NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

OIL AND GAS FLOW ANOMALY DETECTION ON OFFSHORE NATURALLY FLOWING WELLS USING DEEP NEURAL NETWORKS

Guzel Bayazitova

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

OIL AND GAS FLOW ANOMALY DETECTION ON OFFSHORE NATURALLY FLOWING WELLS USING DEEP NEURAL NETWORKS

by

Guzel Bayazitova

Dissertation presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Data Science

Supervisor: Vitor Duarte dos Santos

Supervisor*:* Maria Anastasiadou

June 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Guzel Bayazitova

Lisbon, June 2023

ABSTRACT

The Oil and Gas industry, as never before, faces multiple challenges. It is being impugned for being dirty, a pollutant, and hence the more demand for green alternatives. Nevertheless, the world still has to rely heavily on hydrocarbons, since it is the most traditional and stable source of energy, as opposed to extensively promoted hydro, solar or wind power. Major operators are challenged to produce the oil more efficiently, to counteract the newly arising energy sources, with less of a climate footprint, more scrutinized expenditure, thus facing high skepticism regarding its future. It has to become greener, and hence to act in a manner not required previously.

While most of the tools used by the Hydrocarbon E&P industry is expensive and has been used for many years, it is paramount for the industry's survival and prosperity to apply predictive maintenance technologies, that would foresee potential failures, making production safer, lowering downtime, increasing productivity and diminishing maintenance costs. Many efforts were applied in order to define the most accurate and effective predictive methods, however data scarcity affects the speed and capacity for further experimentations. Whilst it would be highly beneficial for the industry to invest in Artificial Intelligence, this research aims at exploring, in depth, the subject of Anomaly Detection, using the open public data from Petrobras, that was developed by experts.

For this research the Deep Learning Neural Networks, such as Recurrent Neural Networks with LSTM and GRU backbones, were implemented for multi-class classification of undesirable events on naturally flowing wells. Further, several hyperparameter optimization tools were explored, mainly focusing on Genetic Algorithms as being the most advanced methods for such kind of tasks.

The research concluded with the best performing algorithm with 2 stacked GRU and the following vector of hyperparameters weights: [1, 47, 40, 14], which stand for timestep 1, number of hidden units 47, number of epochs 40 and batch size 14, producing F1 equal to 0.97%.

As the world faces many issues, one of which is the detrimental effect of heavy industries to the environment and as result adverse global climate change, this project is an attempt to contribute to the field of applying Artificial Intelligence in the Oil and Gas industry, with the intention to make it more efficient, transparent and sustainable.

KEYWORDS

Anomaly Detection; Multivariate Time Series Classification; Deep Neural Network; Oil and Gas; Genetic Algorithm.

Sustainable Development Goals (SGD):

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ADA	Adaboost
DEAP	Distributed Evolutionary Algorithms in Python
DL	Deep Learning
DT	Decision Tree
GA	Genetic Algorithm
GB	Gradient Boosting
GRU	Gated Recurrent Unit
KNN	K-Nearest Neighbor
LSTM	Long Short-Term memory
LSTM-AE	Long Short-Term memory Autoencoder
ML	Machine Learning
MLP	Multi-Layer Perceptron
NBC	Naïve Bayes Classifier
QDA	Quadratic Discriminant Analysis
RF	Random Forest
RNN	Recurrent Neural Network
SHAP	Shapley Additive Explanations
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
XAI	Explainable Artificial Intelligence
XGBoost	eXtreme Gradient Boosting

1. INTRODUCTION

Artificial Intelligence (AI) has revolutionized the perspectives of industries and businesses, creating value by learning from data, accumulating knowledge from patterns and trends, simulating human logic and automating decision making. It has been used in many industries, such as finance, smart cities, healthcare, cybersecurity, education, criminal justice, etc. The domain knowledge and expertise were the main pillars of decision making, however AI brings huge benefits by automating many processes, thereby improving accuracy and further augmenting human intelligence.

The Oil and Gas industry has been somewhat relatively slow in applying AI to many parts of the E&P Life allowing for a reduction in costs and risks (KUANG et al., 2021) and thereby converting the industry into greener version of itself. Nowadays, AI has entered into all its branches, creating many "intelligent" versions of the sectors, such as intelligent drilling, intelligent development, intelligent exploration, intelligent production, etc. (Sircar et al., 2021). There is still huge potential for further AI development within the industry, such that the E&P industry can reach the point, where many industries enjoy the full scale of revolution 4.0.

1.1. CONTEXT AND PROBLEM IDENTIFICATION

The Petroleum industry is the oldest and the most reliable source of energy, that emerged after coal and kerosine in the late 19th century. Once discovered, it became highly popular and an extremely desired commodity, initializing a surge in oil discoveries and starting an "oil rush" worldwide. As shown in the Figure 1.1, it developed into three main branches:

- Upstream – oil and gas exploration, development and production (E&P)
- Midstream – their transportation and storage
- Downstream – refining, product marketing and retail.

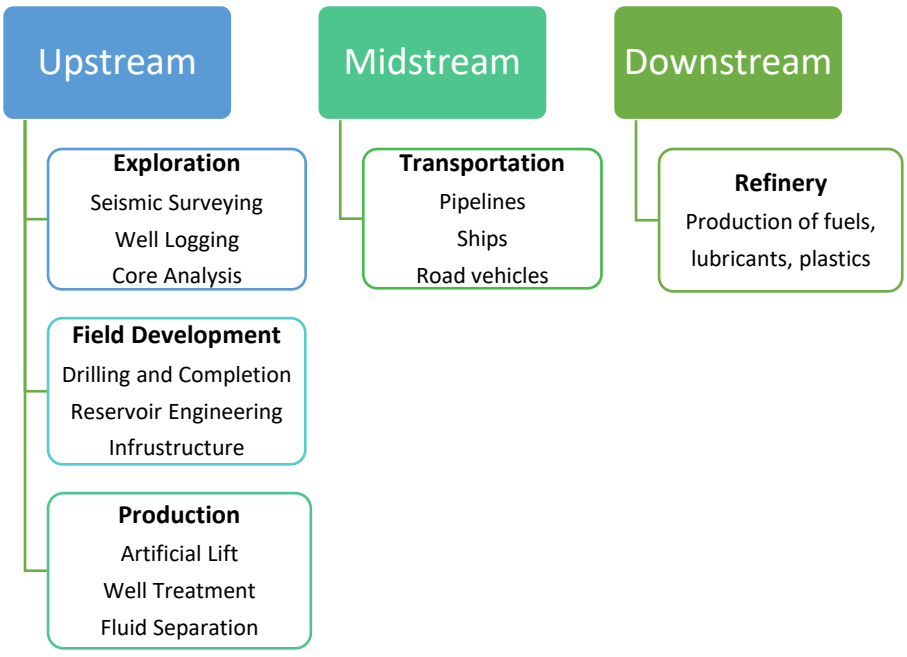


Figure 1.1 - Division of the oil and gas industry into sectors (adapted from Koroteev & Tekic, 2021).

While each of the segments carries its own function, they are all part of a heavy industry, that handles flows of highly hazardous and inflammable materials, which move at high rates, high temperatures, and pressure (Al-Anzi et al., 2022). Such complicated systems require many sensors to control each process, with additional monitoring of the data flow to identify any anomalies thereby allowing for preventive action to be taken. Since there are thousands of sensors on each platform, refinery, pipelines, other tools or machinery, that produce continuous data, the problem turns into a Big Data problem, that requires development of technologies to analyze a massive amount of data (Martí et al., 2015).

There is still a huge lack of personnel, as a whole number of engineers moved out of the industry during the numerous oil industry crises, and as result less experts are available to apply the domain knowledge. This lower number of highly qualified engineers, that can be involved in the real time processes and in analyzing ginormous amounts of data produced by each oil platform, creates a new issue, which is a deficiency in expertise (Amy Chronis & Kate Hardin, 2022). The alarm system, informing rig personnel of a potential failure in early or transient stages, can effectively assist in timely decision making and allow the application of preventive steps. Considering the time window required for diagnosing the issue and the preventive actions that are necessary before the failure occurs, it is essential to spot the transient moment condition as soon as possible to prevent major losses (Marins et al., 2021).

Due to the demanding nature of the Oil and Gas industry, it requires consistent real time monitoring, that is performed by surveillance engineers. This is a formidable task, that has been in existence for many years (Hasan et al., 2017). In the Upstream sector, one engineer might be in charge of not just one well, but multiple within one field, monitoring many aspects of the production from these wells. Considering the limited time window available between start of the issue and the actual failure, an anomaly detection in real time can alleviate the surveillance task to a great extent. In the Downstream sector there is a possibility to perform real time monitoring remotely through Distributed Control Systems (DCS) and Supervisory Control and Data Acquisition (SCADA) systems, measuring an array of variables from the sensors (Athar Khodabakhsh et al., 2018). Yet it is not sufficient for full reliance on the huge data flow, which requires consistent attention from a human being for anomaly identification.

Nowadays the entire production sector of the Petroleum business can be automated, putting the entire cycle of crude oil extraction, transportation and refining under the control of Artificial Intelligence. The Downstream sector has been automated since 1990 with implementation of digital process control networks, as result event-based scheduling and planning software for hydrocarbon movements are quite mature in this part of the petroleum industry (Blancett et al., 2019). However, it can still benefit from AI by implementing new models for predictive maintenance and drone examination of equipment which is inaccessible to personnel. As for the Midstream sector, it is also relatively mature in its AI employment, with software optimizing and implementing actions on loading quantities and transportation routes for each distribution network (Blancett et al., 2019). It can also be further improved by supply chain automation and advancing transmission lines between oil rigs, refineries and final points of petrochemical and fuel sales. Further, surveillance drones and machine-vision algorithms could also diminish the necessity for human's presence and daily monitoring.

According to the McKinsey report a leading offshore oil and gas operator's predictive maintenance system helped to reduce downtime by 20% and increase production by more than 500 barrels of oil annually (Guillaume Decaix et al., 2021). It is not surprising that Artificial Intelligence and Digital Transformation are gaining more popularity within all sectors of the Oil and Gas industry. Schlumberger, which is a top high-tech service company, developed Automated Drilling Solutions, in collaboration with National Oilwell Varco, where oilfield domain knowledge is supported by advanced machine learning applications. This allowed automation of the drilling process to make it safer and more efficient (Schlumberger, 2021). Kongsberg Digital who specializes in the digitalization of the Oil and Gas and maritime industry, have built digital twins and whole ecosystems to advance the drilling process even further (Kongsberg Digital, 2022).

Once Artificial Intelligence was introduced to the Oil and Gas industry, surveillance of numerous processes has undergone through major improvements by combining business or operational intelligence with automated technical calculations (Hasan et al., 2017). The real-time monitoring centers were established to pass the sensor data to remote displays, and advanced Artificial Intelligence models were applied to provide data monitoring and improved control. The introduction of the industrial revolution 4.0 created the new era of industry development by integrating the Internet of Things (IoT), automating cloud computing, real-time data analysis, etc. (Aslam et al., 2022).

The Upstream sector is the most capital-intensive and important sector out of the three industry segments (Koroteev & Tekic, 2021). However, it is the least automated in many aspects, due to the nature of the activity, being often performed in harsh conditions, such as deep-water, in arid deserts, arctic colds, extreme wind, etc. Drilling is a high risk and high capital expenditure operation, that involves fire and explosion risks, operations with radioactive sources, the threat of gas leaks, the movement of personnel by unconventional means of transport, such as helicopters and supply boats, which offer many opportunities for human error due to complexities of the operations, etc. It is characterized by high level of uncertainty due to the unpredictability of the processes, which need to be handled manually, and thus relies heavily on the experts' knowledge (Koroteev & Tekic, 2021).

Nowadays many Artificial Intelligence methods and technologies are incorporated into Upstream branch to resolve its demands. In order to make the Oil and Gas industry more intelligent-based, major operators created many collaborations with IT companies, such as Total and Google Cloud, Chevron and Microsoft, Shell and HP, ExxonMobil and MIT, etc. (KUANG et al., 2021), as shown in Table 1.1.

Table 1.1 - Comparison of AI strategies among global key oil and gas companies and service companies (adapted from KUANG et al., 2021).

No	Companies	Orientation	AI Platform	Partners
1	BP	Horizontal well trajectory control, drilling data processing algorithm	Sandy	Beyond Limits, Belmont technology
2	Shell	Horizontal well trajectory control, drilling data processing algorithm	Geodesic	Microsoft
3	Exxon Mobil	Data collection and integrated solutions	XTO	Microsoft
4	Total	Intelligent solution for E&P,	Cloud Platform	Google

		intelligent seismic imaging processing		
5	Chevron	E&P, storage & transportation projects	DELFI	Microsoft, Schlumberger
6	Schlumberger	E&P, storage & transportation projects	DELFI	Microsoft, Chevron
7	Baker Hughes	Seismic modeling, malfunction prediction and supply chain optimization	Desktop Platform, Azure	NVIDIA, Microsoft
8	Halliburton	Reservoir characterization and simulation	Azure	Microsoft
9	PetroChina	Intelligent basins, intelligent logging, intelligent geophysical exploration, intelligent drilling & completion, intelligent oil production, intelligent fracturing and intelligent equipment	Dream Cloud Platform, Cognitive Computing Platform	Huawei
10	Sinopec	Intelligent factories, intelligent oilfields and intelligent institutes	Oilfield Smart Cloud Industrial Internet Platform	Ali
11	CNOOC	Intelligent oilfields, E&D data management	Intelligent Oilfield Technology Platform	Ali

For well surveillance a new hybrid solution was generated by integrating management by exception (MBE), Business Intelligence (BI) and situational awareness (SA) (Hasan et al., 2017). This MBE method equipped engineers with well control and optimization techniques, that identify anomalies as a deviation from the expected data pattern, allowing personnel to focus just on the challenging wells or processes. Personnel no longer need to drive to the oilfield for visual observation, since it can be done by drones or other visual automation mechanisms, thus decreasing fuel expenditure and expensive human time.

Condition-Based Monitoring (CBE) is another popular approach, a state-of-art technology, in which equipment and machinery is being monitored, involving data-driven analysis to foresee and detect potential failure (Brønstad et al., n.d.). It is applied throughout all the Oil and Gas industry with success, creating systems for anomaly detection and maintenance.

Overall maintenance strategies can be divided into proactive and reactive types, as depicted in Figure 1.2. The conventional Preventive Maintenance, based on failure history analysis and planned scheduled maintenance, can be costly and unworthy. Predictive Maintenance can be performed after the continuous equipment condition monitoring has revealed a high potential for a failure or deteriorating performance (Al-Anzi et al., 2022b).

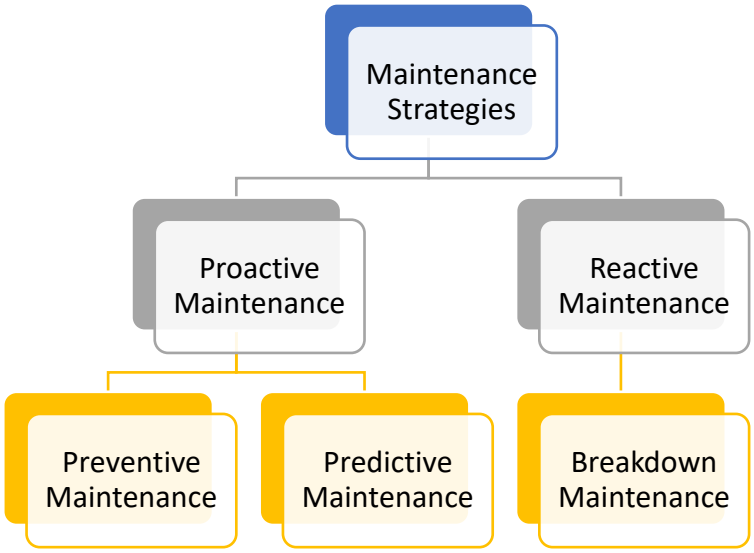


Figure 1.2 - Maintenance strategies (adapted from Al-Anzi et al., 2022b)

Digital Oilfield is a new concept, symbolizing a collection of automation and information technologies, that revolutionizes the way the petroleum industry operates and allows for work processes to be conducted more efficiently (Pandey et al., 2020). It involves data management, automation, integrated production models, predictive maintenance, etc. – a combination of all the emerging technologies that assist in timely reaction and decision making.

According to the United Nation's "The sustainable Development Goals Report 2022", an increased dependence on natural resources exacerbates the pressure on sensitive ecosystems and ultimately affects both human health and the economy(United Nations, 2022). Considering the high global greenhouses gas emissions, increased global average temperature and resultant extreme weather events, it is important to focus on fossil fuel production efficiency, minimizing operating costs, make informed decisions and provide equipment maintenance at a right time, and to eliminate potential failures and catastrophic events. This research will contribute to the application of the Artificial Intelligence in the Oil and Gas industry, with the aim of mitigating its adverse climate effect, whilst enhancing its safety and sustainability.

1.2. OBJECTIVES

Most decisions in the Upstream sector are based on expert knowledge, rather than on the enormous amount of field data, due to high uncertainty of the application, which might result in biased conclusions (Koroteev & Tekic, 2021). An example is geophysical and petrophysical interpretation of seismic surveys, that would enable to produce a reservoir geological model, predict its productivity and allocate locations of appraisal and production wells. This process might take more than a year, and automation using Artificial Intelligence can significantly speed up parts of it and make it more objective.

Drilling itself faces multiple challenges such as shock and vibrations, bit wear, loss of circulation, drill-pipes washout, borehole instability, excessive torque, etc. (Sircar et al., 2021). AI methods were applied to optimize drilling parameters, identify lithology and directional drilling, predict potential tool failure and downtimes.

The same applies to reservoir modelling (Figure 1.3), the process is lengthy and cumbersome, requiring calculations of oil flows via various reservoir development scenarios. The questions related to developing new assets or investing in production enhancement and technologies, can be answered quicker and be expert-independent, with more AI involvement in the process (Koroteev & Tekic, 2021).

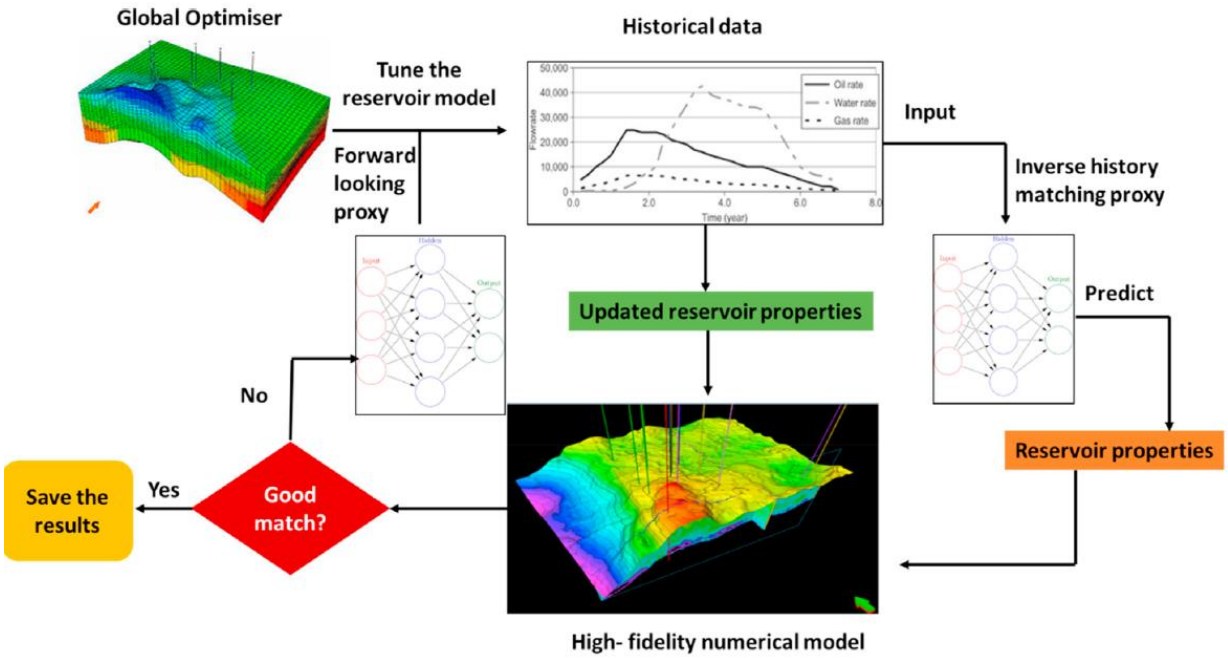


Figure 1.3 - Reservoir modelling outline using artificial neural network (Sircar et al., 2021).

A particular challenge for the Oil and Gas industry is the lack and availability of open data (Vargas et al., 2019a), which hinders further research and AI application advancement. One of the reasons is the confidentiality of high-cost information, whilst another is the difficulty in recognizing and labeling all potential unlikely events from the available data (Soriano-Vargas et al., 2021).

Another issue is the absence of an author's active up to date network, who performed the research in the subject matter. It slows any advancement due to the fact that most papers were not produced by academic researches, but company appointed professionals, applying AI to specific core activities and using proprietary data, which does not encourage further networking (D'Almeida et al., 2022). The lack of collaboration between oil companies, perceiving each other as a competitor, does not help in AI advancement within the petroleum industry (Koroteev & Tekic, 2021). Most companies tend to follow a strategy of developing their own AI projects, without experience and knowledge sharing.

A real Petrobras 3W dataset will be used for this research, which has instances of eight types of undesirable events characterized by eight process variables (Vargas et al., 2019a). It is a unique dataset of naturally flowing oil wells, for which industry experts spend extensive amount of time to validate historical events and produce simulated and hand-drawn instances, that are useful to counterpart data imbalance, under different operating conditions. This research objective is to contribute to the Artificial Intelligence development in the Oil and Gas Upstream area, promoting openness and collaboration between industry players and data scientist, and creating a precedent of adding value for the future safe hydrocarbon energy.

1.3. STUDY IMPORTANCE AND RELEVANCE

The Upstream sector, in comparison to Midstream and Downstream, has caused most of the fatal failures and devastating environmental effects in the industry history.

One of the many disastrous events was the catastrophe of the Macondo well in 2010 due to safety equipment failure which resulted in the death of 11 rig personnel, 17 were also seriously injured, the sinking of the Deepwater Horizon rig, and caused a massive oil spill into the Gulf of Mexico for 87 days, damaging costal and marine environment by releasing 5 million barrels of oil (U.S. Chemical Safety Board, 2016).

There have been other multiple smaller scale events, which also caused enormous harm to the nature and ecosystems. In 2016 the Plugging and Abandonment (P&A) operations in the G-4 well of the Troll Field on the Norwegian Continental Shelf (NCS), performed by Statoil, failed and caused leakage of oil, gas and other flammable substances. Similarly, in 2012 the Elgin P&A operations resulted in an uncontrollable gas release to the seabed (Babaleye et al., 2019).

Less disastrous failures in the Production sector have also caused many smaller scale adverse impacts, such as oil spills, natural gas release, which affected wildlife and polluted nature. This section will cover the basics of oil production theory in order to highlight the importance of identifying potential anomalies and prevent failures.

At the commencement of production, most wells have sufficient pressure to produce oil without having to use pumps. As time progresses, the formation pressure starts to diminish, as less oil and more water are getting produced (Figure 1.4), natural oil lift methods become insufficient, and Artificial lift methods need to be implemented in order to maintain or enhance production (Pandey et al., 2020).

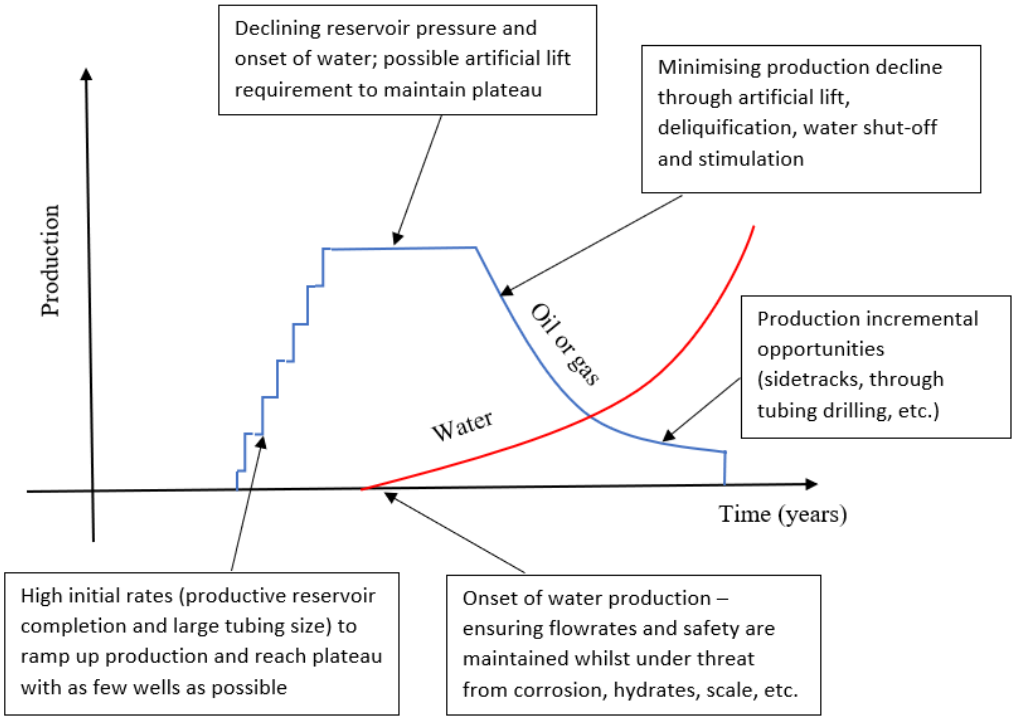


Figure 1.4 - Production stages (adapted from Bellarby, 2009)

Artificial lift is a method of adding energy to the flow of oil within the completion interval to increase the production rate (Bellarby, 2009).

Gas lift, Electric Submersible Pump (ESP), Turbine-Driven Submersible Pumps, Jet Pumps, Progressive Cavity Pump, Beam (Rod) Pump, Hydraulic Piston Pumps are the main methods used to maintain the flow of oil from the well (Bellarby, 2009). During production operations other well intervention techniques might be applied, such as replacement of completion parts, wellbore acid treatment, chemical inhibition, hydraulic fracturing, to name just a few (Cedric Malate, 2003). 90% of all producing oil wells need Artificial lift techniques and equipment to stimulate oil production, which creates very high levels of failures, resulting in downtime with high economical losses and harm to ecosystems (Pandey et al., 2020).

In this study we will focus on malfunctions within naturally flowing wells, since the availability of open data dictates the research. However, most wells require Artificial lift methods, hence more failures are observed due to decline of pumps or turbines. Yet, the selected types of the undesired events are responsible for most of the production losses in the last years (Vargas et al., 2019a) and are relevant for the production optimization.

The United Nations “17 Sustainable Development Goals” calls for a global partnership in collaboration for a better World, with less poverty, raised economic growth, tackled climate change, preserved nature, etc. (United Nations, 2022). The fossil fuels sector emits devastating amount of GHG, affects local ecosystems and environment with each oil spill, produced water and drilling waste discharge. The UN goals № 12 “Responsible consumption and production” and № 13 “Climate action” are the first two, that need to be addressed by the petroleum business, as it is one of the biggest contributors to the global climate change.

