Machine learning model for predictive maintenance of modern manufacturing assets

Abdu Shaala
Faculty of Technology, University of Sunderland
Sunderland, UK
Abdu.shaalan@sunderland.ac.uk

David Baglee
Faculty of Technology, University of Sunderland
Sunderland, UK
David.baglee@sunderland.ac.uk

Derek Dixon
Faculty of Technology, University of Sunderland
Sunderland, UK
Derek.dixon@sunderland.ac.uk

***Abstract*— Predictive maintenance is considered a powerful practice for manufacturing assets health assessment, facilitating the identification of potential failure occurrences. By proactively addressing such failures, manufacturers can avoid unplanned downtime and allocate necessary resources for required maintenance activities. Machine Learning (ML) methods have emerged as a promising tool for preventing equipment failures in Predictive Maintenance applications. However, the effectiveness of Predictive Maintenance applications is largely determined by the Machine Learning techniques utilized and the quality of the data utilised. In this research, we adapted the cross-industry standard process for data mining to develop a predictive maintenance model for a unique, large, and complex manufacturing asset, utilizing various machine learning techniques. Specifically, the research incorporate Random Forest, Support Vector Machines, K-Nearest Neighbors, eXtreme Gradient Boost, and Logistic Regression algorithms to the asset failure records. Following the fitting of all models, Random Forest emerged as the best-performing model based on the recall parameter. However, the algorithm performance was not satisfactory due to the poor data quality. In addition, an exploratory data analysis process was conducted on the data to derive insights into the failure pattern of the machine.**

Keywords—Machine Learning, Data analysis, Predictive Maintenance

I. INTRODUCTION

Predictive maintenance (PdM) is a proactive maintenance strategy that uses data analysis and machine learning algorithms to predict when equipment or machinery is likely to fail [12]. By analysing maintenance records and sensors data, PdM systems can detect patterns and anomalies that may indicate impending equipment failure [4]. PdM supports early scheduling of maintenance activities based on predictive analysis of existing data at optimum times [30]. This approach reduces downtime, extends equipment life, and eliminates the cost of unnecessary repairs and replacements. Modern approaches of PdM relies on a combination of machine learning, artificial intelligence, and data analysis to predict when maintenance activities are expected to be carried out [25]. It utilizes historical data and real-time data from sensors and other sources to identify patterns and anomalies that may indicate impending equipment failure [30]. Predictive maintenance has become increasingly popular in various industries including manufacturing, energy, and transportation, where equipment downtime can have significant costs and impact on productivity (Schmidt et al, 2018).

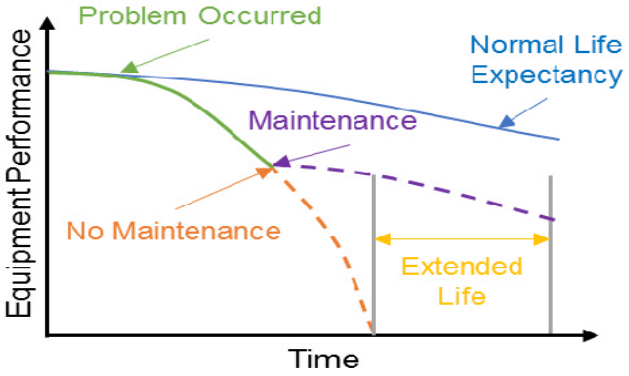

Fig.1. Effect of Life Expectancy with PdM (Lee et al., 2019).

PdM includes prognostic and diagnostics maintenance models. Diagnostics models detect the existence of operational problems and assess the root cause and effect on the functioning equipment [22], while prognostics models forecast the remaining useful life (RUL) or the future state based on present and historical conditions [22]. Knowledge-based approaches rely on knowledge that is constructed by utilizing past experiences to detect defects, analysing the decline of certain systems, and predict future failures.

According to [22], utilizing data to evaluate the deterioration rates of assets can lead to a more precise assessment, which in turn can enable more accurate predictions of degradation. On the other hand, model-driven approach utilize mathematical analysis to explain the process of degradation [1]. In model-driven approach, a mathematical model of the equipment or system is created using collected data to identify the relationship between machines performance and maintenance need in its design and operating conditions. The model is then used to simulate the behaviour of the equipment or system under different conditions and to identify potential failure modes and their associated causes [8].

