National College of Ireland

Predictive Maintenance for Fault Diagnosis and Failure Prognosis in Hydraulic System

MSc Research Project
Data Analytics

Vijit Laxman Chekkala
Student ID: X18199429

School of Computing
National College of Ireland

Supervisor: Dr. Vladimir Milosavljevic

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Vijit Laxman Chekkala		
Student ID:	X18199429		
Programme:	MSc Data Analytics	**Year:**	2019-2020
Module:	Research Project		
Supervisor:	Dr. Vladimir Milosavljevic		
Submission Due Date:	28/09/2020		
Project Title:	Predictive Maintenance for Fault Diagnosis and Failure Prognosis in Hydraulic System		
Word Count:	7405	**Page Count:**	29

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Vijit Laxman Chekkala

Date: 22nd September 2020

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predictive Maintenance for Fault Diagnosis and Failure Prognosis in Hydraulic System

Vijit Laxman Chekkala
X18199429

Abstract

The key role in developing and successfully achieving high-quality results in business is by minimising the cost and maximising the profits. This is possible only if proper resources are optimized and implemented. Failure prognosis is a part of predictive maintenance where data science field is involved in predicting the conditions of a system. With proper machine learning techniques, the monitoring devices can easily replace traditional monitoring devices. In this research, proper fault diagnosis is carried out on the components and the stable conditions of a hydraulic system. A dashboard is created where all the data point is explored by showing their distribution, the correlation matrix, their importance in predicting the conditions and the outliers. Chi-square test of independence is calculated to define the relationship between the categorical values. The scaling and dimensionality reduction step was done by using Quantile Transform Scalar and UMAP technique respectively. The model building and evaluating stages were implemented in Python where RandomisedSearchCV is used for hyperparameter optimization in six classification algorithms. Results showed that using gradient boosting decision trees algorithm helped in achieving greater accuracy than any other machine learning models. The web app was deployed for the research project using Heroku and the dashboard created for exploratory data analysis was published to web using Shiny apps in R.

1 Introduction

A common application found in industries for carrying out large operations like manufacturing and engineering process is a hydraulic system. It works by the transmission of liquid to generate fluid power for carrying out the operations. The working of the system depends on the components inside the hydraulic system. These components are responsible for exchanging the heat generated during the operation, directing the flow of the liquid, to generate the flow according to the need of the operation and a device to act as secondary storage in case of any demand. To properly analyse the conditions of these components, the industries have a monitoring device to record the status of their performance. This is necessary to schedule maintenance if there is any problem with the components or the stability of the hydraulic system. With so much advancement and need of technology, the working of the hydraulic system depends on the environment and the defined operations. It is important to perform diagnosis on the fault conditions and use the prognosis in future to reduce the failure of the hydraulic system (Higgs and Author, 2004). Many industries have accepted data-driven approach for monitoring the state of the system so that there is no unplanned downtime, risk in the employer's life and also crashing the profits of the industries. Based on the needs and improvement in developing a monitoring system, this research study is based on the research

question: **"To what level can integrated framework of different classification algorithms help in the prediction of fault conditions in the hydraulic system?"**

To properly analyse different fault states in the hydraulic system, the data should be of the highest quality. Data plays an important role in training the machine learning algorithms and the whole process depends on the quality of the data. The dataset taken from the UCI machine learning repository for the research study is a combination of sensor and vibrational analysis data (Helwig, Pignanelli and Schutze, 2015). This dataset has the required targeted variables and the stable condition to carry out the modelling and analysis process. The rest of the research study is as follows: Section 2 presents the related work carried out in monitoring conditions of the hydraulic systems. Section 3 and 4 presents the methodology approached and implementation respectively. The results evaluation and discussion are presented in Section 5 and the deployment of the research study is shown in Section 6. In Section 7, there are conclusion and future work.

2 Related Work (2004-2019)

2.1 Introduction

A large number of applications have a hydraulic system for small to large scale business. With a large development of their use, the status monitoring methods have developed from model-based to data-driven approach. Corrective maintenance approach is adopted at a large scale as they are better than the traditional methods used for monitoring the status of the hydraulic system. The close to failure stage in the hydraulic system can cause loss to production, no safety for the employee and decrease in good return of investments.

2.2 Condition Monitoring System

The dataset uploaded by the researchers has developed a framework which is suitable for condition monitoring of a hydraulic system. Approach to the problem is solely dependent on the use of the model for later use. The use of supervised learning with feature extraction from real sensor data is used to suit the changing conditions of the hydraulic system with less effort (Helwig, Pignanelli and Schutze, 2015). Data is taken from real sensors built in the hydraulic system and the values are measured using a physical model. Using multivariate statistics, the problem is approached, and the algorithms used are semantic statical methods and classification algorithms. The algorithms used for classification are support vector machines (SVM), neural networks and decision trees. Features achieved from the sensor data are evaluated to understand their importance in the role of the hydraulic components. To get a warning before the failure of the hydraulic system and their components, the data is analysed in accordance to neglect immediate failure. For stability, it is a binary approach whereas the four component cooler, valve, pump and accumulator have multi-class values. After concluding, there are chances of dealing with the feature selection and important process to increase the accuracy of the machine learning models. Rather than using recent data for monitoring conditions, historical data are considered to be more valuable as they tend to show accurate results (Michael et al., 2005).

Different feature extraction and selection process were introduced to reduce the dimensions for the development of various machine learning model on sensor data (Schneider, Helwig and Schütze, 2017). The important thing properly represents the data which can help in predicting the performance of the components in any system. The adaptive linear approximation (ALA) was first used due to its low computation time for feature extraction methods. It guarantees minimum error while reducing the dimension which can be further used for a reasonable cause. Second, the most widely sued techniques to reduce dimension, principal component analysis is used for dimensionality reduction. In the vibrational analysis, techniques used to better represent the data in density values for signal processing or data compression are the Best Fourier Coefficient (BFC) and best Daubechies-Wavelet Coefficients (BDW).

2.3 Fault Diagnosis in Large Scale Industries

The increasing demand for quality products in large scale industries are increasing. Various monitoring systems are built to scheduled maintenance if there is any problem with any system in the industry (Kano and Nakagawa, 2008). This ensures workplace safety for the employees and also maintain the quality of the industry. It can be any industry, the monitoring system developed to check the conditions are similar. Quality improvement or maintenance results in boosting the production in the industries. In most industries, first principle models were used for monitoring purposes but due to the advancement and slow processing time they are being continuously changed. Therefore there is a need of much efficient monitoring system in large scale industries. Later, the industries started maintaining databases where the operational data is stored and analysed for fault diagnosis. The first approach is to record the sensor data in a way that it can be optimised and used for solving the problem. Based on the data, a proper statistical method is approached. For fault diagnosis, highly approached models to classify multi-class problems are linear discriminant analysis or fisher discriminant analysis. PCA and partial least square (PLS) are useful methodologies in process industries where PLS performs better than PCA in monitoring status.

Most fault diagnosis approach has low efficiency to analyse data and represent the findings, which can lead to false results (Di et al., 2017). They were approached by the expert system and to overcome the problem, decision trees methods are used for fault diagnosis. This method makes proper predictions of the fault tolerance by properly analysing the data. By process large amounts of sensor data in real-time, it becomes difficult for traditional monitoring methods to extract useful information. The use of data mining approach can help in fulfilling the drawbacks set by traditional methods and yield out better results. For fault diagnosis in intelligent equipment system, decision trees are used for predicting the fault conditions in the system. The results were compared to the previous expert system and an increase of 8.6% in predicting the fault conditions. Using decision trees and nested dichotomy algorithms also help in improving accuracy for multi-class classification problems (Jegadeeshwaran and Sugumaran, 2015).

The next research focuses on showing the advantage of using sphere-structured SVM for fault diagnosis (Hu, 2012). SVM is used as it can be applied for solving multi-class classification problems without any limitations. In this research, there are two models used for diagnosing the conditions of the hydraulic system. The backpropagation neural network is also

used to cover loss in every step of the neural network implementation by assigning the weights. Based on the characteristics of the signal values, the output from both the models are compared and the results showed that SVM performed better than BP neural network.

2.4 Condition Monitoring in Wind Turbine

Like in many industries, condition monitoring is used in the wind turbine industry to monitor the status of the components like temperature and detecting faults in the blades (Stetco et al., 2019). Three classification modelling techniques like neural networks, SVM and DT are used for the classification problem. The research mentions different types of monitoring based on vibrational analysis and thermography monitoring analysis. These techniques can be used either for detecting or predicting the faults. Machine learning models used for predicting by finding patterns are the most reliable form of the technique used to analyse the conditions in real-time. Any leads that fall to the failure of the system can be predicted when machine learning techniques are used for fault prognosis. Before approaching the modelling part, outliers are detected and the feature selection process is carried out by wrapper, embedded and filter methods. Both classification and regression-based analysis are done to determine the fault conditions int the wind turbine. The parametric and non-parametric techniques are compared and their learning in approaching a classification problem.

Another approach in the wind turbine industry for internal pump leakage is done by implementing asymmetric SVM (Wu et al., 2015). The efficiency of the algorithm is enhanced using a smaller number of support vectors for classifying the fault conditions. Apart from predicting the outcome, the importance of measuring the conditions is carried out which leads to the failure in the hydraulic pitching system. This is important so that maintenance can be scheduled before the condition gets worse. Another important thing is to note the misclassified error which is acceptable when the model predicts fault condition and there seem to be no fault in the system. It is not acceptable when the model predicts no failure condition and the system fails which leads to unnecessary downtime in the industry. In soil hydraulic conductivity, a similar approach with regression analysis was approached and performing lasso regression which is a supervised learning algorithm performed than SVM (Kotlar, Iversen and Jong van Lier, 2019). This research compares between the parametric and non-parametric techniques that can be used for monitoring hydraulic conditions.

