National College of Ireland

Predictive Maintenance In Industrial Sector Using Machine Learning

MSc Research Project
Data Analytics

Ashutosh Alone
Student ID: X22228381

School of Computing
National College of Ireland

Supervisor: Dr. Ahmed Makki

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Ashutosh Alone
Student ID:	X22228381
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Dr. Ahmed Makki
Submission Due Date:	12/08/2024
Project Title:	Predictive Maintenance In Industrial Sector Using Machine Learning
Word Count:	8293
Page Count:	20

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:	Ashutosh Alone
Date:	12th August 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

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

Predictive Maintenance In Industrail Sector Using Machine Learning

Ashutosh Alone
X22228381

Abstract

In industrial sector there is a need to have a reliable predictive maintenance system as it can help them to reduce downtime, unexpected machine failures can cause significant financial and safety risks. Traditional predictive maintenance methods are not effective enough to manage the increasing complexity and size of the data. Therefore, in this research the use of unsupervised machine learning algorithms is explored. The unsupervised algorithms used are Isolation Forest, One-Class Support Vector Machine and Local Outlier Factor. These models are compared against supervised algorithms like K-NN and Random Forest. The results showed that supervised learning algorithms performed better than unsupervised learning algorithms with perfect accuracy and precision. This high accuracy of K-NN and Random Forest is further justified by performing cross validation on them. On the other hand, the best performing unsupervised algorithm which is Isolation Forest showed high recall but due to low precision it leads to generating false positives. The overall findings of this research show that unsupervised algorithms have potential for anomaly detection in predictive maintenance, but they are currently less effective than supervised learning algorithms.

Keywords: Predictive Maintenance, **Unsupervised Learning**, **Machine Learning**, **Anomaly Detection**, **Isolation Forest**, **Supervised Learning**, **Cross-Validation**.

1 Introduction

1.1 Background

People's day-to-day life has been getting more busy and consuming with each passing day due to the modern lifestyle and fast-growing world. But with this ever-growing world, there has been a continuous revolution in the industrial sector as well. There have been significant technological advancements van Dinter et al. (2022) occurring in the manufacturing industry too. To supply the ever-growing demand of the market, the industrial sector has to work nonstop to meet its needs. Thus, there can't be any unexpected machine failure occurrence in the industrial sector. If such unexpected machine failure were to happen, it would lead to a variety of risks which may include harm to human health, injuries, and an increase in the maintenance cost of the machinery. Also, such failures will result in increased downtime of the machinery, directly affecting the production of the industry, which will again result in further financial loss.

1.2 Motivation

To avoid such issues, predictive maintenance (PdM) represents a good solution as it aims to provide a strategy to forecast machine failures before they occur. This leads to reduced downtime from unplanned maintenance and reduces various expenses associated with it. Throughout the progression of Industry 4.0, there has been a significant transition from traditional predictive maintenance to more reliable approaches using machine learning (ML) and artificial intelligence (AI). Initially, simple statistical models were used for prediction, but with the increasing amount of sensor data and technological progress in the computing sector, it has become necessary to use machine learning techniques to handle big data Lee et al. (2015). These techniques can manage large amounts of data generated by sensors, uncovering patterns that are difficult to detect manually with high accuracy.

1.3 Research Question

This research investigates the various applications of unsupervised machine learning algorithms in predictive maintenance to further increase the accuracy of detecting potential equipment failures. The primary research question is: *Can unsupervised learning algorithms be used for early anomaly detection in equipment for predictive maintenance?* To address this question, the unsupervised machine learning algorithms selected for this research are One-Class Support Vector Machine (SVM), Isolation Forest, and Local Outlier Factor (LOF). The research objectives to achieve this project are as follows:

- Study the field of predictive maintenance thoroughly and focus on the application of unsupervised machine learning algorithms for predictive maintenance.
- Obtain industrial data which has sensor data for equipment failures so that it can be used in predictive maintenance.
- Design frameworks for both supervised and unsupervised machine learning algorithms for anomaly detection.
- Train the models on the industry data.
- Evaluate the effectiveness of all the models in terms of prediction accuracy and efficiency.
- Compare the results of unsupervised machine learning models with supervised machine learning models.

This research aims to use unsupervised learning models for PdM on real-time data; however, instead of using continuous real-time data from sensors, the data used has already been collected and stored in a dataset. Therefore, instead of using real-time monitoring of sensors, this research uses already collected data from them. The adaptability of these models needs to be explored further on different types of machinery in the industry.

1.4 Report Outline

The following sections of this report contain various sections and subsections with thorough discussions. After the introduction, the next section is related work, which discusses various previous research conducted in the field of predictive maintenance using ML. Following that, the third section explains the research methodology to implement this project. The next section is design specification and implementation, where the requirements for this project and how the design is implemented are discussed. It is followed by the evaluation of all the machine learning models, where the results related to these models are shown, and a critical analysis of them is conducted. After the evaluation, the next section is the conclusion and future work, where the findings of this research are summarized with proper answers to the research question.

.

2 Related Work

In the industrial sector, predictive maintenance is one of the most important strategies to reduce the downtime of machinery in various industries, thereby achieving minimal expenses related to sudden maintenance. A variety of studies have been conducted in the field of predictive maintenance across various infrastructures. This section consists of key studies done in this field. This literature review focuses on discussing various existing studies, the methods used for predictive maintenance, their respective achievements, and a few limitations.

The field of predictive maintenance has been studied and developed over time. Dalzochio et al. (2020) have provided an extensive study in the field of predictive maintenance in Industry 4.0. In their paper, they discussed the current status and challenges in predictive maintenance in Industry 4.0. The method they used is a systematic literature review over 38 papers to discuss challenges in predictive maintenance like data integration, handling large data, and ensuring real-time processing for PdM. Their study talks about integrating machine learning models with big data analytics to improve failure prediction accuracy.

As the field of predictive maintenance is ever-growing, different researchers have proposed various methods to address the issue of improving the accuracy of machine failure prediction. The use of predictive maintenance in sustainable manufacturing is shown by Abidi et al. (2022). In their work, the authors introduced an advanced PdM planning model for sustainable manufacturing. The authors performed preprocessing, normalization, and selected optimal features from the data and trained the SVM-based Recurrent Neural Network (RNN) model for prediction. For further improvement, the model was optimized by the Jaya algorithm and Sea Lion Optimization (SLnO). The result of the prediction using SLnO was more advanced than normal ML or Neural Network (NN) models. The RMSE of J-SLnO was 95.3% more advanced than NN. The limitation of this research was that the model building is too complex and not efficient as it requires more computational power.

Predictive maintenance in HVAC (heating, ventilation, and air conditioning) systems in buildings is implemented by Bouabdallaoui et al. (2021). In their paper, the authors developed a smart framework using ML and Internet of Things (IoT) devices to predict the maintenance needs for HVAC systems. The approach they used for prediction involved autoencoders and Long Short Term Memory (LSTM) NN to create the model for

prediction. The shortcoming of this framework was that it failed to predict a few failures and also generated some false positives, showcasing that the model was not accurate. This paper aims to achieve greater accuracy using unsupervised learning algorithms, which Bouabdallaoui et al. (2021) failed to achieve.

This literature review further discusses various research papers by dividing them based on the methods used for predictive maintenance.

2.1 Supervised ML Algorithms

Satwaliya et al. (2023) provided a case study in the manufacturing industry by designing various ML algorithms like Random Forest (RF), Gradient Boosting (GB), and Deep Learning (DL). The authors trained these models on the data and achieved significant results in prediction accuracy. The accuracy of RF was 96%, GB had an accuracy of 95%, and DL had an accuracy of 94%. The average rate of precision, recall, and F-1 Score were all above 85%.

Supervised learning algorithms like Support Vector Machine and logistic regression were used in a study by Gohel et al. (2020). In their paper, they developed a predictive maintenance framework for nuclear infrastructure where accurately predicting machine failure is important. In their research, it was found that the Support Vector Machine outperformed logistic regression in terms of prediction accuracy. The main limitation of this research was that they did not have proper nuclear plant real-time data as they were still collecting data for the analysis.

Another research was done in the field of semiconductor manufacturing for predictive maintenance by Susto et al. (2015), where the authors implemented a multiple classifier approach. In this approach, multiple classification modules are employed with different prediction horizons to improve prediction accuracy. The machine learning models used for this approach are Support Vector Machine and K-Nearest Neighbour (KNN). By implementing this approach, the authors improved the maintenance schedule.

Cakir et al. (2021) discussed an Industrial Internet of Things (IIoT) based conditional monitoring system for predictive maintenance. Their study highly focuses on combining ML and IIoT for improving predictive maintenance to make it more efficient. In their paper, the authors performed classification using popular algorithms like Support Vector Machine, Random Forest, KNN, Decision Trees, and Linear Discriminant Analysis. Among all the applied models, KNN showed the best performance, followed by SVM, Decision Trees, and Random Forest, while Linear Discriminant Analysis performed the worst.

Apart from using machine learning approaches, some research papers show the use of neural network architectures for improved predictive maintenance strategies. A research paper by Theissler et al. (2021) discussed various use cases and challenges in the automotive industry through a machine learning perspective. They further discussed various methods like anomaly detection, deep learning, neural networks, decision trees, and SVM, which are used in predictive maintenance. This paper discusses the need to establish some benchmarks to compare different models effectively.

Another way of predicting unexpected machinery failure detection is by calculating the remaining useful life (RUL) of the equipment. The RUL approach is used by Alfaro-Nango et al. (2022) in an aircraft fleet for predictive maintenance. The dataset used in their research is the N-CMAPSS dataset, on which they applied Principal Component Analysis and monotonicity methods. They then designed a Convolutional Neural Network

(CNN) architecture for prediction with 20 epochs. They achieved a mean RMSE of 10.91, which shows a satisfactory performance of the CNN model. The shortcoming of this research is that no other neural network architecture was designed to compare the model performance for better understanding.

Souza et al. (2021) implemented deep learning strategies for the diagnosis and classification of failures in industrial rotary machines. Their approach focused on the use of predictive maintenance with a convolutional neural network (PdM-CNN). The data used for their research was collected from a single vibration sensor installed on the bearing of the motor drive. They used two failure classification methods for the prediction: one method used all the failure classes as a single class, while in the other method, all individual classes were combined into the same-base failure class. The PdM-CNN model was then trained on two different publicly available datasets. The accuracy of the PdM-CNN model turned out to be 99.58% and 97.25% on the MaFaulDa and CWRU datasets, respectively. Their research shows the effectiveness of CNN in improving the accuracy of fault detection.

2.2 Unsupervised Machine Learning Algorithms

Unsupervised machine learning algorithms are specifically designed for anomaly detection, making them a great choice for unexpected failure prediction in machines. This method focuses on identifying hidden patterns in the data to predict potential equipment failures. The use of unsupervised machine learning models in the manufacturing industry is discussed by Kolokas et al. (2020). In their paper, they proposed the use of an isolation forest classifier as the outlier classification detection model due to its computing efficiency and ability to handle large volumes of data. The evaluation of the model was done using metrics like recall, precision, F-1 score, and accuracy, as these metrics are generally used for model evaluation. They did not select the local outlier factor as the fault detection classifier as it performed poorly in their primary trials. However, in this research paper, the local outlier factor is used to remove the limitations in the research proposed by Kolokas et al. (2020).

After the isolation forest, clustering algorithms are widely used as an unsupervised learning algorithm for predicting unexpected machine failures before they happen. Various clustering algorithms were used by Amruthnath and Gupta (2018) as unsupervised machine learning models for predictive maintenance. In their paper, the dataset chosen contained vibrational sensor data collected from the exhaust fan. Principal Component Analysis (PCA) was then performed for dimensionality reduction. T2 analysis was then conducted for early fault detection. Various clustering algorithms were used, such as C-means clustering, K-means clustering, and hierarchical clustering. For model-based clustering, the Gaussian Mixture Model (GMM) was used as it can be designed well using Gaussian distribution. In their research, they pointed out that PCA T2 statistics are effective for early fault detection. The PCA T2 statistics performed better than the Gaussian Mixture Model, further demonstrating the accuracy of the PCA T2 statistics. The advantage of this method is that when deployed in an unfamiliar system with little to no past knowledge of the domain, it can still identify faults compared to clustering algorithms.

Apart from using unsupervised machine learning algorithms, some researchers have used new approaches for failure prediction in predictive maintenance. The lack of a detailed anomaly detection algorithm is showcased by Carrasco et al. (2021). They pointed

out that normal time series classification does not apply to failure or anomaly detection in the field of predictive maintenance. To address this issue, they created a new framework in which the concept of preceding window ROC is used. They proposed a framework for a temporal unsupervised anomaly detection algorithm, in which the first step is to transform given instances into intervals using time series stamps. They then aggregated the function for early detection of the failures. Finally, the evaluation was done based on preceding window ROC, plotting a graph of sensitivity vs (1-Specificity). The time window for prediction outcome was divided into three intervals: 12 hours, 16 hours, and 20-hour windows. This framework was then used in a case study for further analysis and comparison of the results between different modules. The dataset used for this case study was ArcelorMittal sensor data. The framework performed well on the data, showing significant improvement in the accuracy of fault detection. The main limitation of their approach is that the model did not perform as well as other state-of-the-art research papers and their results. In this research paper, it is aimed that the performance of unsupervised learning algorithms should be as good as supervised learning algorithms.

In conclusion, all the studies collectively highlight the importance of anomaly detection in predictive maintenance in various sectors like manufacturing, the automobile industry, and the energy sector. Predictive maintenance can reduce the downtime of the industry and help reduce the financial burden. Various supervised learning algorithms like KNN, Support Vector Machine, and Random Forest showed high accuracy in detecting machinery failures. Approaches consisting of different neural network architectures like CNN and RNN can also be used for failure prediction as they have practical applications in the industrial sector. Some researchers also showcased the use of various unsupervised learning algorithms like clustering algorithms, Isolation Forest, and Local Outlier Factor for anomaly detection in predictive maintenance. However, the use of unsupervised learning algorithms was limited. Therefore, this paper aims to demonstrate how various unsupervised machine learning algorithms can achieve high accuracy in predicting unexpected machine failures for predictive maintenance.

3 Research Methodology

The research methodology section of this report contains the various systematic processes that were followed for the study of predictive maintenance in the industrial sector using machine learning. The following section contains details of various processes like data collection, data preprocessing, selection of suitable models, training of the models, evaluation of the models, and their various implementation stages. The methodology section is structured by following standard procedures and all the steps are documented properly for a clear understanding of this research.

Figure 1: CRISP-DM Framework

This project follows CRISP-DM framework Schröer et al. (2021) where it consists of various stpes from data collevction to model evaluation.

3.1 Data Collection

Various datasets related to predictive maintenance are available in the market. For this project, the dataset used is *'ai4i2020.csv'*, which is a synthetic dataset containing real data from sensors in the industry. This dataset has various features such as various sensor measurements, *'Product ID'*, *'Type'* and the target variable for prediction, which is *'Machine Failure'*. This dataset is publicly available and taken from a publicly available repository.[1]

3.1.1 Dataset Description

The dataset has around 10,000 rows and 15 feature columns. The dataset has variables containing various sensor readings like *'Air temperature'*, *'Process temperature'*, *'Rotational Speed'*, and *'Torque'*, which consist of the individual sensor readings. Additionally, the target variable for this research paper, *'Machine Failure'*, consists of a record of machine failure. This variable has various types such as tool wear failure (TWF), heat dissipation failure (HDF), power failure (PWF), overstrain failure (OSF), and random failures (RNF). Further, the dataset has a unique identifier named *UID*, which ranges from 1 to 10,000.

3.2 Data Preprocessing

After data collection, the data needs to be processed for further analysis. Preprocessing is an important step in predictive maintenance as it helps maintain the quality of the data and ensures the performance of the model is acceptable.

3.2.1 Data Handling and Cleaning

The *'ai4i2020.csv'* dataset does not contain any missing values, making it an ideal dataset for predictive maintenance as it ensures the integrity of the data. The first step in data cleaning is to remove any unnecessary features. Thus, the *'UDI'* column, which is an identifier, is dropped as it does not contribute to the performance of the predictive model. The next step is to convert any categorical variables into numerical format. The two categorical variables in the dataset named *'Product ID'* and *'Type'* are converted to numerical format using the label encoding method to make them suitable for the machine learning algorithms.

3.2.2 Feature Scaling

Machine learning models are sensitive to the scale of input features, so it is necessary to scale the input variables so that features with a longer range do not dominate the model's behavior, which can lead to biased outcomes. The method used for scaling the features is *'Min-Max Scaling'* and all the feature values were scaled to range between 0 and 1. By performing this normalization step, the features are scaled to a common scale, ensuring that all features contribute equally to the model's performance. This helps improve the performance of the model.

[1]URL: `https://archive.ics.uci.edu/dataset/601/ai4i+2020+predictive+maintenance+dataset`

3.3 Exploratory Data Analysis

A thorough exploratory data analysis is done to understand the relationship between the variables and their relationships within the data. Various visualizations are used to understand the patterns in the data, such as histograms correlation heatmap.

The histogram is used to examine the distribution of variables in the data. The histogram clearly shows the representation of their frequency. The air temperature and the process temperature, which is in kelvin [K], shows the normal distribution of temperature which is around their means of approximately 300 K and 310 K respectively. This distribution shows that the environmental conditions are mostly stable. The rotational speed shows a slightly uneven distribution of RPM values. The values range between 1200 to 1800 RPM which is the typical operating range. Torque value is around 40Nm which shows the consistent load conditions. However, Tool wear has a higher uneven distribution which ranges from 0 to 300 approximately. And the other plots help us to understand the distribution of different classes of failure.

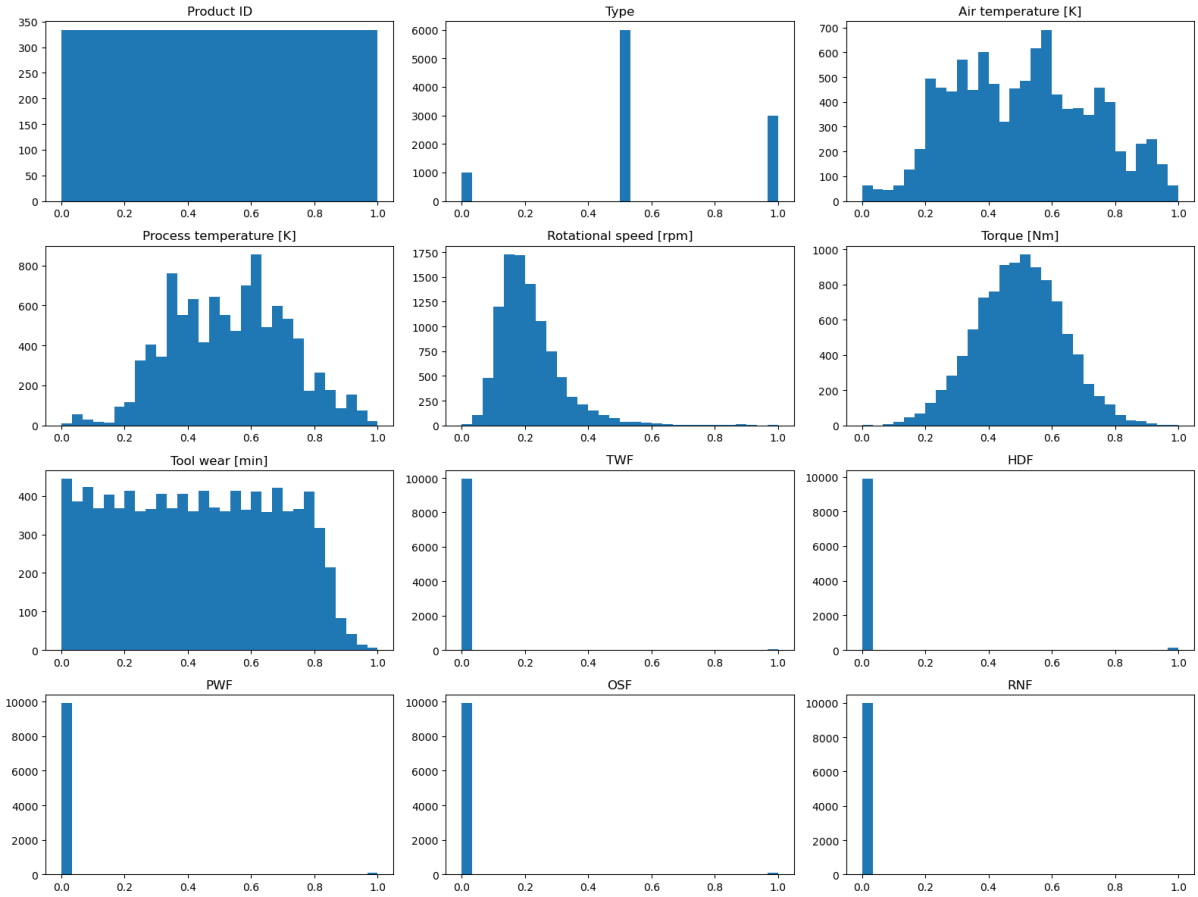

Figure 2: Variable Visualization

The correlation heatmap provides an overview of relationships between different features of the dataset. The colour variation in the map indicates the strength and direction of the relationships of the feature variables with each other. The heatmap values range from -1 to +1. One of the main observations in the above heatmap is that there is a strong positive correlation between air temperature and process temperature. as the air temperature changes the process temperature changes as well. This shows that the process temperature is derived from the air temperature. Rotational speed and torque show

a moderate negative correlation of -0.46 which suggest that higher speed is always associated with lower torque values. Other variables like tool wear show weak relationships and correlation with the other features in the dataset. The other variables have no strong relationships between them, and they contribute individually to machine failures.

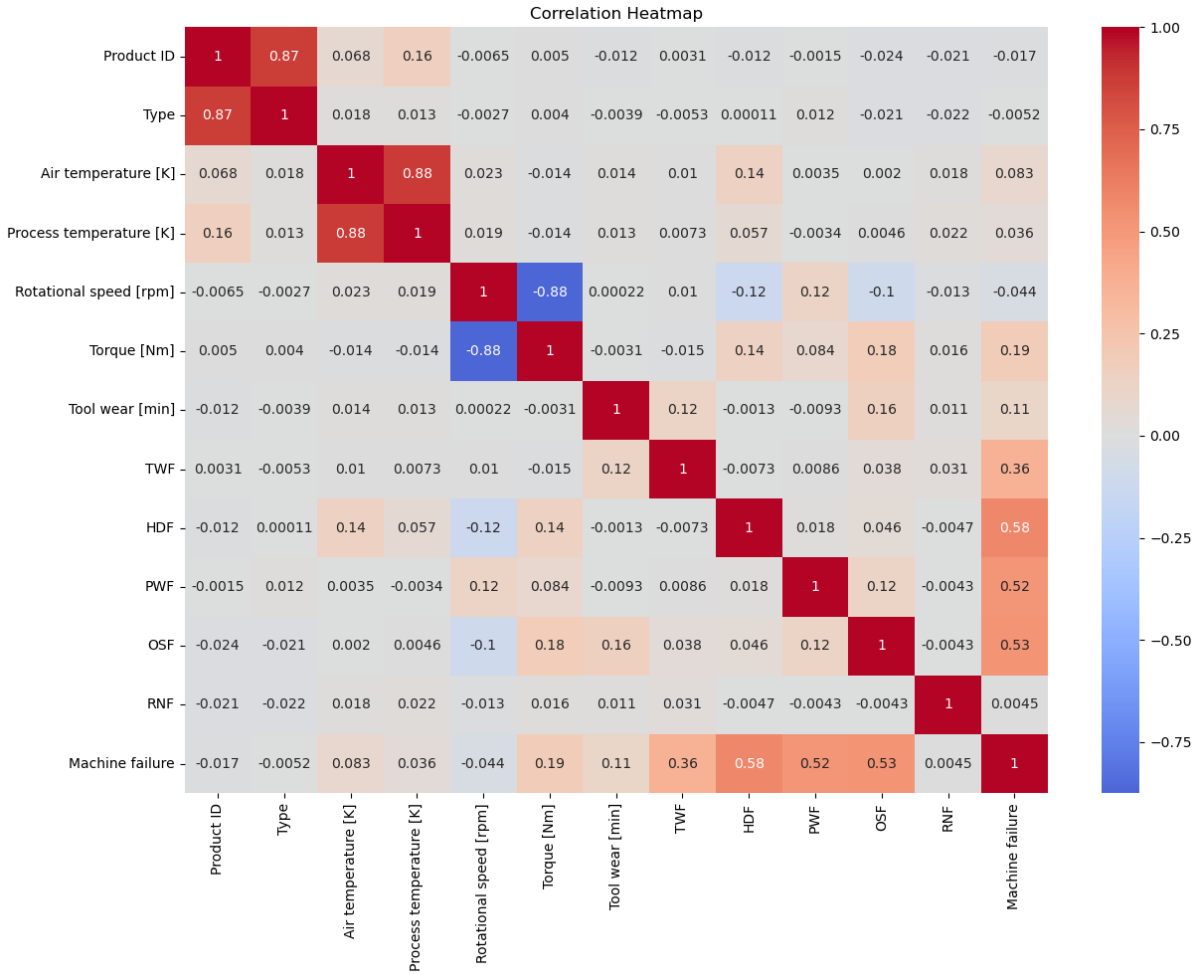

Figure 3: Heatmap

3.4 Data Splitting

After that, the data is split into two parts named training data and testing data. Using the `train_test_split` function, 80% is allocated to training data and 20% is allocated for testing the models. This will help the models to be trained on the majority of the data. The random state is set to 42 to make sure that the data split is reproducible for getting consistent results.

3.5 Model Selection

The most important part of this research is selecting the appropriate machine learning models for predicting failures. If the choice of model is wrong, it can affect the accuracy of the failure prediction. Thorough research has been done to understand which models can be efficient for predictive maintenance. With the help of various previous research

done in the field of predictive maintenance, suitable machine learning models are selected for training and deployment for anomaly detection.

3.5.1 Supervised ML Algorithms

Supervised machine learning algorithms are commonly used in predictive maintenance due to their ability to properly work on labelled data. The dataset used for this project is labelled data which makes supervised learning algorithms one of the ideal choices of machine learning model. Supervised learning algorithms are efficient, highly accurate, and robust for the task of predictive maintenance. The following ML models are used for the training purpose:

K-Nearest Neighbors (K-NN) K-NN is one of the most commonly used supervised learning algorithms in the field of predictive maintenance. K-NN is a simple yet effective machine learning algorithm used for classification tasks. K-NN is straightforward to understand and its interpretability makes it one of the most popular algorithms for classification or regression tasks. As K-NN works well on medium-sized datasets, the dataset used in this project called *'ai4i2020.csv'* lies well within that range. This makes K-NN a suitable algorithm for failure prediction. Numerous previous research studies have used K-NN for its effectiveness in failure prediction. In the study done by Cakir et al. (2021), they used K-NN with other ML models as well, and K-NN performed the best in terms of accuracy prediction. Also, Susto et al. (2015) used K-NN as a classifier in the multiple classifier approach due to its efficiency and accuracy. By thoroughly considering the technical benefits of using K-NN and also looking at previous work done by researchers, K-NN is selected as one of the supervised machine learning models.

Random Forest Random Forest is another supervised learning algorithm which is highly used for the task of failure prediction in predictive maintenance. Random Forest is a robust and effective machine learning model that can be used for high-dimensional data. Random Forest generally achieves high accuracy in prediction due to its ensemble nature. Random Forest aggregates multiple decision trees, which helps reduce the overfitting of the data. Overfitting can reduce the reliability of the model. As Random Forest handles the overfitting of the data by generalizing it, it is a suitable model for tasks like predicting equipment failure in PdM. Many previous research studies have used Random Forest as an ideal choice for failure prediction models in predictive maintenance. Satwaliya et al. (2023) used Random Forest for predictive maintenance in the manufacturing sector and achieved an impressive accuracy of 96%, which was the highest among all the models used. Also, Cakir et al. (2021) used Random Forest as a machine learning model for PdM, which performed well in terms of prediction accuracy. By looking at all the benefits and real-time application uses of the model, Random Forest is selected as another supervised learning model.

3.5.2 Unsupervised ML Algorithms

Unsupervised machine learning algorithms are mainly used for anomaly detection. Anomaly detection is one of the most important aspects in the field of predictive maintenance. Unsupervised machine learning algorithms do not require large-scale data for accurate prediction of faults and anomalies. These factors make unsupervised machine learning

algorithms the main focus of this research. The following unsupervised anomaly detection machine learning models are used in this research:

One-Class Support Vector Machine Standard supervised support vector machine is generally used in tasks like classification and regression for predictive maintenance. However, One-Class Support Vector Machine (SVM) is the unsupervised counterpart of standard SVM. One-Class SVM is specifically designed for anomaly detection tasks Pang et al. (2022), which makes it an ideal choice for this project. As the majority of the dataset represents normal behavior, One-Class SVM is an excellent choice for detecting anomalies. Similar to SVM, One-Class SVM also performs the kernel trick where it transforms the input data to a higher-dimensional space. This makes the algorithm more effective for detecting any anomalies in the dataset. Also, One-Class SVM is very robust for handling outliers in the data. This makes it a robust and efficient machine learning model for anomaly detection. By carefully considering all the merits of One-Class SVM, it is selected as one of the unsupervised machine learning models for anomaly detection.

Isolation Forest An Isolation Forest is an ensemble learning algorithm specifically designed for tasks like anomaly detection. The main idea behind Isolation Forest is that anomalies are few and different, suggesting that anomalies are significantly different from normal instances. Because of this, anomalies are easier to identify than normal points. Due to this, Isolation Forest has become one of the most reliable unsupervised machine learning models for anomaly detection. An Isolation Forest is very efficient for identifying outliers in the data and offers excellent computational efficiency. Because of this, the model becomes easily scalable to any type of dataset. Due to this flexibility, Isolation Forest is an ideal unsupervised ML model. Additionally, in previous research such as the one proposed by Kolokas et al. (2020), Isolation Forest was used as an outlier detection model due to its computational efficiency. By considering all the benefits and usage of this model in previous research, Isolation Forest is selected in this study for anomaly detection.

Local Outlier Factor Local Outlier Factor (LOF) is another excellent unsupervised machine learning algorithm used for anomaly detection by focusing on local outliers Alghushairy et al. (2021). LOF calculates the local density of the data and then compares the local density with the local densities of its neighbors. This is helpful in detecting the degree of anomalies in the data. LOF is excellent for detecting anomalies that may not be detectable by global methods. This approach can especially be used for identifying variations in the industrial sensor data in the dataset used in this project. This will help in detecting early signs of equipment failures if there are any. LOF does not assume any specific distribution of the data, making it flexible and applicable to a variety of data in the industrial field for anomaly detection. LOF has an approach of local density-based analysis which makes it robust to any noise in the data. Also, LOF can handle outliers effectively as it can differentiate between true anomalies in the data and random noise. With all these merits, Local Outlier Factor is considered to be the final unsupervised anomaly detection algorithm for anomaly detection in predictive maintenance.

4 Design Specification and Implementation

4.1 Design Specification

The design specification part of predictive maintenance using machine learning algorithms contains a detailed explanation of the architecture and different techniques used to effectively implement the solution for failure detection in predictive maintenance. This section contains various components and requirements necessary for the successful deployment of this predictive maintenance system.

Table 1: System Specifications

Component	**Specification**
Windows Version	Windows 10
Disk Space	1 TB + 500GB (SSD)
RAM	16GB
Language	Python
Version	3.11.4
IDE	Jupyter Notebook (V. 6.5.4)
Libraries	pandas, sklearn.preprocessing, sklearn.model_selection, sklearn.ensemble, sklearn.svm, sklearn.neighbors, sklearn.metrics, matplotlib.pyplot, numpy, seaborn

4.2 Implementation

Model Implementation:

K-NN: In this research, K-NN is used as a classifier for failure prediction. K-NN works on a principle where it classifies the data points based on the majority class of their nearest neighbor. The K-NN model was initialized in this project by specifying the number of K, which is selected to be 5. Different values of K were tried in this project, but 5 turned out to be the best in terms of performance of the model. A smaller K value can be too sensitive to noisy data, so 5 was selected. After that, the K-NN model is trained on the training data as the K-NN model does not have a separate training phase like other models.

Random Forest: A Random Forest was implemented in this project as an ensemble of decision trees for classification of machine failure. This model works on a principle where it aggregates the predictions of multiple decision trees together. By doing so, the accuracy of the model increases, and it can help reduce overfitting of the data. The number of decision trees fixed for the model is 100 to ensure that the model has enough decision trees to reduce the variance, which is useful for the generalization of the model to new data. Then the Random Forest model is trained on the training set. During the prediction, the Random Forest model combines outputs of the individual decision trees and determines the final prediction for the failures.

Isolation Forest: Isolation Forest is used in this project for anomaly detection using unsupervised learning algorithms. This algorithm works on the principle of recursive partitioning and isolates the data points. This helps in identifying the anomalies quicker than normal points as they have distinctive features. The Isolation Forest model is then initialized with different parameters such as estimators and contamination where we define the expected number of anomalies in the dataset. The contamination was set to 'auto' in order to allow the algorithm to determine the optimal threshold for anomaly detection. Then the model is trained using features of the entire dataset instead of training on split training data. During the training, the Isolation Forest creates multiple trees, and each of them is built by randomly selecting a feature and a split value. This random partitioning identifies the anomalies properly, so the model accuracy increases.

One-Class SVM: One-Class SVM is used for detecting anomalies in the data so that it can be useful for predictive maintenance. In this project, One-Class SVM is initialized with a kernel type parameter. The kernel type used is Radial Basis Function (RBF) as it can handle non-linear relationships between the features. 'Nu', which is the upper bound of training errors, is set to 0.01 to indicate that very few data points are anomalies in the dataset to improve the accuracy of the model. The kernel coefficient 'Gamma' is set to auto, which simplifies the tuning process, and the complexity of the model is controlled. Then the model is trained on the entire pre-processed dataset features. This model helps to differentiate between normal points and anomalies, providing great accuracy for anomaly detection in PdM.

Local Outlier Factor: Local Outlier Factor (LOF) is used for detecting the local density anomalies in the dataset. During the initialization process, the model was initialized with parameters such as the number of neighbors, 'n_neighbors', which is set to 20 to ensure a balance between capturing the local structure of the data and maintaining the model's robustness. The contamination factor was set to auto, allowing the model to determine the optimal threshold for identifying anomalies in the data. This ensures the model's flexibility for an unknown number of anomalies in the dataset. The model is then trained on the features of the entire pre-processed dataset. During the training process, the model calculates the local density of any given points with respect to their neighbors. Data points with lower density compared to their neighbors are considered anomalies.

Tools Used: For data collection and preprocessing, the dataset used in this project is *'ai4i2020.csv'*. The library used for data collection and storage is *'pandas'*, a powerful data manipulation library in Python. For preprocessing, techniques such as label encoding using *'LabelEncoder()'* and normalization using *'MinMaxScaler()'* are employed from the *'sklearn.preprocessing'* library, which is specifically used for data preprocessing.

The next library used is *'sklearn.model_selection'*, mainly used for splitting the data into training and testing sets. This library is also used for evaluating the robustness of the models using *'StratifiedKFold()'*. For the application of machine learning models like *'IsolationForest()'* and *'RandomForestClassifier()'*, the *'sklearn.ensemble'* library is used, providing ensemble methods like Isolation Forest and Random Forest. The *'sklearn.svm'* library is used for applying the unsupervised One-Class Support Vector Machine using the *'OneClassSVM()'* tool. For the implementation of nearest neighbor algorithms like K-NN and LOF, the *'sklearn.neighbors'* library is used, with tools such as *'LocalOutlierFactor()'*

and *'KNeighborsClassifier()'*. For evaluation metrics such as accuracy, precision, recall, F-1 score, and confusion matrix, the *'sklearn.metrics'* library is used, with tools such as *'classification_report()'* and *'confusion_matrix()'*. For visualization purposes, this project includes two libraries: *'matplotlib.pyplot'*, with functions such as *'plt.subplots()'*, *'plt.tight_layout()'*, and *'plt.show()'*; and *'seaborn'*, used for generating boxplots with the tool *'sns.boxplot'*. For numerical operations, the *'numpy'* library is used.

5 Evaluation

After the successful implementation of the selected machine learning models, it is important to evaluate various metrics of the models to understand the model performance in a clear way. The metrics used for evaluating the performance of the models are Accuracy, Precision, Recall, and F-1 score. With the help of these metrics, the overall performance of the model can be understood.

5.1 Unsupervised ML Algorithms

The main motive behind the use of unsupervised learning algorithms is to determine if they can be used for anomaly detection in PdM. This can be confirmed by thoroughly evaluating their metric performances.

Isolation Forest:

Model Performance:

- Accuracy: 0.93
- Precision: 0.31
- Recall: 0.97
- F1 Score: 0.47

Isolation Forest shows a high accuracy of 93% which indicates that it correctly identifies the majority of the instances. However, a high accuracy score doesn't always mean that the model is performing well. The precision of 31% indicates that when the model predicts a failure, it is often wrong, generating many false positives. The recall of 97% shows that the model is sensitive to failures and can almost detect all the failures. The moderate F-1 score reflects the balance between precision and recall, but due to low precision, the F-1 score is lower.

One-Class SVM:

Model Performance:

- Accuracy: 0.97
- Precision: 0.82
- Recall: 0.24

- F1 Score: 0.38

One-Class SVM shows an excellently high accuracy of 97%, indicating that the model is performing well in distinguishing between normal instances and anomalies. The precision of the model is also high, suggesting that the model is successful in reducing false positives. This means that when the model predicts a failure, it is most likely a failure, which can help save on the maintenance cost associated with false PdM. However, the model has a low recall of 24%, indicating that it often fails to predict when the machine is actually going to fail.

Local Outlier Factor:

Model Performance:

- Accuracy: 0.96
- Precision: 0.40
- Recall: 0.10
- F1 Score: 0.16

The model has a high accuracy of 96%, but similar to One-Class SVM, this accuracy can be misleading. The model likely predicts the majority class of non-failure correctly. However, the precision of the model is quite low at 40%, indicating that the model generates many false positives. The recall of the model is also quite low at 10%, showing that the model fails to identify 90% of the actual failures, which can lead to severe operational issues.

5.2 Supervised ML Algorithms

Supervised ML algorithms are commonly used in the task of predictive maintenance. In this project, supervised ML algorithms are implemented to compare the performance of unsupervised ML algorithms with them. This comparison makes it easier to determine the performance of unsupervised ML algorithms with respect to supervised algorithms.

K-NN:

Model Performance:

- Accuracy: 1.00
- Precision: 1.00
- Recall: 0.97
- F1 Score: 0.98

The K-NN model achieves an accuracy of 100%, indicating that the model has perfect accuracy in identifying all instances of failure and non-failure correctly. This demonstrates that K-NN is highly effective in differentiating between failure and non-failure classes. The precision of K-NN is also 100%, meaning that every failure prediction made by the model is correct. K-NN does not produce any false positive failures, which further reduces the maintenance cost associated with false alarms. Even the recall of this model is high at 97%, indicating that the model correctly detects 97% of the actual failures.

Random Forest:

Model Performance:

- Accuracy: 1.00
- Precision: 1.00
- Recall: 0.97
- F1 Score: 0.98

Similar to K-NN, the Random Forest model also achieves a perfect accuracy of 100%. This indicates that the model correctly identifies both failures and non-failures. The precision of Random Forest is also 100%, meaning that every failure predicted by the model is correct. As a result, the model generates no false positives, which is beneficial for reducing maintenance costs. The recall of Random Forest is 97%, indicating that the model correctly detects 97% of the actual failures.

Cross-Validation for K-NN and RF: In the above evaluation, it is seen that K-NN and RF are performing really well on the data. To ensure that the high performance of K-NN and RF is not dependent on the specific train-test split, cross-validation Seraj et al. (2023) is performed on these two models. By using multiple train-test splits, cross-validation helps to generalize the model for unseen data, which can be useful in identifying overfitting or underfitting of the model. Class imbalance is one of the major reasons for a model being overfitted or underfitted. Therefore, to ensure that there is no overfitting of the data, cross-validation is performed.

Cross-Validation Results:

- **K-NN:**
  - Average Precision: 1.00
  - Average Recall: 0.97
  - Average F1-Score: 0.99
  - Average ROC AUC: 0.99
- **Random Forest:**
  - Average Precision: 1.00
  - Average Recall: 0.97

– Average F1-Score: 0.99

– Average ROC AUC: 0.99

After performing cross-validation, both K-NN and RF maintained almost the same performance metrics as before. This shows that even on different subsets of data, the models perform well. If the models had performed poorly on the validation folds, it could have indicated overfitting. However, the models performed similarly on all the data subsets, indicating that they are performing well even in the presence of imbalanced data. In conclusion, it is evident that the models are performing well on the data, and they are not overfitting.

5.3 Results

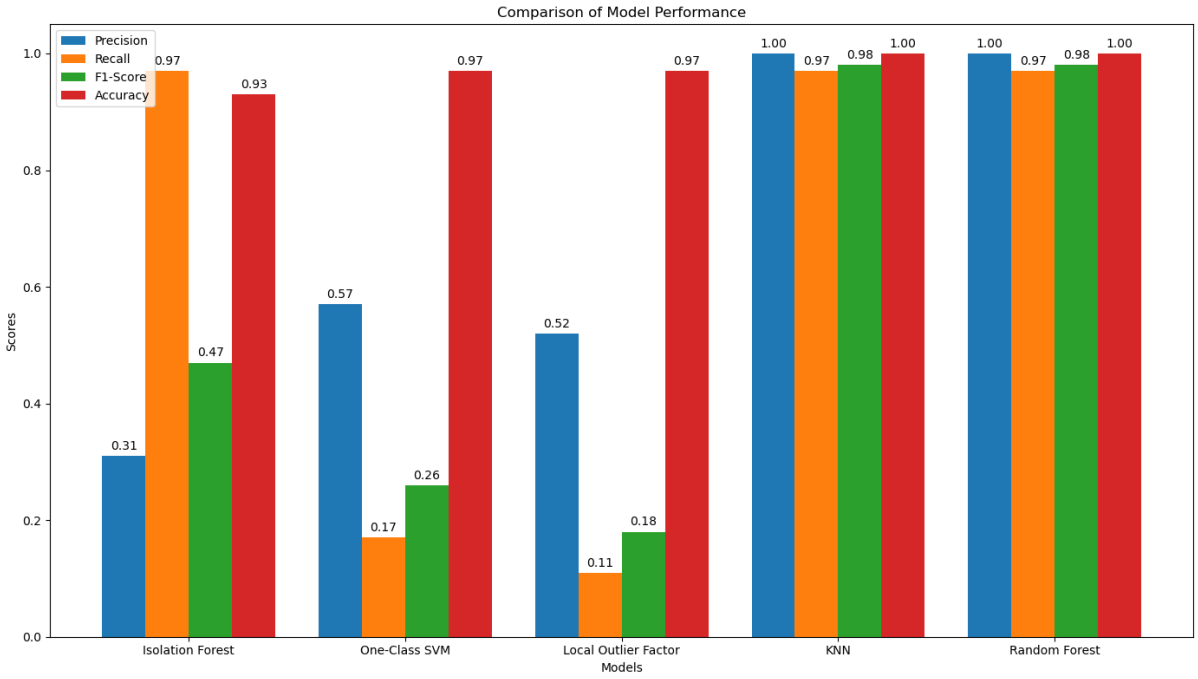

Figure 4: Model performance Comparison

By comparing all the metrics, it can be seen that the supervised ML algorithms showed better performance compared to unsupervised algorithms. Even though the accuracy of unsupervised algorithms was high, due to low precision and recall, these models are not as reliable as the supervised ML algorithms. Among the unsupervised algorithms, Isolation Forest performed better than One-Class SVM and LOF due to its higher recall than the other two, but it still has low precision, which suggests that it generates false positives. In conclusion, K-NN and RF performed the best among all the ML models used for this project.The model performance of K-NN and RF is in line with previous research by Cakir et al. (2021) in which the model showed near perfect accuracy for both the supervised models in their research.

6 Conclusion and Future Work

The main objective of this research is to evaluate whether unsupervised learning algorithms can be used for anomaly detection in machinery for predictive maintenance. To achieve this goal, the implementation design consists of deploying three unsupervised ML algorithms: Isolation Forest, One-Class SVM, and Local Outlier Factor. As a benchmark for comparison, two supervised ML models, K-NN and Random Forest, were also implemented.

After evaluating the performance of the unsupervised algorithms, Isolation Forest performed better than One-Class SVM and LOF with a high recall of 0.97. However, due to its lower precision, it suggests that it generates a significant number of false positives. One-Class SVM showed a more balanced performance in recall and precision compared to Isolation Forest, but it still could not correctly predict the true positives. LOF showed the worst performance among all the models, despite having a high accuracy of 0.96. Due to its poor recall and precision, the model did not perform well in predicting actual failures.

On the other hand, supervised learning algorithms performed exceptionally well. Both K-NN and RF showed perfect accuracy and precision of 1 and near-perfect recall of around 0.97. The findings of this research are consistent with previous research in the field of predictive maintenance, indicating that supervised learning models often perform better than unsupervised learning algorithms. Although unsupervised learning algorithms are designed for anomaly detection, they showed lower precision and recall, which are significantly lower than those of supervised learning algorithms. Furthermore, after performing cross-validation in this study, it is confirmed that the ML algorithms are robust and effective on new data as well. The strength of this study lies in its coverage of both supervised and unsupervised learning algorithms, with proper parameter comparison between them, and the use of techniques like cross-validation to confirm the efficiency of the models.

However, this study also shows some shortcomings. The precision and recall of unsupervised ML models can generate many false positives, which can lead to unnecessary maintenance costs. Additionally, the dataset is limited in size and may not cover all possible operating conditions of the equipment.

In conclusion, the research focused on determining whether unsupervised learning algorithms can be used for anomaly detection in equipment for predictive maintenance. Even though unsupervised algorithms showed high accuracy, due to their low precision and recall, the performance and efficiency of the models fell short compared to supervised machine learning algorithms. Thus, it can be concluded from this research that unsupervised machine learning models alone are not sufficient for predictive maintenance, but they can be used to complement other supervised machine learning models.

Future Work: For future work, the focus of research can shift towards improving the precision and recall of unsupervised models. This might be possible through a hybrid approach where unsupervised algorithms are integrated with supervised algorithms. Additionally, various advanced unsupervised learning algorithms like autoencoders or deep learning-based anomaly detection could be used to improve the performance of unsupervised models. Larger datasets with different parameters and sensor readings could also be used for training the models to provide data diversity. By applying these methods in future research, the performance of unsupervised algorithms can be improved.

References

Abidi, M. H., Mohammed, M. K. and Alkhalefah, H. (2022). Predictive maintenance planning for industry 4.0 using machine learning for sustainable manufacturing, *Sustainability* **14**(6).
URL: *https://www.mdpi.com/2071-1050/14/6/3387*

Alfaro-Nango, A. J., Escobar-Gómez, E. N., Chandomí-Castellanos, E., Velázquez-Trujillo, S., Hernandez-De-León, H. R. and Blanco-González, L. M. (2022). Predictive maintenance algorithm based on machine learning for industrial asset, *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)*.
URL: *https://doi.org/10.1109/CoDIT55151.2022.9803983*

Alghushairy, O., Alsini, R., Soule, T. and Ma, X. (2021). A review of local outlier factor algorithms for outlier detection in big data streams, *Big Data and Cognitive Computing* **5**(1).
URL: *https://www.mdpi.com/2504-2289/5/1/1*

Amruthnath, N. and Gupta, T. (2018). A research study on unsupervised machine learning algorithms for early fault detection in predictive maintenance, *2018 5th International Conference on Industrial Engineering and Applications (ICIEA)*.
URL: *https://doi.org/10.1109/IEA.2018.8387124*

Bouabdallaoui, Y., Lafhaj, Z., Yim, P., Ducoulombier, L. and Bennadji, B. (2021). Predictive maintenance in building facilities: A machine learning-based approach, *Sensors* **21**(4).
URL: *https://www.mdpi.com/1424-8220/21/4/1044*

Cakir, M., Guvenc, M. A. and Mistikoglu, S. (2021). The experimental application of popular machine learning algorithms on predictive maintenance and the design of iiot based condition monitoring system, *Computers Industrial Engineering* **151**: 106948.
URL: *https://doi.org/10.1016/j.cie.2020.106948*

Carrasco, J., López, D., Aguilera-Martos, I., García-Gil, D., Markova, I., García-Barzana, M., Arias-Rodil, M., Luengo, J. and Herrera, F. (2021). Anomaly detection in predictive maintenance: A new evaluation framework for temporal unsupervised anomaly detection algorithms, *Neurocomputing* **462**: 440–452.
URL: *https://doi.org/10.1016/j.neucom.2021.07.095*

Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J. and Barbosa, J. (2020). Machine learning and reasoning for predictive maintenance in industry 4.0: Current status and challenges, *Computers in Industry* **123**: 103298.
URL: *https://doi.org/10.1016/j.compind.2020.103298*

Gohel, H. A., Upadhyay, H., Lagos, L., Cooper, K. and Sanzetenea, A. (2020). Predictive maintenance architecture development for nuclear infrastructure using machine learning, *Nuclear Engineering and Technology* **52**(7): 1436–1442.
URL: *https://doi.org/10.1016/j.net.2019.12.029*

Kolokas, N., Vafeiadis, T., Ioannidis, D. and Tzovaras, D. (2020). A generic fault prognostics algorithm for manufacturing industries using unsupervised machine learning

classifiers, *Simulation Modelling Practice and Theory* **103**: 102109.
URL: *https://doi.org/10.1016/j.simpat.2020.102109*

Lee, J., Bagheri, B. and Kao, H.-A. (2015). A cyber-physical systems architecture for industry 4.0-based manufacturing systems, *Manufacturing Letters* **3**: 18–23.
URL: *https://doi.org/10.1016/j.mfglet.2014.12.001*

Pang, J., Pu, X. and Li, C. (2022). A hybrid algorithm incorporating vector quantization and one-class support vector machine for industrial anomaly detection, *IEEE Transactions on Industrial Informatics* .
URL: *https://doi.org/10.1109/TII.2022.3145834*

Satwaliya, D. S., Thethi, H. P., Dhyani, A., Kiran, G. R., Al-Taee, M. and Alazzam, M. B. (2023). Predictive maintenance using machine learning: A case study in manufacturing management, *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*.
URL: *https://doi.org/10.1109/ICACITE57410.2023.10183012*

Schröer, C., Kruse, F. and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model, *Procedia Computer Science* **181**: 526–534. CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020.
URL: *https://doi.org/10.1016/j.procs.2021.01.199*

Seraj, A., Mohammadi-Khanaposhtani, M., Daneshfar, R., Naseri, M., Esmaeili, M., Baghban, A., Habibzadeh, S. and Eslamian, S. (2023). Chapter 5 - cross-validation, *in* S. Eslamian and F. Eslamian (eds), *Handbook of Hydroinformatics*, Elsevier, pp. 89–105.
URL: *https://doi.org/10.1016/B978-0-12-821285-1.00021-X*

Souza, R. M., Nascimento, E. G., Miranda, U. A., Silva, W. J. and Lepikson, H. A. (2021). Deep learning for diagnosis and classification of faults in industrial rotating machinery, *Computers Industrial Engineering* **153**: 107060.
URL: *https://doi.org/10.1016/j.cie.2020.107060*

Susto, G. A., Schirru, A., Pampuri, S., McLoone, S. and Beghi, A. (2015). Machine learning for predictive maintenance: A multiple classifier approach, *IEEE Transactions on Industrial Informatics* **11**(3): 812–820.
URL: *https://doi.org/10.1109/TII.2014.2349359*

Theissler, A., Pérez-Velázquez, J., Kettelgerdes, M. and Elger, G. (2021). Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry, *Reliability Engineering System Safety* **215**: 107864.
URL: *https://doi.org/10.1016/j.ress.2021.107864*

van Dinter, R., Tekinerdogan, B. and Catal, C. (2022). Predictive maintenance using digital twins: A systematic literature review, *Information and Software Technology* **151**: 107008.
URL: *https://doi.org/10.1016/j.infsof.2022.107008*