A. Cross-Industry Standard Process (CRISP) data mining

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a widely recognized and systematic methodology introduced in the early 2000 was designed to guide organizations through the complex journey of data mining and analytics projects [10]. This well-established framework consists of six distinct phases, each with a specific set of tasks and objectives. sarting with understanding the business problem and data collection, it progresses through data preparation, modelling, evaluation, and ultimately deployment [26]. CRISP-DM places a strong emphasis on

aligning data mining efforts with the organization's objectives and ensures that the data-driven solutions created are practical and valuable. As a versatile and iterative process, CRISP-DM has remained a cornerstone in the field of data science, offering a structured approach to extracting actionable insights and knowledge from data, making it an indispensable tool for businesses and data professionals alike deployment [25].

B. Machine learning, ML, algorithms

ML is the study of computer system statistical models that are used to execute tasks based on patterns and inference rather than explicit instructions and enables computers to learn about a specific topic [2]. Algorithms for machine learning, a subset of artificial intelligence, construct a mathematical model from training or sample data to predict the future or make decisions covertly [26]. With wide range of algorithms that exists, they are classified based on their learning style, which are mainly, Supervised, un-supervised or similarity in function or form [2]. Support Vector Machines (SVM) is one of the most widely used ML techniques, and it is well known for its ability to deal with classification and regression applications of varying complexity with high accuracy [15]. SVMs were initially established as non-probabilistic binary classifiers [30] However, they are now used in multi-class problems [16].

One of the most important characteristics of SVM is its high precision in separating different classes of data, as well as its ability to determine the best point for separating classes of data [26]. Support Vector Regression (SVR) is the name given to SVM when applied to a regression problem. They are then fitted to regression models and used to predict degradation levels and calculate remaining useful life values [30].

Random Forest (RF) originally proposed by Leo Breiman in 2001. RF is an ensemble learning technique which builds multiple decision trees by selecting features and instances at random and then aggregating their predictions using the simple average [29]. This technique is made up of many decision trees that achieve excellent results in a low-dimensional dataset or where the number of variables exceeds the number of observations [31]. The RF algorithm overcomes the overfitting problem in the single decision tree technique by structuring many trees on diverse bootstrapped data to efficiently decrease variance and increase accuracy.

Artificial Neural Network (ANN) is one of the widely used ML technique with several industrial application including predictive control [27]. They are intelligent computation techniques that are inspired by biological neurons [8]. ANN is made up of three layers: an input layer, hidden layers, and an output layer. It has a number of input and output nodes that are not connected to one another. The relationship between them is nonlinear, and it is obtained through the use of weighting functions [15]. ANN decisions are made based on historical data; hence, no expert knowledge is required. Even if the data is inconsistent, the technique does not degrade because it is Robust; and by building an accurate ANN for a specific application, it can be used in real-time without having to change its architecture at every update [15]. The downside to this technique however is that the networks can reach conclusions that contradict the applications' rules and theories. Also, training ANN models takes time; they are "black box" methods meaning that it is impossible to know why the model has reached an output prediction [30].

Linear regression (LR) Linear Regression is the most used prediction model for determining the relationship between variables [13]. It predicts the future of a dependent variable (target) using linear relationships between the target and one or more independent variables (predictors) [20]. The prediction is based on the premise that the target-predictor connection is dependent or causal. There are two types of linear regression: simple regression and multiple regression (MLR) [14].

K-Nearest Neighbour (KNN) is a supervised classification technique that predicts data by identifying similarities to the underlying data. KNN is most commonly used to handle classification issues, although it also may be used in regression applications. The process of determining the nearest neighbours from a given data collection may be characterised as the process of determining the closest point to the input point [30]. The technique operates by calculating the distance between these points. It calculates the distance between each data point and the test data and then determines the likelihood that the points are similar to the test data.

Classification is based on which points have the highest probability of being correct [8].

Xtreme Gradient Boost (XGBoost) was created to have extremely high predictive capabilities. Nonetheless, its usage was restricted since the technique needed just one decision tree to be generated at a time in order to reduce the mistakes of all prior trees in the model. As a result, even small-scale models required a significant amount of time to train XGBoost were developed to minimize the model training time and improve the model performance by changing the way the algorithm works. XGboost is an ensemble learning algorithm that may produce a succession of weak learners through continuous training and then combine these weak learners to produce a strong learner. Its wide application is due to its accuracy, computational speed, ability to deal with missing values and its scalability [7].