2.5 Performance Analysis of the Hydraulic System in Aircraft

The aircraft is a piece of heavy machinery where a hydraulic system is used to provide mechanical power to the entire aircraft. It is difficult to analyse the performance of the hydraulic system due to their working environment (Cui et al., 2019). It is not possible to manually analyse the performance indicators and then raise an alarm for the fault condition in the pump performance of an aircraft. Therefore, it is necessary that the conditions of the system in the aircraft are continuously monitored. In this study, the prediction of the performance analysis is carried out using the particle swarm optimisation (PSO) algorithm. The output from the algorithm is compared to the extreme learning machine (ELM) to verify the effectiveness and analysing outputs. PSO performed better than ELM in terms of predicting the performance

analysis of the pump performance in aircraft. The reason for the use of PSO was the advancement of training the network of ELM to optimize and obtain better and fast results.

In a hydraulic system, the important component is the valve which is used to control the direction of the fluid for transmission (Bykov, Voronov and Voronova, 2019). Therefore, in this research study, the analysis is done on the valve condition. The hydraulic system tends to change its performance under high condition or in a hard environment. If one of the components of the hydraulic system is close to failure, there is a high risk that the stability of the hydraulic system is in danger. To properly predict the failure conditions, machine learning techniques used are SVM, k-nearest neighbor (KNN) and XGBoost. Two hyperparameter techniques like StratifiedKFold and GridSearchCV are used for tuning the hyperparameter and perform cross-validation. From the three algorithms, linear SVM performed better than KNN and XGBoost in the prediction of the failure of valves in the hydraulic system.

Many heavy-duty industrial jobs use the fluid transmission to carry out the operations. This research study focuses on the friction created in the servomechanism hydraulic system. A hybrid approach of neuro-fuzzy control (NFC) which is based on neural networks is used along SVM to predict the response for the friction in the hydraulic servomechanism (Hutamarn, Pratumsuwan and Po-Ngaen, 2012). This is carried out to compensate with the noise, loss and leakage problems when there is a chance of failure. The fuzzy logic control of the input and output parameters are done by SVM and back-propagation. To reduce the failure of the hydraulic servomechanism, this approach was successful in the prediction of the friction produced in the system.

3 Methodology

The “Cross Industry process for data mining” is a problem-solving framework methodology approached in many business domains and their levels are described in Figure 1.

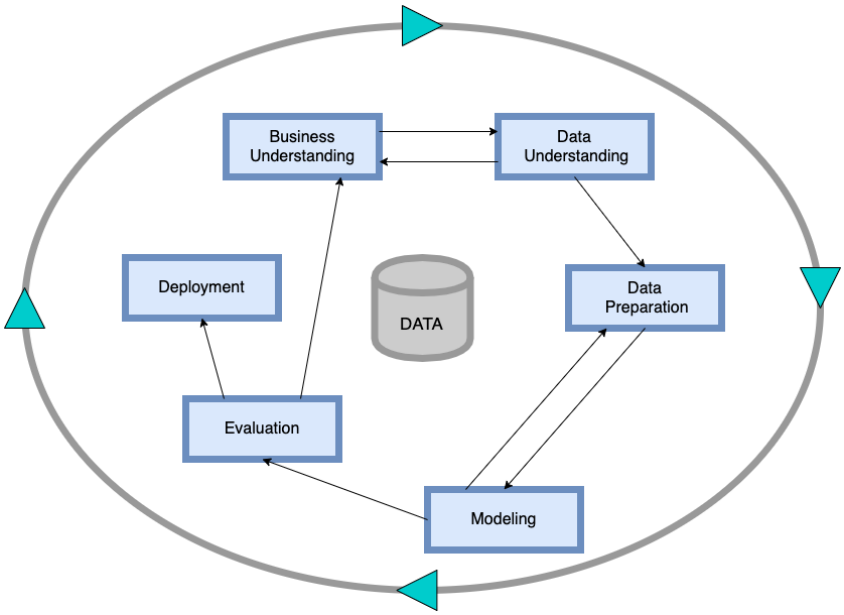

Figure 1: CRISP-DM process

3.1 Business Understanding

In business, success is accomplished if the goals are measured in the early stage. Industrial applications need a hydraulic system to perform heavy duties. This is performed using a pressurized liquid which is generated by the components of the hydraulic system. In every industry, predictive maintenance techniques are developed to ensure safety for their employees, save cost and generate good return of investments, etc. there are multiple components that works together in a hydraulic system and these components determine the stability of the hydraulic system. By properly predicting the conditions of the components in the hydraulic system, proper maintenance can be scheduled to prevent the failure. The dataset chosen for analysis is a multi-class classification problem which defines the overall conditions from optimal behaviour to close to total failure of the hydraulic system and their components. After selecting the dataset, a proper analytical approach was undertaken using machine learning classification algorithms. These algorithms were evaluated using classification metrics to determine the performance on the predictions. Maximising revenue and minimising cost are the two main goals of any business and it can be achieved if proper machine learning algorithms are used for predictive maintenance. The objective and key results (OKR) and key performance indicators (KPI) in terms of understanding the business are discussed in Table 1.

Table 1: Business Understanding

Objective and Key Results (OKR)	Key Performance Indicators (KPI)
Objective: Approaching different dimensionality reduction techniques and classification algorithms for fault diagnosis and failure prognosis in hydraulic system. **Key results:** • Identifying the importance of multiple sensors responsible for predicting the conditions. • Performing statistical technique to find relationship between categorical variables. • Increase model accuracy by applying proper scaling and dimensionality reduction techniques. • Developing a generalised model using evaluation methods.	• Using random forest algorithm to identify important sensors. • Performing Chi-square test to define the relationship between the four conditions i.e. cooler, valve, pump leakage, accumulator with stable conditions. • Scaling the data using quantile transform scalar and using UMAP to reduce dimensions. • Applying proper classification algorithms and optimizing their parameters using RandomisedSearchCV. • Evaluating the models by defining accuracy, confusion matrix and classification reports.

3.2 Data Understanding and Exploratory Data Analysis (EDA)

The second phase is data understanding where data should be collected according to business understanding. For any data-driven approach, an important step is to understand the data which

is used for analysis. The data for this research study is taken from the UCI machine learning repository which is open source and can be downloaded from **https://archive.ics.uci.edu/ml/datasets/Condition+monitoring+of+hydraulic+systems**. To measure the performance of a different component of the hydraulics system, the data was prepared using a test rig machinery piece. It was based on multi-sensor data to cover the four different faults of the hydraulic system. The dataset shows multivariate characteristics as multiple classes are defined to predict different types of faults. Pressure, volume flows and temperature are recorded which is considered as the independent variables and cooler, calve, pump and accumulator are the targeted variables.

Independent Variables: In Figure 2, the independent values processed from the sensor data are recorded with their physical quantity, unit and sampling rate.

Pressure sensor data	Temperature sensor data	Volume flow data	Pump Efficiency, Cooling Efficiency, Vibrations and Efficiency factor
PS1 (bar, 100Hz)	TS1 (∞C, 1Hz)	FS1 (l/min, 10Hz)	EPS1 (W, 100Hz)
PS2 (bar, 100Hz)	TS2 (∞C, 1Hz)	FS2 ((l/min, 10Hz)	CE (%, 1Hz)
PS3 (bar, 100Hz)	TS3 (∞C, 1Hz)		CP (KW, 1Hz)
PS4 (bar, 100Hz)	TS4 (∞C, 1Hz)		VS1 (mm/s, 1Hz)
PS5 (bar, 100Hz)			SE (%, 1Hz)
PS6 (bar, 100Hz)			

Figure 2: Independent Variables

Dependent Variables: The four faults (cooler, valve, pump leakage, accumulator) along with a stable flag to determine the overall stability of the hydraulic systems are shown in Figure 3.

Cooler Condition
3 - Close to failure 20 - Reduced efficiency 100 - Full efficiency

Internal Pump Leakage
0 - No leakage 1 - Weak leakage 2 - Severe leakage

Valve Condition
100 - Optimal switching behaviour 90 - Small lag 80 - Severe lag 73 - Close to total failure

Hydraulic Accumulator/bar
130 - Optimal pressure 115 - Slightly reduced pressure 100 - severely reduced pressure 90 - close to total failure

Stable flag
0 - Conditions were stable 1 - Static conditions might not have been reached yet

Figure 3: Targeted Variables

Correlation: The independent variables are a group of clusters in which different types of sensors represent their values for the targeted variables. The relationship between the independent variables are shown using a correlation matrix.