This research would deepen the knowledge of anomaly detection in the Oil and Gas industry in order to demonstrate the unlimited potential of Data Science application in the hydrocarbon and fossil fuel domain, and encourage all the industry key players to implement AI-based processes more extensively and eagerly. With a better understanding of timely decision making and the impact of predictive maintenance, the industry has the opportunity to contribute to the United Nation’s goals accomplishment, and to make our world a greener and safer place to be.

2. LITERATURE REVIEW

Artificial Intelligence and Internet of Things application in the Oil and Gas industry has spurred increased interest in the research of recent Machine Learning, Data Mining and Deep Learning models to resolve its every day demands. A growing number of publications in the last few years has created a necessity to obtain a better understanding about most popular algorithms and the latest developments in the subject.

This section is dedicated to the literature review of the methods, techniques and algorithms used for resolving Oil and Gas problems and detecting potential faults, failures and anomalies in the Upstream and Midstream sectors.

2.1. OIL AND GAS INDUSTRY

The Oil and Gas industry has a long history, originating from 1848, when the first modern well was drilled in the north-east of Baku on the Absheron Peninsula by Russian engineer Vasily Semenov. Further, the first commercial oil well was drilled in North America Pennsylvania by Erwin Drake in 1859.

The history and development milestones of both Oil and Gas industry is demonstrated in Figure 2.1, together with simultaneously developing Artificial Intelligence from its infancy as mostly statistical methods, to its birth in 1950 with British logician and computer pioneer Alan Turing inventing famous Turing's Learning Machine. Since then both sciences have developed in their own way, and just recently meeting to create a unique opportunity of integrating one into the other for a more efficient and safe practice.

The Oil and Gas industry used Artificial Intelligence and Digital Transformation for decades (Pandey et al., 2020). Many large companies invest now heavily in Research, Development and Innovation looking for opportunities to eliminate accidents, improve decision making, by implementing more actively digital transformation applications (D'Almeida et al., 2022b).

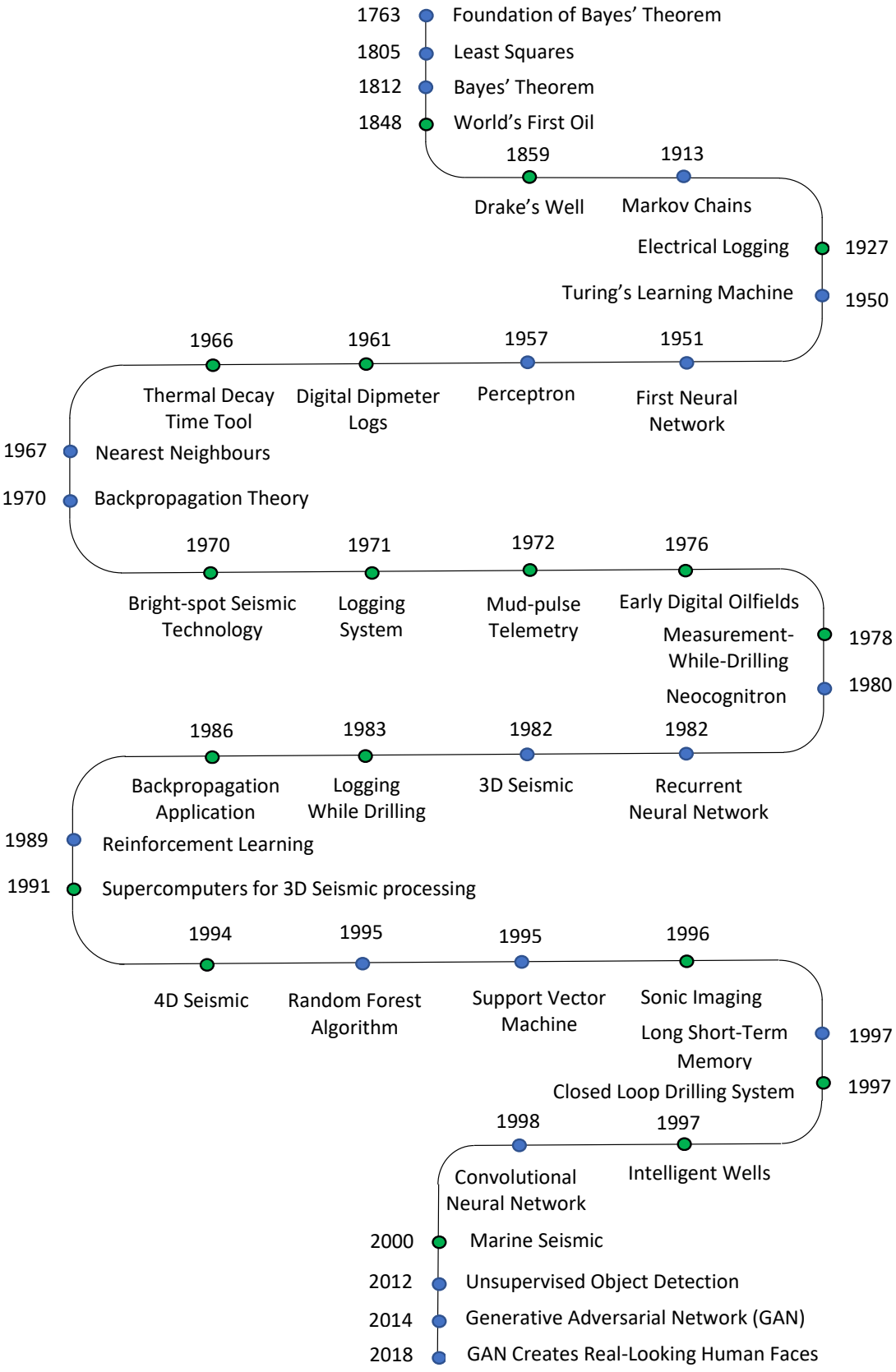


Figure 2.1 – Oil and Gas industry and Machine Learning milestones (adapted from Pandey et al., 2020).

2.2. ANOMALY DETECTION

Anomaly detection is a statistical technique used to identify abnormal patterns in data that deviate from a priori expected behavior (Martí et al., 2015b). It is being applied in many industries: manufacturing, aviation, transportation, banking, health, etc. Oil and Gas has joined recently the trend and started to take the opportunity of identifying anomalies in time and to improve general performance along with reducing potential downtime, minimizing costs and in some occasions saving lives.

2.3. ANOMALY DETECTION IN THE OIL AND GAS INDUSTRY WITH AI METHODS

The aim of the literature review is to explore the methods already applied in the area of the Oil and Gas industry for anomaly detection, using Machine Learning or Deep Learning techniques. The research would help to recognize the recent development and most used methods, that can be further explored and potentially improved in the future.

The framework of this analysis focuses on the Upstream and Midstream sectors of the petroleum industry, which involves Drilling and Exploration, Production and Transportation, where most of the failures occur.

2.3.1. Literature review methodology

The Systematic Literature Review (SLR) was performed following PRISMA methodology (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), which is well known method for establishing the state of knowledge in regards to certain topics. In order to identify and visualize the most significant keywords and terms related to anomaly detection using AI methods in the Petroleum industry, a VOSviewer (VOSviewer - Visualizing Scientific Landscapes, n.d.) , a bibliometric visualization tool was applied.

With intention of narrowing our research on the stated above agenda, we focus on the following questions:

- **RQ1**: What are the most applicable and significant Artificial Intelligence methods that were applied for detection of anomalies and potential failures in the Oil and Gas industry?

- **RQ2**: Which AI methods were applied for 3W dataset anomaly detection and classification?

The research consists of 3 steps: (1) Planning the review: PRISMA search strategy development and initial data selection from the scholar databases, (2) Conducting the review: selection of journals and conference proceedings according to inclusion and exclusion criteria, visualization and bibliometric analysis, (3) analysis of the findings, discussion and conclusion (Figure 2.2).

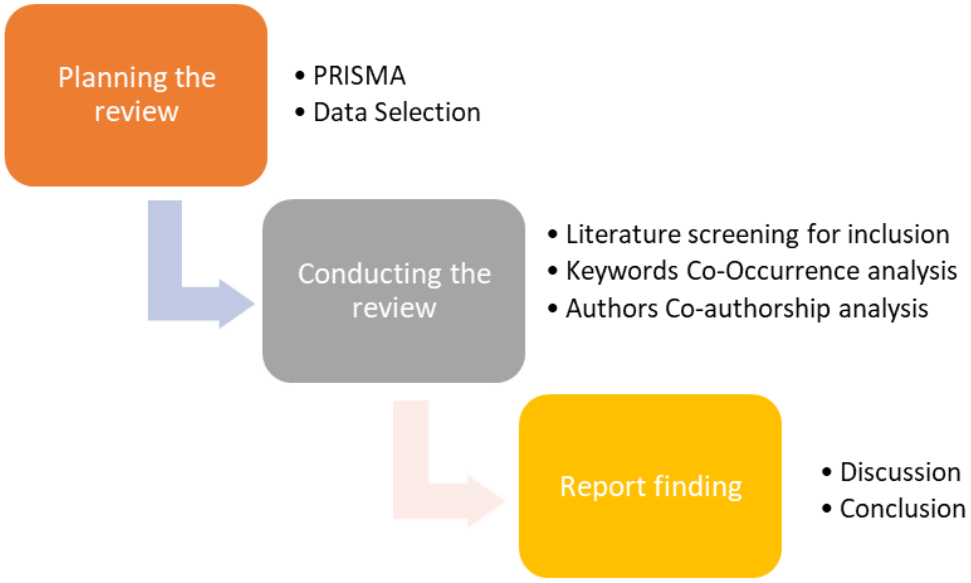


Figure 2.2 - Systematic Literature Review methodology

According to PRISMA guidelines, the data selection was initialized through the publications search, that contain in the titles, abstracts or keywords the following Boolean expressions:

> *("oil and gas" OR "oilfield" OR "oil wells" OR "naturally flowing wells") AND ("artificial intelligence" OR "neural networks" OR "machine learning" OR "deep learning" OR "anomaly" OR "fault" OR "failure" OR "detection" OR "prediction" OR "classification" OR "unsupervised")*

The databases Web of Science, Scopus and Science Direct were analyzed using the above query, on 11th of December 2022. Since the 3W Dataset was created in 2019, we select 2019-2022 time period to limit the research to the recent 4 years.

The OnePetro database, which collects journal articles and conference proceedings from the Society of Petroleum Engineers (SPE), wasn't used for this research, since it doesn't provide free access to its contents. It would be beneficial to include this database for further systematic literature review, since it might provide deeper insight about industry related artificial intelligence application for anomaly detection.

The selected pool of articles was analyzed using Mendeley, an open-source reference manager from Elsevier. The application allowed to perform further data processing, remove duplicates, pull metadata, such as authors, sources, date of publication, citations, etc.

2.3.2. PRISMA results

The quantitative and qualitative analysis was initialized by collecting data through 3 databases: Scopus (1394 results), ScienceDirect (752 results) and Web of Science (668 results). Overall, there were 2814 articles extracted for further analysis using PRISMA framework. General inclusion criteria for all databases were articles having Open Access, written in English, published in the 2019 – 2022 frame, document type Article or Conference Paper.

In order to limit the research to the most relevant material, there were further inclusion criteria applied for each database:

- **Web of Science:**

  Web of Science Index: *(Science Citation Index Expanded ((SCI-EXPANDED)*

  Web of Science Categories: *(Energy Fuels OR Geosciences Multidisciplinary OR Engineering Petroleum OR Geochemistry Geophysics OR Geology OR Computer Science Information Systems OR Remote Sensing OR Computer Science Artificial Intelligence OR Automation Control System)*

- **Scopus:**

  Scopus Subject Area: *(Earth and Planetary Sciences OR Engineering OR Energy OR Environmental Science OR Computer Science OR Decision Sciences)*

  Scopus Exact Source Title: *(Energies OR Journal Of Petroleum Exploration And Production Technology OR Frontiers In Earth Science OR Remote Sensing OR Geofluids OR IEEE Access Petroleum Exploration And Development OR Journal Of Petroleum Science And Engineering OR Marine And Petroleum Geology OR Energy Reports OR Geophysical Journal International OR Energy Science And Engineering OR Frontiers In Energy Research OR Advances In Geo Energy Research OR Applied Energy OR Geophysical Research Letters OR Energy Geoscience OR Geochemistry Geophysics Geosystems OR Petroleum Science OR Journal Of Petroleum Exploration And Production OR Reliability Engineering And System Safety OR Energy Exploration And Exploitation OR Lithosphere OR Natural Gas Industry B OR Oil And Gas Science And Technology OR Petroleum OR Petroleum Research OR Shock And Vibration OR Earth And Planetary Science Letters OR China Geology OR Journal Of The Geological Society OR Engineering Structures OR Journal Of Pipeline Science And Engineering OR Open Geosciences OR Journal Of Geophysics And Engineering OR Energy OR Exploration Geophysics OR Fuel OR Journal Of Geophysical Research Solid Earth OR Journal Of Loss Prevention In The Process Industries OR Geophysics OR Journal Of Natural Gas Science And Engineering OR Petroleum Geoscience OR Computational Intelligence And Neuroscience OR Geology OR Applied Computing And Geosciences OR Earth Sciences Research Journal OR International Journal Of Advanced Computer Science And Applications OR Natural Resources Research OR Process Safety And Environmental Protection OR Results In Engineering OR Energy Strategy Reviews OR IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing OR IEEE Transactions On Geoscience And Remote Sensing OR Computational Geosciences OR Computers And Geosciences OR Energy Engineering Journal Of The Association Of Energy Engineering)*

- **Science Direct:**

  Science Direct Publication Title: *(Journal of Petroleum Science and Engineering OR Energy Strategy Reviews OR International Journal of Applied Earth Observation and Geoinformation OR Applied Energy OR Safety Science OR Journal of Cleaner Production OR Journal of Natural Gas Science and Engineering OR Expert Systems with Applications OR Computers & Geosciences OR Energy Research & Social Science OR Energy OR Reliability Engineering & System Safety OR Marine and Petroleum Geology)*

Once the articles were extracted according to inclusion criteria, an additional 32 articles were added from other sources, which were directly related to the 3W dataset or published in other databases, such as OnePetro, Research Gate or from Oil and Gas conferences, focusing on anomalies detection in Upstream and Midstream sector.

The final set of material was checked for duplicates, and 729 duplicated articles were removed. Further, the titles were screened, and based on the relevance to the subject, 1509 records were excluded. Next, the abstracts were screened, and further 514 records were deleted, since they were not corresponding to the research agenda. As shown in Figure 2.3, 94 articles were selected for the full text screening, according to inclusion and exclusion criteria, from which 36 articles were removed. Finally, 58 articles were used for further analysis.

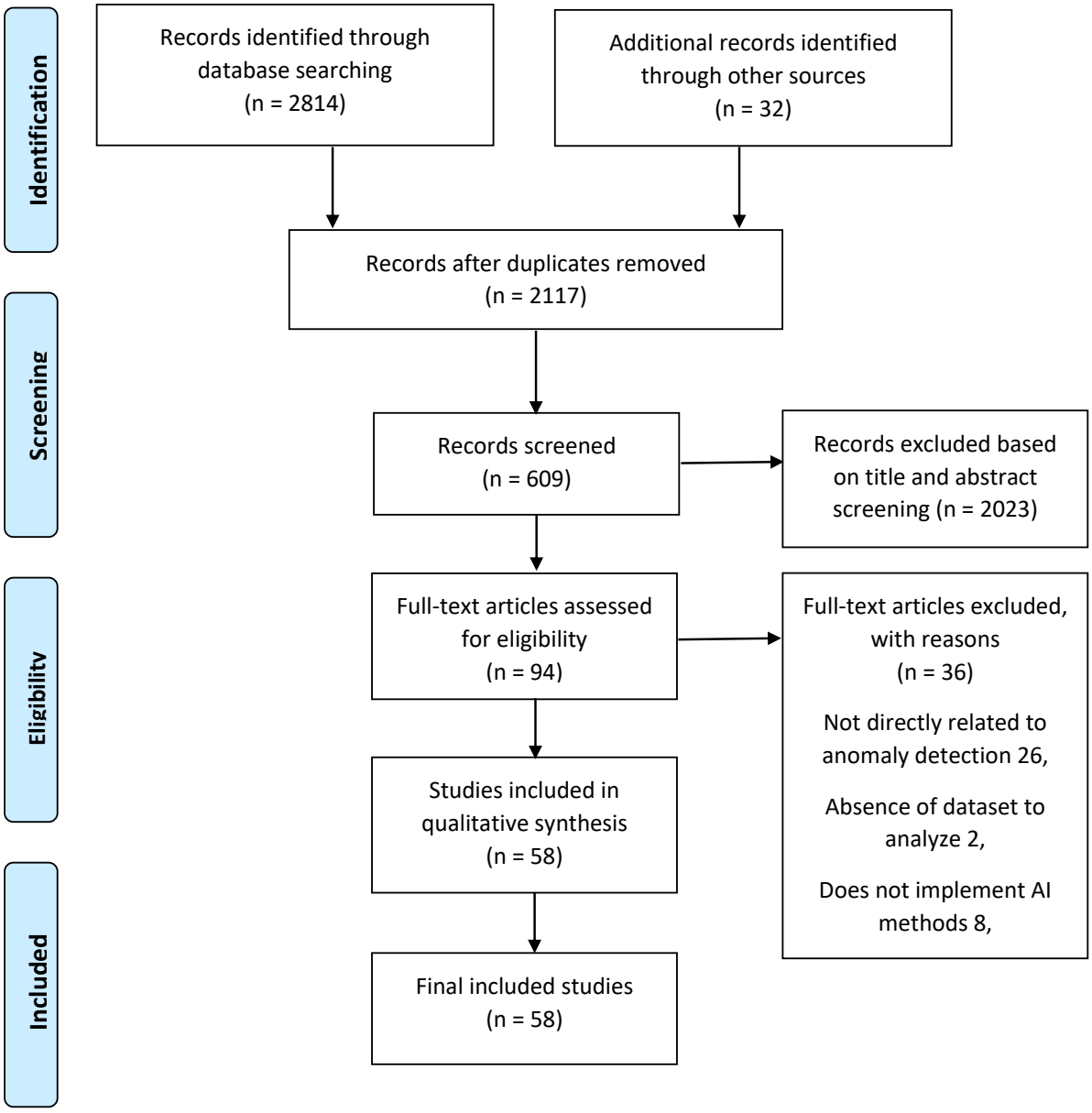


Figure 2.3 - PRISMA flowchart

2.3.3. Results and analysis

The analysis of the final 58 papers reveals that 44 were journal articles and 14 were proceedings from conferences.

As expressed in Table 2.1, the articles were published in 25 journals, majority of which were in Journal of Petroleum Science and Engineering (10). The rest are in IEEE Access (6), Energies (5), Journal of Petroleum Exploration and Production Technology (2), Petroleum Research (1), ACS Omega (1), Advances in Geo-Energy Research (1), American Journal of Operations Research (1), Applied Artificial Intelligence (1), Applied Computational Intelligence and Soft Computing (1), Applied Intelligence (1), Computation (1), Energy Engineering: Journal of the Association of Energy Engineering (1), Expert Systems (1), Frontiers in Earth Science (1), International Journal of Disaster Risk Science (1), International Journal of Greenhouse Gas Control (1), Journal of Applied Logic (1), Journal of Energy Resources Technology, Transactions of the ASME (1), Oil and Gas Science and Technology (1), Petroleum (1), Petroleum Science (1), Sensors (1), SPE Journal (1), Brazilian Journal of Development (BJD) (1).

The corresponding Scimago Ranks and further details are indicated in Table 2.1. Most of the journals are Q1-quartile ranked (7), the next majority are in Q2-quartile (6), and there are 5 cases, in which the journals have mixed ranking Q1 or Q2, depending on the subject area (5). The rest of the journals have Q3-quartile rank (2), Q2/Q3 (2), Q4 (1), and two are not classified (2).

Among the most covered subject areas are Energy (12), Engineering (10), Earth and Planetary Science (9), Computer Science (7) and Mathematics (4). The journals publishers are based mostly in the United States (7), Netherlands (4), Switzerland (4), China (3) and United Kingdom (2), with minority from Germany (1), Hong Kong (1), Egypt (1), France (1) and Brazil (1). The leading publishers are Multidisciplinary Digital Publishing Institute (MDPI) (3), Elsevier (2) and KeAi Communications Co. (2).

Table 2.1 - Journals details and their Scimago Ranks

Journals	Scimago Rank	Number of articles	Publisher	Country	Journal Subject Area
Journal of Petroleum Science and Engineering	Q1	10	Elsevier	Netherlands	Earth and Planetary Sciences, Energy
IEEE Access	Q1	6	Institute of Electrical and Electronics Engineers Inc.	United States	Computer Science, Engineering, Material Science
Energies	Q1/Q2	5	Multidisciplinary Digital Publishing Institute (MDPI)	Switzerland	Energy, Engineering, Mathematics
Journal of Petroleum Exploration and Production Technology	Q2	2	Springer Verlag	Germany	Earth and Planetary Sciences, Energy

Petroleum Research	Q2/Q3	1	KeAi Communications Co.	China	Earth and Planetary Sciences, Energy
ACS Omega	Q1	1	American Chemical Society	United States	Chemical Engineering, Chemistry
Advances in Geo-Energy Research	Q1	1	Yandy Scientific Press	Hong Kong	Earth and Planetary Sciences, Energy, Engineering
American Journal of Operations Research	n/a	1	Scientific Research Publishing	United States	Operations Research and Optimization Theory and Research Technical Approaches, Manufacturing and Service Operations Research, Interfaces with Other Disciplines
Applied Artificial Intelligence	Q3	1	Taylor and Francis Ltd.	United Kingdom	Computer Science
Applied Computational Intelligence and Soft Computing	Q2	1	Hindawi Limited	Egypt	Computer Science, Engineering
Applied Intelligence	Q2	1	Springer Netherlands	Netherlands	Computer Science
Computation	Q2/Q3	1	Multidisciplinary Digital Publishing Institute (MDPI)	Switzerland	Computer Science, Mathematics
Energy Engineering: Journal of the Association of Energy Engineering	Q4	1	Tech Science Press	United States	Energy, Engineering
Expert Systems	Q2	1	Wiley-Blackwell Publishing Ltd	United Kingdom	Computer Science, Engineering, Mathematics
Frontiers in Earth Science	Q1	1	Frontiers Media S.A.	Switzerland	Earth and Planetary Sciences

International Journal of Disaster Risk Science	Q1/Q2	1	Springer Science + Business Media	United States	Environmental Science, Social Sciences
International Journal of Greenhouse Gas Control	Q1	1	Elsevier	Netherlands	Energy, Engineering, Environmental Science
Journal of Applied Logic	Q3	1	Elsevier BV	Netherlands	Mathematics
Journal of Energy Resources Technology, Transactions of the ASME	Q2	1	The American Society of Mechanical Engineers (ASME)	United States	Earth and Planetary Sciences, Energy, Engineering
Oil and Gas Science and Technology	Q2	1	Editions Technip	France	Chemical Engineering, Energy
Petroleum	Q1/Q2	1	KeAi Communications Co.	China	Earth and Planetary Sciences, Energy
Petroleum Science	Q1/Q2	1	China University of Petroleum Beijing	China	Earth and Planetary Sciences, Energy
Sensors	Q1/Q2	1	Multidisciplinary Digital Publishing Institute (MDPI)	Switzerland	Biochemistry, Genetics and Molecular Biology, Chemistry, Computer Science, Engineering, Medicine, Physics and Astronomy
SPE Journal	Q1	1	Society of Petroleum Engineers (SPE)	United States	Earth and Planetary Sciences, Energy
Brazilian Journal of Development (BJD)	n/a	1	Brazilian Journals Publicações de Periódicos e Editora Ltda.	Brazil	Engineering, Biomedical and Clinical Studies, Education, Agricultural, Veterinary and Food Sciences, Language, Communication and Culture

Table 2.2 represents the details of conferences, from which 14 proceedings are related to the current research. Most of the publishers originate from United States (9), the rest are from Greece (1), United Kingdom (1), Slovakia (1), Spain (1) and Brazil (1). The major subjects of research are Engineering (8), Energy (3) and Computer Science (2).

Table 2.2 – Conference details

Conference	Number of proceedings	Publisher Country	Subject
SENSORDEVICES 2021: The Twelfth International Conference on Sensor Device Technologies and Applications (2021)	1	Greece	Sensor Devices
International Joint Conference on Neural Networks (IJCNN) (2020)	1	United Kingdom	Neural Networks
Society of Petroleum Engineers Western North American Regional Meeting 2010 - In Collaboration with the Joint Meetings of the Pacific Section AAPG and Cordilleran Section GSA (2010)	1	United States	Earth and Planetary Sciences
2021 23rd International Conference on Process Control (PC) (2021)	1	Slovakia	Process Control
2017 6th International Symposium on Advanced Control of Industrial Processes, AdCONIP (2017)	1	United States	Chemical Engineering, Engineering, Mathematics
IEEE International Conference on Data Mining, ICDM (2011)	1	United States	Engineering
IEEE International Symposium on Industrial Electronics (2021)	1	United States	Engineering
Annual Offshore Technology Conference (2019)	1	United States	Energy, Engineering
SPE Western Regional Meeting 2015: Old Horizons, New Horizons Through Enabling Technology (2015)	1	United States	Energy, Engineering
Proceedings of the International Joint Conference on Neural Networks (2020)	1	United States	Computer Science
IEEE 16th International Conference on Data Mining (ICDM) (2016)	1	Spain	Engineering
XLII Ibero-Latin-American Congress on Computational Methods in Engineering and III Pan-American Congress on Computational Mechanics, ABMEC-IACM (2021)	1	Brazil	Engineering
Offshore Technology Conference (2021)	1	United States	Energy, Engineering
2021 IEEE 19th International Conference on Industrial Informatics (INDIN) (2021)	1	United States	Computer Science

2.3.4. Keywords co-occurrence

The co-occurrence of keywords was performed using VOSviewer, a text mining software for creating maps based on network data. The analysis was done using the full counting method with two minimum number of keyword occurrences. Out of 150 terms, 26 met the threshold, which are listed in Table 2.3.

Table 2.3 - Keywords co-occurrence and the total link strengths

Keyword	Occurrences	Total link strength
Machine Learning	10	15
Fault Diagnosis	6	10
Classification	3	9
Oil Well Monitoring	3	8
Electrical Submersible Pump	2	7
Fault Detection	3	7
Metric Learning	2	7
Triplet Network	2	7
Fault Detection and Classification	3	6
Convolutional Neural Network	3	5
Dynamometer Card	3	5
Multivariate Time Series Classification	2	5
Pattern Recognition	2	5
Artificial Intelligence	3	4
Flow Instability	2	4
Random Forest Classifier	2	4
Working Condition Diagnosis	2	4
Autoencoder	2	3
Diagnostics	2	3
Sucker Rod Pump	2	3
Unsupervised Machine Learning	3	3
Anomaly Detection	4	2
Deep Learning	2	2
Neural Network	2	2
Support Vector Machine	2	2
Drilling	2	0

The top five keywords that were encountered most often are Machine Learning (9 occurrences, 15 total link strength), Fault Diagnosis (6 occurrences, 10 total link strength), Classification (3 occurrences, 9 total link strength), Oil Well Monitoring (3 occurrences, 8 total link strength) and Electrical Submersible Pump (2 occurrences, 7 total link strength).

