordings of hydrogen concentration are colored green, and the recordings of active power are colored orange. The indications are plotted onto the line plots and are colored red. The power plant uses reversible turbines, and when negative active power is observed, it signifies that water is being pumped from the lower to the higher magazine. Zero active power indicates that no production is taking place.

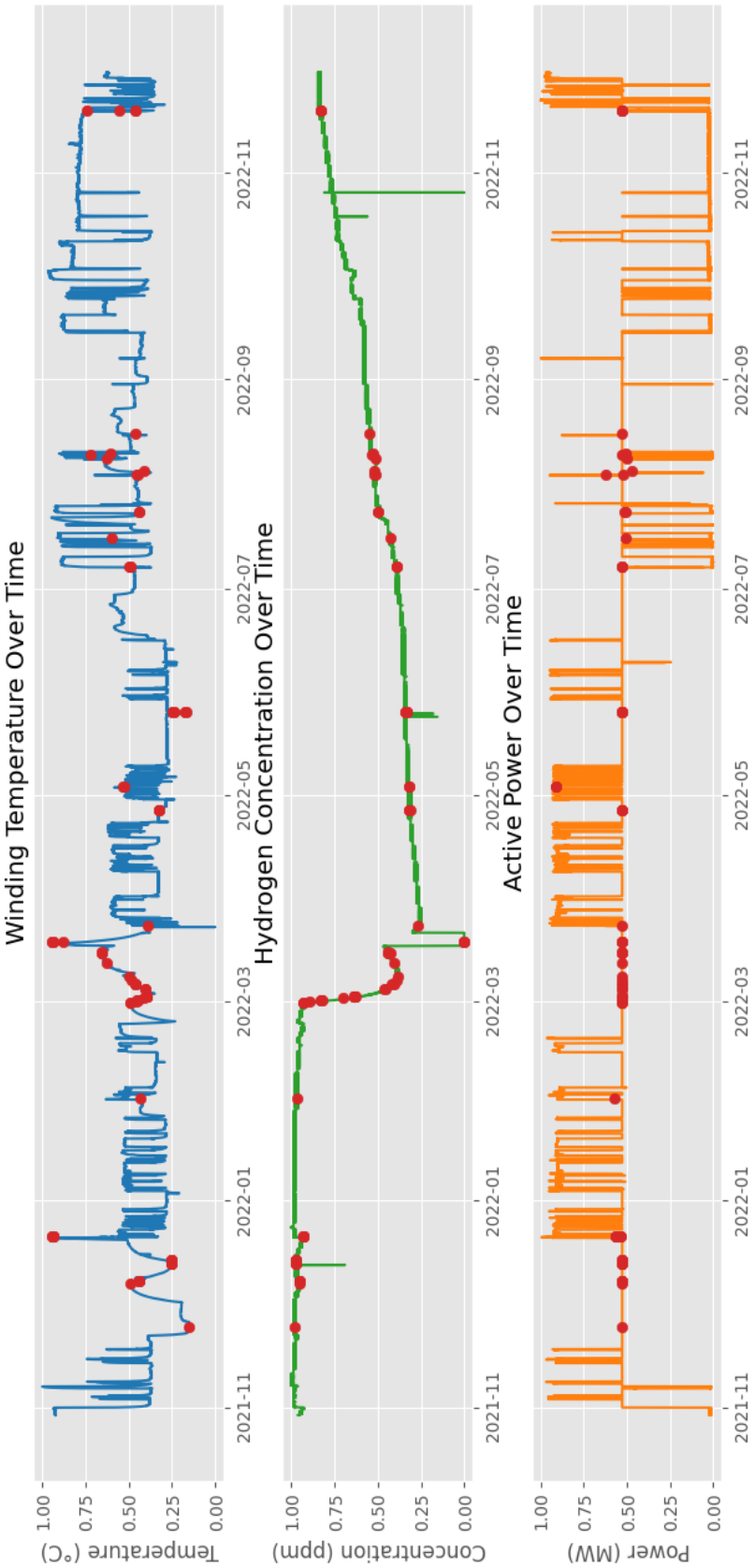


Figure 15: Winding temperature, hydrogen concentration and active power plotted over time, with incidents marked with red.

3.5 Data Preprocessing

Noise, consistency issues, and other issues frequently present in practical applications could reduce the effectiveness of anomaly detection algorithms. The process of data preprocessing is crucial in order to deal with these challenges and enhance the precision and effectiveness of the detection procedure.

The process of data preprocessing is used to improve the overall quality of the data and make it suitable for analysis. This process ensures that the identified anomalies are authentic outliers and not inaccurate results of data inconsistencies or random fluctuations. In addition, the process of preprocessing can improve the effectiveness of the algorithm by decreasing the computational complexity and simplifying the data presentation. Furthermore, it enables easier interpretation of the results and a more accurate comparison of various anomaly detection techniques.

The received data set lacked structure as all its components were listed together in a single column with duplicate timestamps. Consequently, the data set underwent filtration to extract the necessary components. The resulting components were then organized into their respective dataframes and merged based on timestamp, as already shown in Table 3. The log indicating the status of the components was restructured and joined with the rest of the data.

3.5.1 Missing Values

It was crucial to check for the presence of missing values in order to ensure the quality of the test data, as not all components were consistently recorded on a minute-by-minute basis. The absent values are commonly denoted as NaN, which is an acronym for Not a Number. In the columns that record incidents, any missing values were replaced with the number 2. This is because a minute that does not indicate a fault was originally represented by the number 2 in the incident log. Linear interpolation was used to recover missing values in the columns that contained component data[65]. This was done by estimating the values at the missing time points based on the values recorded at the closest two minutes. The method involves filling in the missing data points by constructing a straight line segment that connects the latest known data point prior to the missing values and the next known data point following the missing values in the sequence. Assuming the presence of two data points, (x_1, y_1) and (x_2, y_2), this can be represented mathematically like this:

$$y = y_1 + (x - x_1) \cdot \frac{(y_2 - y_1)}{(x_2 - x_1)} \tag{16}$$

This is easily accomplished using the Pandas function **interpolation()** when programming in Python. The default setting is linear interpolation. The Pandas function **fillna()** is used to replace missing values with a designated value.

The data set also went through a process to identify and remove any duplicate values. It is crucial to check for duplicate values during data preprocessing because such entries can cause distortions in the actual patterns and relationships present in the data. This can result in incorrect conclusions and an improper distribution of data, ultimately impacting the accuracy and reliability of anomaly detection models developed using the data set.

3.5.2 Noise filtration

When analyzing data from power production, it's common to use noise filtration techniques to enhance the precision and reliability of the results. The median filtering technique is a frequently used method for this objective.

Median filtering is a technique used in signal processing to reduce noise[35]. It is a non-linear method. The principle can also be applied to time-series data related to power generation. This technique involves substituting every value with the median value of its neighboring values. The sliding sequence of close data points in a signal is referred to as a "window". This window moves through the entire signal, one entry at a time.

The median filter is a technique used to process a data set X, where $X(n)$ represents the value of a parameter at a particular time step. The filter works by calculating the median value of a subset of the data, which is then used to replace the original value at that time step. This is shown mathematically like this:

$$X_f(n) = \text{median}\{X(n - \frac{M}{2}), X(n - \frac{M}{2} + 1), \ldots, X(n), \ldots, X(n + \frac{M}{2})\} \tag{17}$$

The variable $X_f(n)$ denotes the value of the parameter after filtering, while M denotes the order of the median filter. The order of the filter refers to the number of entries that are considered when calculating the median.

The median filter is a highly efficient method for addressing sharp noise, which refers to instances where certain data points exhibit unpredictable spikes in either direction, either higher or lower than their neighboring data points. The tool's capacity to maintain edges while eliminating noise renders it a valuable asset in tasks related to signals.

The median filter is a preferred method for filtering in power production due to the

frequent measurements taken. It is both simple and effective. The median filtering method has a time complexity of O(n), which means it is fast and efficient for its intended use. The median filter is a viable option for real-time noise filtration in power production systems due to its ability to process large amounts of data efficiently within a short period of time.