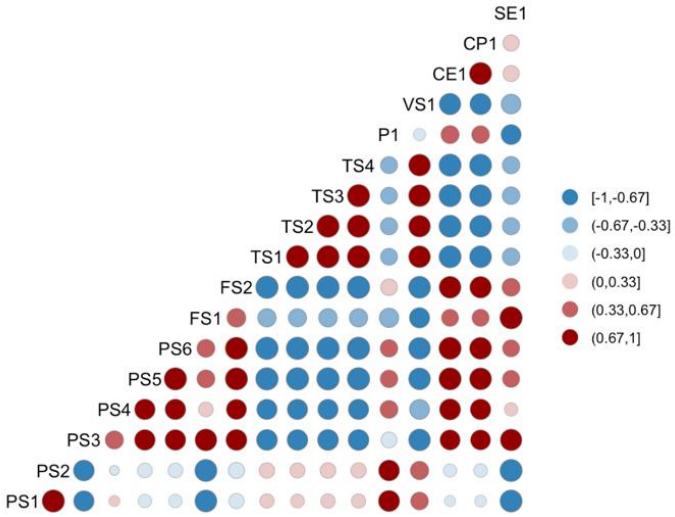

Figure 4: Correlation matrix

It can be seen that different data points that belong to pressure and temperature show a high correlation. Highly correlated variables are not good for machine learning models as a change in one data point can cause a change on the other data point which will overall impact in predicting targeted variables.

3.3 Data Preparation

For predictive modelling task, data preparation is an essential step which is shown in Figure 5.

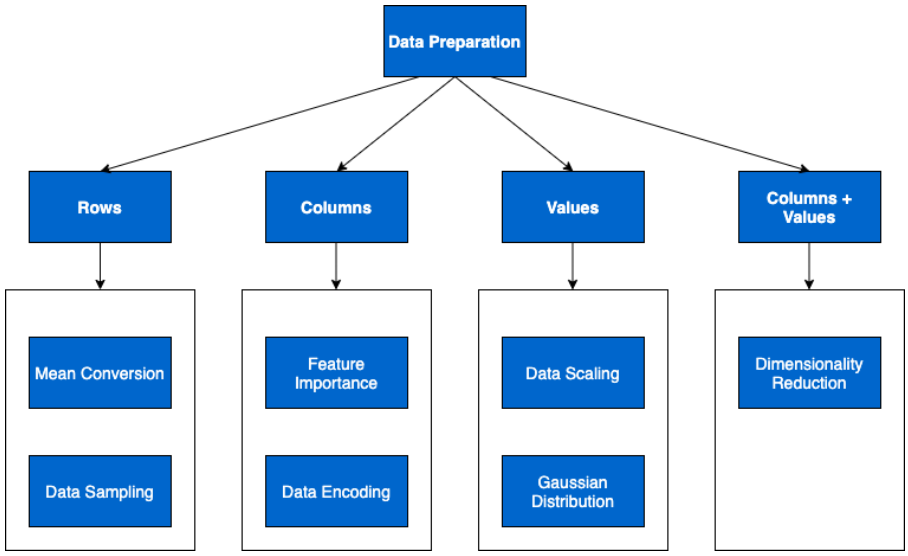

Figure 5: Data Preparation

Mean Conversion: The dataset taken from the UCI machine learning repository was in a text format. All the independent variables were in different text files and the targeted variables were present in a single text file called profile. The data was converted into comma-separated values and then merged together in a single dataframe. Features were extracted in a 60-second duration with 13-time intervals which had 43,680 number of attributes (Helwig, Pignanelli and Schutze, 2015).

Data Sampling: The targeted variables except for coolers shows that the distribution of classes has a high variance. Such high variance leads to an imbalance in the data which can cause machine learning models to make biased predictions. A count plot is shown the classification imbalance problem in valves, pump leakage, accumulator and stable variables.

Feature Importance: In a predictive modelling project, feature importance is calculated to determine the importance of the independent variables in prediction of the targeted variables. This step gives good insight of the data points and with proper implementation can help in increasing efficiency for predictive modelling.

Data Encoding: The independent variables are numeric, but the target variables are in categorical format. However, the individual prediction is taken into considerations for predicting 4 conditions (cooler condition, valve condition, internal pump leakage and hydraulic accumulator) of the hydraulic system. The last target variable Stable will take these 4 faults into consideration. To predict the stability of the hydraulic system, the categorical variables in the dataset should be handled by performing data encoding techniques. This makes the machine learning model easy to perform prediction better than performing on categorical data.

Data Scaling: When there is a need to build a predictive model, the independent variables should be normally distributed in a bell shape called as Gaussian distribution. A histogram is a great visualisation tool to check the distribution of data.

Dimensionality Reduction: After scaling the data, it is important that the dimensions of the input variables are reduced. This is important as there 17 dimensions in the dataset and some of them are highly correlated to each other. It is necessary to project these dimensions to a lower subspace to remove collinearity and also save computation time for processing. Dimensionality reduction techniques help in reducing the dimensions of the data without losing the quality of the data.

3.4 Modelling

The previous studies show various machine learning models built for predicting the conditions of the hydraulic system. In this research study, the model building process is approached in 3 ways:

- **Parametric and Non-parametric technique**
- **Gradient Boosting Decision Tree Algorithms**
- **Bagging (Bootstrap Aggregation)**

For all the above-mentioned approaches, hyperparameters optimisation process is carried out by RandomisedSearchCV. Two more techniques like GridSearchCV and Bayesian Optimization are studied for hyperparameter optimization. Compared to RandomisedSearchCV, the other two techniques require more time to find good hyperparameter combinations.

3.5 Evaluation

In this step, the classification models were evaluated to check the effectiveness and performance. This is done using evaluation techniques like model accuracy, confusion matrix along with classification report where precision, recall, f1-score and support are studied.

3.6 Deployment

After successfully implementing all the classification models and evaluating the results, a web app is deployed to predict the 4 conditions and the stability of the hydraulic system using Streamlit library in Python. On the web app, a link to the dashboard is given where exploratory data analysis is published using R shiny apps.

4 Implementation

With detailed explanation of the CRISP-DM methodology, the series of steps implemented are shown in Figure 6.

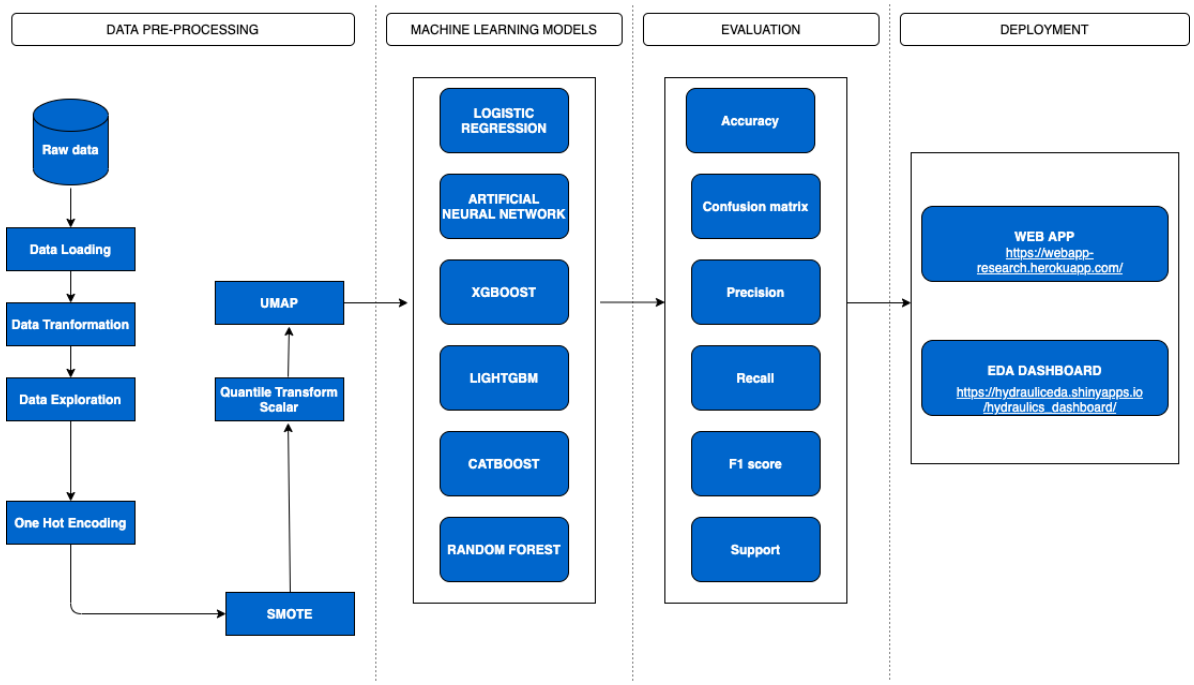

Figure 6: Methodology Flowchart

Exploratory Data Analysis: The dataset downloaded was converted from text to csv format and individual columns were combined together as dependent and independent dataframe. According to the dataset information, the quantities of the sensor data were in a cyclic duration of 60 seconds within 13-time intervals. It becomes difficult to analyse small values of sensor data and therefore the average cycle was taken to transform the independent variables to 2205 number of instances (Bykov, Voronov and Voronova, 2019). The targeted variables Valve and Accumulator consist of 4 class values ranging from optimal behaviour (100 for valve and 130 for accumulator) to close to total failure (73 for valve and 90 for accumulator). When the components of a hydraulic system are in severe lag or have severely reduced pressure then there is a high chance that the component will fail. Therefore, it is important to record every stage of failure in the hydraulic system.

Feature Importance: Machine learning helps in predicting the outcome of any business problem but why that outcome is predicted is dependent on the features responsible. The model built should be interpretable and therefore determining the feature importance is a crucial step. Here, the feature importance is done using a random forest to better understand the logic behind the predictions made by the model.

- **Valve:** Valve helps in handling the flow of liquid into different parts of the hydraulic system so that there is a low chance of the fluid getting pressurised at one part of the system. The pressure, efficiency factor and pump are the main sensors responsible for the failure of valve in the hydraulic system.