As shown in Figure 2.4, the keywords co-occurrence analysis revealed 5 clusters with 25 keywords, 52 links and 66 total line strength. The clusters characterized by colors with the following major nodes:

- Machine Learning – Red
- Fault Diagnosis – Yellow
- Anomaly Detection – Blue
- Unsupervised Machine Learning – Purple
- Convolutional Neural Network – Green.

The keyword co-occurrence network shows that clusters exhibit distinct separation between each other with limited interconnections. Specifically, the blue cluster (major node Anomaly Detection), purple cluster (Unsupervised Machine Learning) and green cluster (Convolutional Neural Network) link just to the yellow (Fault Diagnosis) and red (Machine Learning) clusters. All of them don't have any links between each other. The two biggest clusters, yellow (Fault Diagnostics) and red (Machine Learning) have multiple links in-between and with other clusters.

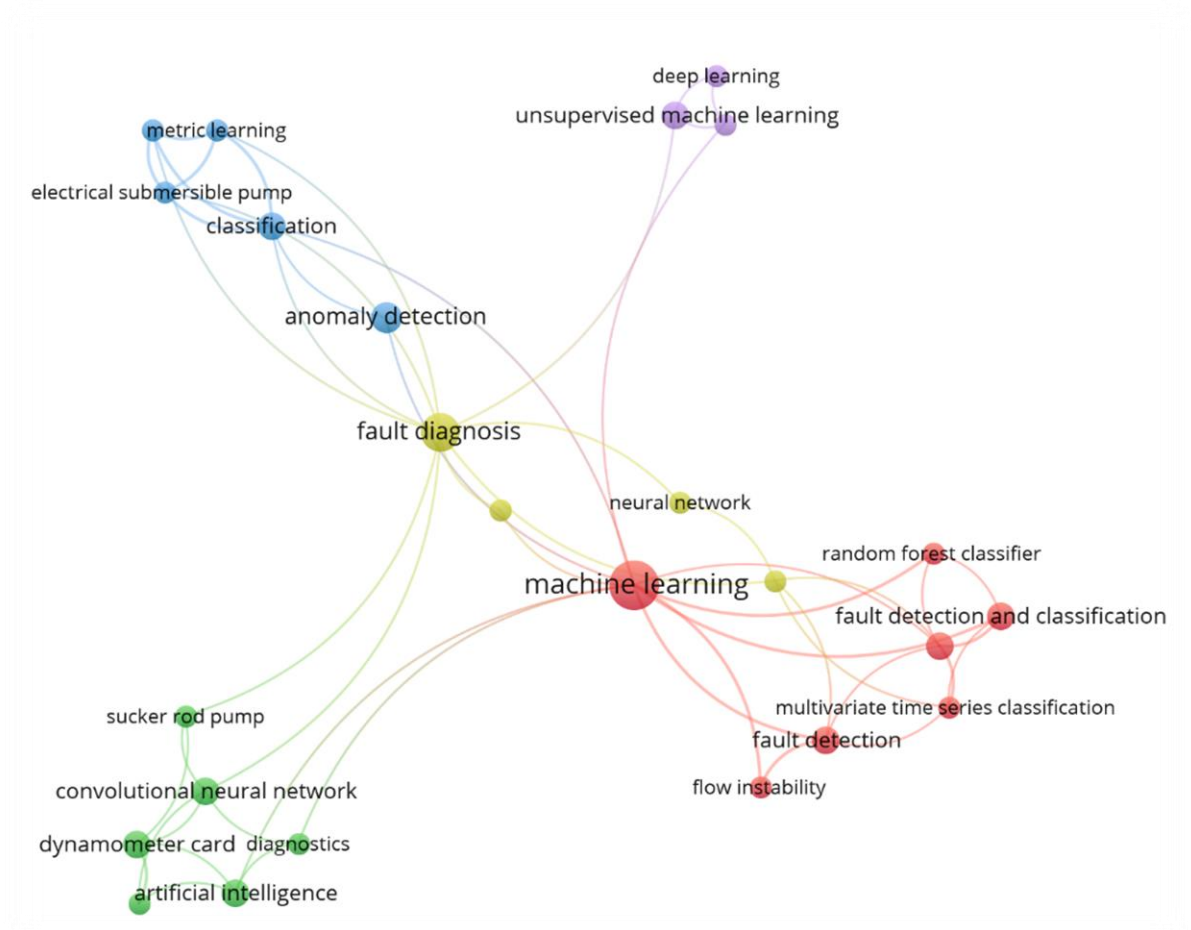


Figure 2.4 - Keywords Co-occurrence Network

The keywords co-occurrence network by year overlay visualization shows that Machine Learning methods gained most popularity from 2021, and there were many methodologies tried and implemented for anomaly detection, such as unsupervised machine learning methods, Random Forest, Support Vector Machine, the most recent being Autoencoder and Neural Network (Figure 2.5).

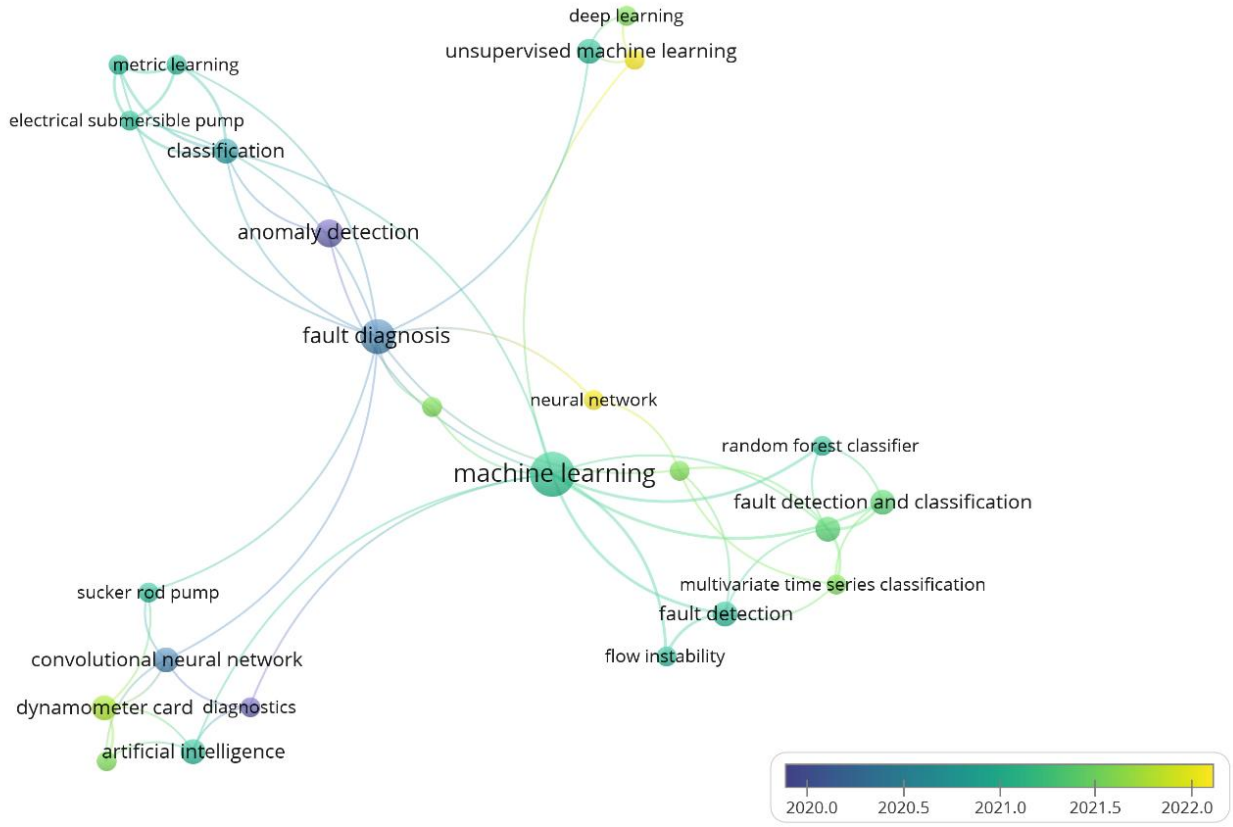


Figure 2.5 - Keywords Co-occurrence by Year

2.3.5. Authors co-authorship

The author's co-authorship was performed in VOSviewer using 25 maximum number of authors per document and minimum number of documents of an author of 2. Out of 282 authors only 20 meet the threshold (Table 2.4).

The authors with the highest rank of total link strength and hence collaborating the most are Varejão Flávio Miguel with a total link strength of 11 ((Mello et al., 2020), (Carvalho, Vargas, Salgado, Munaro, & Varejão, 2021), (Mello et al., 2022)), Antipova Ksenia with a total link strength of 9, Gurina Ekaterina with a total link strength of 9, Klyuchnikov Nikita with a total link strength of 9, Koroteev Dmitry with a total link strength of 9 ((Gurina et al., 2022a), (Gurina et al., 2022b), (Gurina et al., 2020)), Mello Lucas Henrique Sousa with a total link strength of 8, Oliveira-Santos Thiago with a total strength of 8, Ribeiro Marcos Pellegrini with a total strength of 8, Rodrigues, Alexandre Loureiros with a total strength of 8 ((Mello et al., 2022), (Mello et al., 2020)) and Vargas, Ricardo Emanuel Vaz with a total link strength of 7 ((Carvalho, Vargas, Salgado, Munaro, & Varejão, 2021), (Machado et al., 2022), (Vargas et al., 2017), (Carvalho, Vargas, Salgado, Munaro, & Varejao, 2021), (Scoralick et al., 2021), (Marins et al., 2021)).

Table 2.4 - Authors Co-authorship and the total link strengths

Authors	Publications	Total link strength
Varejão Flávio Miguel	3	11
Antipova Ksenia	3	9
Gurina Ekaterina	3	9
Klyuchnikov Nikita	3	9
Koroteev Dmitry	3	9
Mello Lucas Henrique Sousa	2	8
Oliveira-Santos Thiago	2	8
Ribeiro Marcos Pellegrini	2	8
Rodrigues Alexandre Loureiros	2	8
Vargas Ricardo Emanuel Vaz	6	7
Carvalho Bruno Guilherme	2	6
Garcia Ana Cristina Bicharra	2	6
Martí, Luis	2	6
Molina, José Manuel	2	6
Salgado Ricardo Menezes	2	6
Sanchez-Pi Nayat	2	6
Munaro Celso Jose	2	4
Alsaihati Ahmed	2	2
Elkatatny Salaheldin	2	2
Gao X	2	0

In the authors co-authorship network 6 clusters were identified with 20 items, 32 links and total link strength 65, as depicted in Figure 2.6.

Cluster 1 (red) with Varejão, Flávio Miguel as a top author with most total link strength, consists of Oliveira-Santos Thiago, Rodrigues Alexandre Loureiros, Mello Lucas Henrique Sousa and Ribeiro Marcos Pellegrini. Cluster 2 (yellow) with Vargas Ricardo Emanuel Vaz, an author with the most collaborated publications (6), has other three authors: Carvalho Bruno Guilherme, Munaro Celso Jose, Salgado Ricardo Menezes. Cluster 3 (blue) identified four authors, which are Antipova Ksenia, Gurina Ekaterina, Klyuchnikov Nikita, Koroteev Dmitry. Cluster 4 (purple) corresponds to Alsaihati, Ahmed and Elkatatny, Salaheldin ((Alsaihati et al., 2021), (Alsaihati et al., 2022)). Cluster 5 (green) consists of Sanchez-Pi Nayat, Martí Luis, Garcia Ana Cristina Bicharra and Molina, José Manuel ((Martí et al., 2015b), (Martí et al., 2017)). Cluster 6 (azure) has only one member Gao X ((J. Liu et al., 2019), (Wei & Gao, 2020)).

The red and yellow clusters are connected through Varejão, Flávio Miguel, who collaborated the most with other authors.

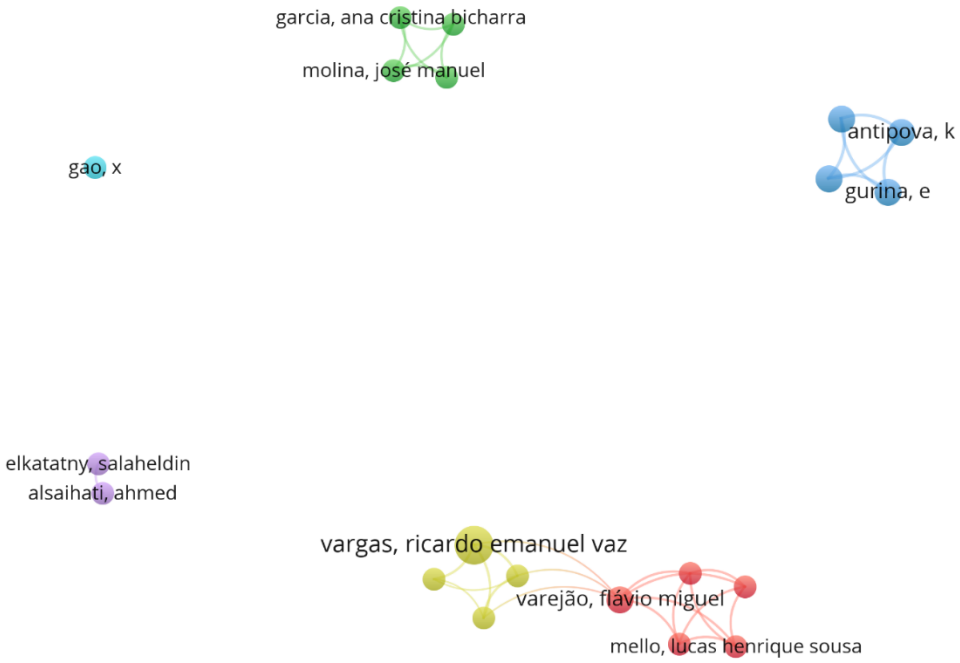


Figure 2.6 - Authors Co-authorship network visualization

Considering the years of publications, clusters 1 (red), 2 (yellow), 3 (blue) and 4 (purple) are the most recent, being produced between 2020 -2022. This group of authors identify the majority of collaboration, which was present in 2021. Cluster 6 (azure) is related to 2019-2020 years. Cluster 5 (green) corresponds to the 2015-2017 interval (Figure 2.7).

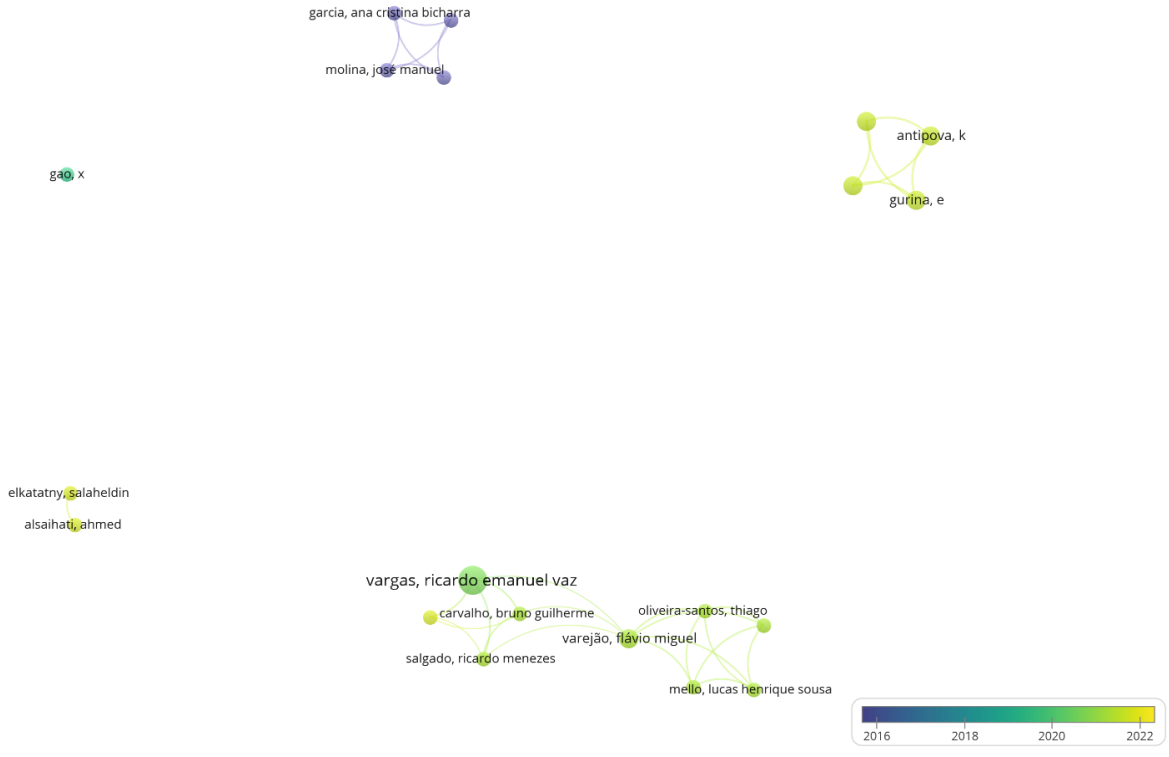


Figure 2.7 - Authors Co-authorship network visualization by Year

2.3.6. Discussion

The aim of the Systematic Literature Review was to identify the journal articles and conference proceedings, that were focused on techniques for anomaly detection in Oil and Gas industry using Machine Learning/ Deep Learning methods.

There were two questions stated in order to narrow our research:

- **RQ1**: What are the most applicable and significant Artificial Intelligence methods that were applied for detection of anomalies and potential failures in the Oil and Gas industry?

- **RQ2**: Which AI methods were applied for 3W dataset research?

All the identified literature was classified according to industry division into 4 main groups: (1) Drilling and Exploration, (2) Oil and Gas pipelines transportation system, (3) Production and reservoir management, (4) 3W Dataset.

Drilling and Exploration anomaly detection

There are 15 articles related to the subject of drilling operations, which focus on such issues as circulation loss, stuck pipe, washout, bit balling, drill pipe breaks, fluid show, potential kick and other downhole abnormalities. For anomaly detection several unsupervised methods were applied in order to identify unusual data records in multivariate time series from downhole and rig floor sensors, such as Regression, K-Nearest Neighbor (KNN), K-Means, t-SNE, dendrograms clustering analysis, Recurrent Neural Networks (RNNs), LSTM Autoencoder (LSTM-AE).

The following supervised Machine Learning methods were implemented to classify the abnormalities: Adaptive Neuro-Fuzzy Inference System (ANFIS), Random Forest (RF), Support vector machine (SVM), K-Nearest Neighbor (KNN), Gradient Boosting (GB), Shapley additive explanations (SHAP), Fully Connected network that has a multi-head attention mechanism (FCMH), eXtreme Gradient Boosting (XGBoost), Adaboost (ADA), Decision tree (DT), Multi-Layer Perceptron (MLP), Naïve Bayes Classifier (NBC) and Quadratic Discriminant Analysis (QDA).

In some publications Deep Learning methods were used either as an unsupervised learning tool to build Autoencoders (Mopuri et al., 2022) or for classification: Convolutional Neural Network (CNN), Artificial Neural Network (ANN), Functional Network (FN), Bag-of-features, the Feed Forward Back Propagation neural network (FFBPN), Recurrent Neural Networks (RNN) with Long Short-Term Memory variant (LSTM-RNN) or Gated Recurrent Unit variant (GRU-RNN) type of architecture.

In a few cases Genetic Algorithm was applied to optimize the multilayer Back Propagation Neural Network, creating GA-BP Neural Network ((Su et al., 2021), (Li et al., 2022)).

For the computer vision algorithm, the following backbones were attempted for image recognition using Regional Convolutional Neural Network (Faster-R-CNN): Single Shot Detector (SSD), You Only Look Once (YOLOv3), ResNet, DarkNet and Inception (Magana-Mora et al., 2021).

Oil and Gas pipelines transportation system anomaly detection

There are 8 articles with research subject related to pipelines transportation system. The major problems highlighted are pipeline leakage due to corrosion and harsh environment, damaged insulation, equipment failure, that provides pressure for oil and gas transportation, such as pumps and compressors, formation of gas hydrate due to low temperature and high pressure, etc. Since most pipelines are unobservable for humans in real time, many remote surveillance algorithms using computer vision are implemented.

The following deep-learning CNN classifiers were applied to detect leakage in underwater pipelines analyzing images: You Only Look Once (YOLO) architectures (YOLOv4, YOLOv4-Tiny, CSP-YOLOv4, YOLOv4@Resnet, YOLOv4@DenseNet), and one on Faster Region-based CNN (RCNN) (Gasparovic et al., 2022). For a DARTS-Drone Technological Solution computer vision algorithm was developed using deep learning neural network DeepLabV3+ and data augmentation (Ravishankar et al., 2022).

Among the unsupervised learning methods that were applied for pattern recognition and clustering were Gaussian mixture model (GMM) and K-Means. The supervised models employed are: Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Gradient Boosting (GB), Decision Tree (DT), Multiple Linear Regression, Neural Network and Multi-layer perceptron (MLP).

The following Deep Learning methods were used in addition to already mentioned computer vision techniques: Long Short-Term Memory (LSTM) and Stacked Auto-Encoder (SAE) (Seo et al., 2021), Convolutional Neural Network (CNN), Inception ResNet V2 and Visual Geometry Group with 16 layers (VGG16) (Vankov et al., 2020).

Production and reservoir management anomaly detection

24 articles are related to production and reservoir management and it is the biggest group from the pool of the identified literature. Apart from naturally flowing wells, in which formation pressure is high enough to provide extraction without additional treatment, most wells require Artificial lift techniques: Beam Pumps (sucker rod system), Electrical Submersible Pumping (ESP), Gas Lift Systems, Hydraulic Pumps, Plungers and Progressive cavity pumps (PCP) (*The Defining Series: Artificial Lift | SLB*, n.d.). Regardless of how robust and well maintained this equipment, they are susceptible to many failures, and most of the research is focused on anomaly detection in this production branch of the industry.

The following supervised learning methods were applied: K-Nearest Neighbor (KNN), Logistic Regression (Logit), Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF), Rule Fit Classifier (RFC), Extreme Learning Machine (ELM), supervised shapelet-based classification algorithm Fast Shapelets, Naive Bayes (NB), Stochastic Gradient Descent (SGD), Quadratic discriminant analysis (QDA), Linear Discriminant Analysis (LDA), boosting techniques e.g., Extreme Gradient Boosting (XGB), Adaptive Boosting (AdaBoost), and Categorical Boosting (CatBoost).

Since for most of the methods the explanation of the decision process is not straightforward, further model analysis for explainability using LIME (Local Interpretable Model-agnostic Explanations) and its interpretability was performed (Alharbi et al., 2022).

A few Genetic Algorithms were applied to optimize the SVM model, due to the problem that the parameters of the Support Vector Machine are difficult to determine when classifying: Chicken Swarm

Optimization (CSO), differential mutation strategy and adaptive inertial strategy (DACSO), Particle Swarm Optimization (PSO) and Bat Algorithm (BA) (J. Liu et al., 2019). Also, Genetic Algorithm optimized Back Propagation neural network (GA-BP) was implemented for offshore submersible motor fault diagnosis (Y. Zhang & Yang, 2022).

One-class Support Vector Machine (SVM) in combination with Yet Another Segmentation Algorithm (YASA), which is designed for time series pattern analysis, and Kalman filters, were applied for anomaly detection of offshore platform turbomachines ((Martí et al., 2015b), (Martí et al., 2017)).

Unsupervised machine learning algorithms that were used are: Cluster based local outlier factor (CBLOF), Histogram-based Outlier Score (HBOS), Isolation Forest (IF), Median Absolute Deviation (MAD), Minimum Covariance Determinant (MCD), Principal Component Analysis (PCA), Gaussian Markov random fields (GMRF), graphical Gaussian model (GGM), sparse Principal Component Analysis (sPCA), sparse Autoencoder, Alternating Decision Tree (ADTree), Support Vector Machine (SVM), Naïve Bayesian Network, Fuzzy C-means algorithm.

A semi-supervised method Random Peek was employed for the case of Artificial Lift systems anomaly detection, where only a small number of samples is labeled, assuming that most of the unlabeled samples should be labeled Normal ((Y. Liu et al., 2010), (Y. Liu et al., 2011)).

The following Deep Learning methods were exercised: Back Propagation Neural Network (BPNN), Convolutional Neural Network (CNN), Triplet network, i.e., an artificial neural network based on a Triplet loss metric, and other metric learning losses, such as Proxy-Anchor loss, Contrastive loss, Lifted Structured loss, CosFace loss (Mello et al., 2022), two stacked Autoencoders (Scoralick et al., 2021), Multilayer Feedforward Neural Network (MFNN), Long Short-Term Memory (LSTM), Convolutional-LSTM (CONV-LSTM) (Sinha et al., 2020), CNN with backbones ResNet50, SE-ResNet50, ResNet50 Ⅱ, SE-ResNet50 Ⅱ, AlexNet (Tan et al., 2022), Deep-Broad Learning System (DBLS), Fast Fourier transform (FFT), Wavelet transformation (Wei & Gao, 2020).

Finally, transfer learning techniques were used for diagnosis of sucker-rod pump working conditions: AlexNet Network, GoogLeNet Network, shallow Convolutional Neural Networks (CNN3 model and CNN2 model) and Fully Connected Neural Network model (FC model) (R. Zhang et al., 2021).

3W Dataset Anomaly Detection and Classification

To the best of our knowledge, there are officially published 11 articles about anomaly detection and classification on offshore naturally flowing wells, using 3W dataset from Petrobras, that was created by combining real, simulated and hand-drawn records, written in English and having open access.

For analysis of the time series data, all authors applied Sliding Windows technique. Some researchers attempted multiclass classification of undesirable events, while others selected one particular abnormality (ex. flow instability) and performed binary classification against all the rest of the classes. (Marins et al., 2021) performed 3 experiments: One-class classifier to identify normal vs abnormal events, thus combining all faults into one unique class, Multiple binary classifiers with several classifiers discriminating each individual fault against normal events, and Single multiclass classifier, identifying each fault against all events, as mentioned earlier.

The next supervised learning methods were applied: KNN (k-Nearest Neighbors), One Nearest Neighbor (1NN), Logistic regression, Support Vector Classifier (SVC), Linear and Quadratic Discriminant Analysis (LDA & QDA), Decision Tree (DT), Random Forest (RF), AdaBoost (ADA), Gaussian Naive Bayes (GNB), Zero Rule (ZR), Extreme Learning Machine (ELM), Multilayer Perceptron (MLP).

A few Genetic Algorithms were again used to optimize the algorithm: (Gatta et al., 2022) created a Convolutional 1D Autoencoder with genetic approach for hyperparameters selection via Biased Random Key Genetic Algorithm (BRKGA), in which different combinations of hyperparameters are regarded as an individual of a population, and each hyperparameter is regarded as a gene of the individual.

Explainability of the classifiers was also researched, and three XAI techniques were applied to interpret black box models to understand the causes of abnormalities: global surrogate model using DT, Shapley Additive Explanation (SHAP), and Local Interpretable-Agnostic Explanation (LIME) (Aslam et al., 2022).

Unsupervised algorithms that were implemented are: t-distributed Stochastic Neighbor Embedding (t-SNE), Principal Component Analysis (PCA), one-class Support Vector Machine (SVM), Cluster-based Algorithm for Anomaly Detection in Time Series Using Mahalanobis Distance (C-AMDATS), Luminol Bitmap, SAX-REPEAT, KNN, Bootstrap, and Robust Random Cut Forest (RRCF).