3.5.3 Data Normalization

In data preprocessing for ML, it is often necessary to standardize the features to ensure that they are on a similar scale. This process can also be beneficial in cases where the data has a normal or near-normal distribution. Standard scaling is a common method used for this purpose.

Standard scaling transforms the data such that it has a mean of 0 and a standard deviation of 1[64]. This scaling method does not set a specific range for the data, as min-max scaling for example does. Instead, it adjusts the distribution of the data based on the mean and standard deviation, which is particularly useful when the data follows a normal distribution.

The formula for standard scaling is as follows:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma} \tag{18}$$

X_{scaled} is the scaled feature, X is the original, with σ and μ being the standard deviation and mean of the original feature, respectively.

After applying standard scaling, the distribution of the data will have the same shape as the original distribution, but with a mean of 0 and a standard deviation of 1. If the original data has a normal distribution, then the scaled data will also have a normal distribution.

3.5.4 Train Test Split

Splitting data into training and testing sets is an important step when working with ML algorithms. This is especially true for supervised learning algorithms. These algorithms learn from the training data, and then their performance is checked with the testing data. This way, we can make sure our models work well with new, unseen data.

Even though unsupervised learning algorithms do not need labeled data, it can still be a good idea to split the data. This is because we often need to adjust certain settings, called hyperparameters, when training these algorithms. By having a test set, we can

check how well the algorithm is doing with different hyperparameters. This helps us find the best settings for the algorithm.

In data sets with a large number of data points and very few anomalies, often referred to as imbalanced data sets, traditional random splitting might not be the best approach for creating training and test sets. This is because there is a risk that some of the few anomalies might end up only in the training set or only in the test set, which could bias the results. In such cases, stratified sampling is necessary. This method involves splitting the data such that the proportions between the normal data and anomalies are the same in both the training and test sets. This ensures that the model gets a representative sample of both classes. This is easily executed by adding the command **stratify=y** in the traditional **train_test_split()** function from *sklearn.model_selection*.

It is also beneficial to add a validation set. The train and test sets are still used to train and test the model, while the validation set is used to tune hyperparameters. Adding a validation set results in a distribution of 60% data in the training set and 20% data in each of the test and validation sets, still maintaining an even distribution of anomalies.

After splitting the data, the total number of data points and how many of them are anomalies are presented in table 5. Since this is a binary classification problem, the data can be divided into classes 0, which represent the non anomalous points, and 1, which represent the anomalous points.

	Data points	**Class 0**	**Class 1**
Train set	343024	342986	38
Validation set	114341	114329	12
Test set	114342	114329	13

Table 5: Proportion of anomalies in each set after data splitting.

3.6 Model Development

Six different algorithms were chosen, each with desired capabilities for anomaly detection, to compare and discuss their performance. Those six algorithms are presented in background information and are: One-Class Support Vector Machines (OCSVM), Isolation Forest (iTrees), Autoencoders (AE), K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), and Local Outlier Factor (LOF). The algorithms are trained and tested one at a time. The Python libraries used to program each of the algorithms are described earlier in the implementation.

3.6.1 Hyperparameter Tuning

Optimizing hyperparameters is a fundamental component of anomaly detection algorithms, as it guarantees the best possible performance and precision of the models in detecting anomalies or atypical patterns within the data. Algorithms for detecting anomalies frequently depend on diverse parameters that regulate their functionality, including the quantity of clusters, distance thresholds, or neighborhood sizes. The selection of these hyperparameters holds great importance in determining the algorithm's efficacy in accurately discriminating between regular and anomalous observations. By fine-tuning the hyperparameters, the algorithm can effectively adjust to the unique features of the data and reduce the occurrence of false positives or negatives. This, in turn, enhances the accuracy and reliability of the results.

It is notable that hyperparameter values can vary among different algorithms, given that each algorithm functions based on unique underlying principles and assumptions. For example, a density-based algorithm may require a parameter to define the minimum number of points within a specific radius to form a dense region, whereas a clustering-based algorithm may require a parameter to set the number of clusters. In turn, the most favorable values for hyperparameters in one algorithm may not be applicable to another algorithm. This underscores the importance of fine-tuning hyperparameters on a per-algorithm basis to guarantee their utmost efficacy in identifying anomalies within the limits of the given data set and problem domain.

The process of hyperparameter tuning was carried out by utilizing the **GridSearchCV()** function from the sklearn library. This particular function employs cross-validation, a resampling technique that partitions the data into multiple train-test sets. The model is then trained on various iterations to assess its performance. Figure 16 provides an illustration of the working process, the red dots being anomalies and purple being normal data. The function initializes a parameter grid, which consists of various values for the hyperparameters of the model. The function **GridSearchCV()** conducts an extensive search of the designated parameter grid and applies a scorer to identify the

optimal parameters for the model.

Figure 16: Illustration of how the K-fold cross validation works[61].

3.6.2 Evaluation

There are several methods available to evaluate the efficacy of the suggested algorithms and their prognostications. When considering anomaly detection, the primary focus of evaluation is its ability to accurately identify anomalies within the data.

Evaluating the success of an anomaly detection algorithm can pose a challenge, especially when confronted with large data sets that include only a small number of anomalies. In situations like these, relying solely on accuracy as an evaluation metric can be misleading. This is because a high accuracy score may only indicate the algorithm's ability to identify the majority class (non-anomalous data points) while disregarding the detection of the rare anomalies. Stated differently, a predictive model that consistently forecasts the most prevalent class might demonstrate a notable level of accuracy, regardless of its poor ability to identify anomalies.

In order to gain a greater understanding of the algorithm's efficacy, it is essential to take into account additional metrics, including precision, recall, and F1-score. Recall quantifies the percentage of true positive anomalies among all the instances the algorithm classified as anomalies, while precision measures the percentage of true positive anomalies among all the actual anomalies in the data set. The F1-score is a statistical measure that represents the harmonic mean of precision and recall. It provides a balanced assessment of the algorithm's capacity to accurately detect anomalies while minimizing the occurrence of false positives. A better level of performance is indicated by a higher F1-score. The scores are computed using the following formula [19]:

$$\begin{aligned} \textbf{Precsion} &= \frac{TruePositives}{TruePostitives + FalsePositives} \\ \textbf{Recall} &= \frac{TruePositives}{TruePositives + FalseNegatives} \\ \textbf{F1-Score} &= \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \end{aligned} \tag{19}$$

Furthermore, the Receiver Operating Characteristic (ROC) curve serves as a valuable tool for evaluating the balance between the true positive rate (also known as sensitivity or recall) and the false positive rate (1-specificity)[82]. This is a graphical representation that illustrates the performance of a binary classifier at various classification thresholds. The curve is generated by plotting the true positive rate against the false positive rate at various thresholds, as depicted in Figure 17.

The scalar value of the Area Under the Curve (AUC) of the ROC curve serves as an

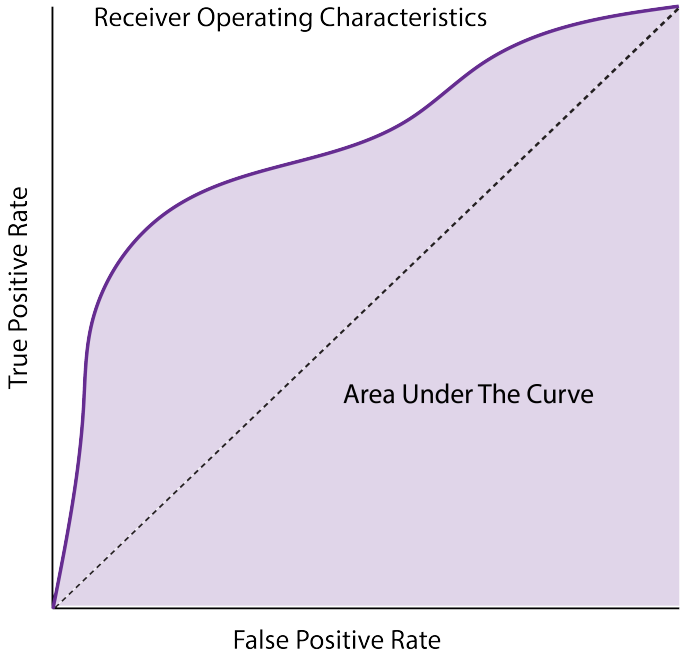


Figure 17: Illustration of how the ROC curve is presented[56]

overall indicator of the classifier's performance. It measures the classifier's capacity to differentiate between positive and negative classes. The AUC measures classification accuracy and ranges from 0 to 1, with higher values indicating better performance. An AUC value of 0.5, denoted by the dotted diagonal line, indicates a random guessing scenario. On the other hand, an AUC value of 1.0 signifies a perfect classifier.