Logistic regression is similar to linear regression, except it uses a binomial response variable. Logistic regression models use a linear combination of input datapoints to address a binary classification problem [29].

Logistic regression is a statistical method that issued to predict the probability of an outcome. It is a special case of linear regression where the target variable is categorical. The algorithm predicts the probability of occurrence of an event by fitting data to a logistic function.

II. RESEARCH APPROACH

The study was carried out by utilizing CRISP_DM as the approach for planning, executing, and evaluating data mining.

The study employed Tableau and the Python programming language as its tools. Specifically, Tableau was utilized for the preliminary stages of data cleansing, exploration, and visualization. Meanwhile, various Python libraries, including Matplotlib, Numpy, Pandas, and Scikit-Learn, were employed for subsequent tasks such as data pre-processing, feature engineering, model training, and validation. The process is represented in the figure below.

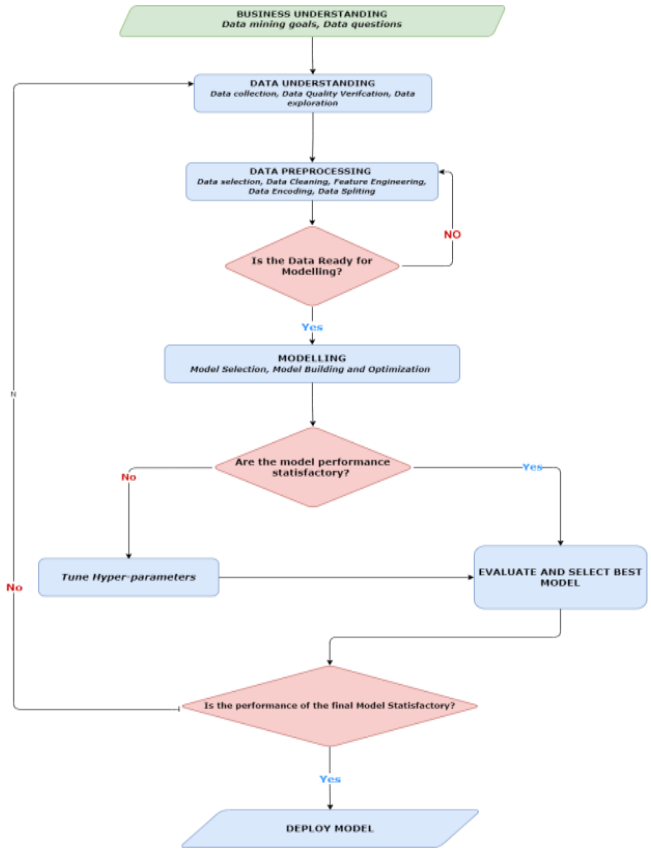

Fig.2. Crips-dm (Huber et al, 2019).

A. *Business Understanding*

This phase involved seeking to understand the business objectives and requirements. To address the frequent breakdowns of the subsea cable manufacturing machine, it was necessary to transform the task into a data mining problem. These breakdowns were resulting in increased downtime, maintenance costs, and operational issues. Hence, to address this issue, data questions such as the following were considered.

i. What is the failure pattern within the period under consideration?

ii. How does the failure pattern impact the total run time of the machine?

iii. Which components are impacting most on the machine's failure?

iv. Which fault types are predominant?

v. What is the time between failures?

vi. How long does it take to repair the machine by the fault type?

vii. Are there specific days, times, or seasons with which failure is prevalent?

The data mining goal of this project is to develop a model to predict failures ahead of time before they occur based on the historical failure data, the results of the fitted ML models are presented in the next chapter.

B. *Data Understanding*

This phase consists of three major stages involving data collection, data quality verification, and data exploration. The study collected historical failure and maintenance data from a tailored Computerized Maintenance Management System, CMMS, system developed specifically for the company under study for a period of 17 months. The data consisted of over 10,000 data points recording the various maintenance operations in the machine over the period of consideration. However, reference [23] noted that to solve a typical predictive maintenance problem through the application of machine learning, three sets of data sources which include fault history, maintenance history, and machine condition data are required to study the degradation pattern in machines. This study focused on the failure records which consisted of over 440 data points after the final data preparation.