Finally, the Deep Learning methods that were attempted are: Long Short-Term Memory (LSTM) Autoencoder and Convolutional Neural Network 1D Autoencoder.

The summary of all the publications with corresponding AI methods is represented in Table 2.5.

Table 2.5 - PRISMA method selected publications

No	Publication	Research Question	Methods
	Drilling and Exploration Anomaly Detection		
1	Application of adaptive neuro-fuzzy inference system and data mining approach to predict lost circulation using DOE technique (case study: Maroon oilfield) (Agin et al., 2020)	Prediction of lost circulation problem during drilling	Data mining (regression) and Adaptive Neuro-Fuzzy Inference System (ANFIS).
2	Deep Learning and Time-Series Analysis for the Early Detection of Lost Circulation Incidents during Drilling Operations (Aljubran et al., 2021)	Detection of lost circulation during drilling	Random Forest as a baseline, Deep Learning methods CNN, ANN and LSTM.
3	Application of Machine Learning Methods in Modeling the Loss of Circulation Rate	Predicting the loss of circulation rate (LCR) while drilling	Support vector machine (SVM), Random Forest (RF), and K-Nearest Neighbor (KNN).

	while Drilling Operation (Alsaihati et al., 2022)		
4	Use of Machine Learning and Data Analytics to Detect Downhole Abnormalities while Drilling Horizontal Wells, with Real Case Study (Alsaihati et al., 2021)	Continuous profile of the surface drilling torque (T&D) prediction to enable the detection of operational problems ahead of time.	Random forest (RF), Artificial Neural Network (ANN), and Functional Network (FN).
5	Forecasting the abnormal events at well drilling with machine learning (Gurina et al., 2022a)	Prediction of six types of drilling accidents probabilities in real-time, using the data from the drilling telemetry representing the time-series.	Bag-of-features, K-Means, Gradient Boosting (GB), Convolution Neural Network (CNN)
6	Making the black-box brighter: Interpreting machine learning algorithm for forecasting drilling accidents (Gurina et al., 2022b)	Interpretability and development of explanatory model of Bag-of-features approach, used for drilling accidents prediction.	Bag-of-features, Shapley additive explanations (SHAP), Fully connected network that has a multi-head attention mechanism (FCMH), T-SNE
7	Application of machine learning to accidents detection at directional drilling (Gurina et al., 2020)	Development of data-driven algorithm for anomaly alarming for directional drilling.	Gradient Boosting (GB), dendrograms clustering analysis
8	AI-Driven maintenance support for downhole tools and electronics operated in dynamic drilling environments (Kirschbaum et al., 2020)	Artificial Intelligence (AI)-driven Condition Based Maintenance (CBM), combining Bottom Hole Assembly (BHA) data with Big Data Analytics (BDA) for downhole electronics failure detection	Random Forest (RF), eXtreme Gradient Boosting (XGBoost)
9	Drilling performance monitoring and optimization: a data-driven approach (Lashari et al., 2019)	Prediction of ROP, drilling performance monitoring and optimization, identifying the bit malfunction or failure, like bit balling.	The feed forward back propagation neural network (FFBPN)
10	A New Method for Intelligent Prediction of Drilling Overflow and Leakage Based on Multi-Parameter Fusion (Li et al., 2022)	Mud overflow and leakage prediction during drilling	Genetic Algorithm to optimize the multilayer Back Propagation Neural Network (GA-BP Neural Network)
11	Well Control Space Out: A Deep-Learning Approach for the Optimization of Drilling	Surveillance method for drilling operations control using cameras and computer vision in real time. The model	Deep Learning methods: Regional Convolutional Neural Network (Faster-R-CNN), Single Shot Detector (SSD), You Only Look

	Safety Operations (Magana-Mora et al., 2021)	for tool joint detection is used to compute the location of the tool joint below the drill floor. In the case of an uncontrolled flow, the Well Control Space Out determines the appropriate measures to take.	Once (YOLOv3), ResNet, DarkNet and Inception backbones
12	Early sign detection for the stuck pipe scenarios using unsupervised deep learning (Mopuri et al., 2022)	Detecting early signs for the stuck events in drilling	Unsupervised learning: Recurrent Neural Networks (RNNs), LSTM Autoencoder (LSTM-AE)
13	Supervised data-driven approach to early kick detection during drilling operation (Muojeke et al., 2020)	Early kick detection during drilling for implementing the appropriate well control strategy to manage kick situations	Supervised models: Artificial Neural Network (ANN), Recurrent Neural Networks (RNN), Long Short-Term Memory variant of RNN, (LSTM-RNN), Gated Recurrent Unit variant of RNN, (GRU-RNN)
14	Prediction of drilling leakage locations based on optimized neural networks and the standard random forest method (Su et al., 2021)	Creating real-time model for predicting leakage layer locations in drilled formations, that cause potential circulation loss	Genetic Algorithm-Back Propagation (GA-BP) neural network, Random Forest (RF)
15	Effective prediction of lost circulation from multiple drilling variables: a class imbalance problem for machine and deep learning algorithms (Wood et al., 2022)	Prediction of lost circulation during drilling	8 Machine Learning methods: Adaboost (ADA), Decision tree (DT), K-Nearest Neighbour (KNN), Multi-Layer Perceptron (MLP), Naïve Bayes Classifier (NBC), Quadratic Discriminant Analysis (QDA), Random Forest (RF) and Support Vector Classifier (SVR). 3 Deep Learning methods: Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM).
	Production anomaly detection and classification		
16	Explainable and Interpretable Anomaly Detection Models for Production Data (Alharbi et al., 2022)	Study of white-box and black-box classifiers for supervised anomaly detection on oil and gas production data.	K-Nearest Neighbor (KNN), Logistic Regression (Logit), Support Vector Machines (SVMs), Decision Tree (DT), Random Forest (RF), and Rule Fit Classifier

			(RFC). Further models analysis for explainability using LIME (local interpretable model-agnostic explanations) and interpretability.
17	Self-Diagnosis of Multiphase Flow Meters through Machine Learning-Based Anomaly Detection (Barbariol et al., 2020)	Method AD4MPFM (Anomaly Detection for Multiphase Flow Meters), enabling the metrology system to detect outliers and to provide a statistical level of confidence in the measures for oil production.	Unsupervised machine learning algorithms: Cluster based local outlier factor (CBLOF), Histogram-based Outlier Score (HBOS), Isolation Forest (IF), Median Absolute Deviation (MAD), Minimum Covariance Determinant (MCD), and Principal Component Analysis (PCA)
18	Sparse Gaussian Markov Random Field Mixtures for Anomaly Detection (Ide et al., 2017)	Anomaly detection of a compressor of offshore oil production, from multivariate noisy sensor data.	Gaussian Markov random fields (GMRF), graphical Gaussian model (GGM), sparse Principal Component Analysis (sPCA), sparse Autoencoder
19	Fault Diagnosis of Rod Pumping Wells Based on Support Vector Machine Optimized by Improved Chicken Swarm Optimization (J. Liu et al., 2019)	Diagnosis the faults of pumping wells by classifying and identifying the indicator diagrams	Support vector machine (SVM), chicken swarm optimization (CSO), differential mutation strategy and adaptive inertial strategy (DACSO), particle swarm optimization (PSO) and bat algorithm (BA)
20	Failure Prediction for Rod Pump Artificial Lift Systems (Y. Liu et al., 2010)	Prediction of Failure for Rod Pump Artificial Lift Systems	Unsupervised methods: Alternating Decision Tree (ADTree), Support Vector Machine (SVM), Naïve Bayesian Network. Semi-supervised: Random Peek.
21	Semi-supervised failure prediction for oil production wells (Y. Liu et al., 2011)	Development of Smart Engineering Apprentice (SEA) framework for Artificial Lift Systems failure prediction	Semi-supervised classification using Random Peek, Support Vector Machines (SVM)
22	Adaptive fault diagnosis of sucker rod pump systems based on optimal perceptron and simulation data (X.-X. Lv et al., 2022)	The improved model of fault diagnosis for the sucker rod production system (SRPS)	Back Propagation Neural Network (BPNN), Extreme Learning Machine (ELM), and Support Vector Machine (SVM) with improved feature extraction
23	An evolutional SVM method based on incremental algorithm and simulated	Fault diagnosis of the sucker rod pumping system (SRPS)	Evolutional SVM method based on incremental algorithm and

	indicator diagrams for fault diagnosis in sucker rod pumping systems(X. Lv et al., 2021)		simulated IDs, ELM, PSO-ELM, BPNN and SVM as baselines
24	Anomaly Detection Based on Sensor Data in Petroleum Industry Applications (Martí et al., 2015b)	Anomaly Detection in Offshore Oil Extraction Turbomachines	One-class support vector machine (SVM), Yet Another Segmentation Algorithm (YASA)
25	On the combination of support vector machines and segmentation algorithms for anomaly detection: A petroleum industry comparative study (Martí et al., 2017)	Anomaly detection of turbomachinery installed in offshore petroleum extraction platforms.	One-class Support Vector Machines (SVM), Kalman filters, Yet Another Segmentation Algorithm (YASA)
26	Metric Learning for Electrical Submersible Pump Fault Diagnosis (Mello et al., 2020)	Electrical Submersible Pump (ESP) fault diagnosis	Convolutional neural network (CNN) trained with a triplet loss learning for extracting relevant features, standard machine learning algorithm such as K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, Quadratic Discriminant Analysis and Naïve Bayes Classifier
27	Ensemble of metric learners for improving electrical submersible pump fault diagnosis (Mello et al., 2022)	Electrical Submersible Pump (ESP) fault diagnosis	Ensembles composed of deep neural networks (convolutional network (ConvNet) with 5 metrics: Triplet network, i.e., an artificial neural network based on a metric called Triplet loss, Proxy-Anchor loss, Contrastive loss, Lifted Structured loss, CosFace loss. Random Forest (RF), majority voting, Principal Component Analysis (PCA)
28	Unsupervised Methods to Classify Real Data from Offshore Wells (Orestes et al., 2021)	Anomalies detection during oil and gas production	Fuzzy C-means algorithm for classification into clusters, Control Chart method, Random Forest (RF)
29	Predicting Compressor Valve Failures from Multi-Sensor Data (Patri et al., 2015)	Ranking sensor dimensions and finding signatures in compressor sensor data, which may aid in the prediction of valve failure	Decision Tree, supervised shapelet-based classification algorithm Fast Shapelets

30	Electric submersible pump broken shaft fault diagnosis based on principal component analysis (Peng et al., 2020)	Identify the cause and the time of ESP shaft fracture, predict the impending breakage time and determine the variable most responsible	Principal Component Analysis (PCA)
31	Machine Learning Models to Predict Gas Hydrate Plugging Risks Using Flowloop and Field Data (Qin et al., 2019)	Evaluating gas hydrate risk based on measurable process parameters	Support vector classifier (SVC) with several kernels, such as linear, polynomial, radial basis functional (RBF), and artificial neural networks (ANN). Feature selection methods SelectKBest and ExtraTreesClassifier.
32	A novel machine learning model for autonomous analysis and diagnosis of well integrity failures in artificial-lift production systems (Salem et al., 2022)	Automated prediction of integrity failures in wells with Artificial Lift gas lift production method	Logistic regression, Naive Bayes (NB), Decision trees (DT), Random Forests (RF), KNN, SVM, Stochastic gradient descent (SGD), Quadratic discriminant analysis (QDA), boosting techniques e.g., Extreme Gradient Boosting (XGB), Adaptive Boosting (AdaBoost), and Categorical Boosting (CatBoost).
33	Fault detection with Stacked Autoencoders and pattern recognition techniques in gas lift operated oil wells (Scoralick et al., 2021)	Detection and classification of failures in oil production wells operated with elevation by gas lift.	Two stacked autoencoders with 9 and 5 neurons, Decision Tree (DT), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), KNN
34	Normal or abnormal? Machine learning for the leakage detection in carbon sequestration projects using pressure field data (Sinha et al., 2020)	Automation of the leakage detection process in carbon storage reservoirs using rates of (CO2) injection and pressure data measured by simple harmonic pulse testing (HPT).	Multilayer Feedforward Neural Network (MFNN), Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), Convolutional-LSTM (CONV-LSTM)
35	A visual analytics approach to anomaly detection in hydrocarbon reservoir time series data (Soriano-Vargas et al., 2021b)	Anomaly detection in time series data of hydrocarbon reservoir using visual analytics approach based on interactive visualizations of time series connected with machine learning approaches.	Isolation Forest (IF)
36	Multi-Scale Normalization Method Combined with a Deep	Development of diagnosis model to identify the working	Four CNN backbones: ResNet50, SE-ResNet50, ResNet50 Ⅱ, SE-

	CNN Diagnosis Model of Dynamometer Card in SRP Well (Tan et al., 2022)	condition of each sucker rod pumping (SRP) well	ResNet50Ⅱ, SVM with radial basis function and particle swarm optimization (PSO), AlexNet model for comparison
37	Fault Diagnosis of Sucker Rod Pump Based on Deep-Broad Learning Using Motor Data (Wei & Gao, 2020)	Fault diagnosis methods of sucker rod pump (SRP)	Convolutional Neural Network (CNN), Deep-Broad Learning System (DBLS), Fast Fourier transform (FFT), Wavelet transformation, Extreme Learning Machine (ELM), Support Vector Machine (SVM), Hidden Markov Model (HMM)
38	Fault Diagnosis of Submersible Motor on Offshore Platform Based on Multi-Signal Fusion (Y. Zhang & Yang, 2022)	Offshore submersible motor fault diagnosis	Back Propagation Neural Network (BP), Genetic Algorithm optimized Back Propagation neural network (GA-BP)
39	An intelligent diagnosis method of the working conditions in sucker-rod pump wells based on convolutional neural networks and transfer learning (R. Zhang et al., 2021)	Diagnosis of sucker-rod pump working conditions	Transfer deep learning methods: AlexNet Network, GoogLeNet Network, shallow convolutional neural networks (CNN3 model and CNN2 model) and Fully Connected Neural Network model (FC model)
	Oil Pipelines and Transportation System Anomaly Detection		
40	An Anomaly Detection Model for Oil and Gas Pipelines Using Machine Learning (Aljameel et al., 2022)	Oil pipeline leakage detection.	Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Gradient Boosting (GB), Decision Tree (DT).
41	A data-driven pipeline pressure procedure for remote monitoring of centrifugal pumps (Giro et al., 2021)	Automated strategy to remotely monitor the status of centrifugal pumps in pipeline transportation systems, when the network of sensors is not available or not present.	Unsupervised clustering techniques: Gaussian mixture model (GMM)
42	Deep Learning Approach for Objects Detection in Underwater Pipeline Images (Gasparovic et al., 2022)	Underwater seafloor pipelines leakage detection, using images, to verify their integrity and determine the need for maintenance	Convolutional Neural Network (CNN), Six different architectures: You Only Look Once (YOLO) architectures (YOLOv4, YOLOv4-Tiny, CSP-YOLOv4, YOLOv4@Resnet, YOLOv4@DenseNet), and one on the Faster Region-based CNN (RCNN) architecture.

43	DARTS-Drone and Artificial Intelligence Reconsolidated Technological Solution for Increasing the Oil and Gas Pipeline Resilience (Ravishankar et al., 2022)	Integrating drone technology and deep learning technique to detect the targeted potential root problems that can cause critical pipeline failures and predict the progress of the detected problems by collecting and analyzing image data periodically	Computer vision algorithm using deep learning neural network DeepLabV3+, data augmentation
44	Development of AI-based diagnostic model for the prediction of hydrate in gas pipeline (Seo et al., 2021)	Diagnose hydrate for flow assurance purposes in gas pipelines	Multi-layer perceptron (MLP), Long Short-Term Memory LSTM, and Stacked Auto-Encoder (SAE)
45	Microwave Nondestructive Testing for Defect Detection in Composites Based on K-Means Clustering Algorithm (Shrifan et al., 2021)	Non-destructive testing (NDT) to detect the underneath defect in composites, used for insulation of steel pipelines in oil and gas industry, based on microwave reflection coefficients	Unsupervised machine learning: K-Means clustering
46	Assessment of the condition of pipelines using convolutional neural networks (Vankov et al., 2020)	Analysis of amplitude-frequency measurements in pipelines to identify the presence of a defect and further clarify its variety	Convolutional Neural Network (CNN), Inception ResNet V2, Visual Geometry Group with 16 layers (VGG16)
47	A minimalist approach for detecting sensor abnormality in oil and gas platforms (Wong et al., 2022)	Detecting abnormality of compressor's shaft's RPM sensor	Multiple Linear Regression, Neural Network
	3W Dataset Anomaly Detection and Classification		
48	Proposal for two classifiers of offshore naturally flowing wells events using k-nearest neighbors, sliding windows and time multiscale (Vargas et al., 2017)	Identification of four anomalous events in oil wells for 3W Dataset: Spurious Closure of DHSV, Rapid Productivity Loss, Hydrates in Production Lines, Choke Valve Closure	KNN (k-Nearest Neighbors), t-distributed Stochastic Neighbor Embedding) (t-SNE)
49	Classification of undesirable events in oil well operation (Turan & Jaschke, 2021)	Multiclass classification of anomalous events in oil wells for 3W Dataset	Decision Tree, as baseline attempted Logistic Regression (LR), Support Vector Classifier (SVC), Linear and Quadratic Discriminant Analysis (LDA & QDA), Random Forest, AdaBoost

			(ADA), Principal Component Analysis (PCA)
50	Statistical analysis of offshore production sensors for failure detection applications (Santos et al., 2021)	Identification of abnormal events in oil wells for 3W Dataset	Principal Component Analysis (PCA) and Logistic Regression (LR)
51	Fault detection and classification in oil wells and production/service lines using random forest (Marins et al., 2021)	Development of CBM system for identification of anomalous events in oil wells for 3W Dataset	Random Forest, Principal component Analysis (PCA), Bayesian non-convex optimization strategy
52	Improving performance of one-class classifiers applied to anomaly detection in oil wells (Machado et al., 2022)	Identification of two types of faults in oil wells for 3W Dataset: Spurious closing of Downhole Safety Valves (DHSV) and Hydrate in Production Line.	Two unsupervised learning methods: Long Short-Term Memory (LSTM) autoencoder and one-class Support Vector Machine (OCSVM), trained on faulty events as a target class.
53	Predictive maintenance for offshore oil wells by means of deep learning features extraction (Gatta et al., 2022)	Multiclass classification of anomalous events in oil wells for 3W Dataset	Deep learning method for feature extraction: 1D AutoEncoder using Convolutional Neural Network. Machine learning classifiers: Random Forest, Nearest Neighbors, Gaussian Naive Bayes and Quadratic Discriminant Analysis, hyperparameters selection via Biased Random Key Genetic Algorithm (BRKGA).
54	Data-driven Detection and Identification of Undesirable Events in Subsea Oil Wells (Brønstad et al., 2021)	Development of CBM system for identification of anomalous events in oil wells for 3W Dataset	Random Forest (RF), Principal Component Analysis (PCA)
55	Flow Instability Detection in Offshore Oil Wells with Multivariate Time Series Machine Learning Classifiers (Carvalho, Vargas, Salgado, Munaro, & Varejao, 2021)	3W Dataset Flow Instability prediction	Binary machine learning classifiers: One Nearest Neighbor (1NN), Gaussian Naïve Bayes (GNB), Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), Random Forest (RF). As a baseline - the Zero Rule (ZR) classifier.
56	Hyperparameter Tuning and Feature Selection for Improving Flow Instability Detection in Offshore Oil Wells. (Carvalho, Vargas, Salgado, Munaro, & Varejão, 2021)	Improvement of previous 3W Dataset Flow Instability prediction	Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Adaptive Boosting (ADA), Extreme Learning Machine (ELM) and Multilayer Perceptron (MLP), Zero-rule (ZR)

			classifier, Sequential Feature Selection SFS-F (forward), SFS-B (backward) and Genetic Algorithm for feature selection.
57	Detecting Interesting and Anomalous Patterns in Multivariate Time-Series Data in an Offshore Platform Using Unsupervised Learning (Figueirêdo et al., 2021)	A comparative evaluation performance of unsupervised learning algorithms for pattern recognition in 3W Dataset undesirable events, such as Spurious closure of DHSV and Quick restriction in PCK	Six unsupervised machine learning algorithms: Cluster-based Algorithm for Anomaly Detection in Time Series Using Mahalanobis Distance (C-AMDATS), Luminol Bitmap, SAX-REPEAT, KNN, Bootstrap, and Robust Random Cut Forest (RRCF).
58	Anomaly Detection Using Explainable Random Forest for the Prediction of Undesirable Events in Oil Wells (Aslam et al., 2022)	Identification of anomalous events in oil wells for 3W Dataset and model interpretation	Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbor (K-NN), SMOTE, Explainable Artificial Intelligence (XAI). Three XAI techniques: global surrogate model using DT, Shapley Additive Explanation (SHAP), and Local Interpretable-Agnostic Explanation (LIME).

The systematic literature review following PRISMA methodology allowed an insight into the current state of knowledge in the area of anomaly detection in the Petroleum industry. Specifically, most of the publications are related to the Production sector, and the least - to the Oil and Gas pipeline and transportation equipment. The specter of applied methods is wide and many advanced techniques are implemented to enhance the result, such as using Genetic algorithms for Machine Learning and Deep Learning model optimization, creating stacked Autoencoders algorithms, applying Convolutional Neural Networks for improved feature extraction, explaining black box models using Explainable Artificial Intelligence techniques, etc.

Regarding the detection of undesirable events in naturally flowing wells for 3W Dataset, there were many experiments attempted in setting up multiclass or binary classifications. An array of supervised and unsupervised learning methods was applied with some outstanding results. The recent publications, made in 2022, focused more on Deep Learning algorithms, since they provide higher accuracy of classification. It would be suggested to develop further the latest contributions by attempting other Recurrent Neural Network configurations with LSTM and GRU architectures, or focus on identifying other particular types of faults, which were not analyzed before.

3. METHODOLOGY

The research methodology is composed of 3 main phases, which are Exploration, Analytical and Conductive Phases.

As the objective of the literature review process was to obtain the insight about the state of the knowledge in the area of anomaly detection, identifying potentially applicable but not yet exploited algorithms, the result will be implemented for designing the Methodology of the project.

The Exploration Phase starts with a literature review of the recent development in the field of Artificial Intelligence in the Oil and Gas industry and the 3W dataset particularly (Figure 3.1). It was recognized, that Deep Learning methods were implemented only in 2 out of 11 official publications (taking into account those only written in English):

- "Improving performance of one-class classifiers applied to anomaly detection in oil wells" (Machado et al., 2022), in which LSTM Autoencoder was used for binary classification of two types of anomalies, such as Spurious closing of Downhole Safety Valves (DHSV) and Hydrate in Production Line, and then compared to the one-class Support Vector Machine (OCSVM) classifier,
- "Predictive maintenance for offshore oil wells by means of deep learning features extraction" (Gatta et al., 2022), where instead of applying statistical methods for feature engineering, the Autoencoder (AE) using the Convolutional Neural Network (CNN) is created in order to decrease the dimensionality of the feature space. Then the extracted features are loaded to four Machine Learning algorithms for multi-class classification. The hyperparameters of the classifiers were optimized using Biased Random Key Genetic Algorithm (BRKGA).

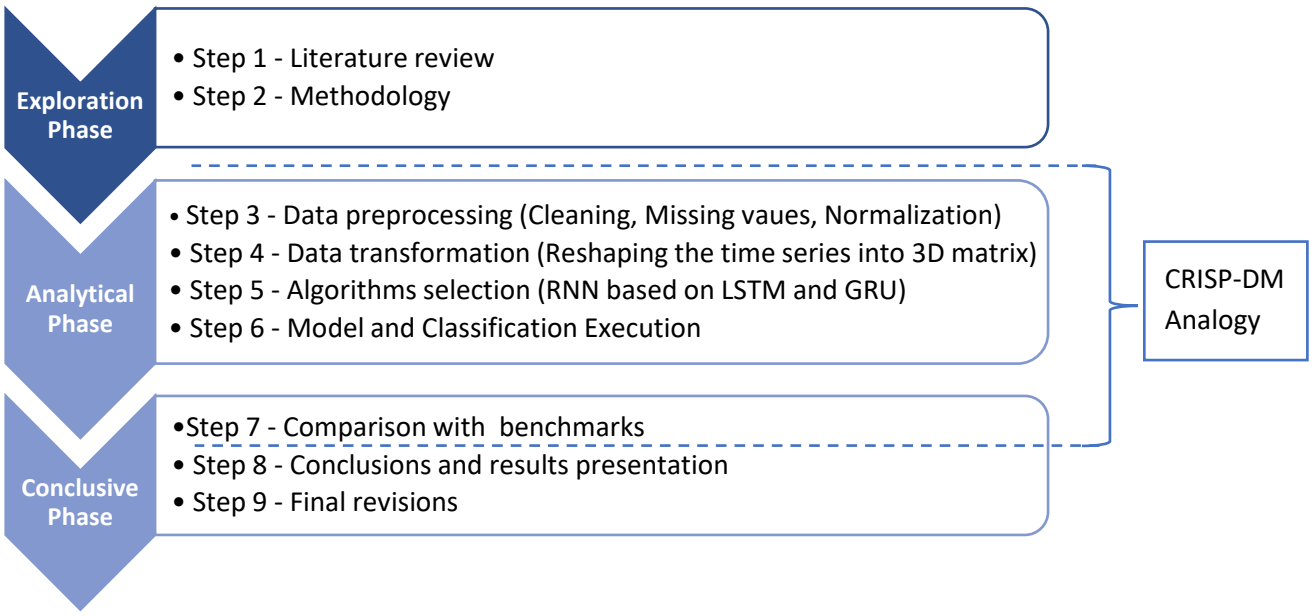


Figure 3.1 - Phases of the research

The suggested Methodology of the research is based on further exploration of Deep Learning techniques for multi-class classification, creating Recurrent Neural Network (RNN) configurations with Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures.

As shown in Figure 3.1, the Analytical and partially Conclusive Phases of the research methodology are greatly inspired by the Cross-Industry Standard Process for Data Mining (CRISP-DM), which consists of 6 phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment (Nick Hotz, 2022). Despite this being a Data Mining approach, the main pillars of the process model are highly relevant for the project, and can be implemented in the research parts of it.

The Analytical Phase would include the following intermediate objectives:

- Pre-processing the 3W Dataset by data cleaning, imputing or removing missing values, standardizing the data for better performance of the deep neural networks
- Data transformation by converting into the 3D matrix expected by LSTM and GRU backbones: [samples, timesteps, features]
- Based on Recurrent Neural Network (RNN), developing the algorithm with LSTM and GRU architectures to perform the abnormal events multi-class classification
- Evaluating the results by comparing with benchmarks from the previous researches.

The suggested workflow is presented in Figure 3.2, starting with the overview of the 3W Dataset. Its descriptive introduction and analysis are essential and detailed in the Data Processing part of the research, since it is a challenging dataset, and requires a thorough grasp for the task of anomaly detection in this project.

The Research Design runs in parallel, including already performed literature review, overview of the suggested algorithms and an on-going process of improving the model with the consideration of potentially adding unsupervised algorithms for dimensionality reduction or genetic algorithms for hyperparameters tuning.