To obtain a more accurate and nuanced evaluation of the performance of an anomaly detection model, one can use metrics such as precision, recall, F1-score, and ROC curves. These measures can provide valuable insights into the model's ability to detect anomalies.

4 Results

In order to evaluate the efficacy of the models based on the chosen features, a single test case has been developed. During the test, all indicators are used. As previously discussed in the Methodology section, the highly unbalanced nature of the data set may lead to accuracy scores that are misleading in terms of the model's actual performance. In order to enhance the assessment, the evaluation metrics of precision, recall, and F1 scores are presented in conjunction with the ROC curve.

4.1 ROC Curves

As explained in the methodology, ROC curves are used to demonstrate a model's ability to detect anomalies. To assess the model's capacity to distinguish between two points, one can evaluate the true positive rate versus the false positive rate through plotting. A higher AUC score signifies a better ability of the model to distinguish between the classes. An AUC score of 0.5 indicates that the model's performance is equivalent to that of random guessing. This is depicted in the graph through the diagonal blue line. An ideal curve would ascend vertically along the y-axis, then tracing the function $y = 1$ until it intersects with the blue line denoting an AUC score of 1.

It is crucial to keep in mind that although the ROC curve offers valuable information regarding the model's ability to detect anomalies, a high AUC score does not necessarily ensure that the model is successful in detecting anomalies. The ROC curves for each model are depicted in Figures 18 to 23. The AUC score of each illustration can be found in the lower right-hand corner.

The AUC scores across the board are not outstanding. Some models, including the LOF model, demonstrate a performance that is marginally superior to chance, registering a score of 0.56. The models that demonstrate the best performance are OCSVM, with a score of 0.76, and iForest, with a score of 0.72. The KNN algorithm demonstrated above-average performance with a score of 0.69, whereas both Artificial Neural Network (ANN) and Autoencoder (AE) delivered a score of 0.61. Despite the fact that none of these scores are particularly strong, it is noteworthy that a score of 0.76 or 0.72 is significantly superior to a random model.

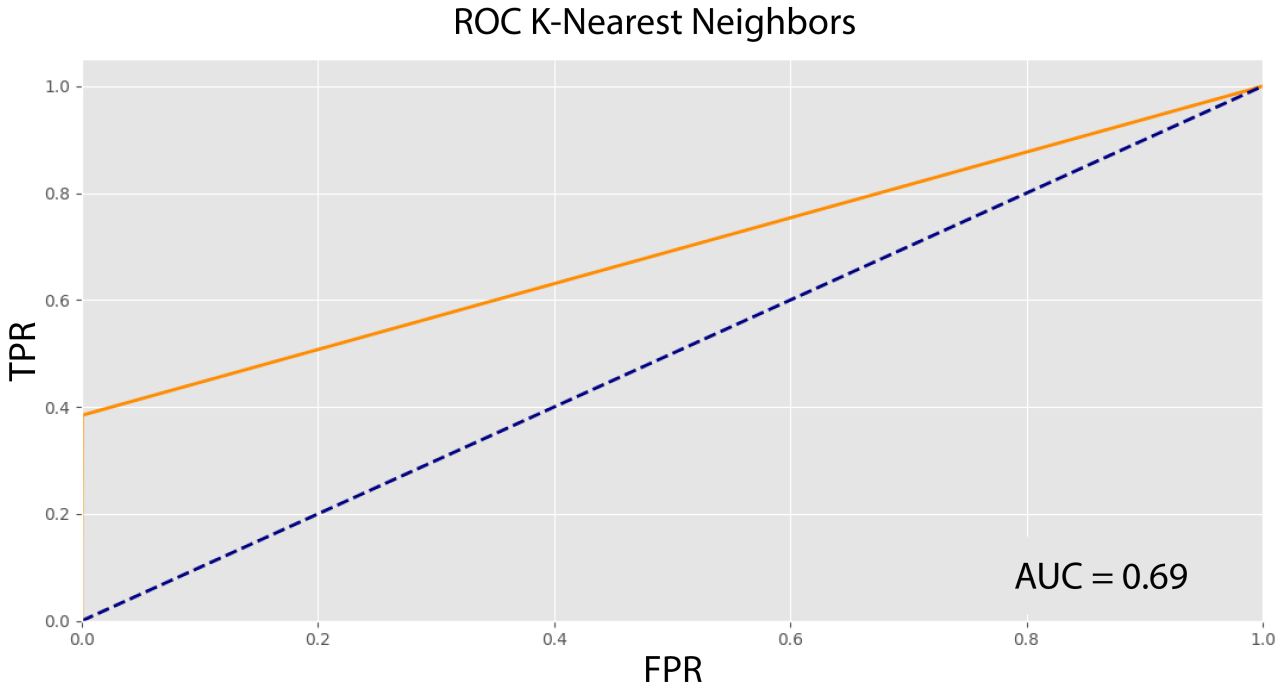


Figure 18: ROC curve of K-Nearest Neighbors model.

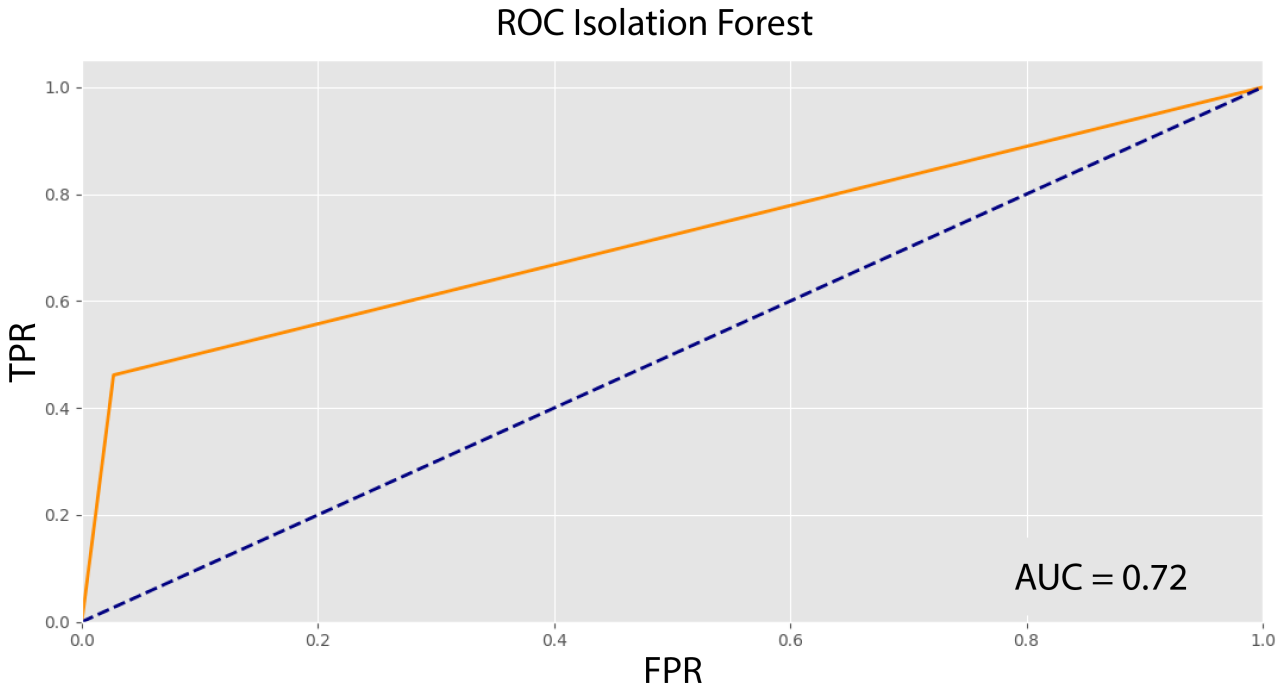


Figure 19: ROC curve of Isolation Forest model.