TABLE 1 VARIOUS FIELDS IN THE DATA UNDERSTUDY.

S/N	Field Name	Description
	Asset	The various asset level in the machine
	Object	The various component at each asset level
	Status	The status of work down
	Final TRT	The final run time of the machine
	Actual Start/Completion	The actual start and completion of machine repairs.
	Maintenance Operations	The various maintenance operations are undertaken by the machine
	Registration Date	The registration date of failures and other maintenance activities.
	Fault Description	The description of each fault
	Work Down	Details of work down
	Fault Type	The fault type occurring in each machine.

The analysis focused on failure records and aimed to predict the failures occurrences based on historical dates data and their dates. Analysing the data presented that the data had quality issues including inconsistency in records, repetition of same tasks under different applications, using different acronyms for similar tasks and incomplete information for same records, which were addressed during the data preparation. The final step involved exploring the failure records and gaining insights through descriptive analytics and studying failure patterns over time.

C. *Data Preparation*

This step involved improving the quality of data by selecting and pre-processing relevant data to create a final dataset. Data quality validation was performed, and missing data and labelling inconsistencies were addressed to maintain the verifiability of data's fusion attributes [15]. The goal of feature engineering is to improve the prediction capability of Machine Learning algorithms by developing new features from existing data [5].

TABLE 2 NEW DATA FRAME.

	Date	Occurrence	Tot. F	Month	Year	Week	Day	Week of Month	Operational day
Count	504	504.000000	504.000000	504.000000	504.000000	504.000000	504	504	504
Unique	504	NaN	NaN	NaN	NaN	NaN	7	NaN	NaN
Top	8-7-16	NaN	NaN	NaN	NaN	NaN	Tues.	NaN	NaN
Freq	1	NaN	NaN	NaN	NaN	NaN	72	NaN	NaN
First	3-1-16	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Last	20-5-17	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Mean	NaN	0.454365	0.875000	5.515984	2016.2777778	22.121032	NaN	3.343254	14.549603
Std	NaN	0.499840	1.252532	3.434071	0.448348	15.020305	NaN	1.304235	8.724951
Min	NaN	0.000000	0.000000	1.000000	2016.000000	1.000000	NaN	1.000000	0.000000
25%	NaN	0.000000	0.000000	3.000000	2016.000000	9.750000	NaN	2.000000	7.000000
50%	NaN	0.000000	0.000000	5.000000	2016.000000	18.500000	NaN	3.000000	14.000000
75%	NaN	1.000000	6.000000	12.000000	2017.000000	53.000000	NaN	6.000000	30.000000
Max	NaN	1.000000	6.000000	12.000000	2017.000000	53.000000	NaN	6.000000	30.000000

TABLE 3 STATISTICAL REPRESENTATION OF THE FINAL DATA FRAME.

Date	Occ.	Tot. F	Fri	Mon	Sat	Sun	Thu	Tues	Wed
1-3	1.0	2.0	0	0	0	1	0	1	0
1-4	1.0	2.0	0	1	0	0	1	0	0
1-5	1.0	2.0	1	0	0	0	0	1	0
1-6	1.0	1.0	0	0	0	0	0	0	1
1-7	1.0	1.0	0	0	0	0	1	0	0

The final dataset was used to generate new features, and two relevant features were chosen for the study, Occurrence and Total failures. The Occurrence feature ranges from 0 to 1, where 0 represents failure and 1 represents a normal operation. The distribution of "Occurrence" and "Total Failure" variables were analysed using the Seaborn library in Python. To avoid the model assuming false relationships between numerical variables, all variables were encoded and assigned binary values using one-hot encoding. One-hot encoding is a technique used in data preprocessing to convert categorical variables into a numerical format that machine learning algorithms can work with effectively [10].

D. Modelling

The next stage of the CRISP-DM approach consisted of model selection, building, and optimization. The algorithm chosen for predictive maintenance in machine learning depends on the problem to be solved. In order to predict the remaining useful life of a machine, regression models are generally recommended, while classification models are suitable for predicting machine failure. In this application, the objective is to predict whether a failure will occur within a given timeframe, making classification models the preferred choice. When selecting the best machine learning strategy, there is a need to balance the strengths and weaknesses of each approach. Hence, several techniques were analysed, and the hyper-parameters of the algorithms were improved to achieve an optimized model. The following strategies were considered for this application:

i. Random Forest (RF)

ii. K-nearest neighbors (KNN)

iii. Logistics Regression

iv. XGBoost

v. Support Vectors Machine (SVM)

E. Final Model Selection

The final model was selected by comparing the outcomes of all the models considered and selecting the best. The study found that Random Forest (RF) was the best model for predicting actual failures based on recall value, followed by SVM, which performed better than Logistic Regression. RF was also better at predicting actual negatives compared to SVM, as shown by the precision metric. This implies that RF is a more robust model for predicting actual failures while also having fewer alarms compared to SVM and Logistic Regression.

Several studies have compared different Machine Learning algorithms for predictive maintenance, including Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), k-Nearest Neighbour (KNN), Naïve

Bayes, Artificial Neural Network (ANN), and XGBoost. Results varied depending on the specific application, but RF was generally found to have higher performance, even with poor data quality. However, there is no universally accepted model for predictive maintenance, and access to more data sources is crucial for further improvement. In the research work of reference [6], the performance of a predictive maintenance model for automobile maintenance was improved by using additional GIS data, with both DNN and RF performing well.

Even with such results, the utilisation of the different models presented proximity in un-reliable prediction analysis that did not each sufficient percentage. Such outcome relies mainly on the quality of the data provided and the in ability of the models to provide efficient predictions. Such outcomes indicate the need for reliable sources of data to be able to have high ML applications efficiency. Multiple approaches could be applied including MonteCarlo applications to improve the data quality, but such initiative has not been applied on the data yet.

III. Conclusion

The current research explored the use of big data analytics, specifically machine learning, in predictive maintenance as a viable maintenance strategy to reduce downtime and associated costs. The study used failure data from a subsea cable manufacturing machine and applied the CRISP_DM data mining process to develop a predictive maintenance model. The analysis revealed that most failures occurred in the zone 1 turn table of the machine, with electrical faults being the major cause of failure. The study also found that the Random Forest model achieved the highest accuracy of 63% in predicting true failures when compared to other models. However, performance could have been improved if access to relevant and reliable data sources in developing a robust model had been made available.

REFERENCES

[1] Aivaliotis, P. *et al.* (2021) 'Prediction assessment methodology for maintenance applications in manufacturing', *Procedia CIRP*, 104, pp. 1494–1499.

[2] Baghoolizadeh, M., Nasajpour-Esfahani, N., Pirmoradian, M. and Toghraie, D., 2023. Using different machine learning algorithms to predict the rheological behavior of oil SAE40-based nano-lubricant in the presence of MWCNT and MgO nanoparticles. *Tribology International*, *187*, p.108759.

[3] Bona, G.D. *et al.* (2021) 'Implementation of Industry 4.0 technology: New opportunities and challenges for maintenance strategy', *Procedia Computer Science*, 180, pp. 424–429. Available at: https://doi.org/10.1016/j.procs.2021.01.258.

[4] Bousdekis, A. *et al.* (2021) *A Review of Data-Driven Decision-Making Methods for Industry 4.0 Maintenance Applications - ProQuest*.

[5] Cardoso, D. and Ferreira, L. (2021) 'Application of Predictive Maintenance Concepts Using Artificial Intelligence Tools', *Applied Sciences*, 11(1), p. 18.

[6] Chen, C. *et al.* (2020) 'Automobile Maintenance Modelling Using gcForest', in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*. *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, pp. 600–605.

[7] Chen, T. & Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[8] Chen, X., Van Hillegersberg, J., Topan, E., Smith, S. and Roberts, M., 2021. Application of data-driven models to predictive maintenance: Bearing wear prediction at TATA steel. *Expert systems with applications*, *186*, p.115699.

[9] Galetsi, P., Katsaliaki, K. and Kumar, S. (2020) 'Big data analytics in the health sector: Theoretical framework, techniques and prospects', *International Journal of Information Management*, 50, pp. 206–216.

[10] Hadjicostis, C.N., 2004. Periodic and non-concurrent error detection and identification in one-hot encoded FSMs. *Automatica*, *40*(10), pp.1665-1676.

[11] Huber, S., Wiemer, H., Schneider, D. and Ihlenfeldt, S., 2019. DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. *Procedia Cirp*, *79*, pp.403-408.

[12] Hurtado, J., Salvati, D., Semola, R., Bosio, M. and Lomonaco, V., 2023. Continual learning for predictive maintenance: Overview and challenges. *Intelligent Systems with Applications*, p.200251.

[13] Kavitha, S., Varuna, S. & Ramya, R. (2016). A comparative analysis on linear regression and support vector regression. 2016 Online International Conference on Green Engineering and Technologies (IC-GET).

[14] Krishnan, S. *et al.* (2016) 'ActiveClean: interactive data cleaning for statistical modeling', *Proceedings of the VLDB Endowment*, 9(12), pp. 948–959.

[15] Lee, W.J. *et al.* (2019) 'Predictive Maintenance of Machine Tool Systems Using Artificial Intelligence Techniques Applied to Machine Condition Data', *Procedia CIRP*, 80, pp. 506–511. Available at: https://doi.org/10.1016/j.procir.2018.12.019.

[16] Lu, B., Chen, Z. and Zhao, X. (2021) 'Data-driven dynamic predictive maintenance for a manufacturing system with quality deterioration and online sensors', *Reliability Engineering & System Safety*, 212, p. 107628. Available at: https://doi.org/10.1016/j.ress.2021.107628.

[17] Liu, B., Pang, J., Yang, H. and Zhao, Y., 2023. Optimal condition-based maintenance policy for leased equipment considering hybrid preventive maintenance and periodic inspection. *Reliability Engineering & System Safety*, p.109724.

[18] Liu, Q., Chen, Y., Liu, Y., Lei, Y., Wang, Y. and Hu, P., 2023. A review and guide on selecting and optimizing machine learning algorithms for daylight prediction. *Building and Environment*, p.110822.

[19] Maulud, D. & Abdulazeez, A. (2020). A Review on Linear Regression Comprehensive in Machine Learning. Journal of Applied Science and Technology Trends. 1 (4). pp. 140-147.

[20] Montero Jimenez, J.J. *et al.* (2020) 'Towards multi-model approaches to predictive maintenance: A systematic literature survey on diagnostics and prognostics', *Journal of Manufacturing Systems*, 56, pp. 539–557.

[21] Naji, A. *et al.* (2016) 'Maintenance Management and Innovation in Industries: A Survey of Moroccan Companies', *International Journal of Innovation*, 4(2), pp. 188–197.

[22] Paolanti, M. *et al.* (2018) 'Machine Learning approach for Predictive Maintenance in Industry 4.0', in *2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)*. *2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)*, pp. 1–6.

[23] Rathore, S.S. *et al.* (2019) 'An overview of diagnostics and prognostics of rotating machines for timely maintenance intervention', *IOP Conference Series. Materials Science and Engineering*, 691(1).

[24] Saeed Farahani *et al.* (2022) 'A data-driven predictive maintenance framework for injection molding process', *Journal of Manufacturing Processes*, 80, pp. 887–897.

[25] Schmidt, B. and Wang, L., 2018. Predictive maintenance of machine tool linear axes: A case from manufacturing industry. *Procedia manufacturing*, *17*, pp.118-125.

[26] Schröer, C., Kruse, F. and Gómez, J.M., 2021. A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, pp.526-534.

[27] Shin, J., Jun, H. & Kim, J. (2018). Dynamic control of intelligent parking guidance using neural network predictive control. Computers & Industrial Engineering. 120. pp. 15-30.

[28] Soesatijono, S. and Darsin, M. (2021) 'Literature Studies on Maintenance Management', *JEMMME (Journal of Energy, Mechanical, Material, and Manufacturing Engineering)*, 6(1), pp. 67–74.

[29] Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia Medica. pp. 12-18.

[30] Taşcı, B., Omar, A. and Ayvaz, S., 2023. Remaining useful lifetime prediction for predictive maintenance in manufacturing. *Computers & Industrial Engineering*, *184*, p.109566.

[31] Wen, Y., Fashiar Rahman, M., Xu, H. & Tseng, T. (2022). Recent advances and trends of predictive maintenance from data-driven machine prognostics perspective. Measurement. 187. p. 110276.