The core of the research is the project itself, where the pipeline of the algorithms will be setup, and the classification performed to detect the undesirable events. Finally, a comparison with the previous publications results will be implemented, with the main focus on the papers, which performed multi-class classification.

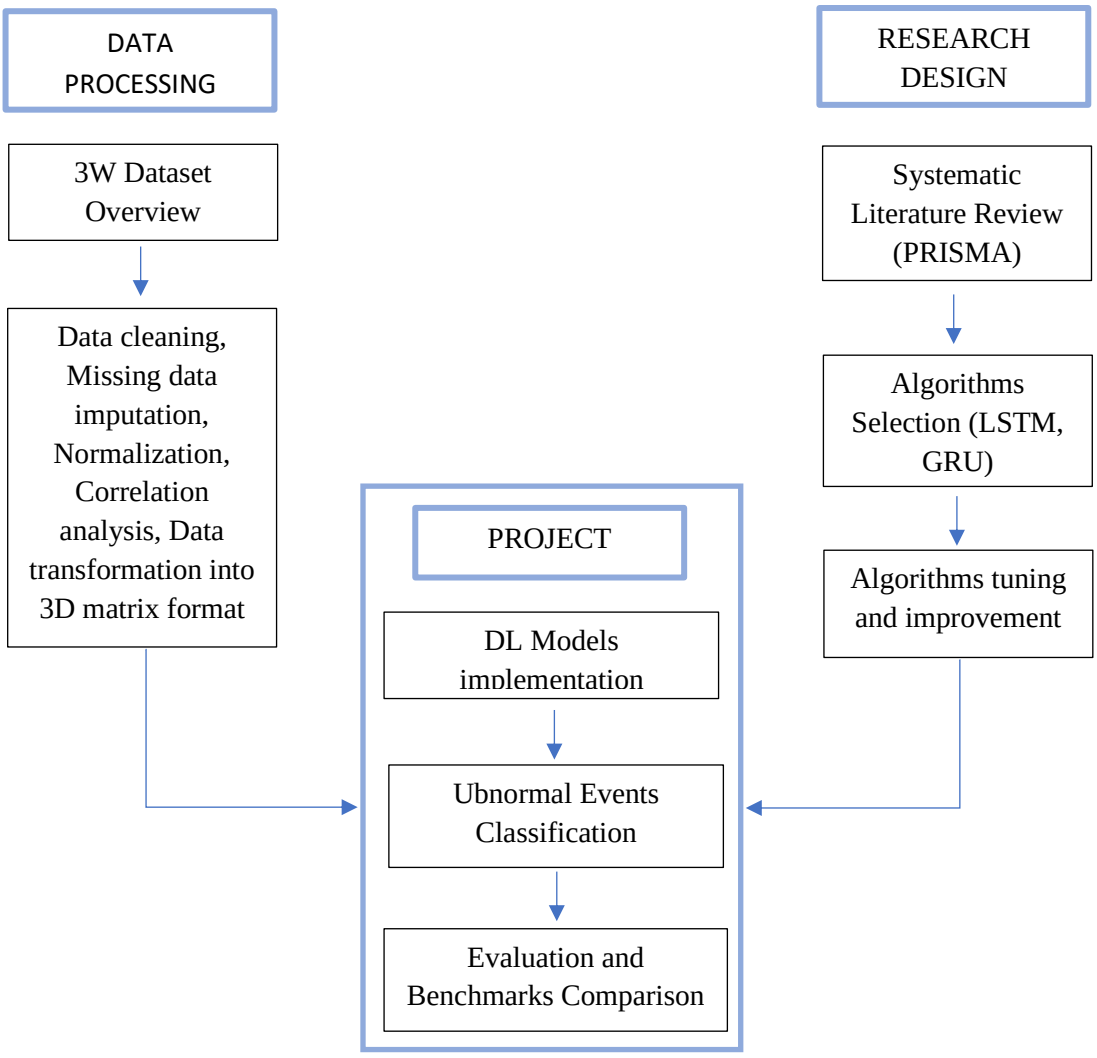


Figure 3.2 - Methodology overview

4. PROJECT

The major goal of the 3W Petrobras project is development of a new automated AEM (Abnormal Event Management) with machine learning algorithms, for which the 3W dataset was created by compiling real data from 21 wells during actual operations from 2012 to 2018 (Vargas et al., 2019b). The naturally flowing wells were selected as being less complex and more suitable for research and innovation in predictive maintenance. As displayed in Figure 4.1, the types of undesirable events, which are a focus of the project, account for most of the production losses, hence it is highly desirable to detect their start at the earliest opportunity and to take appropriate measures to mitigate or eliminate adverse scenarios.

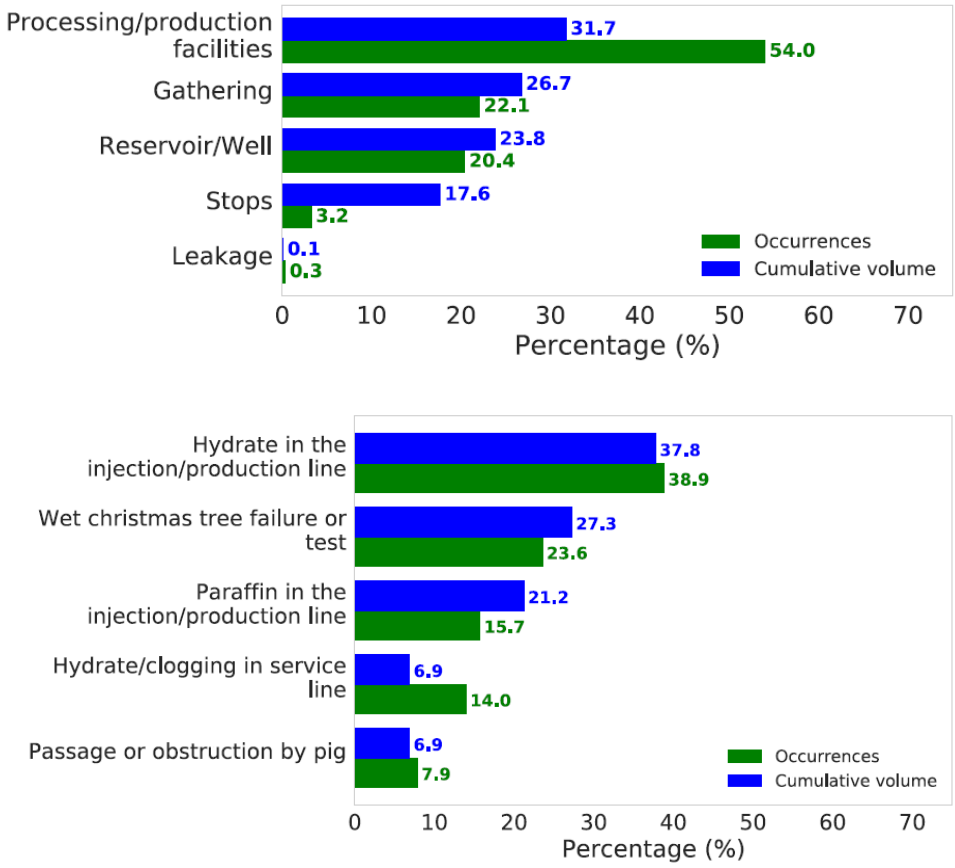


Figure 4.1 - Breakdown of cumulative oil volume loss (blue bars) and corresponding number of failures (green bars) between 2014 and 2017 for Petrobras (Marins et al., 2021).

4.1. 3W DATASET

Naturally flowing wells are those in which the formation pressure is sufficient to produce oil at a commercial rate without requiring a pump. Most reservoirs at the initial stage of development have sufficient pressure for a natural flow and thus require less equipment and automation for control and also for successful oil and gas production. Figure 4.2 presents the basic schema of an offshore platform connecting to a subsea Christmas tree through a production line and subsequently to production tubing and the reservoir itself. Subsea Christmas trees are a complex assembly installed on top of

wellhead to monitor and control the production, whilst being operated through an electro-hydraulic umbilical.

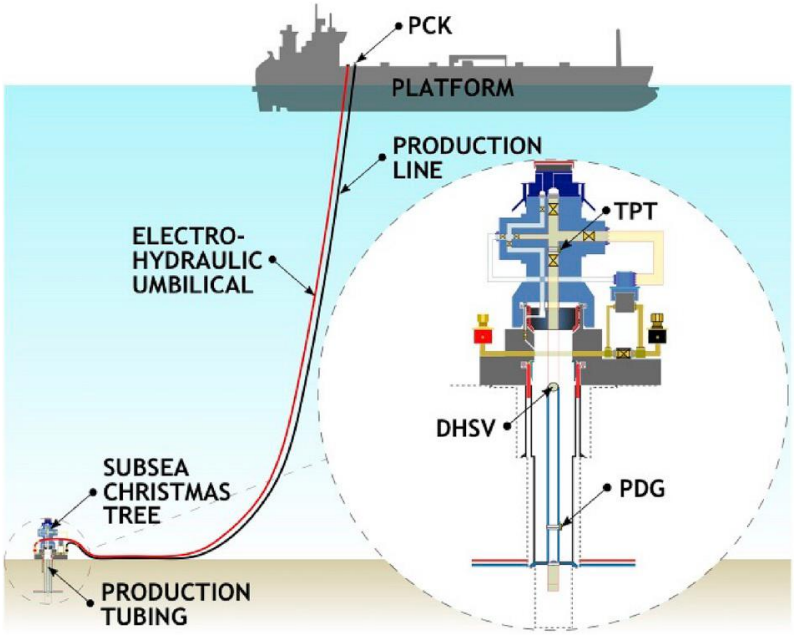


Figure 4.2 - Simplified schematic of a typical offshore naturally flowing well (Vargas et al., 2019b).

The 3W dataset combines measurements from topside and subsea sensors: located in the production tubing (P-PDG), on the subsea Christmas tree (P-TPT and T-TPT), the production line (P-MON-CKP and T-JUS-CKP), and the gas lift line (P-JUS-CKGL, T-JUS-CKGL, and QGL) (Santos et al., 2021) (Table 4.1).

Table 4.1 - 3W dataset variables

Number	Tag	Name	Unit
1	P-PDG	Pressure at the PDG	Pa
2	P-TPT	Pressure at the TPT	Pa
3	T-TPT	Temperature at the TPT	deg°C
4	P-MON-CKP	Pressure upstream of the PCK	Pa
5	T-JUS-CKP	Temperature downstream of the PCK	deg°C
6	P-JUS-CKGL	Pressure downstream of the GLCK	Pa
7	T-JUS-CKGL	Temperature downstream of the GLCK	deg°C
8	QGL	Gas lift flow rate	sm^3/s

The 3W dataset is organized into folders according to the type of fault, with each event progressing from normal operation to transient condition, following through to a steady-state anomaly. The 8 types of recognized and labeled events are:

- Class 0 – Normal operation
- Class 1 – Abrupt Increase of Basic Sediment and Water (BSW) – suspended water, sediments and other impurities in the production measured as a percentage of the production stream (The SLB Energy Glossary | Energy Glossary, n.d.). The lifecycle of each well contains periods of increasing level of BSW, however an unexpected rise of it indicates a developing production issue, which needs to be remedied quickly.
- Class 2 – Spurious Closure of the Downhole Safety Valve (DHSV) – the valve isolates wellbore fluids in the event of a catastrophic failure of surface equipment (The SLB Energy Glossary | Energy Glossary, n.d.). In case the valve fails in a spurious manner without any surface signs, it needs to be reopened, hence the automatic event identification is essential.
- Class 3 – Severe Slugging – an event in which a sequence of liquid slugs is followed by large gas bubbles. It is a cyclical phenomenon, that can lead to wellhead and pipeline damage, hence it is considered as a critical type of abnormality (Vargas et al., 2019b).
- Class 4 – Flow Instability – pressure changes within acceptable thresholds, with differences due to slugging which represent absence of cyclicity. This event can transform into slugging and then a severe variant, which requires imminent actions (Vargas et al., 2019b).
- Class 5 – Rapid Productivity Loss – flow loss due to changes in reservoir static pressure, with alternating BSW percentage, production viscosity and changes in production line diameter, etc. (Vargas et al., 2019b).
- Class 6 – Quick Restriction in the Production Choke (PCK) – a term used by Petrobras to indicate issues with a PCK valve, which is installed at the beginning of the production line. When it is operated manually, short restrictions might be observed due to operational problems, which need to be identified and reversed (Vargas et al., 2019b).
- Class 7 – Scaling in PCK – a mineral deposit, which can create a significant restriction or even a plug in the production tubing (The SLB Energy Glossary | Energy Glossary, n.d.). Thus, monitoring the production choke is helpful for recognizing the event and taking appropriate actions, such as scale inhibitor injections (Vargas et al., 2019b).
- Class 8 – Hydrate in Production Line – compounds of complex ions formed by water and other substances, at reduced temperatures and high pressure, which might lead to plugging of the pipelines (The SLB Energy Glossary | Energy Glossary, n.d.). It is one of the biggest issues in oil and gas production and which can stop flow for a long period, hence it needs to be recognized immediately.

These faults might interact with each other, which might create a difficulty in identifying one of them, or one fault might trigger another one from a different class (Marins et al., 2021).

Two types of labelling are implemented on two levels: first by instance (which is a file within each folder, be it real, simulated or hand-drawn) and second by observation (each row within each file also has a label according to the event).

The real ones were obtained from the real wells, the simulated were generated by Schlumberger through OLGA system (OLGA Dynamic Multiphase Flow Simulator, n.d.), the hand-drawn were produced by the 3W database creators using expert knowledge, so that the data mimics a typical sensor reading of the particular event type (Figure 4.3). However, for anomaly detection only real instances with an undesirable event of a normal period (1, 2, 5, 6, 7 and 8) longer or equal to 20 minutes must be used (Vargas et al., 2019c).

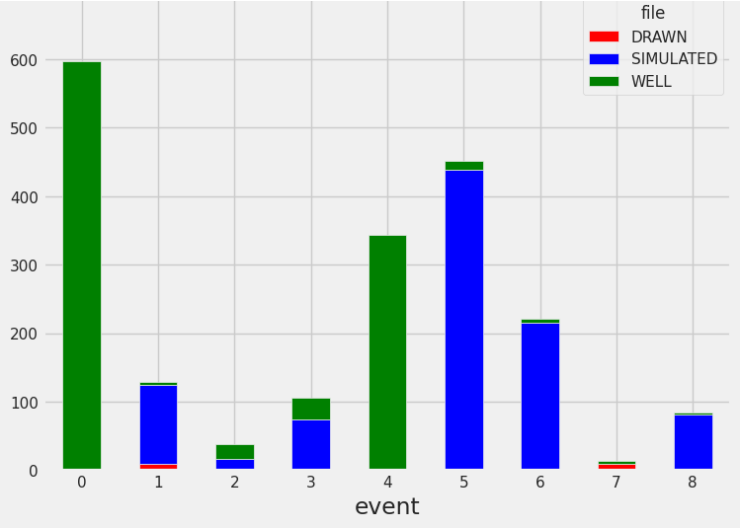


Figure 4.3 – The number of instances in the 3W Dataset

As depicted in Figures 4.4 and 4.5, each observation is labelled according to the three periods as normal, faulty transient and faulty steady state. The faulty transient state is characterized by the development of undesirable events, but still not reaching a failure condition, and is labelled by three digits with the last one corresponding to the event label (for example, 105 as faulty transient and 5 as steady state fault).

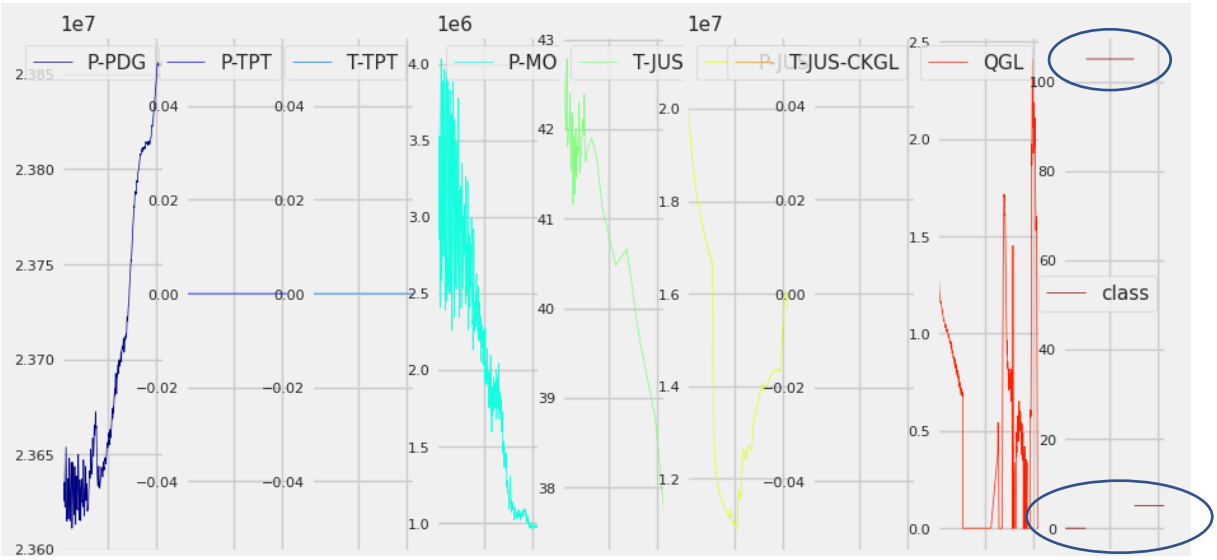


Figure 4.4 - Class 5 time series of the WELL-00015 instance

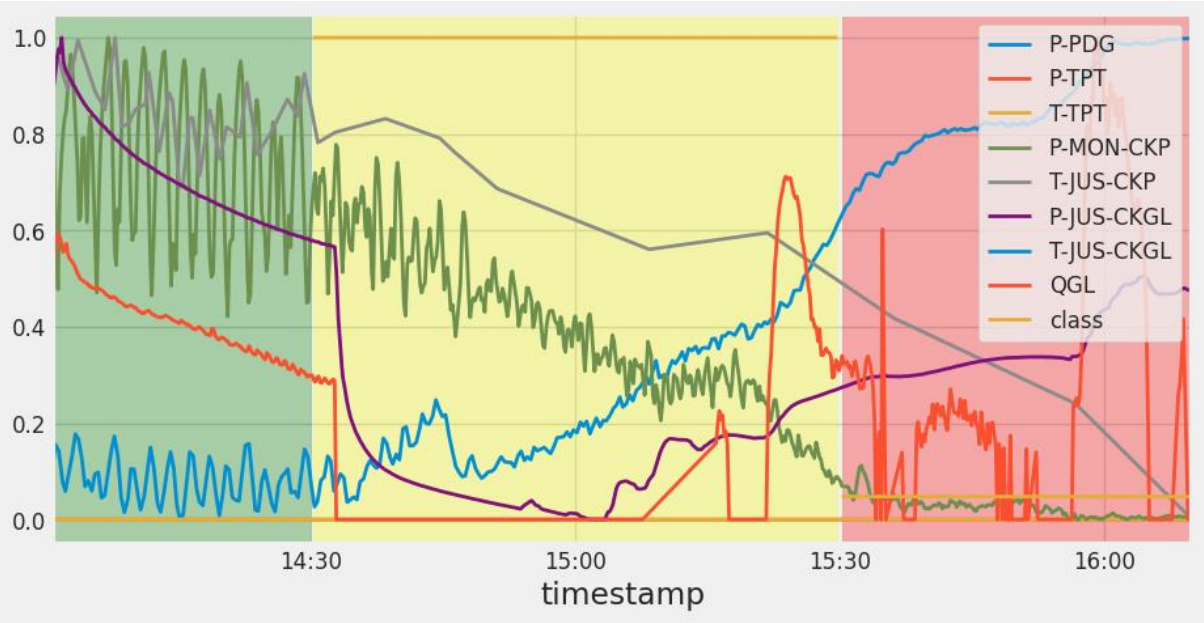
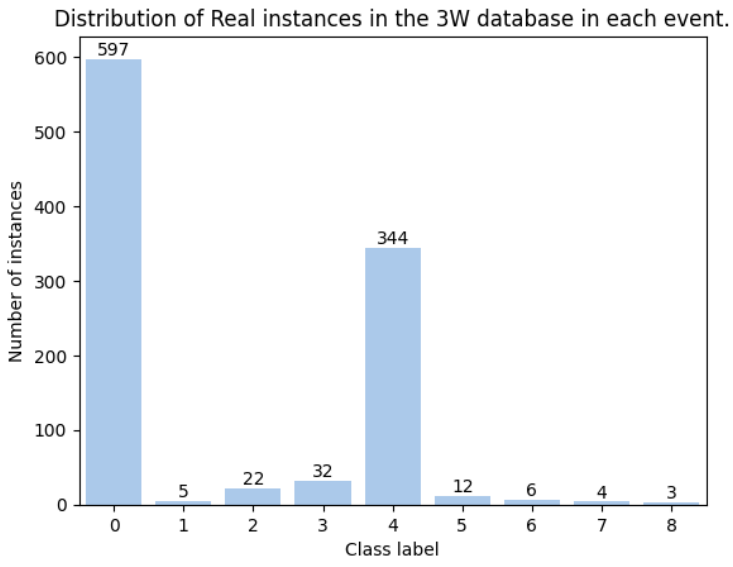


Figure 4.5 - Class 5 time series of the WELL-00015 instance with observations Normal (green), Faulty Transient (yellow) and Faulty Steady State (red)

4.2. DATA PREPROCESSING

Since the objective of this project is anomaly detection, only real instances were considered for the analysis. All the Simulated and Drawn instances were ignored, leaving only files, that start with “Well”.

With the removal of synthetic data, it turns into a very imbalanced dataset, classes 0 and 4 being the majority classes and the rest a minority (Figure 4.6).



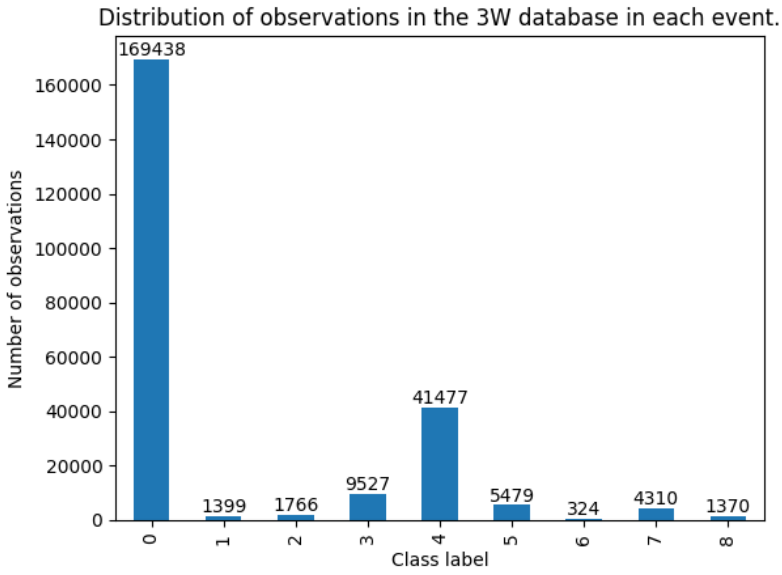


Figure 4.6 - Real instances and observations distribution according to fault events

For this project it is decided to treat faulty transient as faulty events, since they naturally progress into failure and this way they will be recognized sooner. First, all the csv files were combined into one file, converting all faulty transient classes into corresponding steady state faulty. Considering them as faulty events, the classification becomes multiclass classification with 9 classes identified. Also, while concatenating the files, they were down sampled to 1 minute to decrease the calculation time. As demonstrated in Figure 4.7, the final combined file shows the presence of multiple spikes and noise as well as frozen and missing data.

Since there are observations without class, they were deleted, and “class” type was converted to categorical.

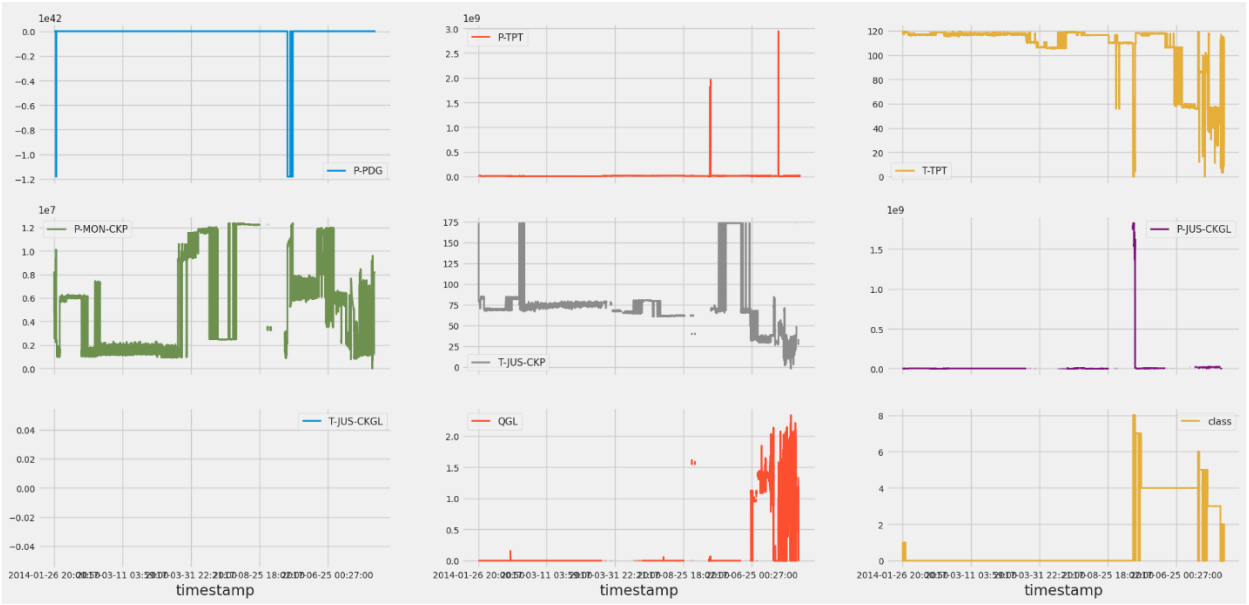


Figure 4.7 - Combined data visualization

The box plots identify that P-PDG has negative values, and other channels encounter many outliers. Their nature might be due to sensors readings being affected by jumps in temperature and pressure, which is common for oilwell operations. Despite that, some of them might indicate the start of failure, and care needs to be taken with their treatment or removal (Figure 4.8).