Figure 20: ROC curve of One-class SVM model.

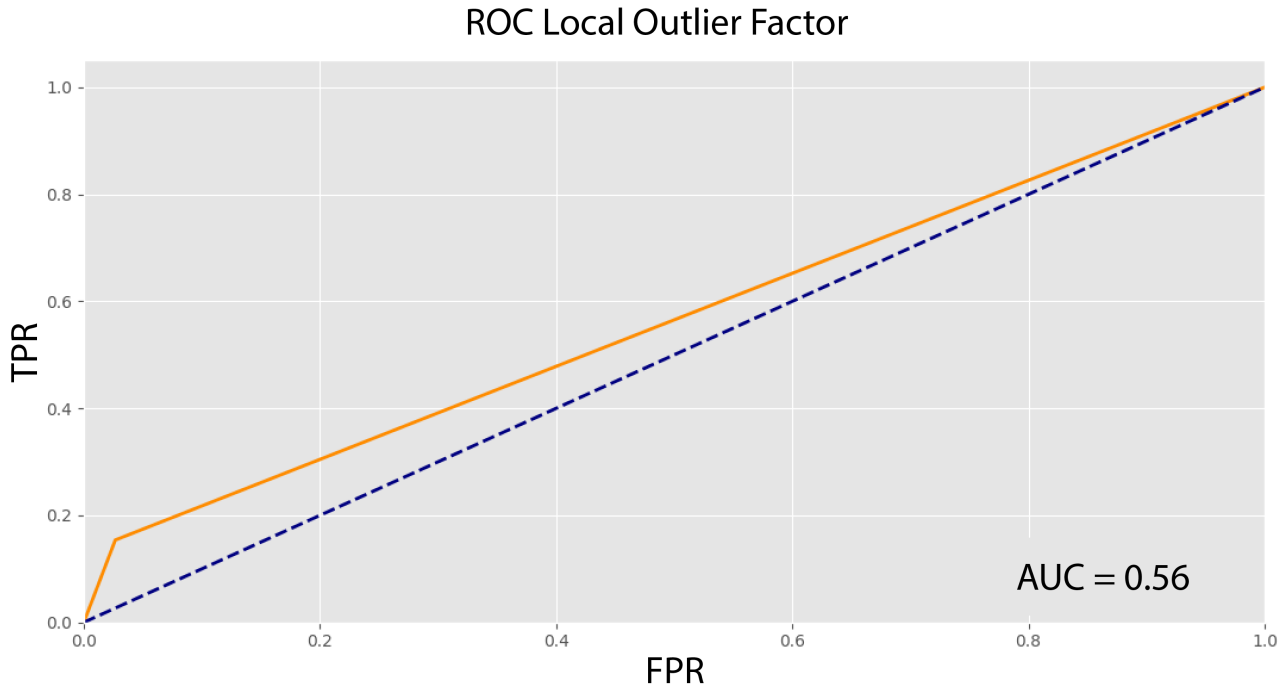


Figure 21: ROC curve of Local Outlier Factor model.

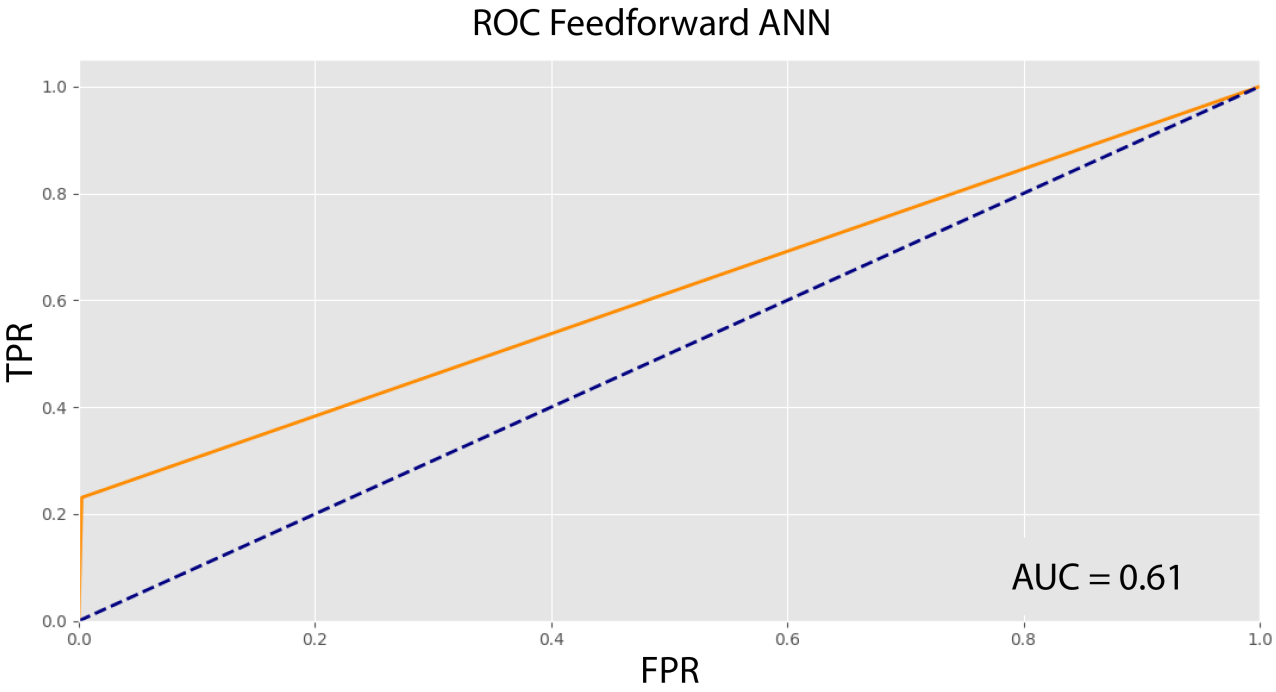


Figure 22: ROC curve of Feedforward ANN model.

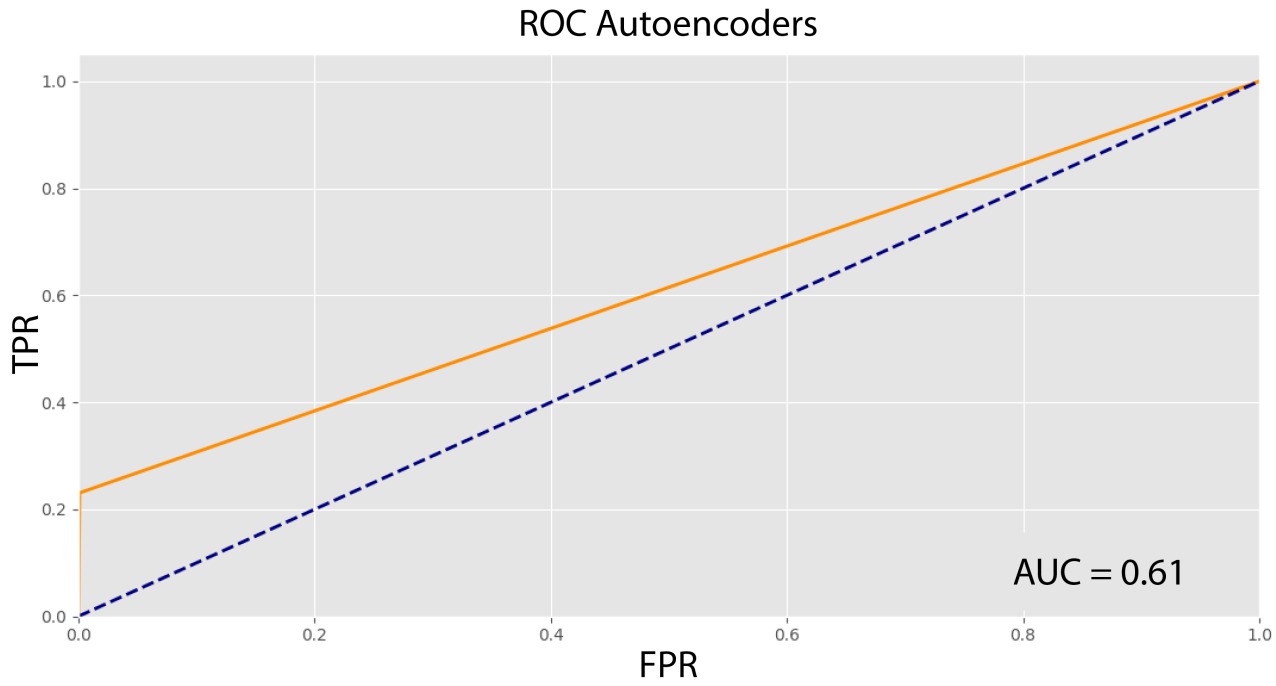


Figure 23: ROC curve of Autoencoders model.