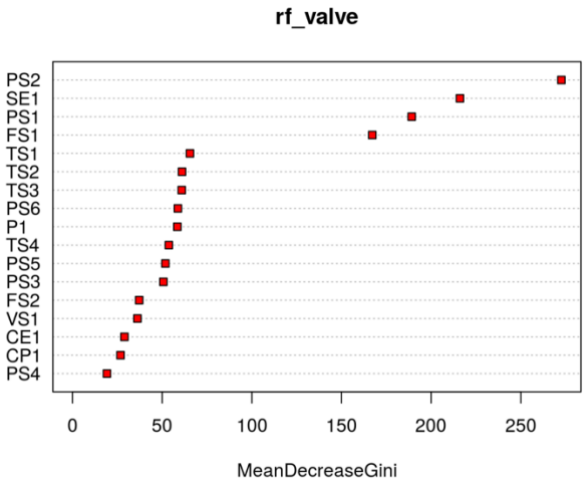

Figure 7: Variable Importance for Valve Condition

- **Pump:** Pump is used to generate flow for the fluid to performs tasks within the hydraulic system. If a hydraulic pump fails, there is a high chance that the hydraulic system will not be stable. The SE1 (Efficiency factor) is the most important variable responsible for predicting the pump leakage condition shown in Figure 8, as leakage will have a direct effect on the hydraulic system.

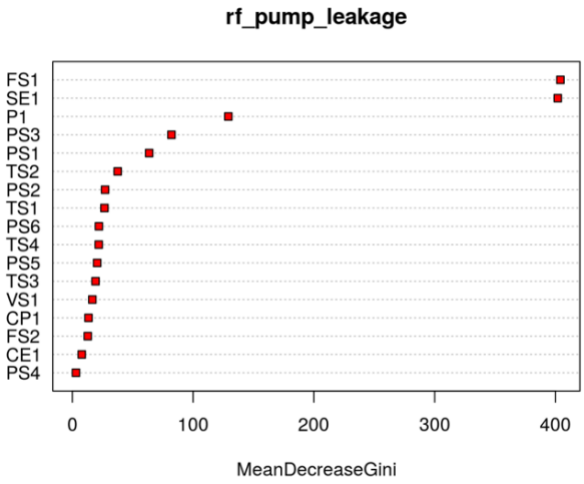

Figure 8: Variable Importance for Pump Condition

- **Stable:** Valves, volume flow and efficiency factor are the important features to the overall condition of the hydraulic system. From the Figure 9, it can be seen that coolers are not important in predicting stability as coolers and stable variables are independent of each other.

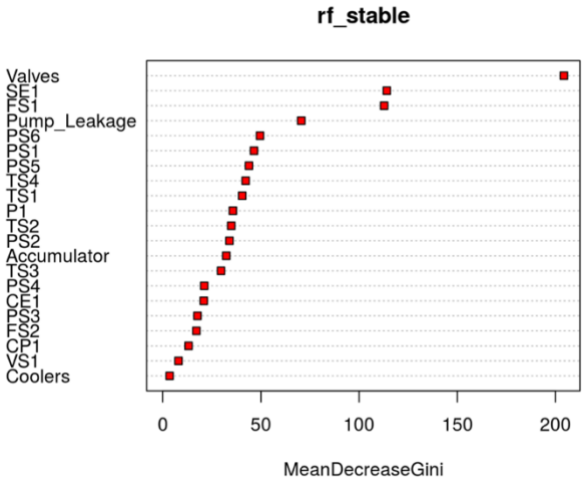

Figure 9: Variable Importance for Stable Condition

Quantile Transform Scalar: The distribution of the independent variables is highly skewed and randomly distributed with outliers. Using quantile transform scalar, it changes the probability of the variable value to another probability distribution using cumulative probability distribution (CDF). The pressure-flow (PS1) data transformed from randomly distributed to normally distributed is shown in Figure 10. Similarly, all the independent variables were transformed using quantile transform scalar.

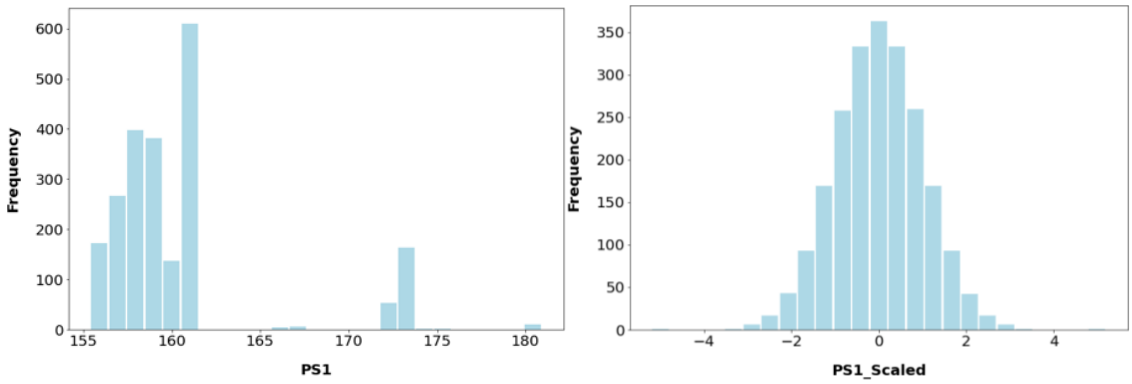

Figure 10: Quantile Transform Scalar

One-Hot Encoding: Machine learning algorithm works well on numerical data than on categorical data. All the independent variables are in numeric format, but to predict the stability of the hydraulic system there are 4 categorical values which should be encoded. Thus, one-hot encoding is used to transform the categorical variables like coolers, valves, pump leakage and accumulator.

Classification Imbalance using Synthetic Minority Oversampling Technique (SMOTE): Imbalanced dataset will cause poor performance in the model building process. The best way to handle such dataset is by creating synthetic data which will help in balancing the dataset. Oversampling technique with SMOTE is used to create synthetic data to help minority classes have equal value with other classes while training the model. The Figure 11 explains the dealing of classification imbalance in the valve component.

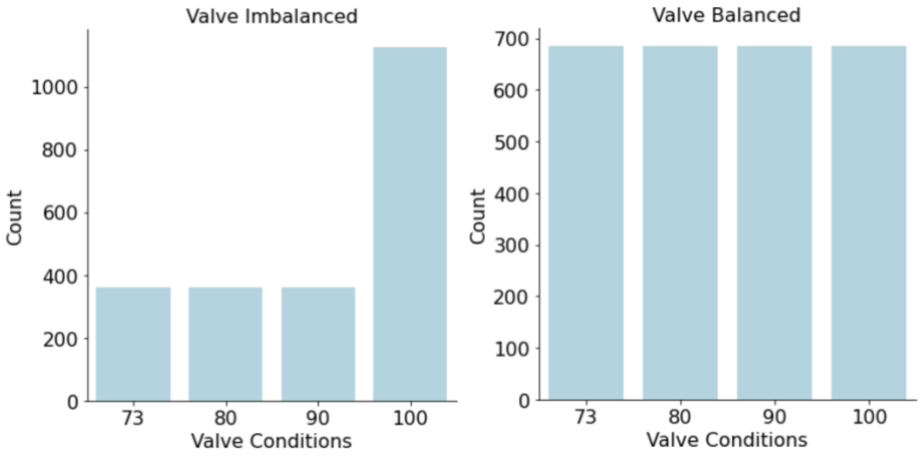

Figure 11: Dealing with Classification Imbalance

Dimensionality Reduction: The independent variables in its raw form are non- linear and consist of a large number of features. A higher number of features may have data points that may overlap with each other. When data points are overlapped, it creates an error during the training of the machine learning model process. Three techniques were applied to reduce the input dimensions as shown in Table 2.

Table 2: Comparison of Dimensionality Reduction Techniques

Dimensionality Reduction Technique	Time	Drawbacks
PCA	30ms	Does not separates the classes
t-SNE	18.2s	Slow and global properties are not captured
UMAP	8.06s	Slow than PCA

- **Principal Component Analysis (PCA):** In Figure 12, Principal Component Analysis is used to reduce the dimensions and the scatter plot is used to represent the class distribution of the data points. PCA uses the concept of variance to calculate the important dimension, it may help in reducing the dimensions, but the data points are overlapping with each other and thus information are lost.

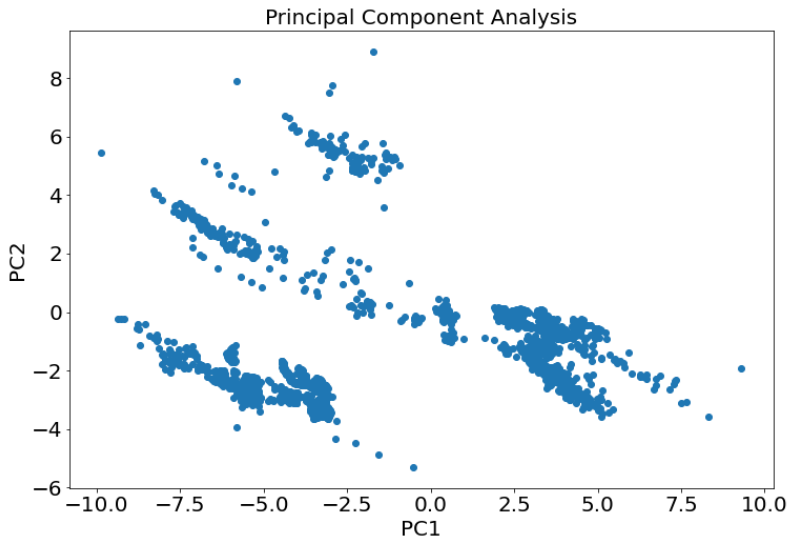

Figure 12: PCA

- **t-Distributed Stochastic Neighbor Embedding (t-SNE):** Due to the higher number of features and drawback of PCA, the 2nd dimensionality reduction technique used is t-SNE. It helps in identifying similar data points and in Figure 13, it can be seen that these data points are grouped together. Unlike PCA, the scatter plot shows that the cluster information is not lost, and all the local structures of the independent variables are clustered separately. The local structures are preserved when dimensions are reduced using t-SNE, but the global properties of the data are not preserved. Also, the computation time to reduce dimension using t-SNE technique is high.