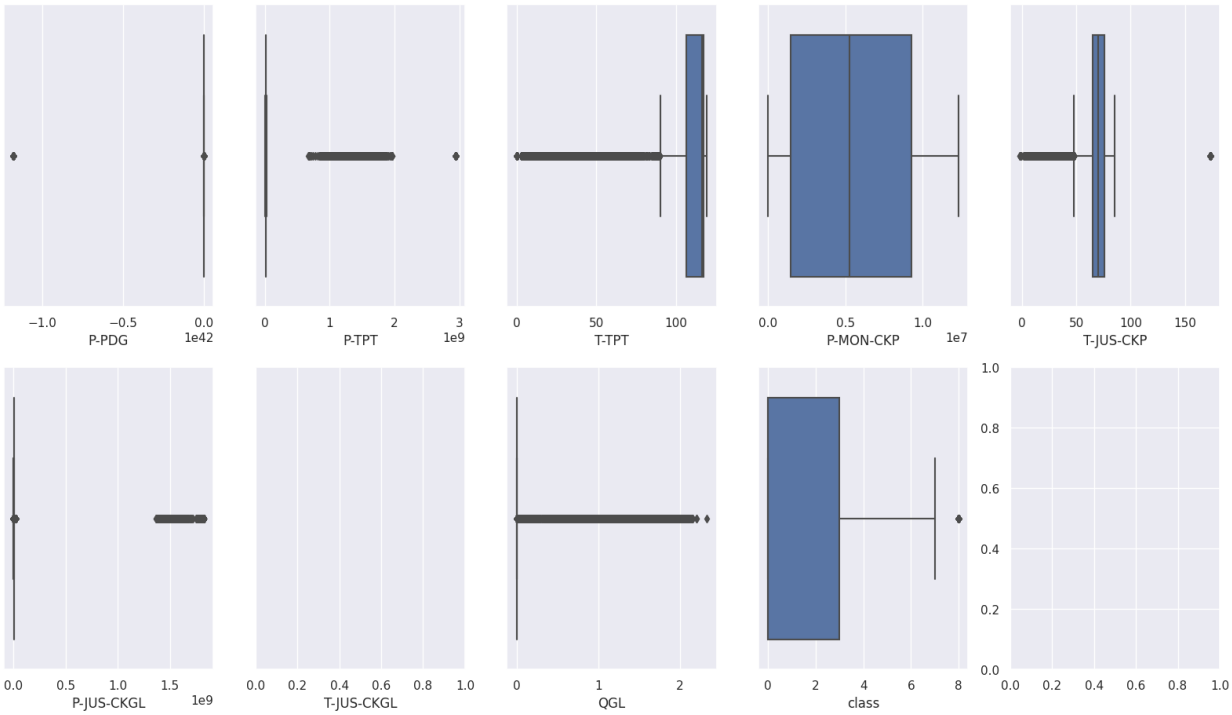


Figure 4.8 - All features box plots

For the initial attempt, the outliers were replaced by their corresponding lower and upper limits using quantile ranges of 0.1 and 0.9. Then, since variable T-JUS-CKGL has no data, and QGL is a frozen channel with value 0, they were both dropped.

The correlation heatmap of remained data shows, that two pairs of variables have high correlation (Figure 4.9):

- P-MON-CKP and P-TPT (0.83) - Pressure upstream of the PCK and Pressure at the TPT
- T-JUS-CKP and T-TPT (0.9) - Temperature downstream of the PCK and Temperature at the TPT.

These correlations are valid, since they represent pressure and temperature at the subsea Christmas tree and at the Production Choke (PCK), which are connected by the production line. One of each channel can be dropped, or dimensionality reduction methods could be implemented, however, since the number of remaining variables is just 6, they are all retained for further analysis.

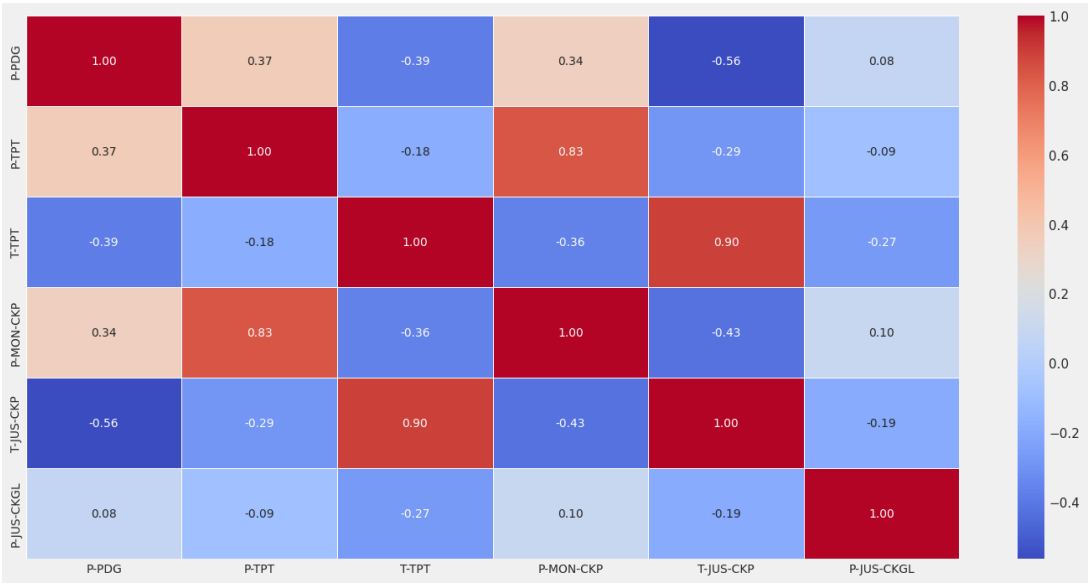


Figure 4.9 - A heatmap visualization of the correlation matrix

The missing values were filled by the “forward fill” method, in which the last valid observation is propagated forward. The final processed dataset is saved and divided into training and testing sets, stratifying by y to ensure that relative class frequencies are approximately preserved in each train and test split. The histogram-distribution, box plots and relationship between variables and classes of the final processed data are shown in Figures 4.10, 4.11 and 4.12 respectively.

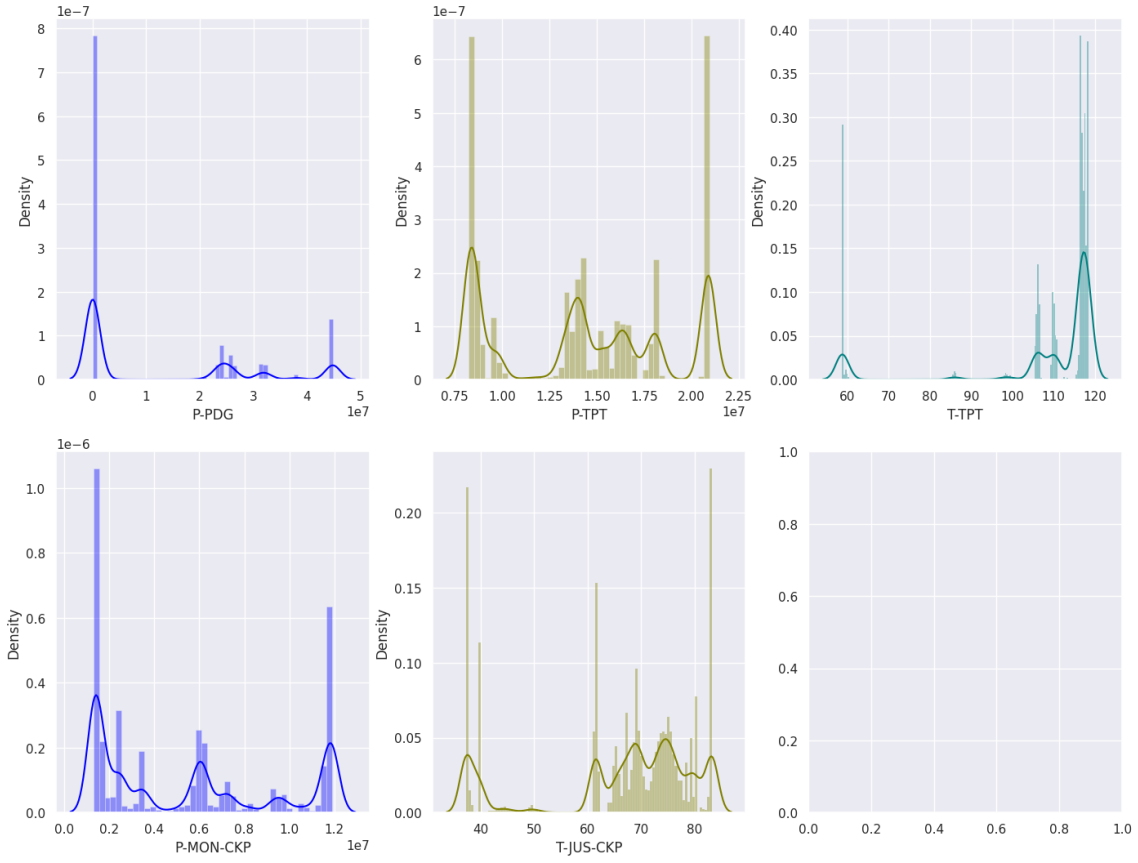


Figure 4.10 - Histogram - distributions of the final processed data

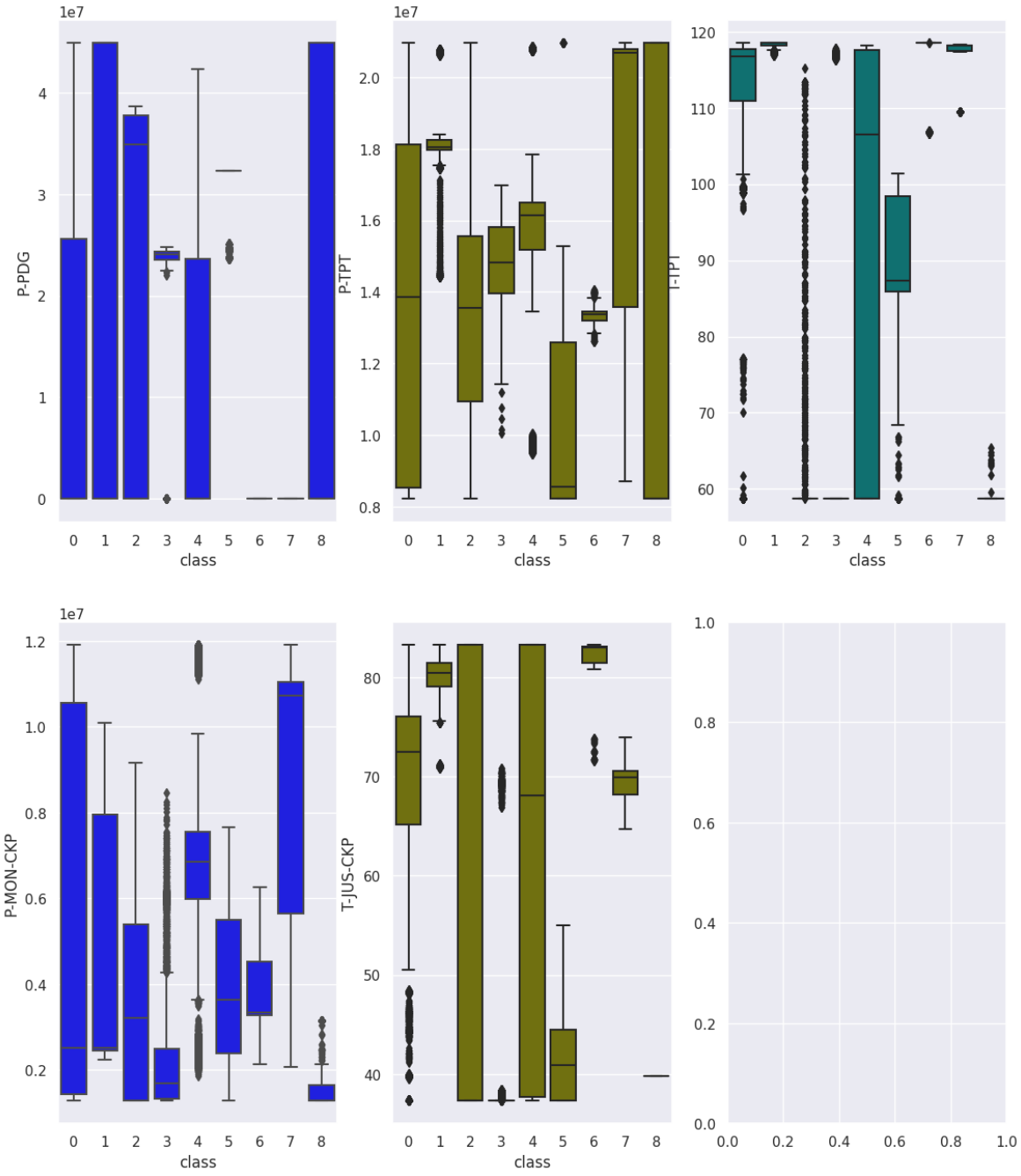


Figure 4.11 - Box Plots of the final processed data

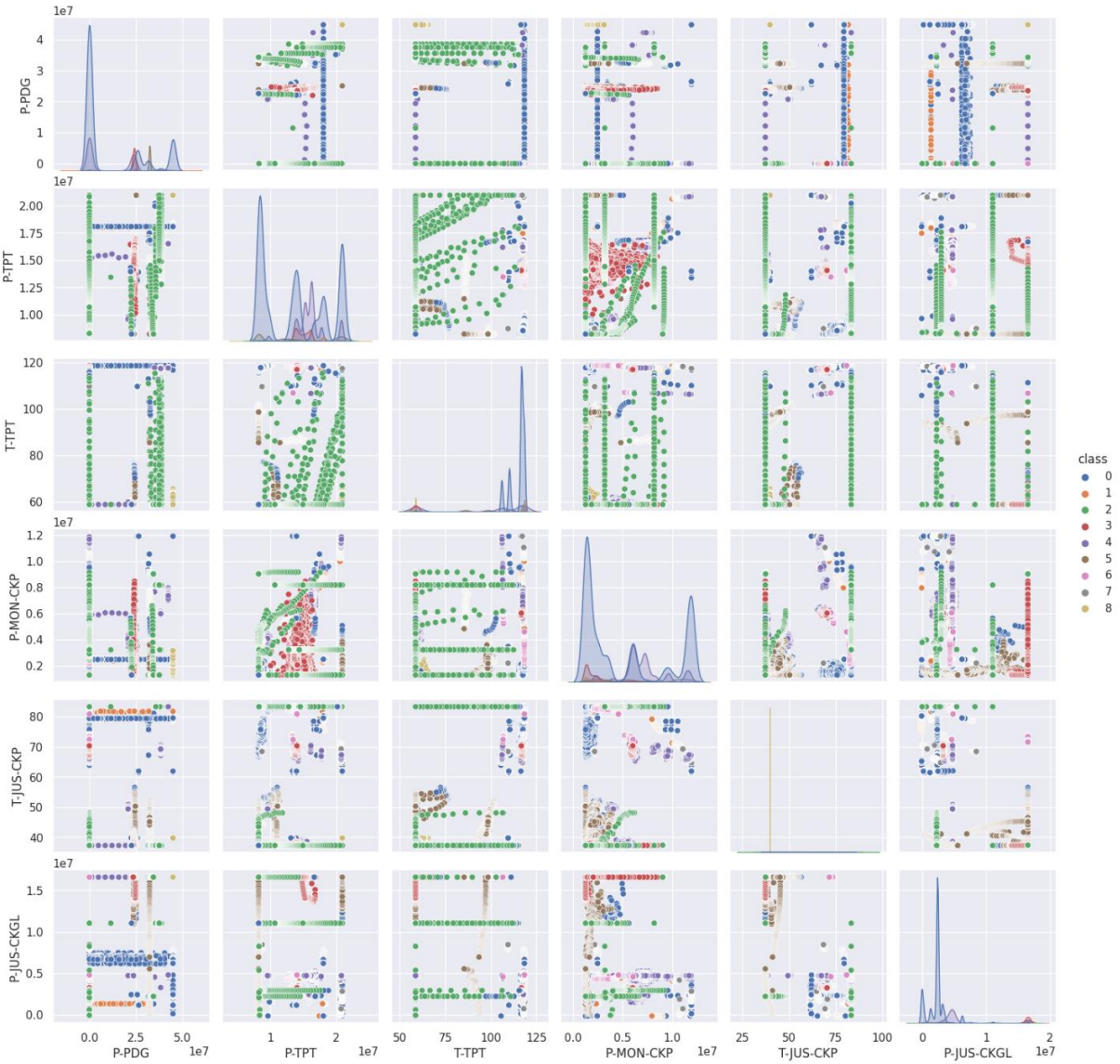


Figure 4.12 - Relationship between variables and classes

There are many methods to handle imbalanced data, such as Random Oversampling, SMOTE, BorderLine SMOTE, KMeans SMOTE, SVM SMOTE, ADASYN, SMOTE-NC, etc.(Satyam Kumar, 2020). One of the newest techniques, which was just recently developed, is CLUBS, standing for Clustering of Lower and Upper Boundaries standardization, based on examining dissimilarity correlations between classes and creating synthetic samples for the minority classes (Michele Lanni et al., 2020).

For this project, to tackle the imbalanced data issue, the train set was augmented using SMOTE method, since it is the most basic and simplest method, and all the further data processing can easily become computationally expensive. If it significantly improves the classification, then potentially other methods can be attempted too. With SMOTE the data was augmented to the size of the X-train and y-train from 164563 rows to 1067463. Then the data was scaled with StandardScaler and passed for further data transformation.

Since Recurrent Neural Network (RNN) requires a 3D format [samples, timesteps, features], the data was converted into a 3D matrix with the window size equal to 30 as an initial experiment.

4.3. ALGORITHMS

For the project two algorithms were selected as being potentially powerful enough for resolving such tasks and were currently un-explored, according to the literature review, namely Recurrent Neural Network configurations with LSTM and GRU architectures.

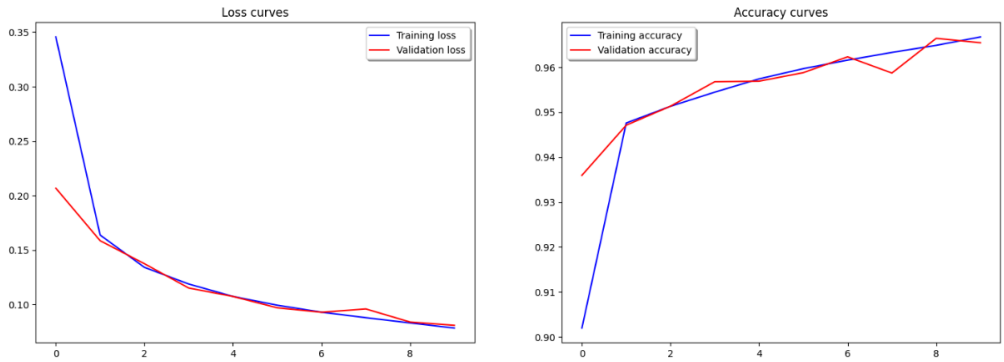
The first attempt of an LSTM algorithm was run with the original processed data (without train set being transformed by SMOTE) to evaluate the initial classification. The network had just 2 stacked LSTM layers and one Dense layer, with the number of hidden units equal to 10, an activation function "tanh", a batch size equal to 30 and 10 epochs, which resulted in a macro average F1 score equal to 0.75%.

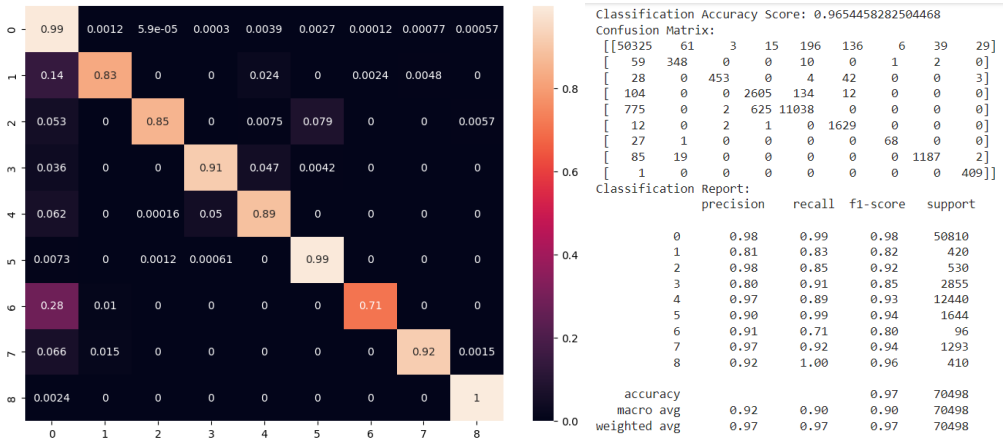
To understand the general response of different backbones and their parameters, the algorithms were run with the same batch size, timestep and number of epochs, but with activation functions "relu", "softmax", "LeakyReLU" or "swish", with 10 or 20 number of hidden units, and a different number of LSTM or GRU layers, applied both before and after SMOTE oversampled train data (Table 4.2).

Table 4.2 - Deep Neural Networks model architectures F1 scores

Model architecture	Before SMOTE F1	After SMOTE F1
Stacked 2 LSTM and 1 Dense, 10 hidden units, activation "tanh"	0.75	0.90
Stacked 2 LSTM and 1 Dense, 10 hidden units, activation "softmax"	0.19	0.71
Stacked 2 LSTM and 1 Dense, 10 hidden units, activation "relu"	0.85	0.88
Stacked 2 LSTM and 1 Dense, 20 hidden units, activation "relu"	0.90	0.92
Stacked 2 LSTM and 1 Dense, 10 hidden units, activation "LeakyReLU"	0.83	0.91
Stacked 2 LSTM and 1 Dense, 20 hidden units, activation "LeakyReLU"	0.85	0.85
Stacked 2 LSTM and 1 Dense, 10 hidden units, activation "swish"	0.87	0.88
Stacked 3 LSTM, 1 RepeatVector, 1 Dense, 10 hidden units, activation "relu"	0.82	0.90
Stacked 2 GRU and 1 Dense, 10 hidden units, activation "LeakyReLU"	0.86	0.87
Stacked 2 GRU and 1 Dense, 20 hidden units, activation "LeakyReLU"	0.90	0.92

The best result was achieved by models with LSTM and GRU backbones of 2 layers with 20 hidden units each, giving a F1 score equal to 0.90% before SMOTE (Figure 4.13) and 0.92% after SMOTE.






```
Classification Accuracy Score: 0.9654458282504468
Confusion Matrix:
 [[50325    61     3    15   196   136     6    39    29]
 [   59   348     0     0    10     0     1     2     0]
 [   28     0   453     0     4    42     0     0     3]
 [  104     0     0  2605   134    12     0     0     0]
 [  775     0     2   625 11038     0     0     0     0]
 [   12     0     2     1     0  1629     0     0     0]
 [   27     1     0     0     0     0    68     0     0]
 [   85    19     0     0     0     0     0  1187     2]
 [    1     0     0     0     0     0     0     0   409]]
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.99      0.98     50810
           1       0.81      0.83      0.82       420
           2       0.98      0.85      0.92       530
           3       0.80      0.91      0.85      2855
           4       0.97      0.89      0.93     12440
           5       0.90      0.99      0.94      1644
           6       0.91      0.71      0.80        96
           7       0.97      0.92      0.94      1293
           8       0.92      1.00      0.96       410

    accuracy                           0.97     70498
   macro avg       0.92      0.90      0.90     70498
weighted avg       0.97      0.97      0.97     70498
```

Figure 4.13 - LSTM with “relu” activation and 20 hidden units

To summarize the observations from the initial algorithms settings, it is recognized that an increased number of layers doesn’t improve the results. To the contrary, it makes them worse. A higher number of hidden units in each layer helps in increasing the classification metrics. However, it is not clear, whether there is an optimal number, that would significantly affect the results, or whether 20 units is a plateau value, which has already produced the maximum possible F1 score. To investigate further effects of hyperparameter settings, their optimization was performed using Random Search, Hyperopt and Genetic Algorithms in order to identify the best model for classification.

4.4. HYPERPARAMETERS OPTIMIZATION

The following hyperparameters were selected for the model optimization:

- timestep (or window size)
- number of hidden units of each layer
- number of epochs
- batch size.

There are many other parameters that could be optimized, such as learning rate, activation function, optimizer type, number of layers. Some of them have already been attempted for visibility and transparency of the model performance: both LSTM and GRU backbones were implemented with various activations, and several network layers were also built, which showed poorer results. For the purpose of making computations less expensive, only quantitative parameters were estimated, and RNN was run with just the GRU backbone, since it is much faster and produced similar to the LSTM results. All attempts were made on original without oversampling with SMOTE data for the abovementioned reasons.

The traditional GridSearch CV was not selected, since it works by trying every possible combination of parameters and can get very resource intensive. However, another standard method Random Search was attempted, despite its drawbacks of potentially missing important points in the search space.

4.4.1. Random Search

For the Random Search method, the parameters search space was selected without limiting the step between the minimum and maximum range, thus giving more freedom of choice. The algorithm was run with the objective of identifying parameters with a resultant minimum validation loss, a batch size of randomly selected parameters set to equal 10, and a number of iterations to 1. The identified optimum parameters are a timestep equal to 23, number of units equals 40, epochs of 16 and a batch size 24, which produced an improved result of F1 = 0.94% (Figure 4.14).

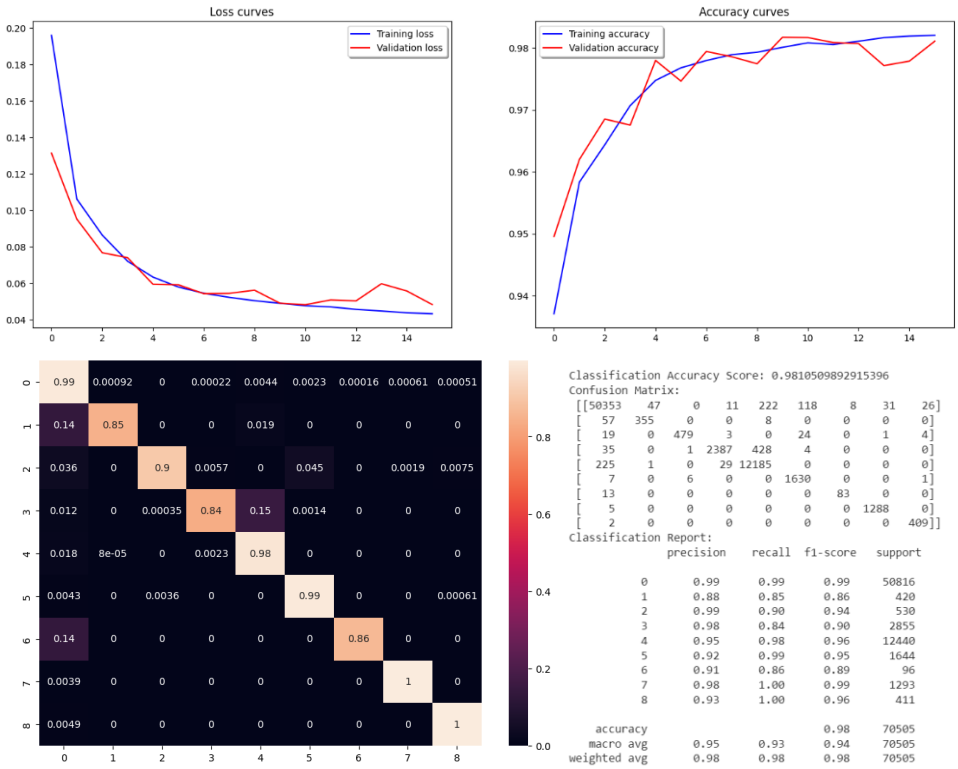


Figure 4.14 - Random Search optimized model with F1 = 0.94%

4.4.2. Hyperopt Optimization

Next, a hyperparameters tuning technique Hyperopt was applied, which uses a form of Bayesian optimization. Only 3 trials were implemented, since it takes a very long time to run the process (more than 8 hours), and the search space was set limited by the steps for each parameter within the selected range, in order to decrease the scope of settings to evaluate (Figure 4.15). The objective was again to minimize the validation loss, and the best parameters were run for the final evaluation.