4.2 Performance Scores

The following tables list the performance scores. Given that this is a binary classification problem, it is possible to partition the scores into two distinct classes. Specifically, Class 0 denotes the non-anomalous points, while Class 1 corresponds to the anomalous points. Table 6 has the performance scores of Class 0, whereas Table 7 presents the performance scores of Class 1.

Algorithm	**Precision**	**Recall**	**F1-score**
K-nearest neighbors	1	1	1
Local outlier factor	1	0.97	0.99
Isolation forest	1	0.97	0.99
One-class support vector machines	1	0.98	0.99
Artificial neural networks	1	1	1
Autoencoders	1	1	1

Table 6: Classification report on the models for class 0.

In Class 0, the performance scores are consistently close to 1, which indicates excellent performance in this class. It is anticipated because the majority of data points fall within this category, making it easier for the model to identify them. It is critical to keep in mind that a drop of even one percent in this class denotes the presence of over a thousand incorrectly identified points. The models that use LOF, OCSVM, and iForest demonstrate a slightly lower performance in comparison to the other models that achieve a flawless score of 1 for both precision and recall. However, obtaining an F1-score of 0.99 denotes a near-perfect score.

Algorithm	**Precision**	**Recall**	**F1-score**
K-nearest neighbors	0.67	0.15	0.25
Local outlier factor	0	0.15	0
Isolation forest	0	0.46	0
One-class support vector machines	0	0.54	0.1
Artificial neural networks	0.20	0.15	0.17
Autoencoders	0.03	0.23	0.05

Table 7: Classification report on the models for class 1

.

Performance scores in Class 1 are frequently lower than those in Class 0. It is not surprising that this result has been produced given that Class 1 has a limited number of points. It is worth noting that even a single misidentified point can have a major impact on the overall score in this class. The most important metric in this class is recall. This particular metric denotes the number of anomalies that have been accurately

recognized as such. The models employing KNN, LOF, and ANN demonstrate minimal efficacy, as they only identify 15% of the anomalies. With a rate of 23%, AE is ranked second-lowest. The OCSVM and iForest algorithms have recall scores of 54% and 46%, respectively, which places them at the top.

The precision score for class 1 is variable, and determining its value in this class is not simple. Precision is the measure of the accuracy of identifying anomalies, specifically the ratio of correctly identified anomalies to all the detected anomalies. Nonetheless, due to the uneven distribution of the data set, the score faces a significant drop in cases where additional non-anomalous points are incorrectly categorized. The precision score for AE for class 1 is almost 0, while the recall score for AE for class 0 is perfect, serving as an illustration of this. While there are enough misclassified points in class 1 that could affect the precision score, the number of misclassified points is not enough to influence the recall score of class 0. It is interesting to mention that KNN and ANN models demonstrate the highest precision scores for class 1, with respective values of 0.67 and 0.20. This suggests that although these models only detected a few anomalies, they demonstrated higher confidence in those particular findings.

The OCSVM model demonstrates the best performance by attaining the highest AUC and recall score when combining all observations. The performance shown by iTrees is comparable in terms of both AUC and recall score. Although these two models demonstrate a high degree of efficacy in detecting anomalies, they also have a tendency to misclassify a small proportion of non-anomalous data points. The KNN model is notable for its accurate performance, as it ranks third in terms of AUC score, and even though the recall score may have been below-average, the precision score was significantly higher.

Another notable aspect of the results is the length of the running time. Compared to other algorithms, the OCSV algorithm demonstrated a considerably longer processing time. The ANN and Autoencoder AE models demonstrated an interminable duration of execution. The iTree model demonstrated the fastest running time, whereas the KNN and LOF models also exhibited good computational time.

5 Discussion and Conclusion

In the context of hydropower plants, there has been a historical reliance on planned periodic maintenance. However, as we see significant transformations taking place within the energy market, there is a concurrent shift in the landscape of energy production. The main focus is shifting from methods that prioritize equipment to those that prioritize profitability. This shift poses a new obstacle: forecasting the point at which parts will deteriorate due to usage. Predictive maintenance involves evaluating the equipment's condition and trends, whereas scheduled maintenance has a set time frame. Failures do not often happen without any prior cause or warning. Typically, they advance from a state of normality to a state of collapse characterized by various symptoms, culminating in eventual failure. It is therefore critical to assess and predict the actual conditions and trends throughout this deterioration process, taking appropriate maintenance measures before any failure occurs. This is exactly the objective of preventive maintenance.

Even small errors can quickly become significant ones. Components of power plants are typically custom-made, making immediate off-the-shelf replacements impractical. As an example, repairing a malfunctioning transformer component can take more than twelve months. In the event of a severe generator failure, the entire unit may need to be lifted off the power plant's roof and transported to a different location for repairs. In instances of extreme severity, it may take several years to fully restore the power plant's operations. Detecting small errors early on can prevent a considerable number of incidents from becoming more serious issues. Anomaly detection is a crucial aspect to consider in this context. By employing diverse methods, one can identify slight alterations in different areas within a power facility. Such hidden patterns might otherwise go unnoticed by a limited number of measuring instruments or the human eye.

This thesis delved into the field of predictive maintenance and analytics, in specific anomaly detection in the hydropower domain. The research involved a review of relevant literature and an examination of present trends. A test case was set up using data from the winding temperature and hydrogen concentration of the power transformer, along with active power. Six distinct models for detecting anomalies were created using various ML algorithms. These models were fine-tuned using a validation set and then evaluated on a separate test set.

The presented results indicate that some of the models face challenges in accurately predicting all anomalies. However, they all exhibit greater performance compared to random chance. One of the main reasons for this challenge is the significant imbalance present in the data set. Several measures were implemented, including the stratification of the training and test sets as well as the inclusion of a validation set for hyperparameter tuning. However, additional tactics could be taken into account. There are more strategies that can be employed, such as oversampling the minor class,

utilizing combined algorithms, and extracting time series features.

Detecting anomalies involves more than simply applying ML algorithms to data sets. The nature of the data can greatly benefit from statistical approaches as a part of data preprocessing. The data obtained from hydroelectric power plants presents a challenge due to the large amount of normal condition data and the limited availability of non-anomalous data. It is challenging to establish the pattern of an incident before it occurs due to the complexity and rarity of potential errors.

Time series data, particularly those related to a specific domain such as hydroelectric power production, contain trends and patterns that can be taken advantage of. In Norway, a country with well-defined seasons, production tends to follow a certain pattern. As previously described for the Duge power plant, when electricity prices are low and magazines are filled, turbines can be reversed to pump water, or production may cease for a longer period of time. Another example is that during the winter season, it may be necessary to increase the temperature of the water cooling system in the generator to avoid the possibility of a cold start. Time series analysis can uncover various patterns that may contribute to an overall trend. Employing these techniques can aid in extracting additional features from the data set and contribute to achieving a more balanced distribution of points by minimizing the input.

Nevertheless, some of the models were capable of performing to some degree. The OCSVM and iForest models were able to detect roughly half of the anomalies, which is a reasonable outcome considering the preliminary nature of the study. When examining the scores for class 0, it was observed that OCSVM and iForest exhibited the lowest level of accuracy among the models. It is necessary to assess what holds the highest significance. In certain situations, it may be preferable to tolerate a higher number of misclassified points as anomalies if it results in the detection of more anomalies. However, if this number becomes excessively high, it would become impractical to track which points are truly anomalous. One clear distinction between these two models is their respective running times. iForest is considerably faster than OCSVM, which is an important factor to consider when selecting a model. Especially when iForest nearly matches the performance of OCSVM. However, this variable may not be significant if the input is altered as proposed through the utilization of time series analysis.