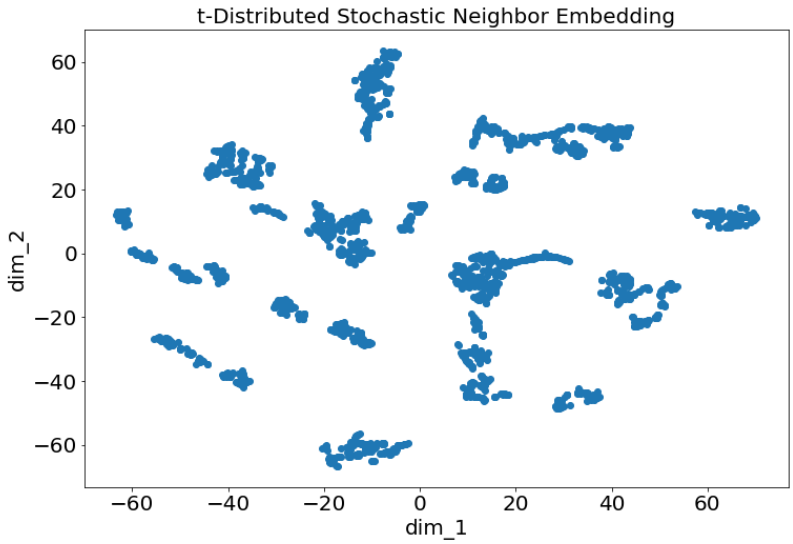

Figure 13: t-SNE

- **Uniform Manifold Approximation and Projection (UMAP):** Using PCA and t-SNE technique, the dimensions are reduced but there are drawbacks which are handled by UMAP. The UMAP technique performs clustering of data like t-SNE technique and reduces the dimensions like PCA. In Figure 14, it can be seen that clusters are formed much better than t-SNE and are separated far away from each other. This technique takes less time to compute than t-SNE and both the local and global structures of the data are preserved.

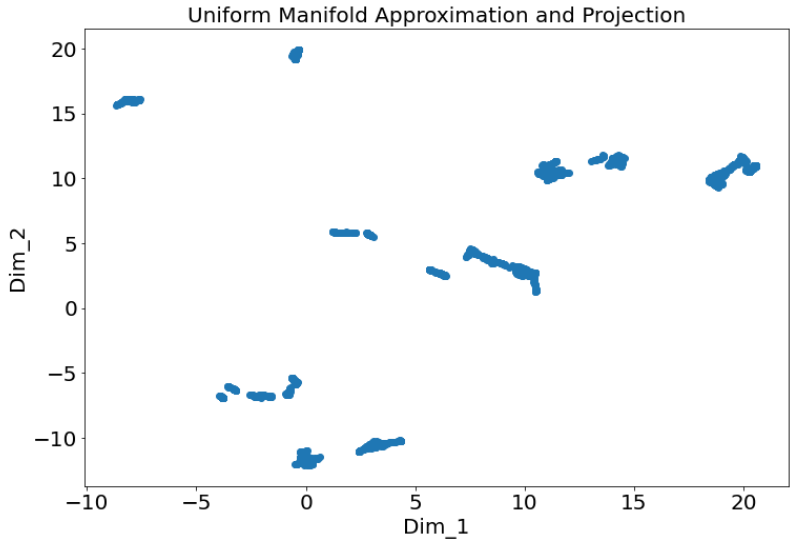

Figure 14: UMAP

Model Implementation: The stable condition is a binary value whereas the components of the hydraulic system have multi-class values. There are different set of classes and they are implemented in three approaches.

- **Approach 1 – Parametric and Non-parametric Technique:** The first approach is the implementation of Logistic Regression and Artificial Neural Network model. The parameters used in both the techniques are described by maximum likelihood estimation. Both algorithms use functional form *f* to determine the relationship between dependent and independent variables. Due to differences in the functional form *f*, Logistic Regression and ANN are defined as parametric and non-parametric method respectively (Dreiseitl and Ohno-Machado, 2002). The complexity in LR compared to ANN is low which makes ANN a more flexible model but is vulnerable to overfitting. Parameter chosen are shown in Table 3.

Table 3: Parameters for Approach 1

Function	Logistic Regression	Artificial Neural Network
Regularisation parameter	C	----------
Optimisation for training algorithm	max_iter, solver	batch_size, epochs, optimizer, kernel_initializer, units

- **Approach 2 - Gradient Boosting Decision Trees Algorithms:** The first gradient boosting machine (GBM) was developed in 2002 where subsequent trees are used for prediction with the help of base learners (Friedman, 2002). After 14 years, a better and scalable machine learning model eXtreme Gradient Boosting (XGBoost) was developed using the GBM algorithm. The drawbacks in GBM like the ability to give high variance result and not cover linear combinations were solved by XGBoost (Chen and Guestrin, 2016). Microsoft in 2017 developed LightGBM which was again based on GBM algorithm to reduce the computational complexity. The performance of LightGBM was compared to different GBM algorithms on 5 different datasets and the results evaluated shows that LightGBM was faster than other GBM algorithms (Ke et al., 2017). A year later, CatBoost was launched by Yandex Technologies to better handle categorical variables and boost the performance without any exceptional treatments (Dorogush, Ershov and Gulin, 2018). All the parameters taken for approach 2 are shown in Table 4.

Table 4: Parameters for Approach 2

Function	XGBoost	LightGBM	CatBoost
Parameters to control overfitting	learning_rate max_depth min_child_weight	learning_rate max_depth	max_depth max_feature criterion
Parameters for controlling speed	colsample_bytree	colsample_bytree	n_estimators
Regularisation parameter	gamma	----------	----------
Objective and evaluation	----------	objective boosting_type metric	bootstrap

- **Approach 3 – Bagging Technique:** In the third approach, the Random Forest algorithm is used for prediction of the binary and multi-class classification problem. The initial stage consists of multiple decision trees which are provided to different base learner models. This is the bootstrap stage and the outputs from different base learner models are aggregated using a voting classifier. the majority votes from the classifier define the output (Glocker, Haynor and Recognition, 2016). Thus, it known as Bootstrap aggregation technique.

Table 5: Parameters for Approach 3

Function	Random Forest
Objective and evaluation	bootstrap
Parameter to control overfitting	max_depth max_features min_samples_leaf criterion
Parameter for controlling the speed	n_estimators

5 Result Evaluation

The entire study of the research is analysed and evaluated. For individual components, random forest algorithm is used to check the important variables. The independent variables help in determining their importance with the targeted variables. A chi-square test is performed between the categorical variables to define their relationship with each other. The test is performed on the basis of individual variables corresponding to the stability of the hydraulic system. Finally, 6 classification models were built and evaluated.

Pearson's Chi-squared test:

- **Null Hypothesis:** Assuming that the cooler, valve, pump and accumulator condition are not related to the stable condition of the hydraulic system.
- **Alternated Hypothesis:** Assuming that the components are related with the stable condition.

Table 6: Pearson's Chi-squared test

Component	X-squared	df	p-value
Cooler	0.038177	2	0.9811
Valve	1104.4	3	< 2.2e-16
Pump	799.91	2	< 2.2e-16
Accumulator	365.28	3	< 2.2e-16

- **Solution:**
  - The level of significance in Cooler is greater than 0.05, therefore we accept the null hypothesis and state with 95% confidence interval that there is no relationship between the cooler and stable condition of the hydraulic system.
  - The level of significance in valve, pump and accumulator components are less than 0.05. This means that we reject the null hypothesis and conclude with 95% confidence interval there is relationship between valve, pump and accumulator condition with the stable condition.

Result 1 – Fault Diagnosis in Stable Condition: A pie chart is used to classify the variables important for the stable conditions.