As shown in Figure 4.16, the maximum F1 score achieved was 0.94% with a window size of 10, the number of hidden units 20, number of epochs 20 and a batch size of 10, which was an improvement from the after SMOTE result of 0,92%, but is similar to Random Search method achievement.

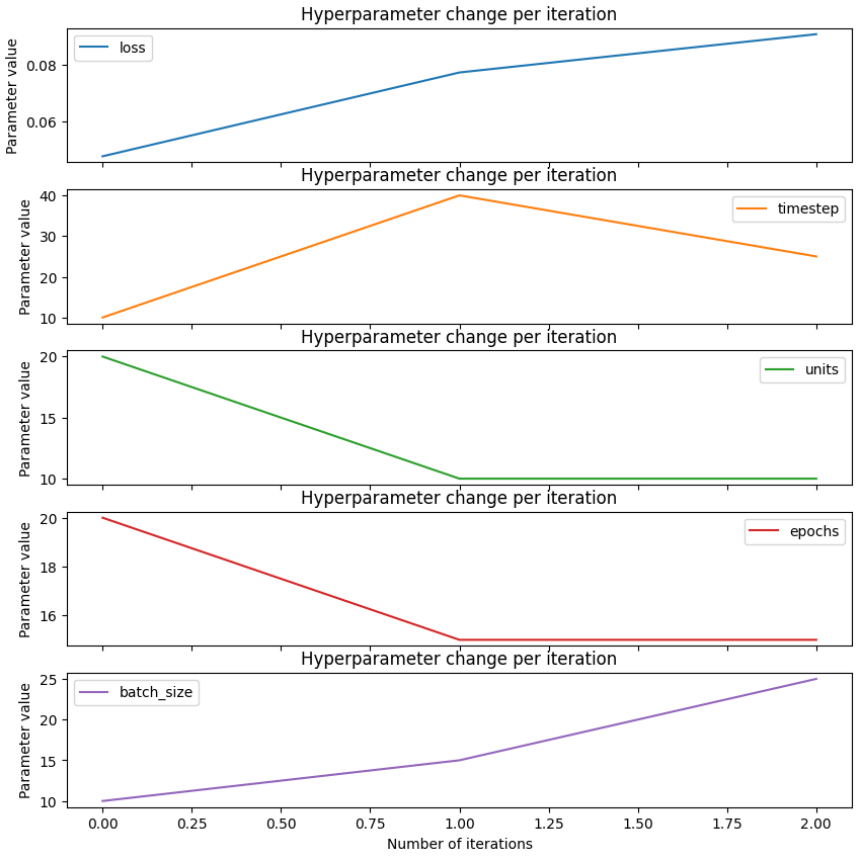


Figure 4.15 - Hyperopt hyperparameters change per each iteration

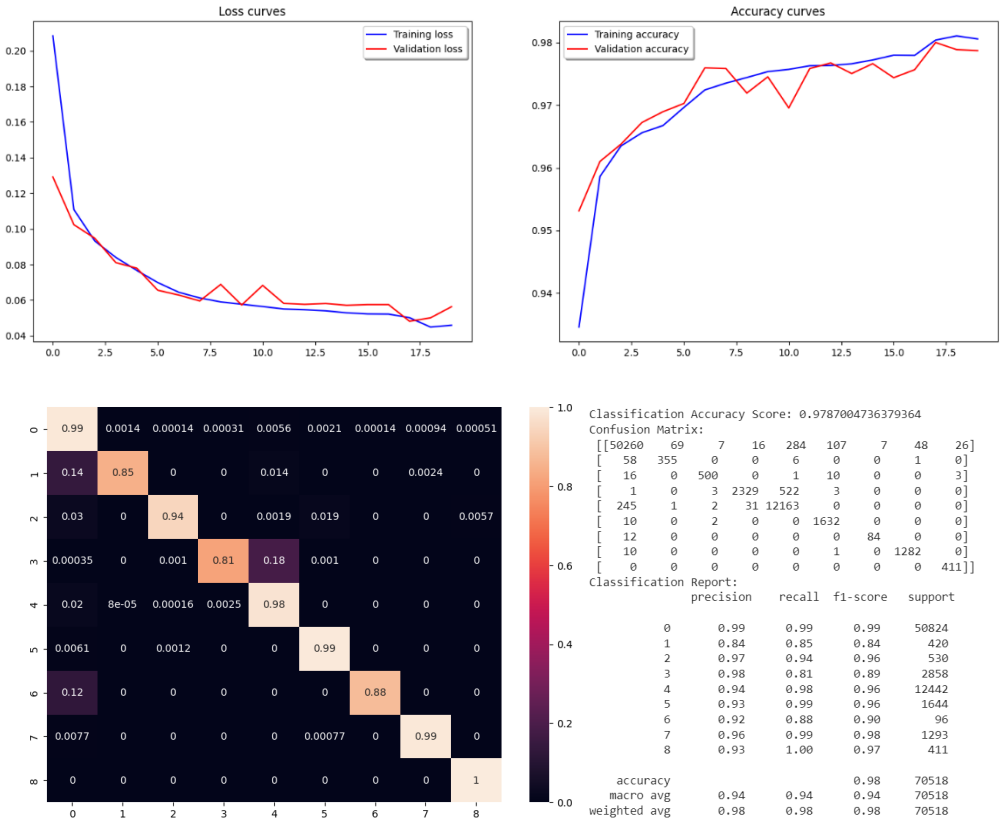


Figure 4.16 - Hyperopt optimized model with F1= 0.94%

4.4.3. Genetic Algorithms

Genetic Algorithm is an extensively used method for hyperparameter optimization, which applies evolution by natural selection and is highly inspired by Charles Darwin's theory of evolutionary biology. The idea that individuals with higher survival potential and better adaptation to the surrounding environment conditions will have a higher probability of existence and passing their genes to further generations, is implemented by creating individuals, representing hyperparameters selection. Children with better qualities will pass their chromosomes to their children, and so in generation after generation only the strongest and fittest candidates will survive and become a global optimum.

The advantage of Genetic Algorithm over other standard methods is the absence of the requirement for an exhaustive analysis of the search space (Vanneschi & Silva, 2023), which is impractical, considering the huge size of the potential combinations of variables. An iterative process of improving individuals' initial random population, utilizing nature-inspired concepts, such as selection, crossover and mutation, creates a population with the highest fitness, which would be the best solution and can be utilized for further model tuning.

For hyperparameter optimization, each individual is a collection of decimals, representing phenotypes, based on which fitness is evaluated and, as it happens in nature, further selection is performed. To apply genetic operators, each phenotype is converted to a genotype, that is completely independent of fitness and has a binary representation of 0 and 1.

4.4.3.1. Genetic Algorithm 1

There are a few pre-built Genetic Algorithm methods, that are extensively used, such as TPOT, PyGAD, DEAP (Distributed Evolutionary Algorithms in Python), Neuroevolution optimization, etc.

For this project, the DEAP framework was selected, since it provides a unique evolutionary algorithm, that simplifies each step with its toolbox – a container of tools for all sorts of initializers and genetic operators (Overview — DEAP 1.3.3 Documentation, n.d.).

Each individual was encoded into binary string of bit length 26, with the timestep equal to 8 bits, the number of hidden units 6, epoch 5 bits and a batch size of 7 bits. The gene initialization values are chosen as the most appropriate for representing the decimal values of the hyperparameters.

To identify the best model settings, the DEAP toolbox was run for 5 generations with a population size of 5 each (Figure 4.17). The hyperparameters range was not set to have limits, as was done with Random Search and Hyperopt, since this method doesn't require an exhaustive search by iteration through the entire search space. As depicted in Figure 4.18, the variables could have extreme values, and the best model achieved F1 score 0.96% is with a window size equal to 1, the number of hidden units 53, with 29 epochs and a batch size of 25.

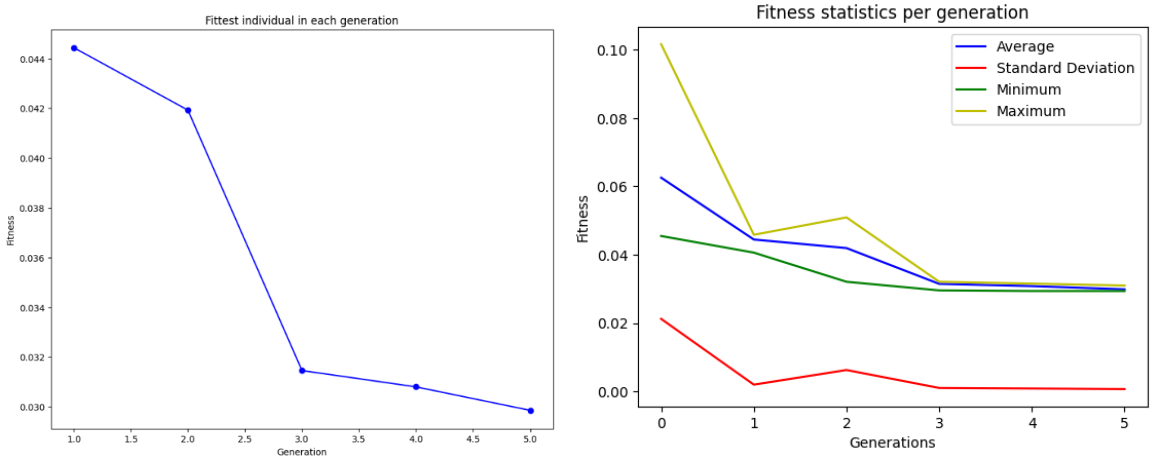


Figure 4.17 - DEAP Genetic Algorithm fitness and statistics per generation

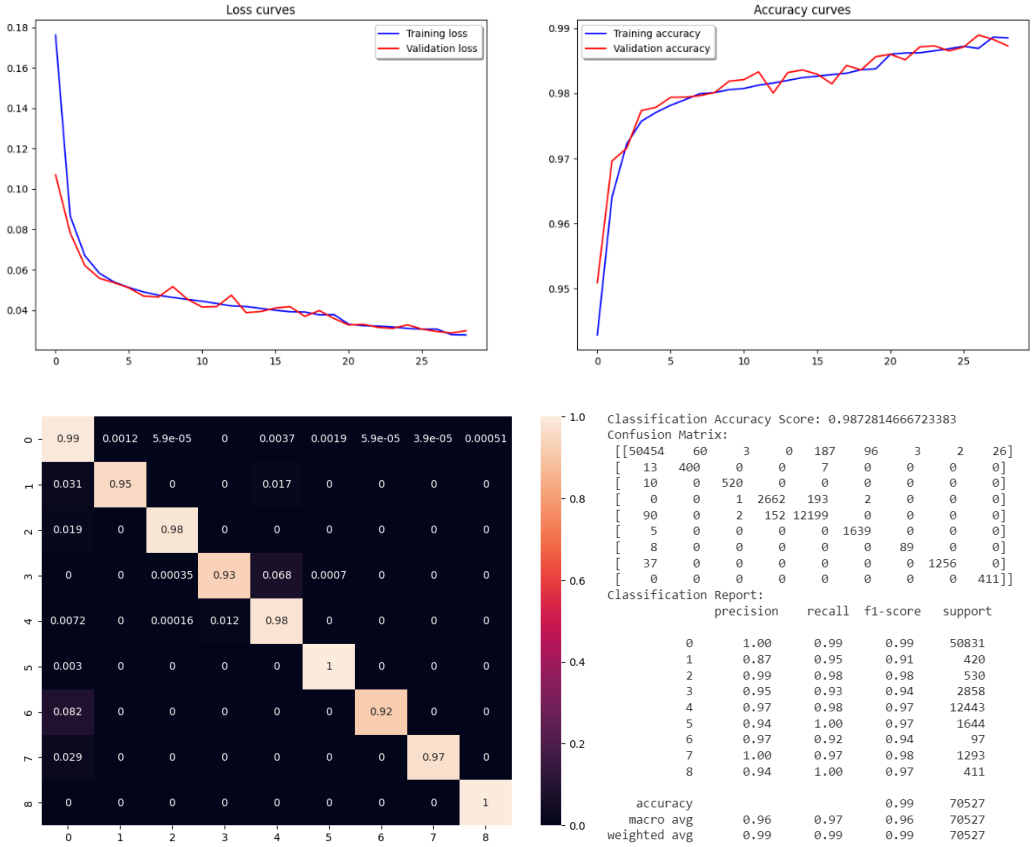


Figure 4.18 - DEAP Genetic Algorithm optimized model with F1= 0.96%

4.4.3.2. Genetic Algorithm 2

Another approach is creating the Genetic Algorithm by assigning all operators manually, which would give more transparency to the optimization process and an opportunity to tailor the process for the task. The initial set up is similar to previous experiment, with binary encoding, with each chromosome length equal to 26 bits. This time, the fitness function was selected as the F1 score of each model. The

evolutionary algorithm included selection of the fittest individual with maximum fitness in each population, one point crossover and mutation by flipping bits at the random change point.

The algorithm was run for 3 generations with 3 chromosomes in each population (Figure 4.19). It was highly desired to experiment with a higher number of individuals and iterations, however, due to technical limitations it was not feasible. As shown in Figure 4.20, the best result was achieved with timestep 125, the number of hidden units of 61, epochs 15 and a batch size of 99, resulting in a final F1 of 0.94%.

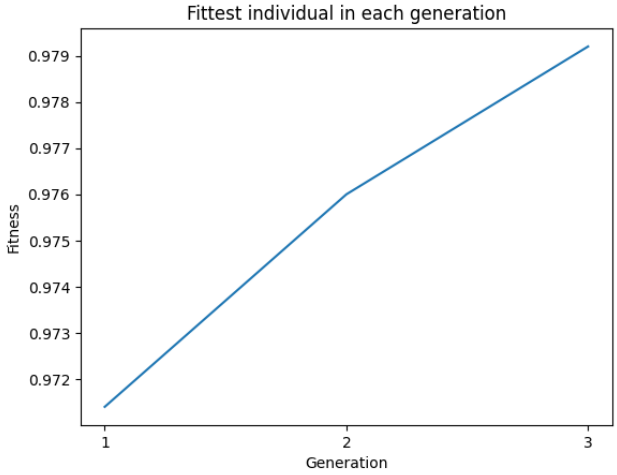


Figure 4.19 - Genetic Algorithm 2 fitness evolution per generation

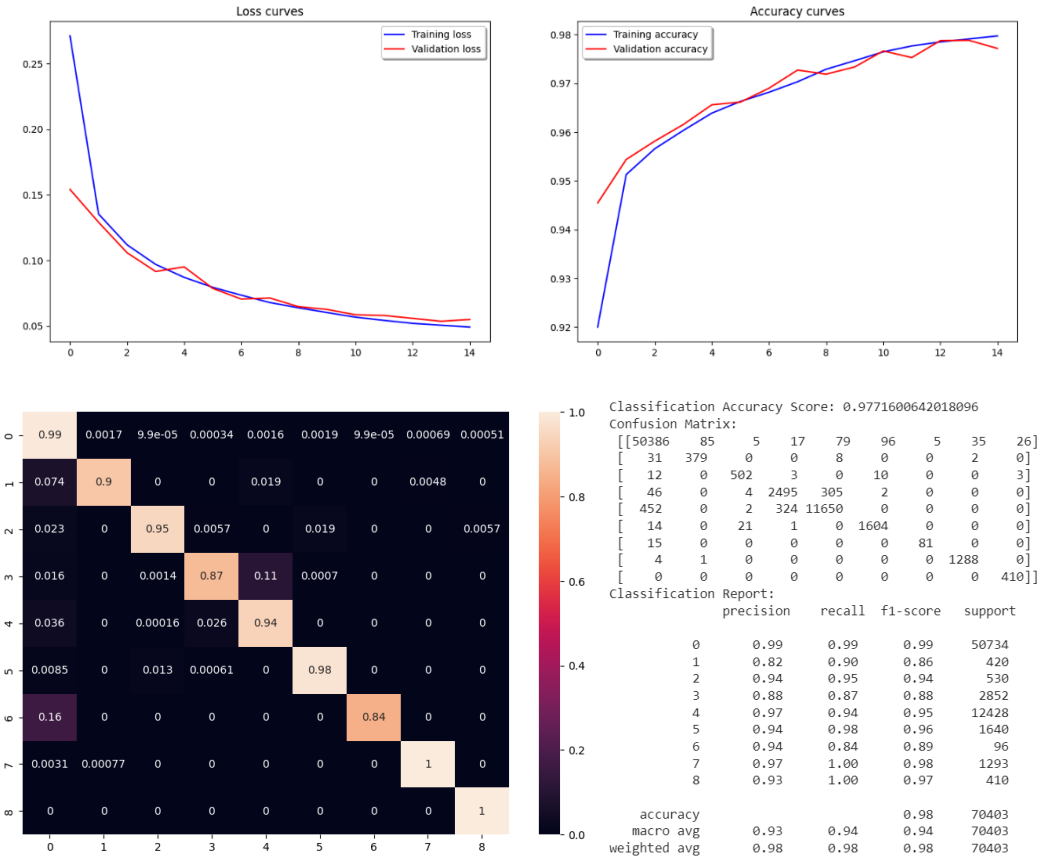


Figure 4.20 - Genetic Algorithm 2 optimized model with F1= 0.94%

The experiment results were worse than the first Genetic Algorithm, which could be due to the small number of iterations and individuals. As result, the evolution process was extremely limited, converging on the best individual among 3 chromosomes after 2 generations.

4.4.3.3. Genetic Algorithm 3

The third Genetic Algorithm experiment was performed with value chromosome encoding, in which each phenotype is represented as a string of direct hyperparameter decimal values. The fitness function was set as a validation loss of each model. The evolutionary operators included tournament selection with size 3, one point crossover and random resetting mutation methods. The algorithm was attempted to run many times, with 5 generations and 3 generations and a corresponding population size; however, it was revealed, that for successful evolution, the number of individuals should be sufficient for an increased chance of crossover and mutation. With a small number of chromosomes, the algorithm was selecting the same best individual and converging early, without attempting any other variations of hyperparameters. Increasing the number of generations and population size was not feasible, as it became extremely computationally expensive.

To overcome this issue, and considering Genetic Algorithm 1 results with DEAP toolbox, it was decided to decrease the timestep size to 1 and the number of epochs to 1 for the initial set up, which would speed up the process and allow for an increasing population size and the running of more generations. In this case the objective was to optimize the number of hidden units and batch size, which could be later implemented for the final model evaluation with an arbitrary number of selected of epochs.

The algorithm converged with the best model [1, 47, 1, 14] and which actually was the best model for all 5 generations. The variation in the best model fitness is due to the stochastic nature of RNN algorithms, however, the best fitness was still the smallest in each generation (Figure 4.21).

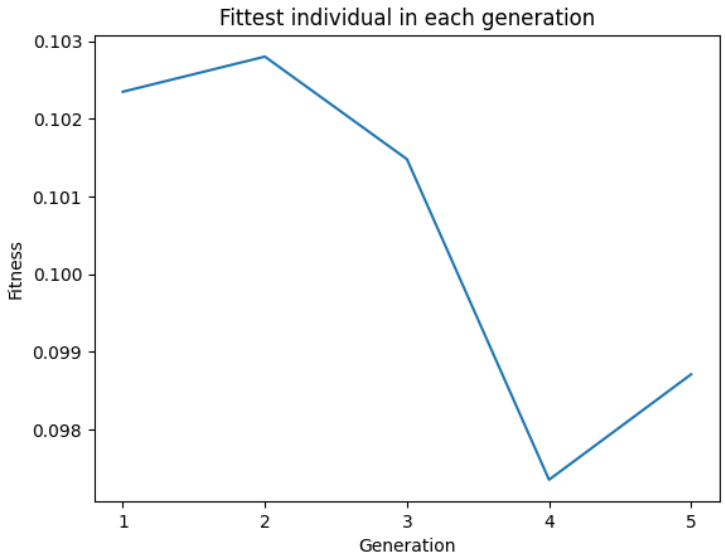


Figure 4.21 - Genetic Algorithm 3 fitness evolution per generation

To evaluate the final best model, it was run for an arbitrary 40 epochs, assigning the best model with a timestep of 1, the number of hidden units equal 47 and a batch size of 14.

As shown in Figure 4.22, the F1 score achieved is 0.97%, and it is the best result for all the algorithms.

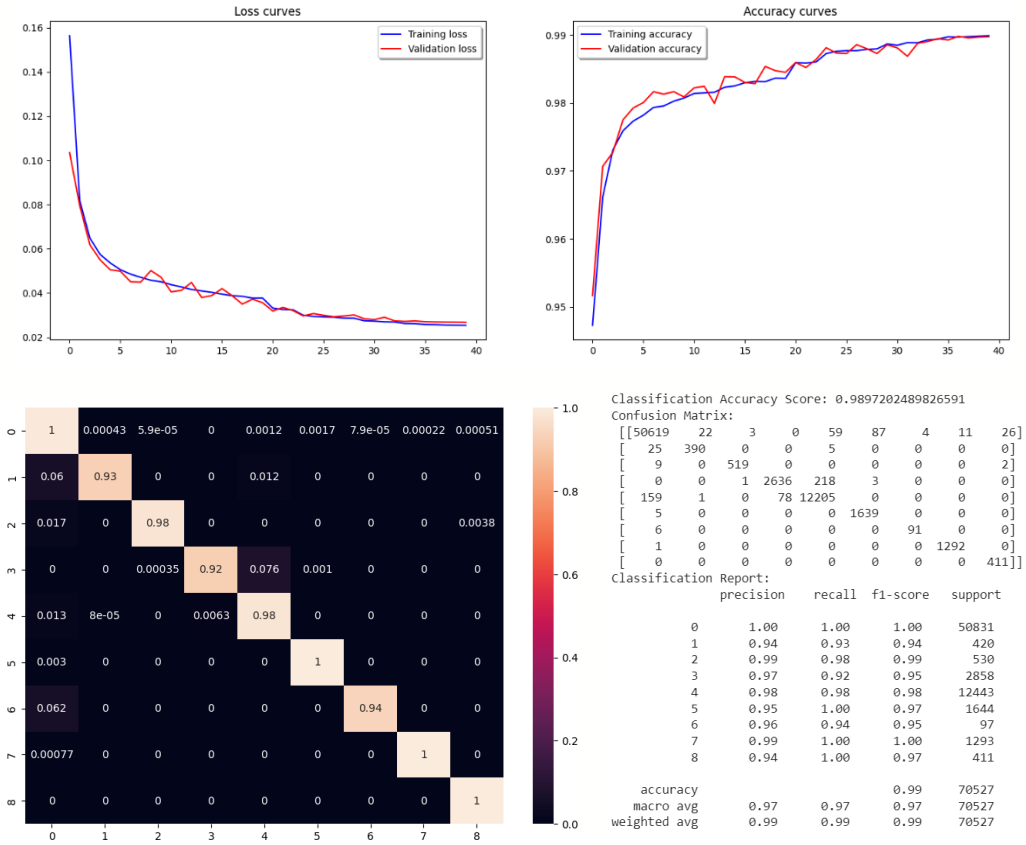


Figure 4.22 - Genetic Algorithm 3 optimized model with F1= 0.97%

4.5. DISCUSSION

The algorithms proposed in this study showed good performance for multiclass classification. Although only real instances were selected for the analysis, since this was a requirement for realistic benchmarking with other papers. The data also appeared to be heavily imbalanced, the RNN model variations with LSTM and GRU backbones identified all the undesirable events with F1 score 0.90% before oversampling and 0.92% after SMOTE.

Since there were many parameters, that could affect the deep neural networks performance, a few trials were initiated to evaluate their effect, such as number of layers, activation function and the number of hidden units. It was revealed, that LSTM and GRU produce similar results, with “relu” and “LeakyReLU” being the most efficient, but GRU performs much faster, which is an advantage for hyperparameter optimization. To speed up the process and make it less computationally expensive, all the algorithms were run with GRU backbone and “relu” activation function. The results are presented in Table 4.3, where the final F1 score per best model is calculated.

Table 4.3 - Hyperparameters optimization best results

Algorithm	Best model	F1 score
Random Search	[23, 40, 16, 24]	0.94
Hyperopt	[10, 20, 20, 10]	0.94
Genetic Algorithm 1	[1, 53, 29, 25]	0.96
Genetic Algorithm 2	[125, 61, 15, 99]	0.94
Genetic Algorithm 3	[1, 47, 40, 14]	0.97

The best F1 score was achieved with Genetic Algorithms, however, it is worth mentioning, that there were a few limitations, that potentially affected results. The first two Genetic Algorithms were performed with binary chromosome representation, so the upper and lower limits of each parameter was not set, but configured by selecting the number of bits for each gene. This allowed for an "accidental" application of the most extreme version of 3D GRU matrix selection, with a timestep equal to 1, i.e., selecting each single observation for classification, rather than a window of several observations. Despite producing good F1 score, it could potentially create an "overfitting" issue, if applied on new timeseries data.

Another limitation was the difficulty to run all Genetic Algorithms for more than 5 generations, and while DEAP allowed this to occur due to internal shortcuts through not running all individual evaluations, the second and third experiments clearly struggled to run till the end. Genetic Algorithm 2 particularly could not be run for more than 3 generations, which lead to poor individual crossover and mutation, and a convergence with best model on the 2nd generation. It could only be speculated, that with higher computational resources, this algorithm could have been run for many more generations and individuals in each population, possibly resulting in a higher F1 score.

The Genetic Algorithm 3, in which each chromosome was represented as decimal values, was also attempted to run for more than 3 generations. However, since it also failed to run until the end, the decision was made to amalgamate the DEAP finding, where the best timestep was equal to 1, and rerun the Genetic Algorithm 3 for just 1 epoch for each chromosome, but with an increased number of generations and individuals. This allowed for the simulation of the desired wide-range of chromosome evaluation, with an assumption of an increasing number of epochs on the most successful final model. As result, the algorithm was allowed to create many variations of individuals by mutating and creating new children with crossover operators, and produced the best F1 equal to 0.97%.

This result might be not strictly comparable to the outcomes in other papers, since each research applied different assumptions for evaluations, such as, choice of training and testing sets, treatment of faulty transient observations (in this project they were combined with faulty steady state events and considered as faults). Despite that, the following papers also performed multiclass classification for 3W dataset and can be approached as a reference:

- "Classification of undesirable events in oil well operation", by Turan & Jaschke. The best F1 macro average achieved is 0.85% with Decision Tree algorithm

- “Fault detection and classification in oil wells and production/service lines using Random Forest” by Marins et al. The F1 score was not calculated, but this had Accuracy of 0.94%

- “Predictive maintenance for offshore oil wells by means of deep learning features extraction” by Gatta et al. A number of Machine Learning algorithms were applied after Convolutional Neural Network 1D AutoEncoder was implemented for feature extraction, which resulted in F1 equal to 0.898%.

This research has a different approach to the above-mentioned papers, but shows a comparable result for multiclass classification of undesirable events for the dataset. It can form the basis for developing further Deep Learning algorithms as a precedent with a confirmed good response to the task and a high attained outcome.

5. CONCLUSION

5.1. SYNTHESIS OF THE DEVELOPED WORK

In this project the detection of anomalies in the production of oil and gas was addressed using Deep Neural Networks and Genetic Algorithms for their optimization. The 3W dataset from Petrobras was taken as an example of labelled time series data, which can be applied to the pre-build algorithm and identification of anomalies with high accuracy. Ten initial RNN models with LSTM and GRU architectures were tested and the best one was optimized using Random Search, Hyperopt and three Genetic Algorithms. This identified and classified abnormal events with an outstanding 0.97% F1 score, and can be implemented for other labeled time series anomalies detection and classification.