Many studies in the literature review utilize multiple algorithms by integrating them into a single, more comprehensive model. The KNN algorithm demonstrated superior precision compared to the OCSVM and iForest algorithms. By combining certain models, there is a possibility of improving performance. For example, OCSVM has the capability to establish a threshold during prediction, which is not a characteristic of KNN. OCSVM could potentially derive advantages from utilizing KNN's data clustering capabilities.

The study's use of a single test case posed limitations, which must be addressed. It is possible that these models could produce different outcomes for other data sets. Despite being trained, fine-tuned, and evaluated on distinct data partitions, this alone does not ensure their robustness.

Additionally, the analysis indicated that relying solely on the winding temperature and hydrogen concentration of the transformer may not be adequate for a comprehensive evaluation of its condition. This is probably a contributor to why the models failed to identify all anomalies. Nevertheless, the test case demonstrates that it is possible to detect anomalies using these features, and it highlights the need for further research.

The implementation of predictive maintenance in the hydropower sector is still in its early stages. Although the learning process may be challenging, the potential benefits are significant. Efforts to optimize processes, increase production efficiency, and prolong equipment lifespans are key motivators for further exploration and adoption of these technologies. Although the technology is still in its early stages, the literature demonstrates its potential.

The hydropower sector should collaborate to establish a comprehensive database of various types of malfunctions. It is necessary to support initiatives such as Smartkraft and MonitorX that work toward this objective. As production habits evolve, maintenance practices should also adapt accordingly. It is possible that predictive maintenance may emerge as the leading approach to maintenance. The adoption of condition monitoring, predictive maintenance, and anomaly detection is widespread in other industries and should be incorporated into hydroelectric power generation. The ongoing energy crisis necessitates inventive solutions. It is imperative to not only increase production but also optimize and streamline the process in every possible way.

6 Further Work

This thesis demonstrates initial research on predictive maintenance and anomaly detection in hydropower plants using ML and AI algorithms, with a test case on a hydropower transformer. There are multiple opportunities for future research that can expand upon the groundwork established in this study.

Firstly, it is beneficial to explore an even broader spectrum of ML and AI models. This study involved the construction and testing of six models. Nonetheless, there is a wide range of algorithms in the fields of ML and AI that could potentially provide better outcomes or reveal novel perspectives on the issue being addressed. These models may provide alternative viewpoints and be useful in enhancing the dependability and efficacy of predictive maintenance strategies.

Furthermore, this thesis suggests investigating the extraction of time series features to detect anomalies. This has the potential to reveal patterns that the current models might fail to detect. The utilization of time series analysis has demonstrated its efficacy in detecting trends, cycles, and other patterns in data that may suggest anomalies, particularly in intricate systems such as hydropower plants.

Additionally, it would be highly valuable to explore how sensitive the machine learning models are to various categories of incidents. An effective approach would be to develop targeted test cases that concentrate on a particular case or cases related to the same section of the transformer. By conducting a thorough analysis of the model's performance in particular scenarios, one can enhance its ability to detect various anomalies and improve its overall sensitivity.

Finally, it is recommended that the models showing potential in this study undergo further testing in diverse scenarios. Combining these models can enhance the system's reliability and robustness in detecting anomalies. Furthermore, it is imperative to assess the efficacy of these models in comparison with other data sets. Gaining a more complete comprehension of their performance in various situations would enhance the validation of their practicality and efficacy in all scenarios.

References

[1] Ali Reza Abbasi. *Fault detection and diagnosis in power transformers: a comprehensive review and classification of publications and methods*. Aug. 2022. DOI: 10.1016/j.epsr.2022.107990.

[2] Mennatallah Amer, Markus Goldstein, and Slim Abdennadher. "Enhancing one-class Support Vector Machines for unsupervised anomaly detection". In: *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description, ODD 2013*. 2013, pp. 8–15. ISBN: 9781450323352. DOI: 10.1145/2500853.2500857.

[3] *anomali - vitenskapsfilosofi – Store norske leksikon*. URL: https://snl.no/.search?query=anomali+-+vitenskapsfilosofi.

[4] S. Anoop and N. Naufal. "Monitoring of winding temperature in power transformers - A study". In: *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies, ICICICT 2017*. Vol. 2018-January. Institute of Electrical and Electronics Engineers Inc., Apr. 2018, pp. 1334–1337. ISBN: 9781509061068. DOI: 10.1109/ICICICT1.2017.8342763.

[5] S. Anoop et al. "Thermal stress monitoring and pre-fault detection system in power transformers using fibre optic technology". In: *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies, ICICICT 2017*. Vol. 2018-January. Institute of Electrical and Electronics Engineers Inc., Apr. 2018, pp. 886–891. ISBN: 9781509061068. DOI: 10.1109/ICICICT1.2017.8342682.

[6] Alessandro Betti et al. "Condition monitoring and predictive maintenance methodologies for hydropower plants equipment". In: *Renewable Energy* 171 (June 2021), pp. 246–253. ISSN: 18790682. DOI: 10.1016/j.renene.2021.02.102.

[7] Bjørn Tore Furnes. *Sviktmodell for krafttransformatorer*. Tech. rep. 2016.

[8] Lasse Brekke. "Teknisk-økonomisk analyse av reinvesteringsbehov i vannkraftverk". In: *195* (2015). URL: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2368111.

[9] Lasse Brekke. *Teknisk-økonomisk analyse av reinvesteringsbehov i vannkraftverk*. Tech. rep.

[10] Vilde Brennmoen, Supervisor Ümit, and Cali Co-Supervisor. *Predictive Maintenance Hydropower Generator*. Tech. rep. 2022.

[11] Isara Buaphan and Suttichai Premrudeepreechacharn. "Development of expert system for fault diagnosis of an 8-MW bulb turbine downstream irrigation hydro power plant". In: *2017 6th International Youth Conference on Energy, IYCE 2017*. Institute of Electrical and Electronics Engineers Inc., Aug. 2017. ISBN: 9781509064090. DOI: 10.1109/IYCE.2017.8003740.

[12] Umit Cali et al. "Digitalization of Power Markets and Systems Using Energy Informatics". In: *Digitalization of Power Markets and Systems Using Energy Informatics* (2021). DOI: 10.1007/978-3-030-83301-5.

[13] Pablo Calvo-Bascones, Miguel A. Sanz-Bobi, and Thomas M. Welte. "Anomaly detection method based on the deep knowledge behind behavior patterns in industrial components. Application to a hydropower plant". In: *Computers in Industry* 125 (Feb. 2021). ISSN: 01663615. DOI: 10.1016/j.compind.2020.103376.

[14] Candice Hudson. *Optimizing Operations and Maintenance with Predictive Analytics - Schneider Electric Blog*. 2015. URL: https://blog.se.com/industry/machine-and-process-management/2015/07/17/optimizing-operations-maintenance-predictive-analytics/.

[15] Ryan Carlson, Josh Hancox, and Wade Johnson. *CONSIDERATIONS FOR UPGRADING AND REPLACING TRANSFORMERS AT HYDROELECTRIC GENERATION FACILITIES*. Tech. rep. 2020.

[16] Yu Chen et al. "Fault anomaly detection of synchronous machine winding based on isolation forest and impulse frequency response analysis". In: *Measurement: Journal of the International Measurement Confederation* 188 (Jan. 2022). ISSN: 02632241. DOI: 10.1016/j.measurement.2021.110531.

[17] Circuit Globe. *What is Power System? Definition & Structure of Power System*. 2022. URL: https://circuitglobe.com/power-system.html.

[18] Kapil Dev et al. *Failure Mode and Effect Analysis (FMEA) Implementation: A Literature Review Bottling of Biogas-A Renewable Approach View project*. Tech. rep. 2018. URL: https://www.researchgate.net/publication/333209894.

[19] Zachary DeVries et al. "Using a national surgical database to predict complications following posterior lumbar surgery and comparing the area under the curve and F1-score for the assessment of prognostic capability". In: *Spine Journal* 21.7 (July 2021), pp. 1135–1142. ISSN: 18781632. DOI: 10.1016/j.spinee.2021.02.007.

[20] *Digitalisering av energisektoren Et mulighetsrom*. Tech. rep.

[21] Ibrahim Dincer and Haris Ishaq. "Hydro Energy-Based Hydrogen Production". In: *Renewable Hydrogen Production* (2022), pp. 191–218. DOI: 10.1016/B978-0-323-85176-3.00012-3.

[22] *Duge kraftverk - Sira-Kvina kraftselskap*. URL: https://www.sirakvina.no/duge-kraftverk/duge-kraftverk-article256-922.html.

[23] *Executive summary – Hydropower Special Market Report – Analysis - IEA*. URL: https://www.iea.org/reports/hydropower-special-market-report/executive-summary.

[24] Chuang Fu et al. "Predictive maintenance in intelligent-control-maintenance-management system for hydroelectric generating unit". In: *IEEE Transactions on Energy Conversion* 19.1 (Mar. 2004), pp. 179–186. ISSN: 08858969. DOI: 10.1109/TEC.2003.816600.

[25] Shuyuan Gan, Zhifang Song, and Lei Zhang. "A maintenance strategy based on system reliability considering imperfect corrective maintenance and shocks". In: *Computers & Industrial Engineering* 164 (Feb. 2022), p. 107886. ISSN: 0360-8352. DOI: 10.1016/J.CIE.2021.107886.

[26] J.H. Harlow. *Electric Power Transformer Engineering*. CRC Press, 2017.

[27] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. "The Elements of Statistical Learning Data Mining, Inference, and Prediction". In: (2009).

[28] *HRV | virakkraft.com*. URL: https://virakkraft.com/hrv.

[29] *Hydrogen (H2) in Power Transformers – Power Transformer Health*. URL: https://powertransformerhealth.com/2022/09/06/hydrogen-h2-in-power-transformers/.

[30] Naimul Islam et al. "Power transformer health condition evaluation: A deep generative model aided intelligent framework". In: *Electric Power Systems Research* 218 (May 2023). ISSN: 03787796. DOI: 10.1016/J.EPSR.2023.109201.

[31] Trevor A Kletz. *Hazop-past and future*. Tech. rep. 1997, pp. 263–266.

[32] Krishna Kumar and R. P. Saini. "A review on operation and maintenance of hydropower plants". In: *Sustainable Energy Technologies and Assessments* 49 (Feb. 2022). ISSN: 22131388. DOI: 10.1016/j.seta.2021.101704.

[33] H D Kuna, R García-Martinez, and F R Villatoro. "Outlier detection in audit logs for application systems". In: (2014). DOI: 10.1016/j.is.2014.03.001. URL: http://dx.doi.org/10.1016/j.is.2014.03.001.

[34] Pengzhi Li, Yan Pei, and Jianqiang Li. "A comprehensive survey on design and application of autoencoder in deep learning". In: *Applied Soft Computing* 138 (2023), p. 110176. DOI: 10.1016/j.asoc.2023.110176. URL: https://doi.org/10.1016/j.asoc.2023.110176.

[35] Yike Liu. "Noise reduction by vector median filtering". In: *Geophysics* 78.3 (2013), pp. V79–V86. ISSN: 19422156. DOI: 10.1190/GEO2012-0232.1.

[36] Marco Peixeiro. *The Complete Guide to Time Series Analysis and Forecasting | by Marco Peixeiro | Towards Data Science*. 2019. URL: https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775.

[37] *Matplotlib — Visualization with Python*. URL: https://matplotlib.org/.

[38] Merethe Ruud. *Fem spørsmål og svar om digital transformasjon - Tu.no*. 2022. URL: https://www.tu.no/artikler/fem-sporsmal-og-svar-om-digital-transformasjon/518601.

[39] Michael Matzer. *So deckt der Local Outlier Factor Anomalien auf*. 2019. URL: https://www.bigdata-insider.de/so-deckt-der-local-outlier-factor-anomalien-auf-a-803652/.

[40] *Milestones: 1969–1976 - Office of the Historian*. URL: https://history.state.gov/milestones/1969-1976/oil-embargo.

[41] R. Keith Mobley. "Benefits of Predictive Maintenance". In: *An Introduction to Predictive Maintenance* (Jan. 2002), pp. 60–73. DOI: 10.1016/B978-075067531-4/50004-X.

[42] R. Keith Mobley. "Financial Implications and Cost Justification". In: *An Introduction to Predictive Maintenance* (2002), pp. 23–42. DOI: 10.1016/B978-075067531-4/50002-6.

[43] R. Keith Mobley. "Impact of Maintenance". In: *An Introduction to Predictive Maintenance* (2002), pp. 1–22. DOI: 10.1016/B978-075067531-4/50001-4.

[44] R. Keith Mobley. "Predictive Maintenance Techniques". In: *An Introduction to Predictive Maintenance* (Jan. 2002), pp. 99–113. DOI: 10.1016/B978-075067531-4/50006-3.

[45] *MonitorX*. URL: https://www.sintef.no/projectweb/monitorx/.

[46] *Nord Pool*. URL: https://www.nordpoolgroup.com/.

[47] *NumPy*. URL: https://numpy.org/.

[48] Pallavi Pandey. *Outlier Detection using Isolation Forests – Machine Learning Geek*. 2020. URL: https://machinelearninggeek.com/outlier-detection-using-isolation-forests/.

[49] Liliane Pintelon and Alejandro Parodi-Herz. "Maintenance: An Evolutionary Perspective". In: *Springer Series in Reliability Engineering*. Vol. 8. Springer London, 2008, pp. 21–48. DOI: 10.1007/978-1-84800-011-7{_}2.

[50] Johannes Pöppelbaum, Gavneet Singh Chadha, and Andreas Schwung. "Contrastive learning based self-supervised time-series analysis". In: *Applied Soft Computing* 117 (Mar. 2022). ISSN: 15684946. DOI: 10.1016/J.ASOC.2021.108397.

[51] *Predictive Maintenance vs. Predictive Analytics, What's the Difference? - AUTOMATION INSIGHTS*. 2022. URL: https://automation-insights.blog/2022/07/06/predictive-maintenance-vs-predictive-analytics-whats-the-difference/.

[52] *Produkter - ISY JobTech - Norconsult Informasjonssystemer*. URL: https://www.nois.no/produkter/fdv-og-eiendomsforvaltning/isy-jobtech/.

[53] *PyCharm: the Python IDE for Professional Developers by JetBrains*. URL: https://www.jetbrains.com/pycharm/.

[54] *Python Release Python 3.9.0 | Python.org*. URL: https://www.python.org/downloads/release/python-390/.

[55] Sandi Ritlop. *Using Biodegradable Lubricants*. 2008. URL: https://www.hydroreview.com/world-regions/using-biodegradable-lubricants/#gref.

[56] Sarang Narkhede. *Understanding AUC - ROC Curve | by Sarang Narkhede | Towards Data Science*. 2018. URL: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[57] Iqbal H. Sarker. "Deep Cybersecurity: A Comprehensive Overview from Neural Network and Deep Learning Perspective". In: *SN Computer Science* 2.3 (May 2021). ISSN: 26618907. DOI: 10.1007/S42979-021-00535-6.

[58] *scikit-learn: machine learning in Python — scikit-learn 1.2.2 documentation*. URL: https://scikit-learn.org/stable/.

[59] Luka Selak, Peter Butala, and Alojzij Sluga. "Condition monitoring and fault diagnostics for hydropower plants". In: *Computers in Industry* 65.6 (2014), pp. 924–936. ISSN: 01663615. DOI: 10.1016/j.compind.2014.02.006.

[60] Serafeim Loukas. *What is Machine Learning: Supervised, Unsupervised, Semi-Supervised and Reinforcement learning methods | by Serafeim Loukas, PhD | Towards Data Science*. 2020. URL: https://towardsdatascience.com/what-is-machine-learning-a-short-note-on-supervised-unsupervised-semi-supervised-and-aed1573ae9bb.

[61] Rahul Shah. *GridSearchCV |Tune Hyperparameters with GridSearchCV*. 2022. URL: https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/.