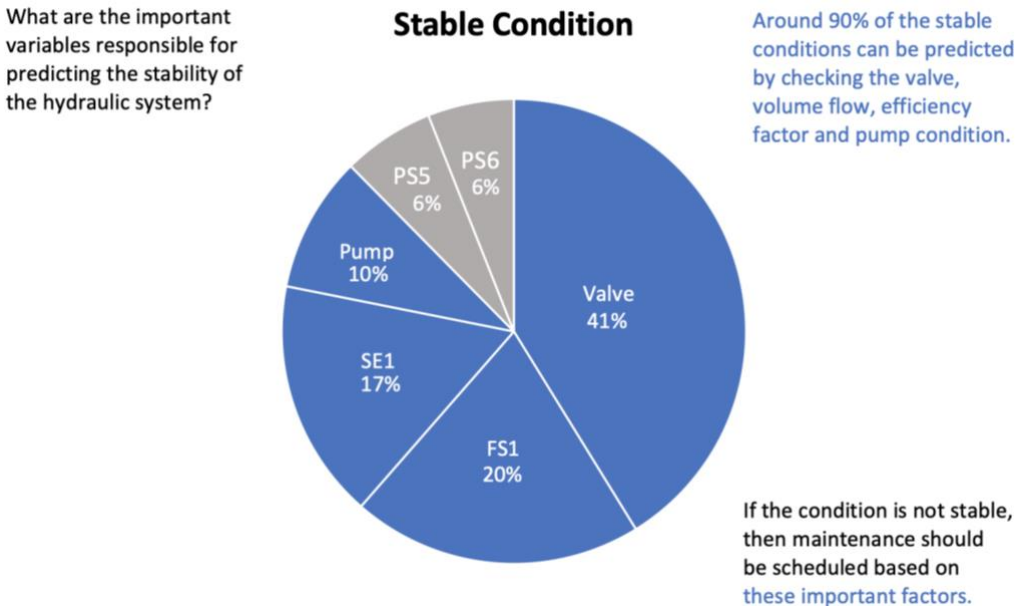

Figure 15: Fault Diagnosis - Stable Condition

From the Figure 15, it can be observed that the valve and the pump are the two components that should be observed to determine stable conditions of the hydraulic system.

Failure Prognosis for Stable Condition: Highest performance for overall stability of the system can be predicted using CatBoost and LightGBM model. Though their accuracy is the same, the ability to properly classify the true class denoted by recall in the Catboost model.

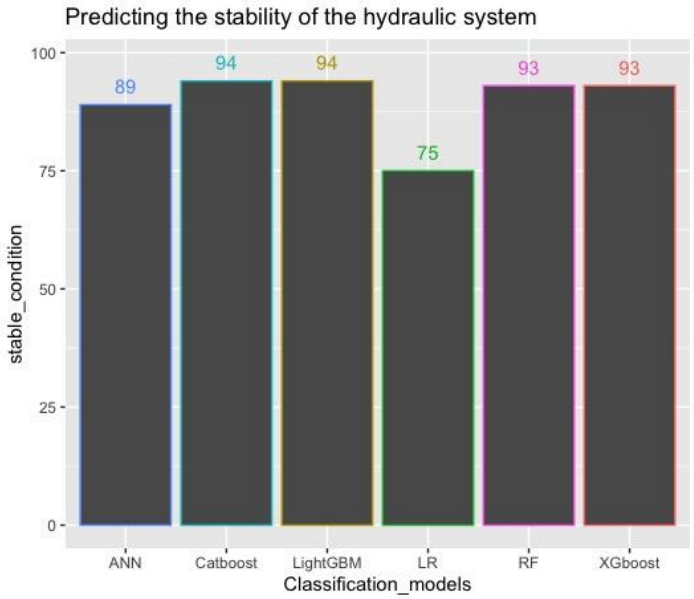

Figure 16: Failure Prognosis for Stable Condition

Catboost Classification Report for Stable Condition: The confusion matrix displays the corrected predicted values of binary class. Recall shows the correctly predicted class and precision shows the positive classes predicted correctly. The recall and precision values show that there were a high number of positive examples predicted correctly but also predicting lot of false-positive values. F1-score denotes the harmonic mean and support mentions the true predicted values in the class.

```
Confusion Matrix :
 [[539  24]
 [ 32 287]]
              precision    recall  f1-score   support

           0       0.94      0.96      0.95       563
           1       0.92      0.90      0.91       319

    accuracy                           0.94       882
   macro avg       0.93      0.93      0.93       882
weighted avg       0.94      0.94      0.94       882
```

Figure 17: CatBoost Classification Report for Stable condition

Result 2 – Fault Diagnosis for Valve Condition: According to the Pearson's Chi-square test and Result 1, stable condition is 41% dependent on the valve condition. Figure 18 explains the responsible factors for valve condition.

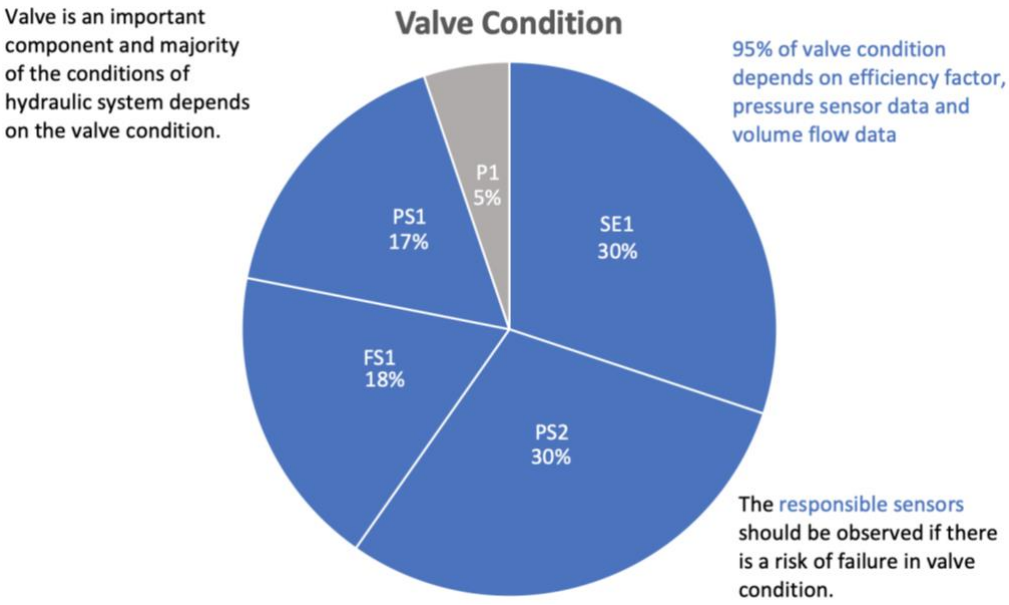

Figure 18: Fault Diagnosis for Valve Condition

In a hydraulic system, the direction of the flow of liquid is controlled by the valve. Therefore, the pressure PS1, PS2 and volume flow FS1 are the important sensors for the valve condition. If the efficiency of the system is low, then it will affect the valve condition.

Failure Prognosis for Valve Condition: Valve is the important component of the hydraulic system and based on the classification report, the best performing model is ANN.

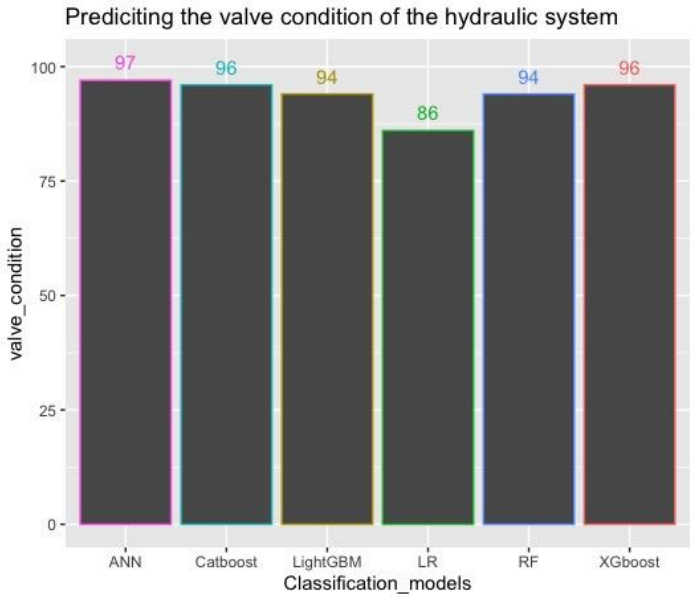

Figure 19: Failure Prognosis for Valve Condition

ANN Classification Report for Valve Condition: The multi-class problem shows the confusion matrix with the correct and incorrect predicted values. Macro average values show

that it has a high recall and low precision values indicating a lot of true predicted values classified correctly but with false-positive predicted values.

```
Confusion Matrix :
 [[145   8   0   0]
 [  1 136   0   0]
 [  0   1 135   4]
 [  0   3   7 442]]
              precision    recall  f1-score   support

          73       0.99      0.95      0.97       153
          80       0.92      0.99      0.95       137
          90       0.95      0.96      0.96       140
         100       0.99      0.98      0.98       452

    accuracy                           0.97       882
   macro avg       0.96      0.97      0.97       882
weighted avg       0.97      0.97      0.97       882
```

Figure 20: ANN Classification Report for Valve Condition

Result 3 – Fault Diagnosis for Pump Condition: After valve, stable condition is 10% dependent on the pump performance. The important sensors are highlighted in the pie chart Figure 21.