The research was developed in response to a global awareness of heavy industries adverse effect on global climate change and a demand for enhanced efficiency and sustainability. The United Nation's "The sustainable Development Goals Report 2022" states it is paramount to address the growing issues of greenhouse gas emissions and irresponsible consumption without timely equipment maintenance, which might lead to failures and catastrophic events. Anomaly detection in the Oil and Gas industry is a huge step forward for action in support of the global collaboration for these United Nations goals, with aspiration to increase the chances of preserving our planet for the next generations.

5.2. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

The main limitation of the project, as with any Deep Learning and Big Data projects, is technical capacity of performing the algorithms for the desired number of epochs and iterations. Since the main objective was creating a pipeline of models for the detection and classification of anomalies, Google Colab was sufficient to obtain the primary results and an F1 score for quality justification. However, in the case of more robust calculation required by Genetic Algorithms hyperparameter optimization, more computational resources will be needed.

For future work it would be of high interest to employ Explainable Artificial Intelligence (XAI) algorithms to interpret the Deep Learning algorithms, as being a black box models due to hidden nature of layers and neurons actions. With better understanding which parameters reveal potential faults and the need to diagnose it sooner with higher precision, within the scope of the suggested Deep Learning algorithm, the anomaly detection can become a straightforward task for reservoir and production engineers in the Oil and Gas industry.

REFERENCES

Agin, F., Khosravanian, R., Karimifard, M., & Jahanshahi, A. (2020). Application of adaptive neuro-fuzzy inference system and data mining approach to predict lost circulation using DOE technique (case study: Maroon oilfield). *Petroleum*, *6*(4), 423–437. https://doi.org/10.1016/j.petlm.2018.07.005

Al-Anzi, F. S., Lababidi, H. M. S., Al-Sharrah, G., Al-Radwan, S. A., & Seo, H. J. (2022a). Plant health index as an anomaly detection tool for oil refinery processes. *Scientific Reports 2022 12:1*, *12*(1), 1–18. https://doi.org/10.1038/s41598-022-18824-2

Alharbi, B., Liang, Z., Aljindan, J. M., Agnia, A. K., & Zhang, X. (2022). Explainable and Interpretable Anomaly Detection Models for Production Data. *SPE Journal*, *27*(01), 349–363. https://doi.org/10.2118/208586-PA

Aljameel, S. S. ;, Alomari, D. M. ;, Alismail, S. ;, Khawaher, F. ;, Alkhudhair, A. A. ;, Aljubran, F. ;, Aljameel, S. S., Alomari, D. M., Alismail, S., Khawaher, F., Alkhudhair, A. A., Aljubran, F., & Alzannan, R. M. (2022). An Anomaly Detection Model for Oil and Gas Pipelines Using Machine Learning. *Computation 2022, Vol. 10, Page 138*, *10*(8), 138. https://doi.org/10.3390/COMPUTATION10080138

Aljubran, M., Ramasamy, J., Albassam, M., & Magana-Mora, A. (2021). Deep Learning and Time-Series Analysis for the Early Detection of Lost Circulation Incidents during Drilling Operations. *IEEE Access*, *9*, 76833–76846. https://doi.org/10.1109/ACCESS.2021.3082557

Alsaihati, A., Abughaban, M., Elkatatny, S., & Shehri, D. al. (2022). Application of Machine Learning Methods in Modeling the Loss of Circulation Rate while Drilling Operation. *ACS Omega*, *7*(24), 20696–20709. https://doi.org/10.1021/ACSOMEGA.2C00970/ASSET/IMAGES/LARGE/AO2C00970_0009.JPEG

Alsaihati, A., Elkatatny, S., Mahmoud, A. A., & Abdulraheem, A. (2021). Use of Machine Learning and Data Analytics to Detect Downhole Abnormalities while Drilling Horizontal Wells, with Real Case Study. *Journal of Energy Resources Technology, Transactions of the ASME*, *143*(4). https://doi.org/10.1115/1.4048070/1086232

Amy Chronis, & Kate Hardin. (2022). *2023 Oil and Gas Industry Outlook | Deloitte US*. https://www2.deloitte.com/us/en/pages/energy-and-resources/articles/oil-and-gas-industry-outlook.html

Aslam, N., Khan, I. U., Alansari, A., Alrammah, M., Alghwairy, A., Alqahtani, R., Alqahtani, R., Almushikes, M., & Hashim, M. A. L. (2022). Anomaly Detection Using Explainable Random Forest for the Prediction of Undesirable Events in Oil Wells. *Applied Computational Intelligence and Soft Computing*, *2022*. https://doi.org/10.1155/2022/1558381

Athar Khodabakhsh, Ismail Ari, Mustafa Bakir, & Ali Ozer Ercan. (2018). *Multivariate Sensor Data Analysis for Oil Refineries and Multi-mode Identification of System Behavior in Real-time.* https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8501917

Babaleye, A. O., Kurt, R. E., & Khan, F. (2019). Safety analysis of plugging and abandonment of oil and gas wells in uncertain conditions with limited data. *Reliability Engineering & System Safety*, *188*, 133–141. https://doi.org/10.1016/j.ress.2019.03.027

Barbariol, T., Feltresi, E., & Susto, G. A. (2020). Self-Diagnosis of Multiphase Flow Meters through Machine Learning-Based Anomaly Detection. *ENERGIES*, *13*(12). https://doi.org/10.3390/en13123136

Bellarby, J. (2009). Well completion design, Volume 56. *Developments in Petroleum Science*, 726.

Blancett, J., Pocker, S., Ranjan, A., & Technology Solutions, C. (2019). Automating the Petroleum Industry, from Wells to Wheels. *Cognizant*. https://www.cognizant.com/us/en/pages/whitepapers/automating-the-petroleum-industry-from-wells-to-wheels-codex4114.pdf

Brønstad, C., Netto, S. L., & Ramos, A. L. L. (n.d.). *Data-driven Detection and Identification of Undesirable Events in Subsea Oil Wells*.

Brønstad, C., Netto, S. L., & Ramos, A. L. L. (2021). Data-driven Detection and Identification of Undesirable Events in Subsea Oil Wells. *SENSORDEVICES 2021 : The Twelfth International Conference on Sensor Device Technologies and Applications*.

Carvalho, B. G., Vargas, R. E. V., Salgado, R. M., Munaro, C. J., & Varejao, F. M. (2021). Flow Instability Detection in Offshore Oil Wells with Multivariate Time Series Machine Learning Classifiers. *IEEE International Symposium on Industrial Electronics*, *2021-June*. https://doi.org/10.1109/ISIE45552.2021.9576310

Carvalho, B. G., Vargas, R. E. V., Salgado, R. M., Munaro, C. J., & Varejão, F. M. (2021). Hyperparameter Tuning and Feature Selection for Improving Flow Instability Detection in Offshore Oil Wells. *IEEE International Conference on Industrial Informatics (INDIN)*, *2021-July*. https://doi.org/10.1109/INDIN45523.2021.9557415

Cedric Malate, R. M. (2003). *Well intervention techniques*.

D'Almeida, A. L., Bergiante, N. C. R., de Souza Ferreira, G., Leta, F. R., de Campos Lima, C. B., & Lima, G. B. A. (2022a). Digital transformation: a review on artificial intelligence techniques in drilling and production applications. *The International Journal, Advanced Manufacturing Technology*, *119*(9–10), 5553–5582. https://doi.org/10.1007/S00170-021-08631-W

D'Almeida, A. L., Bergiante, N. C. R., de Souza Ferreira, G., Leta, F. R., de Campos Lima, C. B., & Lima, G. B. A. (2022b). Digital transformation: a review on artificial intelligence techniques in drilling and production applications. *International Journal of Advanced Manufacturing Technology*, *119*(9–10), 5553–5582. https://doi.org/10.1007/S00170-021-08631-W/TABLES/7

Figueirêdo, I. S., Carvalho, T. F., Silva, W. J. D., Guarieiro, L. L. N., & Nascimento, E. G. S. (2021, August 9). Detecting Interesting and Anomalous Patterns In Multivariate Time-Series Data in an Offshore Platform Using Unsupervised Learning. *Offshore Technology Conference*. https://doi.org/10.4043/31297-MS

Gasparovic, B., Lerga, J., Mausa, G., & Ivasic-Kos, M. (2022). Deep Learning Approach For Objects Detection in Underwater Pipeline Images. *APPLIED ARTIFICIAL INTELLIGENCE*, *36*(1). https://doi.org/10.1080/08839514.2022.2146853

Gatta, F., Giampaolo, F., Chiaro, D., & Piccialli, F. (2022). Predictive maintenance for offshore oil wells by means of deep learning features extraction. *Expert Systems*, e13128. https://doi.org/10.1111/EXSY.13128

Giro, R. A., Bernasconi, G., Giunta, G., & Cesari, S. (2021). A data-driven pipeline pressure procedure for remote monitoring of centrifugal pumps. *Journal of Petroleum Science and Engineering*, *205*, 108845. https://doi.org/https://doi.org/10.1016/j.petrol.2021.108845

Guillaume Decaix, Matthew Gentzel, Andy Luse, Patrick Neise, & Joel Thibert. (2021). *A smarter way to digitize maintenance and reliability | McKinsey*. https://www.mckinsey.com/capabilities/operations/our-insights/a-smarter-way-to-digitize-maintenance-and-reliability

Gurina, E., Klyuchnikov, N., Antipova, K., & Koroteev, D. (2022a). Forecasting the abnormal events at well drilling with machine learning. *APPLIED INTELLIGENCE*, *52*(9), 9980–9995. https://doi.org/10.1007/s10489-021-03013-x

Gurina, E., Klyuchnikov, N., Antipova, K., & Koroteev, D. (2022b). Making the black-box brighter: Interpreting machine learning algorithm for forecasting drilling accidents. *JOURNAL OF PETROLEUM SCIENCE AND ENGINEERING*, *218*. https://doi.org/10.1016/j.petrol.2022.111041

Gurina, E., Klyuchnikov, N., Zaytsev, A., Romanenkova, E., Antipova, K., Simon, I., Makarov, V., & Koroteev, D. (2020). Application of machine learning to accidents detection at directional drilling. *Journal of Petroleum Science and Engineering*, *184*. https://doi.org/10.1016/j.petrol.2019.106519

Hasan, M. H., Malik, A. A., & Jasamai, M. (2017). A Review on Anomaly Detection Methods for Optimizing Oil Well Surveillance. *IJCSNS International Journal of Computer Science and Network Security*, *17*(11).

Ide, T., Khandelwal, A., & Kalagnanam, J. (2017). Sparse Gaussian Markov Random Field Mixtures for Anomaly Detection. *IEEE Xplore*, 955–960. https://doi.org/10.1109/ICDM.2016.0119

Kirschbaum, L., Roman, D., Singh, G., Bruns, J., Robu, V., & Flynn, D. (2020). AI-Driven maintenance support for downhole tools and electronics operated in dynamic drilling environments. *IEEE Access*, *8*, 78683–78701. https://doi.org/10.1109/ACCESS.2020.2990152

Kongsberg Digital. (2022). *Kognitwin® - Say hello to your digital twin*. https://kongsbergdigital.com/products/kognitwin/

Koroteev, D., & Tekic, Z. (2021). Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future. *Energy and AI*, *3*, 100041. https://doi.org/10.1016/J.EGYAI.2020.100041

KUANG, L., LIU, H., REN, Y., LUO, K., SHI, M., SU, J., & LI, X. (2021). Application and development trend of artificial intelligence in petroleum exploration and development. *Petroleum Exploration and Development*, *48*(1), 1–14. https://doi.org/10.1016/S1876-3804(21)60001-0

Lashari, S. Z., Takbiri-Borujeni, A., Fathi, E., Sun, T., Rahmani, R., & Khazaeli, M. (2019). Drilling performance monitoring and optimization: a data-driven approach. *Journal of Petroleum Exploration and Production Technology*, *9*(4), 2747–2756. https://doi.org/10.1007/s13202-019-0657-2

Li, M., Zhang, H. R., Zhao, Q., Liu, W., Song, X. Z., Ji, Y. Y., & Wang, J. S. (2022). A New Method for Intelligent Prediction of Drilling Overflow and Leakage Based on Multi-Parameter Fusion. *ENERGIES*, *15*(16). https://doi.org/10.3390/en15165988

Liu, J., Feng, J., & Gao, X. (2019). Fault Diagnosis of Rod Pumping Wells Based on Support Vector Machine Optimized by Improved Chicken Swarm Optimization. *IEEE Access*, *7*, 171598–171608. https://doi.org/10.1109/ACCESS.2019.2956221

Liu, Y., Yao, K., Liu, S., Raghavendra, C. S., Lenz, T. L., Olabinjo, L., Seren, B., Seddighrad, C. S., & Babu, C. G. D. (2010). Failure Prediction for Rod Pump Artificial Lift Systems. *Society of Petroleum Engineers Western North American Regional Meeting 2010 - In Collaboration with the Joint Meetings of the Pacific Section AAPG and Cordilleran Section GSA*, *2*, 845–852. https://doi.org/10.2118/133545-MS

Liu, Y., Yao, K. T., Liu, S., Raghavendra, C. S., Balogun, O., & Olabinjo, L. (2011). Semi-supervised failure prediction for oil production wells. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 434–441. https://doi.org/10.1109/ICDMW.2011.151

Lv, X., Wang, H., Zhang, X., Liu, Y., Jiang, D., & Wei, B. (2021). An evolutional SVM method based on incremental algorithm and simulated indicator diagrams for fault diagnosis in sucker rod pumping systems. *Journal of Petroleum Science and Engineering*, *203*. https://doi.org/10.1016/j.petrol.2021.108806

Lv, X.-X., Wang, H.-X., Xin, Z., Liu, Y.-X., & Zhao, P.-C. (2022). Adaptive fault diagnosis of sucker rod pump systems based on optimal perceptron and simulation data. *Petroleum Science*, *19*(2), 743–760. https://doi.org/10.1016/j.petsci.2021.09.012

Machado, A. P. F., Vargas, R. E. V., Ciarelli, P. M., & Munaro, C. J. (2022). Improving performance of one-class classifiers applied to anomaly detection in oil wells. *Journal of Petroleum Science and Engineering*, *218*, 110983. https://doi.org/10.1016/J.PETROL.2022.110983

Magana-Mora, A., Affleck, M., Ibrahim, M., Makowski, G., Kapoor, H., Otalvora, W. C., Jamea, M. A., Umairin, I. S., Zhan, G., & Gooneratne, C. P. (2021). Well Control Space Out: A Deep-Learning Approach for the Optimization of Drilling Safety Operations. *IEEE Access*, *9*, 76479–76492. https://doi.org/10.1109/ACCESS.2021.3082661

Marins, M. A., Barros, B. D., Santos, I. H., Barrionuevo, D. C., Vargas, R. E. V., de M. Prego, T., de Lima, A. A., de Campos, M. L. R., da Silva, E. A. B., & Netto, S. L. (2021). Fault detection and classification

in oil wells and production/service lines using random forest. *Journal of Petroleum Science and Engineering*, *197*, 107879. https://doi.org/10.1016/J.PETROL.2020.107879

Martí, L., Sanchez-Pi, N., Molina, J. M., & Garcia, A. C. B. (2015a). Anomaly Detection Based on Sensor Data in Petroleum Industry Applications. *Sensors (Basel, Switzerland)*, *15*(2), 2774. https://doi.org/10.3390/S150202774

Martí, L., Sanchez-Pi, N., Molina, J. M., & Garcia, A. C. B. (2017). On the combination of support vector machines and segmentation algorithms for anomaly detection: A petroleum industry comparative study. *Journal of Applied Logic*, *24*, 71–84. https://doi.org/10.1016/J.JAL.2016.11.015

Mello, L. H. S., Oliveira-Santos, T., Varejão, F. M., Ribeiro, M. P., & Rodrigues, A. L. (2022). Ensemble of metric learners for improving electrical submersible pump fault diagnosis. *Journal of Petroleum Science and Engineering*, *218*, 110875. https://doi.org/10.1016/J.PETROL.2022.110875

Mello, L. H. S., Ribeiro, M. P., Oliveira-Santos, T., Varejão, F. M., & Rodrigues, A. L. (2020, July 1). Metric Learning for Electrical Submersible Pump Fault Diagnosis. *Proceedings of the International Joint Conference on Neural Networks*. https://doi.org/10.1109/IJCNN48605.2020.9207133

Michele Ianni, Elio Masciari, Giuseppe M. Mazzeo, Mario Mezzanzanica, & Carlo Zaniolo. (2020). Fast and effective Big Data exploration by clustering. *Future Generation Computer Systems*, *102*, 84–94.

Mopuri, K. R., Bilen, H., Tsuchihashi, N., Wada, R., Inoue, T., Kusanagi, K., Nishiyama, T., & Tamamura, H. (2022). Early sign detection for the stuck pipe scenarios using unsupervised deep learning. *Journal of Petroleum Science and Engineering*, *208*. https://doi.org/10.1016/j.petrol.2021.109489

Muojeke, S., Venkatesan, R., & Khan, F. (2020). Supervised data-driven approach to early kick detection during drilling operation. *Journal of Petroleum Science and Engineering*, *192*. https://doi.org/10.1016/j.petrol.2020.107324

Nick Hotz. (2022). *What is CRISP DM? - Data Science Process Alliance*. https://www.datascience-pm.com/crisp-dm-2/

OLGA Dynamic Multiphase Flow Simulator. (n.d.). Retrieved March 7, 2023, from https://www.software.slb.com/products/olga

Orestes, A., Castro, D. S., De, M., Santos, J. R., Rodrigues Leta, F., Benevenuto, C., Lima, C., Brito, G., Lima, A., Jesus, D., Santos, R., Leta, M., Lima, F. R., & Lima, C. B. C. (2021). Unsupervised Methods to Classify Real Data from Offshore Wells. *American Journal of Operations Research*, *11*(5), 227–241. https://doi.org/10.4236/AJOR.2021.115014

Overview — DEAP 1.3.3 documentation. (n.d.). Retrieved May 9, 2023, from https://deap.readthedocs.io/en/master/overview.html

Pandey, Y. N., Rastogi, A., Kainkaryam, S., Bhattacharya, S., & Saputelli, L. (2020). Machine Learning in the Oil and Gas Industry. In *Machine Learning in the Oil and Gas Industry*. Apress. https://doi.org/10.1007/978-1-4842-6094-4

Patri, O. P., Reyna, N., Panangadan, A., & Prasanna, V. (2015). Predicting Compressor Valve Failures from Multi-Sensor Data. *SPE Western Regional Meeting 2015: Old Horizons, New Horizons Through Enabling Technology*, 725–735. https://doi.org/10.2118/174044-MS

Peng, L., Han, G., Landjobo Pagou, A., & Shu, J. (2020). Electric submersible pump broken shaft fault diagnosis based on principal component analysis. *Journal of Petroleum Science and Engineering*, *191*, 107154. https://doi.org/10.1016/J.PETROL.2020.107154

Qin, H., Srivastava, V., Wang, H., Zerpa, L. E., & Koh, C. A. (2019). Machine Learning Models to Predict Gas Hydrate Plugging Risks Using Flowloop and Field Data. *Proceedings of the Annual Offshore Technology Conference*, *2019-May*. https://doi.org/10.4043/29411-MS

Ravishankar, P., Hwang, S., Zhang, J., Khalilullah, I. X., & Eren-Tokgoz, B. (2022). DARTS-Drone and Artificial Intelligence Reconsolidated Technological Solution for Increasing the Oil and Gas Pipeline Resilience. *INTERNATIONAL JOURNAL OF DISASTER RISK SCIENCE*, *13*(5), 810–821. https://doi.org/10.1007/s13753-022-00439-w

Salem, A. M., Yakoot, M. S., & Mahmoud, O. (2022). A novel machine learning model for autonomous analysis and diagnosis of well integrity failures in artificial-lift production systems. *Advances in Geo-Energy Research*, *6*(2), 123–142. https://doi.org/10.46690/ager.2022.02.05

Santos, M. de J. R., Castro, A. O. de S., Leta, F. R., de Araujo, J. F. M., Ferreira, G. de S., Santos, R. de A., Lima, C. B. de C., & Lima, G. B. A. (2021). Statistical analysis of offshore production sensors for failure detection applications / Análise estatística dos sensores de produção offshore para aplicações de detecção de falhas. *Brazilian Journal of Development*, *7*(8), 85880–85898. https://doi.org/10.34117/BJDV7N8-681

Satyam Kumar. (2020). *7 Over Sampling techniques to handle Imbalanced Data | Towards Data Science*. Towards Data Science. https://towardsdatascience.com/7-over-sampling-techniques-to-handle-imbalanced-data-ec51c8db349f

Schlumberger. (2021). *Schlumberger and NOV Announce Collaboration to Accelerate Adoption of Automated Drilling Solutions | Schlumberger*. https://www.slb.com/newsroom/press-release/2021/pr-2021-0510-slb-nov-collaboration

Scoralick, R., Scoralick Fontoura do Nascimento, R., Henrique Groenner, B., Vargas, R. E. V., & Humberto Ferreira dos Santos, I. (2021). Fault detection with Stacked Autoencoders and pattern recognition tech-niques in gas lift operated oil wells. *XLII Ibero-Latin-American Congress on Computational Methods in Engineering AndIII Pan-American Congress on Computational Mechanics, ABMEC-IACMRio de Janeiro, Brazil, November 9-12, 2021*. https://www.researchgate.net/publication/363279803

Seo, Y., Kim, B., Lee, J., & Lee, Y. (2021). Development of ai-based diagnostic model for the prediction of hydrate in gas pipeline. *Energies*, *14*(8). https://doi.org/10.3390/en14082313

Shrifan, N. H. M. M., Jawad, G. N., Isa, N. A. M., & Akbar, M. F. (2021). Microwave Nondestructive Testing for Defect Detection in Composites Based on K-Means Clustering Algorithm. *IEEE Access*, *9*, 4820–4828. https://doi.org/10.1109/ACCESS.2020.3048147

Sinha, S., de Lima, R. P., Lin, Y. Z., Sun, A. Y., Symons, N., Pawar, R., & Guthrie, G. (2020). Normal or abnormal? Machine learning for the leakage detection in carbon sequestration projects using pressure field data. *INTERNATIONAL JOURNAL OF GREENHOUSE GAS CONTROL*, *103*. https://doi.org/10.1016/j.ijggc.2020.103189

Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., & Oza, H. (2021). Application of machine learning and artificial intelligence in oil and gas industry. *Petroleum Research*, *6*(4), 379–391. https://doi.org/10.1016/J.PTLRS.2021.05.009

Soriano-Vargas, A., Werneck, R., Moura, R., Mendes Júnior, P., Prates, R., Castro, M., Gonçalves, M., Hossain, M., Zampieri, M., Ferreira, A., Davólio, A., Hamann, B., Schiozer, D. J., & Rocha, A. (2021a). A visual analytics approach to anomaly detection in hydrocarbon reservoir time series data. *Journal of Petroleum Science and Engineering*, *206*, 108988. https://doi.org/10.1016/J.PETROL.2021.108988

Su, J., Zhao, Y., He, T., & Luo, P. (2021). Prediction of drilling leakage locations based on optimized neural networks and the standard random forest method. *Oil and Gas Science and Technology – Revue d'IFP Energies Nouvelles*, *76*. https://doi.org/10.2516/ogst/2021003

Tan, C., Chen, P., Feng, Z., Ai, X., Lu, M., Zhou, Q., & Feng, G. (2022). Multi-Scale Normalization Method Combined with a Deep CNN Diagnosis Model of Dynamometer Card in SRP Well. *Frontiers in Earth Science*, *10*. https://doi.org/10.3389/feart.2022.852633

The Defining Series: Artificial Lift | SLB. (n.d.). Retrieved January 30, 2023, from https://www.slb.com/resource-library/oilfield-review/defining-series/defining-artificial-lift

The SLB Energy Glossary | Energy Glossary. (n.d.). Retrieved March 7, 2023, from https://glossary.slb.com/

Turan, E. M., & Jaschke, J. (2021). Classification of undesirable events in oil well operation. *Proceedings of the 2021 23rd International Conference on Process Control, PC 2021*, 157–162. https://doi.org/10.1109/PC52310.2021.9447527

United Nations. (2022). *THE 17 GOALS | Sustainable Development*. https://sdgs.un.org/goals

U.S. Chemical Safety Board. (2016). *Macondo Blowout and Explosion | CSB*. https://www.csb.gov/macondo-blowout-and-explosion/

Vankov, Y., Rumyantsev, A., Ziganshin, S., Politova, T., Minyazev, R., & Zagretdinov, A. (2020). Assessment of the condition of pipelines using convolutional neural networks. *Energies*, *13*(3). https://doi.org/10.3390/en13030618

Vanneschi, L., & Silva, S. (2023). *Lectures on Intelligent Systems*. https://doi.org/10.1007/978-3-031-17922-8

Vargas, R. E. V., Munaro, C. J., Ciarelli, P. M., & de Araujo, J. C. D. (2017). Proposal for two classifiers of offshore naturally flowing wells events using k-nearest neighbors, sliding windows and time multiscale. *2017 6th International Symposium on Advanced Control of Industrial Processes, AdCONIP 2017*, 209–214. https://doi.org/10.1109/ADCONIP.2017.7983782

Vargas, R. E. V., Munaro, C. J., Ciarelli, P. M., Medeiros, A. G., Amaral, B. G. do, Barrionuevo, D. C., Araújo, J. C. D. de, Ribeiro, J. L., & Magalhães, L. P. (2019a). A realistic and public dataset with rare undesirable real events in oil wells. *Journal of Petroleum Science and Engineering*, *181*, 106223. https://doi.org/10.1016/J.PETROL.2019.106223

VOSviewer - Visualizing scientific landscapes. (n.d.). Retrieved January 8, 2023, from https://www.vosviewer.com//

Wei, J., & Gao, X. (2020). Fault Diagnosis of Sucker Rod Pump Based on Deep-Broad Learning Using Motor Data. *IEEE Access*, *8*, 222562–222571. https://doi.org/10.1109/ACCESS.2020.3036078

Wong, P., Wong, W. K., Juwono, F. H., Gopal, L., & Yusoff, M. A. (2022). A minimalist approach for detecting sensor abnormality in oil and gas platforms. *Petroleum Research*, *7*(2), 177–185. https://doi.org/10.1016/j.ptlrs.2021.09.007

Wood, D. A., Mardanirad, S., & Zakeri, H. (2022). Effective prediction of lost circulation from multiple drilling variables: a class imbalance problem for machine and deep learning algorithms. *Journal of Petroleum Exploration and Production Technology*, *12*(1), 83–98. https://doi.org/10.1007/s13202-021-01411-y

Zhang, R., Wang, L., & Chen, D. (2021). An intelligent diagnosis method of the working conditions in sucker-rod pump wells based on convolutional neural networks and transfer learning. *Energy Engineering: Journal of the Association of Energy Engineering*, *118*(4), 1071–1082. https://doi.org/10.32604/EE.2021.014961

Zhang, Y., & Yang, K. (2022). Fault Diagnosis of Submersible Motor on Offshore Platform Based on Multi-Signal Fusion. *Energies*, *15*(3). https://doi.org/10.3390/en15030756

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa