Application of Machine Learning Models for Predictive Analytics on AI4I 2020

Hard Rakesh Shah
School of Computer Science & Engineering
Sacred Heart University
Fairfield, CT
shahh7@mail.sacredheart.edu

Brian Haley
School of Computer Science & Engineering
Sacred Heart University
Fairfield, CT
haleyb@mail.sacredheart.edu

Eman Abdelfattah
School of Computer Science & Engineering
Sacred Heart University
Fairfield, CT
abdelfattahe@sacredheart.edu

***Abstract*— One of the critical aspects in modern industrial operations is predictive maintenance. It involves leveraging advanced technologies such as machine learning to anticipate equipment failures before they occur. The research not only contributes to the broader understanding of machine learning applications in industrial contexts but also addresses the imperative need for predictive maintenance in optimizing operational efficiency and minimizing downtime and which machine learning model serves best over the dataset selected, incorporating how it can be made more accurate for better performance. The results of the experiments demonstrate that Random Forest and Support Vector Machines achieved comparable, high performance in case of High-quality and Medium-quality machines where the F_1 score reached 85.71% and 90.32% respectively.**

I. Introduction

The integration of advanced technologies has become paramount in optimizing manufacturing processes in industry. In this study, we assess the effectiveness of three machine learning algorithms on the AI4I 2020 dataset, comprising machines of varying quality levels [1]. By focusing on accuracy, precision, F_1 score, recall, and AUC-ROC metrics, we aim to evaluate the predictive maintenance capabilities of these algorithms.

In this study, we delve into the application of three prominent machine learning algorithms—Random Forest (RF), Logistic Regression (LR), and Support Vector Machines (SVM)—on the AI4I2020 dataset. This dataset serves as a valuable repository, encompassing machines of low, medium, and high quality, allowing for a comprehensive evaluation of these algorithms' efficacy in predictive maintenance scenarios.

The cornerstone of our investigation lies in understanding how each machine learning algorithm contributes to the predictive maintenance landscape. Random Forest, known for its ensemble learning capabilities. It combines the predictive power of multiple decision trees that are constructed during training. Each tree is trained on a random subset of the dataset, and the final prediction is determined by aggregating the individual predictions of the constituent trees. The model excels in handling complex relationships within data, providing high accuracy and mitigating overfitting.

Logistic Regression, a classic yet powerful algorithm as it calculates the probability of an instance belonging to a particular class. It is particularly valuable for its interpretability and efficiency.

SVM is a supervised learning algorithm used for either classification or regression tasks.

By rigorously assessing the performance of these models using key performance metrics such as accuracy, precision, F_1 score, recall, AUC-ROC, and confusion matrix our study aims to discern their strengths and limitations across the spectrum of machine quality levels.

The significance of our research extends beyond academic exploration, aiming to bridge the gap between theory and practical implementation in the industrial domain. As we navigate through the intricate landscape of machine learning models, our findings aim to provide actionable insights that can guide the development of robust predictive maintenance strategies. In doing so, we envision contributing to the ongoing evolution of manufacturing practices, fostering resilience, sustainability, and efficiency in the face of dynamic operational challenges.

This paper is organized as follows: section II presents the related work. Section III includes the description of the dataset. Section IV, the results, and analysis are presented. Finally, section V offers not only the conclusion but also the future work.

II. Related Work

In the realm of predictive maintenance, recent studies have embraced innovative machine learning methodologies to effectively address machinery wear and tear. For instance, [2] introduced a groundbreaking predictive maintenance approach, leveraging machine learning on a cutting machine, while [3] demonstrated the practical utility of their

2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) | 979-8-3315-4090-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/UEMCON62879.2024.10754668

979-8-3315-4090-6/24/$31.00 ©2024 IEEE

comprehensive machine learning methodology through a simulated example. [4] underscored the potency of machine learning as a robust solution for analyzing and interpreting collected data, and [5] provided a comprehensive summary of machine learning algorithms employed for predictive maintenance. [6] contributed a detailed analysis of three different motors using predictive maintenance techniques, and [7] conducted a comparative study between time-based and predictive maintenance strategies. [8] demonstrated the practical application of machine learning in predictive maintenance for oil and gas drilling installations, and [9] employed AI algorithms to monitor cutting tools and motors within the context of predictive maintenance. [10] offered insights through a case study, applying AI techniques to predictive maintenance research using a synthetic dataset, while [11] evaluated two methods providing reasoning for complex classifier results on a predictive maintenance dataset. [12] aimed to detect and verify defects across various industries using five machine learning algorithms, and [13] enhanced interpretability in predictive maintenance datasets using the LIME algorithm. [14] demonstrated the efficiency and reliability of machine learning in reducing maintenance costs within industrial settings, and [15] validated the model development of predictive maintenance using artificial neural networks for HVIM, establishing connections among its significant factors. [16] illustrated how the PdMA AI System utilizes electrical data for motor measurement via online and offline testing, and [17] achieved an 80% accuracy in detecting machine failures using the Support Vector Machines model. In addition, [18] provided a comprehensive summary of AI techniques enhancing TMP to address preventive, predictive, and prescriptive maintenance challenges, while [19] proposed a method to display accurate failure prognostics. Finally, [20] proposed a parametric predictive maintenance decision framework that fixes an issue in the traditional framework, and [21] compares different machine learning methods using performance metrics such as precision and recall. These collective studies contribute to the dynamic evolution of predictive maintenance methodologies, showcasing diverse applications and innovative approaches within the field.

Furthermore, machine learning approaches demonstrated high performance in variety of different applications such energy management for electric vehicles, sustainable energy, energy consumption and production, and renewable energy generation [22, 23, 24, 25, 26].

III. Dataset Description

The dataset utilized for this research is the "AI4I 2020" dataset, comprising information related to machine performance in an industrial setting [1]. This dataset has 10,000 instances and fourteen features such as air temperature, process temperature, and machine failure status. A list of all features along with their brief description is presented in Table 1.

IV. Experimental Results and Analysis

To comprehensively analyze machine performance, this study is conducted in two phases. Phase 1 involves training and testing Random Forest, Logistic Regression, and Support Vector Machines models using 80:20 split.

On the other hand, in phase 2 the dataset was divided into three distinct subsets based on the quality levels of the machines. These subsets were categorized as High, Medium, and Low-Quality machines. The categorization was achieved

Table 1. Features along with their description

Feature	**Description of the feature**
Unique ID	Unique identifier ranging from 1 - 10000
Product ID	Consists of a letter L for low (50%), M for medium (30%), or H for high (20%)
Type	Determines whether the Product ID has a letter L, M, or H
Air temperature [Kelvin]	Temperature is normalized to a standard deviation of 2 Kelvin around 300 Kelvin
Process temperature [Kelvin]	Normalized to a standard deviation of 1 Kelvin, added to Air temperature plus 10 Kelvin
Rotational speed [revolutions per minute]	Calculated from a power of 2860 W (Watts), overlaid with a normally distributed noise
Torque [Nanometers]	Normally distributed around 40 nanometers (Nm) with f = 10 Nm. Negative values are not allowed
Tool wear [minutes]	The quality variants H/M/L add 5/3/2 minutes of tool wear
Machine failure	Indicates whether the machine has failed or not
Tool Wear Failure	The tool will be replaced of fail
Heat Dissipation Failure	Causes a process failure under certain condition
Power Failure	The process fails if the power is below 3500 W or above 9000 W
Overstrain Failure	The process fails due to overstrain under certain condition
Random Failures	A chance of 0.1% for each process to fail

through the evaluation of product ID and specific quality indicators present in the dataset. The dataset was preprocessed to handle missing values and ensure uniformity in data representation. Subsequently, it was divided as in phase 1 into 80%:20% for training and for testing sets respectively. During phase 2, the experiment was conducted for all three quality subsets using RF and SVM models as they demonstrated good performance during phase 1.

The performance of the machine learning models to predict machine failures during the two phases was assessed using a comprehensive set of six performance metrics. The following six-performance metrics were computed to gauge the effectiveness of each model in predicting machine failures.

- **Accuracy:** It is calculated using the formula:
  Accuracy = (TP + TN) / (TP + TN + FP + FN)
  Where TP are True Positives, TN are True Negatives, FP are False Positives, and FN are False Negatives.
- **Precision:** The ratio of correctly predicted positive observations to the total predicted positives. The following formula is used to calculate it:
  Precision = TP / (TP + FP)
- **Recall:** The below formula is used for its calculation:
  Recall = TP / (TP + FN)
- **F_1 score:** It is calculated using the formula:
  F_1 score = 2*P*R / (P + R)
  Where P is the precision and R is the recall. F_1 score ranges from 0 to 1, where higher values indicate a better balance between both precision and recall.
- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** Its values range from 0 to 1. One is the highest and best value which means that the model has strong discriminative ability. The curve plots the true positive rate (TPR) against the false positive rate (FPR). The formula for TPR is:
  TPR = TP / (TP + FN)
  and the formula for FPR is:
  FPR = FP / (FP + TN)
- **Confusion Matrix:** It provides a detailed summary of a model's performance by breaking down its predictions into four components: true positives, true negatives, false positives, and false negatives. These components are defined as follows:

Fig. 1 and Fig. 2 provide valuable insights into the Random Forest model's performance for the entire dataset as only two instances out of 2,000 instances are misclassified as negatives during testing. Fig. 3 and Fig. 4 present the performance of Support Vector Machine model where sixty one out of 2,000 instances are misclassified as negatives during testing. Fig. 5 presents the summary of the performance of Logistic Regression model. These five figures demonstrate that Random Forest model during phase 1 outperforms Support Vector Machine and Logistic Regression when it is trained and tested for the entire dataset.

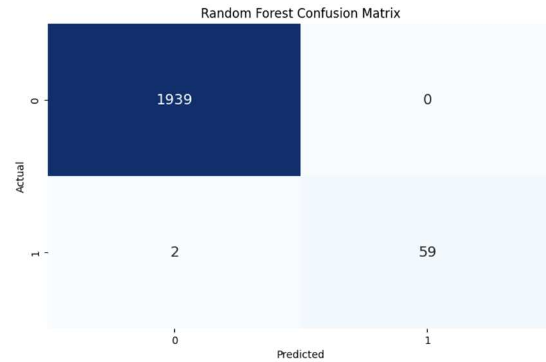

Fig 1. Confusion Matrix for Random Forest

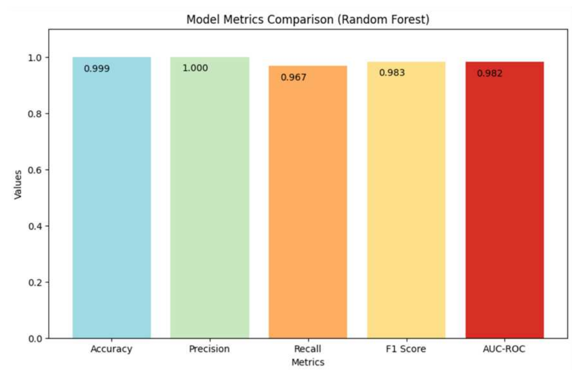

Fig 2. Performance Metrics for Random Forest.

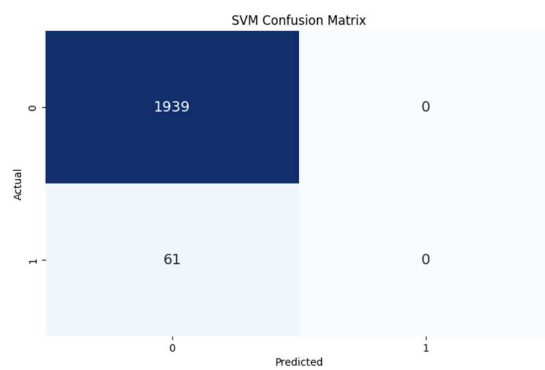

Fig 3. Confusion Matrix for Support Vector Machine.

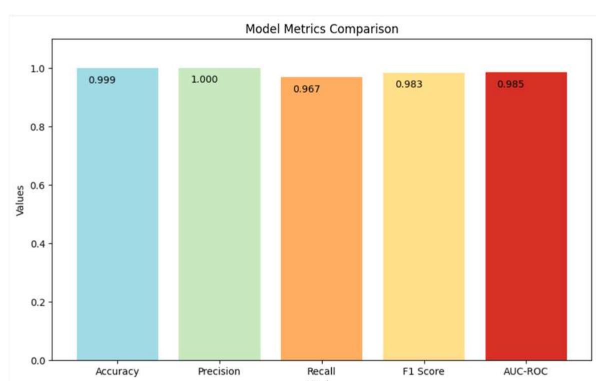

Fig 4. Performance Metrics for Support Vector Machine.

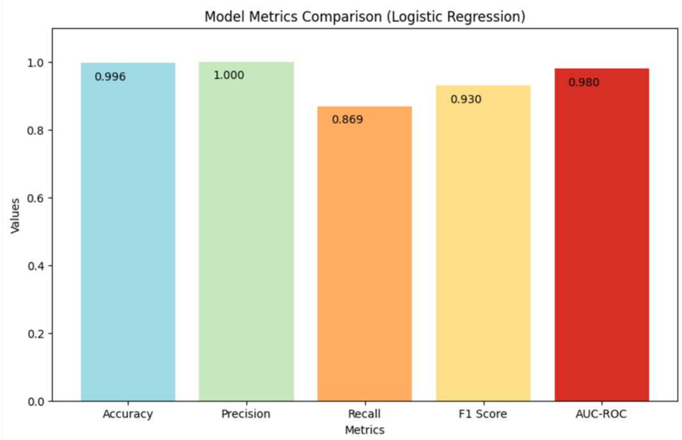

Fig 5. Performance Metrics for Logistic Regression.

As shown in Fig. 6, for High-Quality machines, Random Forest model achieved an impressive accuracy of 99.50%, indicating its high overall predictive capability. Precision of 100.00% implies that the model correctly identified all machine failures without any false positives. A recall of 75.00% suggests that the model successfully captured three-fourths of the actual machine failures. The F_1 score reached 85.71%. AUC-ROC score of 100.00% underscores the model's robust discrimination ability between positive and negative instances.

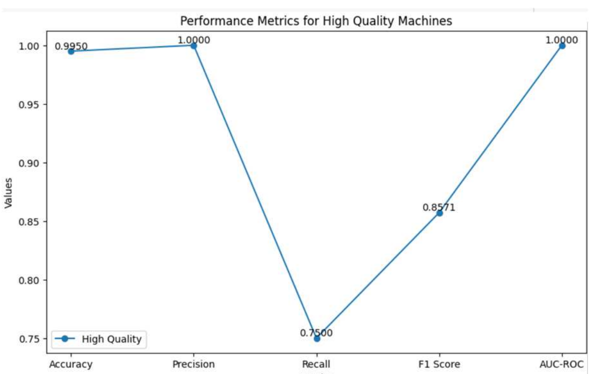

Fig 6. Performance Metrics for Random Forest of High-Quality machines.

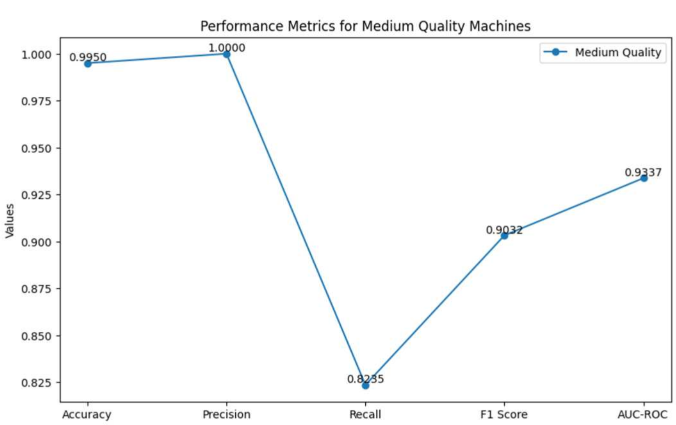

Fig 7. Performance Metrics for Random Forest of Medium-Quality machines.

The Random Forest model demonstrated consistent high performance across all metrics for Medium-Quality machines as demonstrated in Fig. 7. Accuracy remained at 99.50%, emphasizing the model's reliability in correctly classifying instances. Precision of 100.00% signifies the absence of false positives. With a Recall of 82.35%, the model effectively identified a significant portion of machine failures in this category. F_1 score reached 90.32%. AUC-ROC score of 93.37% suggests robust discrimination ability, though slightly lower than High-Quality machines.

As shown in Fig. 8, Random Forest model is performing ideal which seems to need further investigation and may be need of data preprocessing in predicting failures for Low-Quality machines.

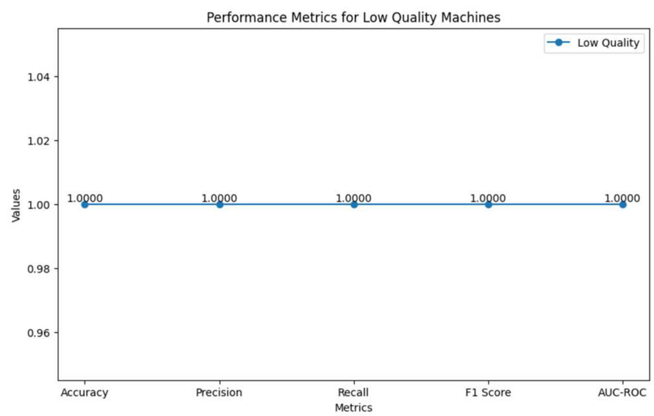

Fig 8. Performance Metrics for Random Forest of Low-Quality machines.

During phase 2, we applied also SVM model to predict machine failures across three distinct qualities of machines: High-Quality, Medium-Quality, and Low-Quality.

The SVM model achieved an impressive accuracy of 99.50%, indicating its high overall predictive capability as presented in Fig. 9. Precision of 100.00% implies that the model correctly identified all machine failures without any false positives. A recall of 75.00% suggests that the model captured three-fourths of the actual machine failures. The F_1 score reached 85.71. AUC-ROC score of 83.40% underscores the model's robust discrimination ability between positive and negative instances.

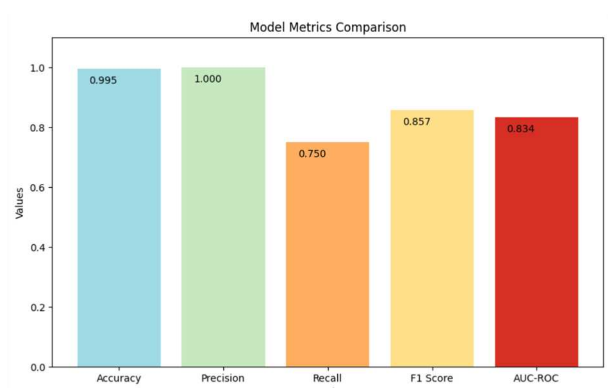

Fig 9. Performance Metrics for SVM of High-Quality machines.

The SVM model demonstrated consistent high performance across all metrics for Medium-Quality machines as shown in Fig. 10. Accuracy remained at 99.50%, emphasizing the model's reliability in correctly classifying instances. Precision of 100.00% signifies the absence of false positives. With a recall of 82.40%, the model identified a significant portion of machine failures in this category. F_1 score reached 90.30%. AUC-ROC score of 92.8% suggests robust discrimination ability as it is higher than the score achieved in case of High-Quality machines for this model.

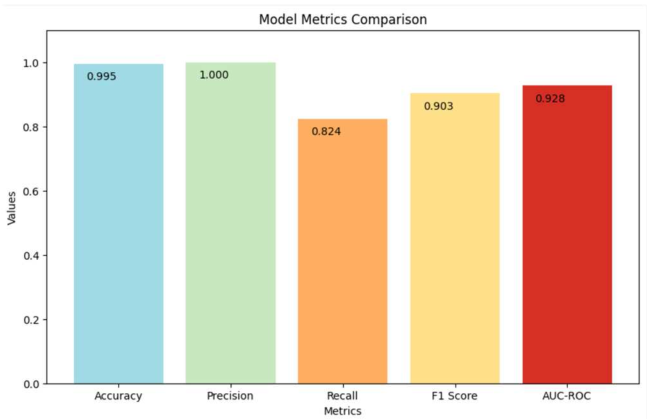

Fig 10. Performance Metrics for SVM of Medium-Quality machines.

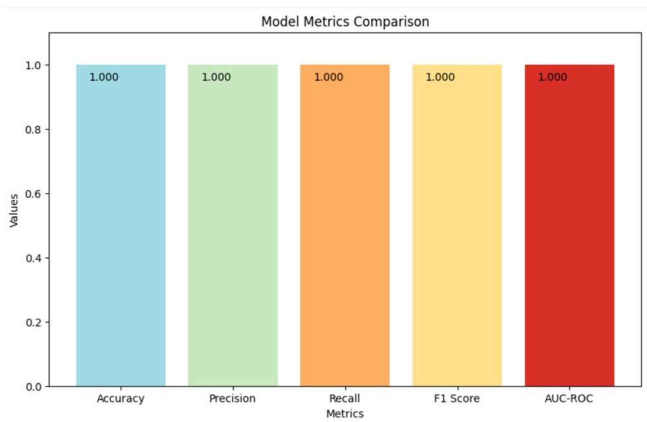

Fig 11. Performance Metrics for SVM of Low-Quality machine.

As demonstrated in Fig. 11, SVM model is performing ideal for Low-Quality machines same as reported in case of RF model.

V. CONCLUSION AND FUTURE WORK

In this study, we explored the effectiveness of machine learning models, specifically RF, SVM, and LR in predicting machine failures using the AI4I 2020 dataset. Random Forest outperforms the other models.

Furthermore, we assess the performance of RF and SVM across three different qualities of machines: High-Quality, Medium-Quality, and Low-Quality. Our results reveal that both RF and SVM have comparable performance. They exhibit remarkable predictive capabilities across all quality categories. They are showcasing consistent excellence in terms of all performance metrics used in the experiment. The models demonstrated their prowess in discerning patterns and predicting machine failures accurately, making them robust choices for predictive maintenance tasks.

In Low-quality machines, where every metric reached a value of 1.0, caution must be exercised in interpreting these results. This scenario may suggest potential issues with the dataset. As we can confirm via all the models giving 100 percent results for every performance metrics in only Low-Quality machines, this derives that Low-Quality machines have a potential to be investigated. For future work, further investigation and preprocessing steps are recommended to obtain a more accurate representation of the model's capabilities.

In conclusion, RF and SVM demonstrate good performance when it comes to predictive maintenance for High-quality and Medium-quality machines.

VI. REFERENCES

[1] UCI Irvine Machine Learning Repository, “AI4I 2020 Predictive Maintenance Dataset,”

https://archive.ics.uci.edu/dataset/601/ai4i+2020+predictive+maintenance+dataset

[2] Paolanti, Marina, Luca Romeo, Andrea Felicetti, Adriano Mancini, Emanuele Frontoni, and Jelena Loncarski. "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), pp. 1-6, 2018.

[3] Susto, Gian Antonio, Andrea Schirru, Simone Pampuri, Seán McLoone, and Alessandro Beghi. "Machine learning for predictive maintenance: A multiple classifier approach," *IEEE transactions on industrial informatics*, vol. 11, no. 3, pp. 812-820, 2014.

[4] Ran, Y., X. Zhou, P. Lin, Y. Wen, and R. Deng. "A survey of predictive maintenance: Systems, purposes and approaches,” arXiv 2019, *arXiv preprint arXiv:1912.07383*, 2019.

[5] Zhang, Weiting, Dong Yang, and Hongchao Wang. "Data-driven methods for predictive maintenance of industrial equipment: A survey," *IEEE systems journal*, vol. 13, no. 3, pp. 2213-2227, 2019.

[6] Lu, Bin, David B. Durocher, and Peter Stemper. "Predictive maintenance techniques," *IEEE Industry Applications Magazine*, vol. 15, no. 6, pp. 52-60, 2009.

[7] Hashemian, Hashem M. "State-of-the-art predictive maintenance techniques." *IEEE Transactions on Instrumentation and measurement*, vol. 60, no. 1, pp. 226-236, 2010.

[8] Cline, Brad, Radu Stefan Niculescu, Duane Huffman, and Bob Deckel. "Predictive maintenance applications for machine learning," In *2017 annual reliability and maintainability symposium* (RAMS), pp. 1-7, 2017.

[9] Lee, Wo Jae, Haiyue Wu, Huitaek Yun, Hanjun Kim, Martin BG Jun, and John W. Sutherland. "Predictive maintenance of machine tool systems using artificial intelligence techniques applied to machine condition data," *Procedia Cirp 80*, pp. 506-511, 2019.

[10] Kamel, H. "Artificial intelligence for predictive maintenance," *In Journal of Physics: Conference Series*, vol. 2299, no. 1, pp. 012001, IOP Publishing, 2022.

[11] Matzka, Stephan. "Explainable artificial intelligence for predictive maintenance applications," *In 2020 third international conference on artificial intelligence for industries* (AI4I), pp. 69-74, 2020.

[12] Sharma, Neelam, Tejasv Sidana, Saransh Singhal, and Sachin Jindal. "Predictive maintenance: Comparative study of machine learning algorithms for fault diagnosis," Available at SSRN 4143868, 2022.

[13] Torcianti, Andrea, and Stephan Matzka. "Explainable artificial intelligence for predictive maintenance applications using a local surrogate model," *In 2021 4th International Conference on Artificial Intelligence for Industries* (AI4I), pp. 86-88, 2021.

[14] Satwaliya, Deepanshu Singh, H. Pal Thethi, Anubhuti Dhyani, G. Ravi Kiran, Mustafa Al-Taee, and Malik Bader Alazzam. "Predictive Maintenance using Machine Learning: A Case Study in Manufacturing Management," *In 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering* (ICACITE), pp. 872-876, 2023.
[15] Zainol, Ahmad Masry Bin, and Nurul Rawaida Ain Bt Burhani. "Failure Prediction for High Voltage Induction Motor using Artificial Neural Network (ANN)," *In 2023 International Conference on Cyber Management and Engineering* (CyMaEn), pp. 33-38, 2023.
[16] Bundasak, Supaporn, and Pawin Wittayasirikul. "Predictive maintenance using AI for Motor health prediction system," *In 2022 International Electrical Engineering Congress* (IEECON), pp. 1-4, 2022.
[17] Assagaf, Idrus, Agus Sukandi, Abdul Azis Abdillah, Samsul Arifin, and Jonri Lomi Ga. "Machine Predictive Maintenance by Using Support Vector Machines," *Recent in Engineering Science and Technology,* vol. 1, no. 01 pp. 31-35, 2023.
[18] Mattioli, Juliette, Paolo Perico, and Pierre-Olivier Robic. "Improve total production maintenance with artificial intelligence," *In 2020 Third International Conference on Artificial Intelligence for Industries* (AI4I), pp. 56-59, 2020.
[19] Chen, Chuang, Cunsong Wang, Ningyun Lu, Bin Jiang, and Yin Xing. "A data-driven predictive maintenance strategy based on accurate failure prognostics," Eksploatacja i Niezawodność, vol. 23, no. 2, pp. 387-394, 2021.
[20] Huynh, Khac Tuan, Antoine Grall, and Christophe Bérenguer. "A parametric predictive maintenance decision-making framework considering improved system health prognosis precision." *IEEE Transactions on Reliability,* vol. 68, no. 1, pp. 375-396, 2018.
[21] Akanni, Feranmi. "Application of Machine Learning Methods on Predictive Maintenance," Master's thesis, 2019.
[22] Jui, Julakha Jahan, Mohd Ashraf Ahmad, MM Imran Molla, and Muhammad Ikram Mohd Rashid, "Optimal Energy Management Strategies for Hybrid Electric Vehicles: A Recent Survey of Machine Learning Approaches," *Journal of Engineering Research*, 2024.
[23] Yao, Zhenpeng, Yanwei Lum, Andrew Johnston, Luis Martin Mejia-Mendoza, Xin Zhou, Yonggang Wen, Alán Aspuru-Guzik, Edward H. Sargent, and Zhi Wei She, "Machine learning for a sustainable energy future," *Nature Reviews Materials* 8, no. 3, pp. 202-215, 2023.
[24] Mhlanga, David, "Artificial intelligence and machine learning for energy consumption and production in emerging markets: a review," *Energies*, vol.16, no. 2, pp. 745, 2023.
[25] Benti, Natei Ermias, Mesfin Diro Chaka, and Addisu Gezahegn Semie, "Forecasting renewable energy generation with machine learning and deep learning: Current advances and future prospects," *Sustainability*, vol. 15, no. 9 pp. 7087, 2023.
[26] Kapp, Sean, Jun-Ki Choi, and Taehoon Hong, "Predicting industrial building energy consumption with statistical and machine-learning models informed by physical system parameters," *Renewable and Sustainable Energy Reviews* vol.172, pp. 113045, 2023.