[62] *SmartKraft – Støttet prosjekt | Enova*. URL: https://www.enova.no/om-enova/om-organisasjonen/teknologiportefoljen/smartkraft/.

[63] Rahul Soni and Bhinal Mehta. "Review on asset management of power transformer by diagnosing incipient faults and faults identification using various testing methodologies". In: *Engineering Failure Analysis* 128 (Oct. 2021). ISSN: 13506307. DOI: 10.1016/J.ENGFAILANAL.2021.105634.

[64] *Standardization vs Normalization. Distinguishing between two common... | by Aashish Nair | Towards Data Science*. 2022. URL: https://towardsdatascience.com/standardization-vs-normalization-dc81f23085e3.

[65] Angelina Steffens. "Efficient Methods for Handling Missing Data". In: (1994).

[66] *Strømnettet - Energifakta Norge*. URL: https://energifaktanorge.no/norsk-energiforsyning/kraftnett/.

[67] *Summary for Policymakers — Global Warming of 1.5 ºC*. URL: https://www.ipcc.ch/sr15/chapter/spm/.

[68] M. A. Taghikhani and A. Gholami. "Prediction of hottest spot temperature in power transformer windings with non-directed and directed oil-forced cooling". In: *International Journal of Electrical Power and Energy Systems* 31.7-8 (Sept. 2009), pp. 356–364. ISSN: 01420615. DOI: 10.1016/J.IJEPES.2009.03.009.

[69] *TensorFlow*. URL: https://www.tensorflow.org/.

[70] Thomas Leypoldt Marthinsen. *Hva er vannkraft og hvordan produseres det? | Tjenestetorget*. 2023. URL: https://tjenestetorget.no/blogg/vannkraft.

[71] Srikanth Thudumu et al. "A comprehensive survey of anomaly detection techniques for high dimensional big data". In: *Journal of Big Data* 7.1 (Dec. 2020). ISSN: 21961115. DOI: 10.1186/s40537-020-00320-x.

[72] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation Forest". In: ().

[73] *UNITED STATES DEPARTMENT OF THE INTERIOR BUREAU OF RECLAMATION*. Tech. rep. 2003.

[74] Jorge M. Uribe, Stephanía Mosquera-López, and Montserrat Guillen. "Characterizing electricity market integration in Nord Pool". In: *Energy* 208 (Oct. 2020). ISSN: 03605442. DOI: 10.1016/J.ENERGY.2020.118368.

[75] *User Guide — pandas 2.0.1 documentation*. URL: https://pandas.pydata.org/docs/user_guide/index.html.

[76] *vannkraftverk – Store norske leksikon*. URL: https://snl.no/vannkraftverk.

[77] *vannturbin – Store norske leksikon*. URL: https://snl.no/vannturbin.

[78] Hongteng Wang et al. "Anomaly detection for hydropower turbine unit based on variational modal decomposition and deep autoencoder". In: *Energy Reports* 7 (Nov. 2021), pp. 938–946. ISSN: 23524847. DOI: 10.1016/j.egyr.2021.09.179.

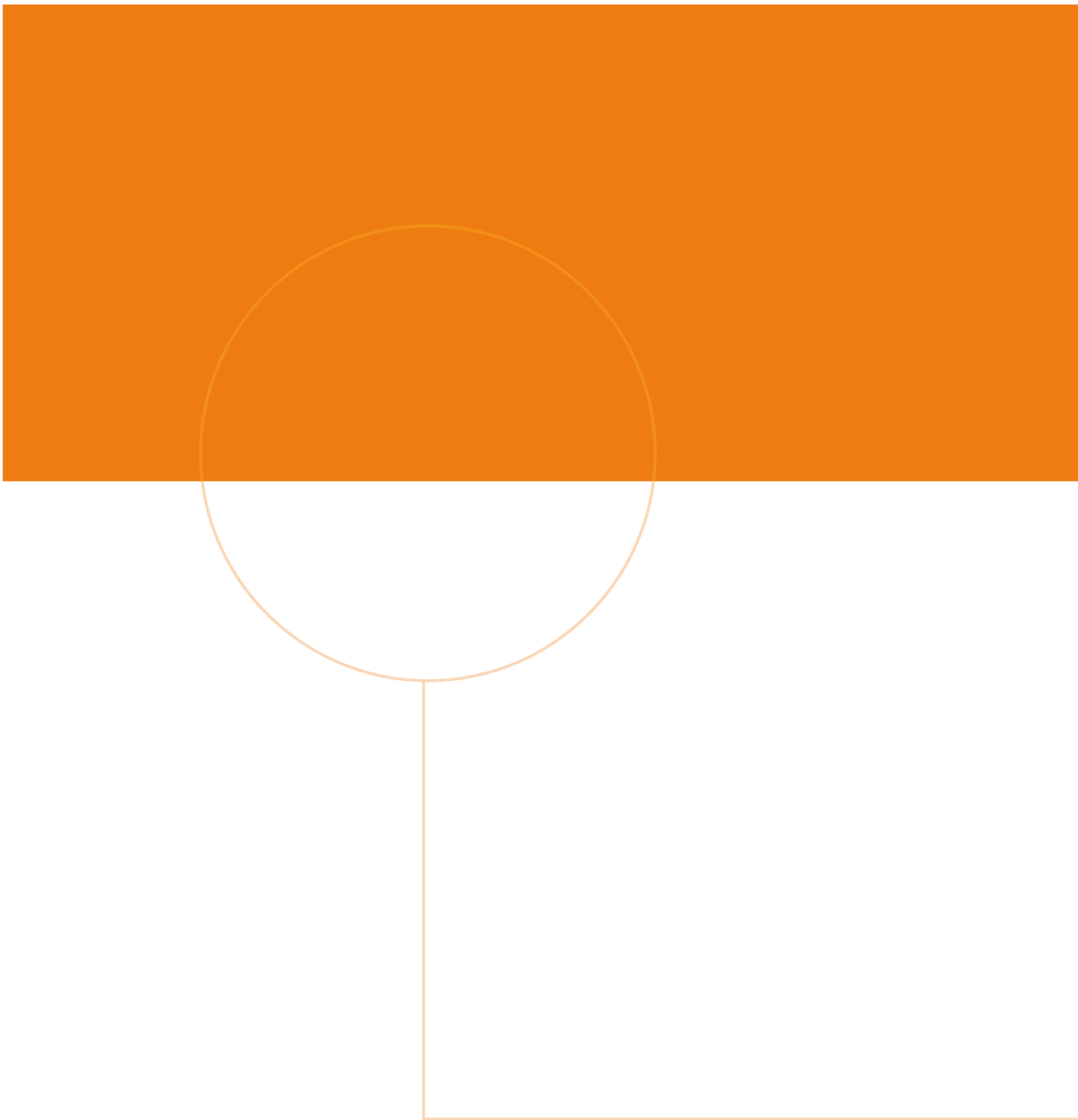
[79] *What Is Anomaly Detection? | RapidMiner*. URL: https://rapidminer.com/glossary/anomaly-detection/.

[80] Zhikai Xing and Yigang He. "Multi-modal information analysis for fault diagnosis with time-series data from power transformer". In: *International Journal of Electrical Power and Energy Systems* 144 (Jan. 2023). ISSN: 01420615. DOI: 10.1016/j.ijepes.2022.108567.

[81] Zhikai Xing et al. "Health evaluation of power transformer using deep learning neural network". In: *Electric Power Systems Research* 215 (Feb. 2023). ISSN: 03787796. DOI: 10.1016/J.EPSR.2022.109016.

[82] Guixin Zhang, Zhenlei Wang, and Hua Mei. "Sensitivity clustering and ROC curve based alarm threshold optimization". In: *Process Safety and Environmental Protection* 141 (Sept. 2020), pp. 83–94. ISSN: 09575820. DOI: 10.1016/j.psep.2020.03.029.

[83] Dexu Zou et al. “Outlier detection and data filling based on KNN and LOF for power transformer operation data classification”. In: *Energy Reports* 9 (Sept. 2023), pp. 698–711. ISSN: 23524847. DOI: 10.1016/J.EGYR.2023.04.094.



NTNU

Kunnskap for en bedre verden