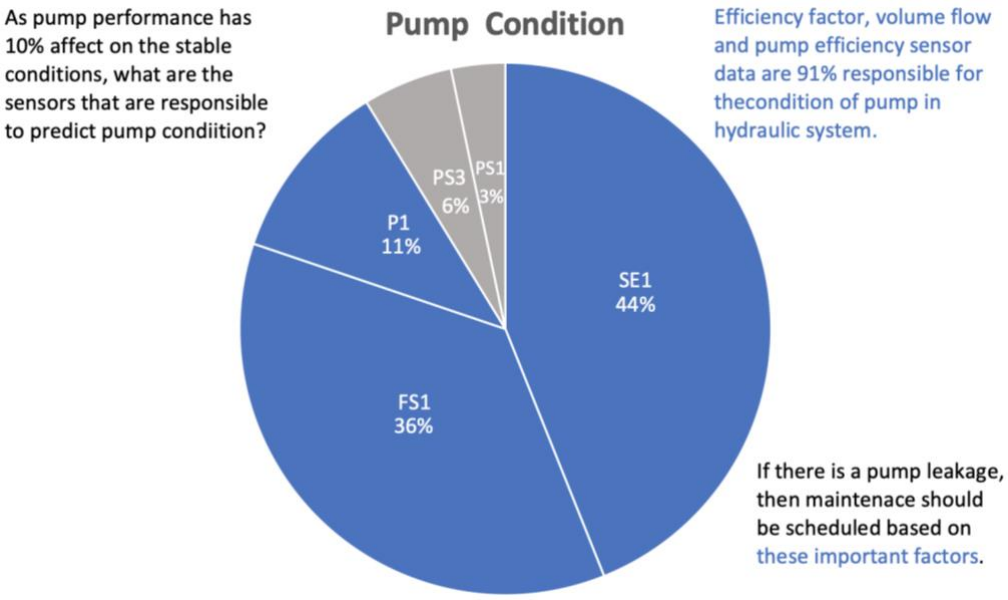

Figure 21: Fault Diagnosis for Pump Condition

The volume flow FS1 is 36% important as the pump determines the flow of the liquid in the hydraulic system. If the pump condition is severe, then it will have a direct effect on the efficiency of the hydraulic system.

Failure Prognosis for Pump Condition: For predicting the pump condition, LightGBM and Random forest models are performing better than other classification models.

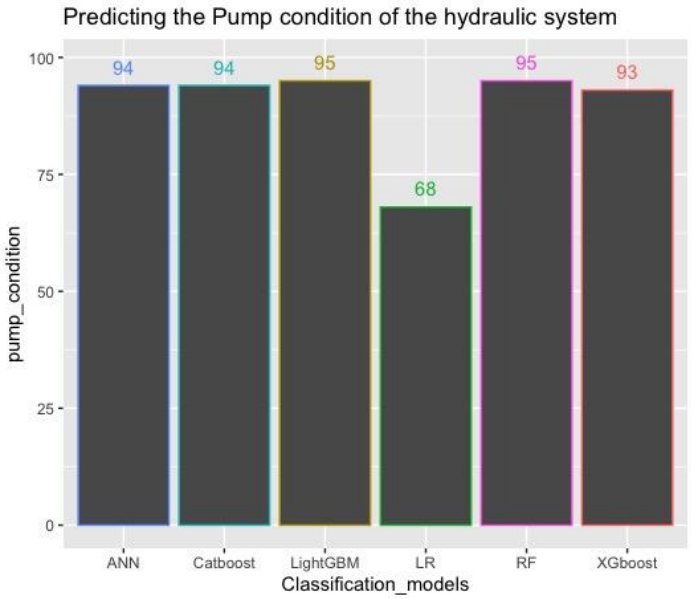

Figure 22: Failure Prognosis for Pump Condition

LightGBM Classification Report for Pump Condition: The values of precision and recall differs by 1% and that shows with high precision and low recall that some of the predicted values which were positive were not predicted correctly. Based on the classification report, LightGBM is better in predicting the classes based on precision and recall values.

```
Confusion Matrix :
 [[471   5   4]
 [  3 190  14]
 [  0  14 181]]
              precision    recall  f1-score   support

           0       0.99      0.98      0.99       480
           1       0.91      0.92      0.91       207
           2       0.91      0.93      0.92       195

    accuracy                           0.95       882
   macro avg       0.94      0.94      0.94       882
weighted avg       0.96      0.95      0.95       882
```

Figure 23: LightGBM Classification Report for Pump Condition

6 Discussion

A lot of study has gone into improving the condition monitoring system in industries. The shift from model-based to data-driven approach has been successful in fault analysis. Approaches in the previous study and this research study differ only on the factors of the data processing step. With numerous methods, presenting data and transforming them without losing the quality is challenging. All these factors depend on the data which is extracted with the sensors installed inside the hydraulic system.

The three approaches are compared with the previous study. The final results show that using ensemble methods in multi-classification problems are outstanding. Random forest and the gradient boosting decision trees algorithms with parameter tuning perform better in

predicting the multi-class values. Previous study mentions that there are chances of performing better in predicting values in the pump leakage and accumulator conditions by 25% and 40% respectively (Helwig, Pignanelli and Schutze, 2015). This was achieved by (Hu, 2012), where the highest accuracy achieved was 93% in Pump and 92% in Accumulator using sphered support vector machines. Our results were improved from 93% to 95% in Pump and 92% to 98% in Accumulator as shown in Table 7. The improvements achieved in the past and this research depends on the pre-processing techniques. The advancements in the Gradient boosting decision trees algorithms are also responsible to perform better than previous machine learning models.

Table 7: Discussion

Variable	**Previous Approach**	**Approach 1**	**Approach 2**	**Approach 3**
Cooler	96% (SVM)	100% (ANN)	100% (Catboost)	100% (RF)
Valve	95% (ANN)	97% (ANN)	96% (Catboost)	94% (RF)
Pump	93% (SVM)	94% (ANN)	95% (LightGBM)	95% (RF)
Accumulator	92% (SVM)	92% (ANN)	97% (Catboost)	98% (RF)
Stable	97% (SVM)	90% (ANN)	95% (Catboost)	93% (RF)

7 Deployment

Figures 24 and 25 in the deployment section are deployed and published to web using Heroku and R Shiny apps respectively. The link for Figure 24: **https://webapp-research.herokuapp.com/**. All the dataset used in the research project is displayed in the Select Dataset section like cooler, valve, pump leakage, accumulator and stable condition. Except for LightGBM classifier, all the models can be used in the web app. The selected classifier shows the model implementation along with parameter tuning on the sidebar. It shows the shapes, number of classes and the classifier used for fault prognosis in the hydraulic system. Finally, a confusion matrix is displayed to show the correct and false predicted values by the selected algorithms. Different parameters can be tried on the web app for fault prognosis in the hydraulic system.

To show the fault diagnosis of the hydraulic system, a dashboard is created and published to web using Shiny apps in R. The link for Figure 25: **https://hydrauliceda.shinyapps.io/hydraulics_dashboard/**, which shows the histogram of the independent variables in the dataset. However, the screenshot is just showing one part of the dashboard and the dashboard shows other options like data exploration, variable importance using Random Forest and outliers in the dataset.

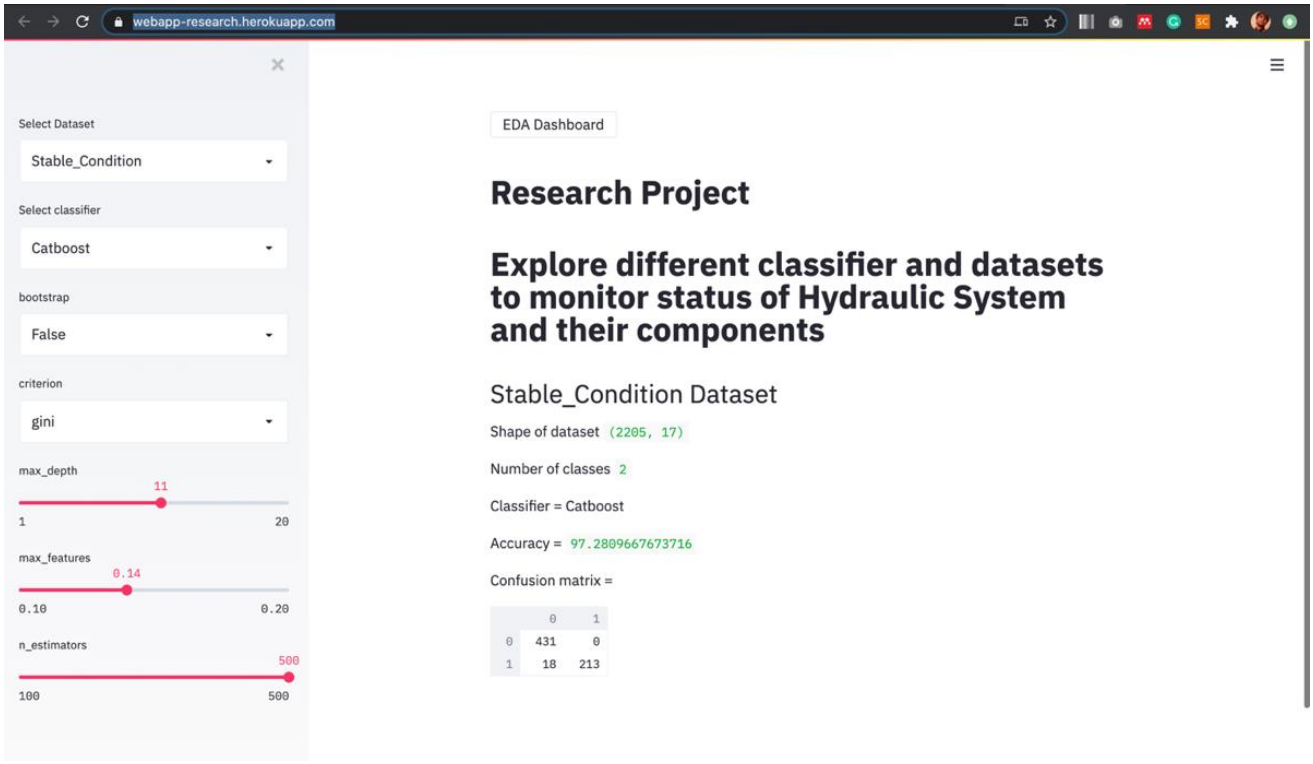

Figure 24: WebApp

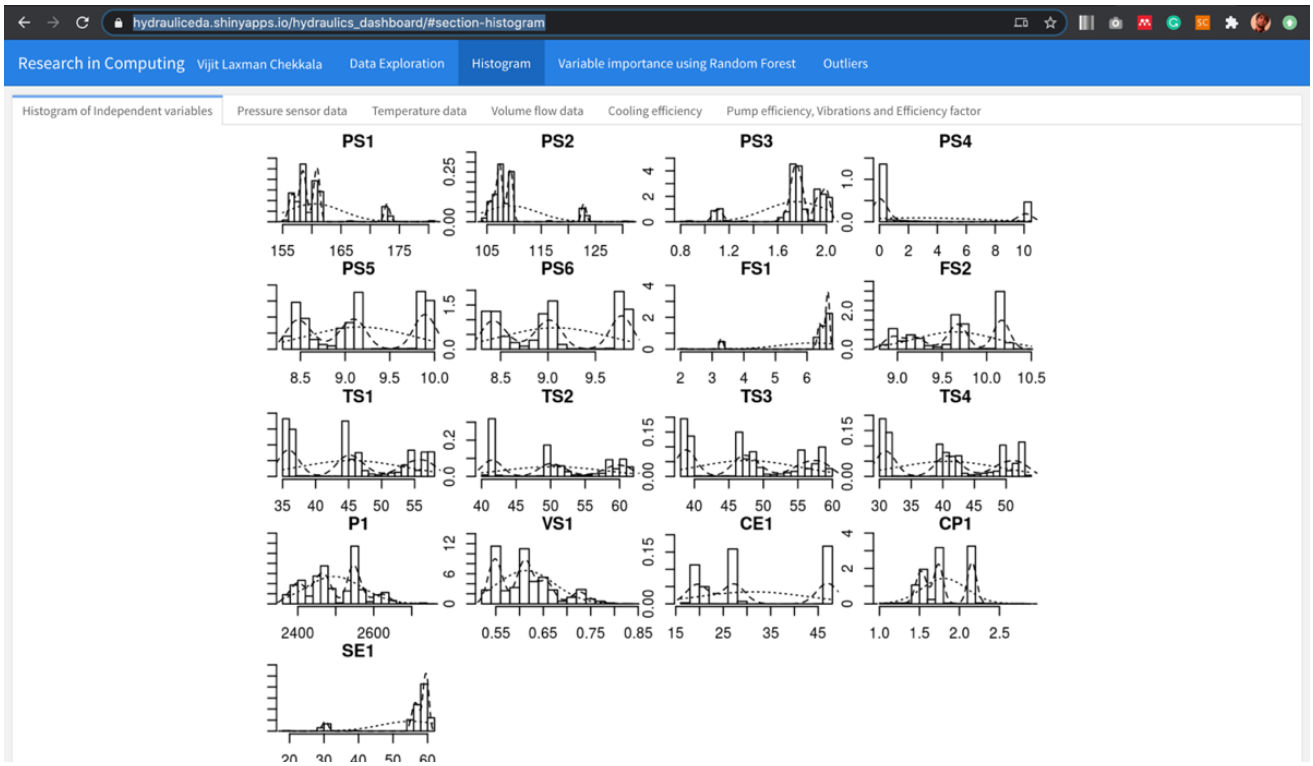

Figure 25: EDA Dashboard

8 Conclusion and Future Work

The entire research was carried out to properly diagnose the conditions in the hydraulic system for preventive maintenance and prognosis using machine learning techniques for predictive maintenance. All the approach in the research study were done in terms of business purpose where these results can be implemented to schedule maintenance. Different scaling techniques and the importance of outliers in the data were studied. Three dimensionality reduction techniques like PCA, t-SNE and UMAP were implemented and UMAP was selected for further implementation as it was faster than t-SNE and no data were lost during the reduction process like PCA. After representing the data into the highest quality for the classification modelling and evaluation, the best results were obtained by the gradient boosting decision trees algorithms. CatBoost and LightGBM are robust models that help in predictive modelling problems with little complexity. The classification reports were similar in both the model's evaluation part, these models can be developed for fault prognosis in monitoring systems. An advanced version of predictive modelling for binary and multi-classification problem with less computational complexity was developed for the fault diagnosis and failure prognosis in the hydraulic system.

With so much advancement, there are intelligent devices which do not need to be monitored continuously and remote devices that can be monitored at a far location away from the installed system. The difference between all these systems can be studied and the same approach can be used to build a predictive model. For this research, only some of the data were recorded using vibrational analysis. With so much growth in the monitoring system, the sensors should record the data in real-time using vibrational analysis. Performing real-time analysis will be better than any maintenance program as every problem can be observed and proper maintenance can be scheduled.

References

Bykov, A. D., Voronov, V. I. and Voronova, L. I. (2019) 'Machine Learning Methods Applying for Hydraulic System States Classification', *2019 Systems of Signals Generating and Processing in the Field of on Board Communications, SOSG 2019*. IEEE, pp. 1–4. doi: 10.1109/SOSG.2019.8706722.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu, pp. 785–794. doi: 10.1145/2939672.2939785.

Cui, J. *et al.* (2019) 'System', *2019 Chinese Control And Decision Conference (CCDC)*. IEEE, pp. 6093–6097.

Di, Y. *et al.* (2017) 'A data mining approach for intelligent equipment fault diagnosis', *Proceedings of 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2017*, pp. 1082–1086. doi: 10.1109/IAEAC.2017.8054180.

Dorogush, A. V., Ershov, V. and Gulin, A. (2018) 'CatBoost: gradient boosting with categorical features support', pp. 1–7. Available at: http://arxiv.org/abs/1810.11363.

Dreiseitl, S. and Ohno-Machado, L. (2002) 'Logistic regression and artificial neural network classification models: A methodology review', *Journal of Biomedical Informatics*, 35(5–6), pp. 352–359. doi: 10.1016/S1532-0464(03)00034-0.

Friedman, J. H. (2002) 'Stochastic gradient boosting', *Computational Statistics and Data Analysis*, 38(4), pp. 367–378. doi: 10.1016/S0167-9473(01)00065-2.

Glocker, B., Haynor, D. R. and Recognition, I. (2016) 'Random Decision Forest Random Forests for Localization of Spinal Anatomy Data Fusion Methodology and Applica- tions'.

Helwig, N., Pignanelli, E. and Schutze, A. (2015) 'Condition monitoring of a complex hydraulic system using multivariate statistics', *Conference Record - IEEE Instrumentation and Measurement Technology Conference*, 2015-July, pp. 210–215. doi: 10.1109/I2MTC.2015.7151267.

Higgs, P. A. and Author, L. (2004) 'ESDA2004-58216', pp. 1–16.

Hu, X. (2012) 'Study on fault diagnosis of hydraulic pump based on sphere-structured support vector machines', *2012 2nd International Conference on Consumer Electronics, Communications and Networks, CECNet 2012 - Proceedings*. IEEE, pp. 2894–2896. doi: 10.1109/CECNet.2012.6201946.

Hutamarn, S., Pratumsuwan, P. and Po-Ngaen, W. (2012) 'Neuro-fuzzy based on support vector machine friction compensator in servo hydraulic system', *Proceedings of the 2012 7th IEEE Conference on Industrial Electronics and Applications, ICIEA 2012*. IEEE, pp. 2118–2122. doi: 10.1109/ICIEA.2012.6361080.

Jegadeeshwaran, R. and Sugumaran, V. (2015) 'Health monitoring of a hydraulic brake system using nested dichotomy classifier – A machine learning approach', *International Journal of Prognostics and Health Management*, 6(1), pp. 1–10.

Kano, M. and Nakagawa, Y. (2008) 'Data-based process monitoring, process control, and quality improvement: Recent developments and applications in steel industry', *Computers and Chemical Engineering*, 32(1–2), pp. 12–24. doi: 10.1016/j.compchemeng.2007.07.005.

Ke, G. *et al.* (2017) 'LightGBM: A highly efficient gradient boosting decision tree', *Advances in Neural Information Processing Systems*, 2017-Decem(Nips), pp. 3147–3155.

Kotlar, A. M., Iversen, B. V. and Jong van Lier, Q. (2019) 'Evaluation of Parametric and Nonparametric Machine-Learning Techniques for Prediction of Saturated and Near-Saturated Hydraulic Conductivity', *Vadose Zone Journal*, 18(1), pp. 1–13. doi: 10.2136/vzj2018.07.0141.

Michael, W. J. *et al.* (2005) 'Integrating data sources to improve hydraulic head predictions: A hierarchical machine learning approach', *Water Resources Research*, 41(3), pp. 1–14. doi: 10.1029/2003WR002802.

Schneider, T., Helwig, N. and Schütze, A. (2017) 'Automatic feature extraction and selection for classification of cyclical time series data', *Technisches Messen*, 84(3), pp. 198–206. doi:

10.1515/teme-2016-0072.

Stetco, A. *et al.* (2019) 'Machine learning methods for wind turbine condition monitoring: A review', *Renewable Energy*. Elsevier Ltd, 133, pp. 620–635. doi: 10.1016/j.renene.2018.10.047.

Wu, X. *et al.* (2015) 'Internal leakage detection for wind turbine hydraulic pitching system with computationally efficient adaptive asymmetric SVM', *Chinese Control Conference, CCC*. Technical Committee on Control Theory, Chinese Association of Automation, 2015-Septe, pp. 6126–6130. doi: 10.1109/ChiCC.2015.7260599.