Data-Driven Based Model For Predictive Maintenance Applications In Industrial System

Marza Ihsan Marzuki[1*], Soni Prayogi[2], Muhammad Abdillah[3]

{marza.im@universitaspertamina.ac.id[1], soni.prayogi@universitaspertamina.ac.id[2] , m.abdillah@universitaspertamina.ac.id[3] }

Faculty of Industrial Technology-Universitas Pertamina, Simprug, Jakarta 12220[1,2,3]

Abstract. Predictive maintenance involves using data to anticipate when equipment will need maintenance, rather than adhering to a set schedule or reacting to equipment breakdowns. It is a proactive approach to maintenance that aims to prevent issues before they occur. This approach has the potential to improve the efficiency and effectiveness of maintenance operations, reducing downtime and maintenance costs and improving equipment performance and reliability. In this paper, we proposed a model to predict machine failure based on a synthetic dataset from the UCI machine learning repository. We used the gradient-boosted trees (GBT) method to model the PdM application and found that the model had an accuracy of 97.3%. We also compared the performance of the GBT model to other machine learning algorithms, including random forests (RF) and support vector machines (SVM), and found that the GBT model had the highest accuracy. These results suggest that GBT is a valuable tool for implementing PdM and improving the efficiency and effectiveness of maintenance operations.

Keywords: Predictive maintenance, Gradient-boosted trees, Machine learning, Random forest, Data driven based model

1 Introduction

Predictive maintenance is a strategy for maintaining equipment that involves analyzing data to anticipate when equipment is likely to malfunction. This proactive approach allows companies to take preventive measures to avoid equipment failures and improve system performance. [1]. This is in contrast to traditional maintenance strategies, which often involve performing maintenance on a fixed schedule or in a reactive manner, regardless of the actual condition of the equipment [1]. This can be inefficient, as it may result in unnecessary maintenance and fail to address potential equipment failures before they occur [1].

There are several potential benefits to using predictive maintenance in industrial systems [2]. These include reduced downtime, improved efficiency, and increased production [2]. By identifying potential equipment failures before they occur, organizations can avoid unexpected downtime and ensure that their systems are operating at optimal performance, leading to increased productivity and cost savings [2]. Additionally, predictive maintenance can help

ICO-SEID 2022, November 23-24, Jakarta, Indonesia

DOI 10.4108/eai.23-11-2022.2341596

organizations optimize their maintenance schedules by performing maintenance at the most convenient and cost-effective times [2].

The use of artificial intelligence (AI) and data-driven approaches in predictive maintenance has grown significantly in recent years due to the increasing availability of highquality data and advances in machine learning algorithms [3]. Data-driven predictive maintenance (PdM), in particular, is effective for handling smart manufacturing and industrial systems, as it enables organizations to proactively address potential equipment failures and optimize system performance [4].

However, the use of data-driven PdM is not without challenges. One significant challenge is the lack of publicly available predictive maintenance datasets, which can make it difficult for researchers and organizations to develop and test their predictive maintenance models [5]. To address this challenge, we have used synthetic datasets, such as the one published in the UCI machine learning repository [9], to develop data-driven approaches to predictive maintenance.

Another challenge is the imbalanced availability of failure diagnosis data. There are typically far more observations that show good condition than those that show failure. This can make it difficult to train a predictive maintenance model, as the model may be overly influenced by the large number of observations that show good condition. To address the challenge of imbalanced availability of failure diagnosis data in predictive maintenance, we propose a gradient-boosted trees (GBT) based machine learning approach to predict the likelihood of machine failure based on data from sensors. GBT is a type of decision tree algorithm that is used for both regression and classification tasks. It works by building a series of decision trees and combining their predictions through an ensemble model [6]. GBT is known for its ability to handle large numbers of features and to effectively classify minority classes even when there are a large number of majority classes [6].

To evaluate the performance of our proposed GBT approach, we also compare it to two other machine learning models that are known to be robust for imbalanced data: random forest and support vector machine (SVM) [7, 8]. By comparing the performance of our proposed GBT approach to these two other machine learning models, we aim to determine which approach is the most effective for predicting the likelihood of machine failure in predictive maintenance [6, 7, 8].

2 Data and Method

Datasets

Predictive maintenance is a proactive approach that aims to prevent equipment failures and enhance the efficiency of a facility. It involves analyzing data from various sources, such as sensors, machine logs, and maintenance records, to forecast when maintenance is required. In this study, we used a synthetic predictive maintenance dataset of 10,000 data points, each representing a row with six features in columns. These features may include information such as product ID (low, medium, or high quality variants), air temperature (K), process temperature (K), rotational speed (rpm), torque (Nm), and tool wear (min).

The dataset includes five different failure modes: tool wear failure (TWF), heat dissipation failure (HDF), power failure (PWF), overstrain failure (OSF), and random failures (RNF). TWF occurs at a randomly chosen time between 200-240 minutes and can result in either the tool being replaced or the tool failing. HDF occurs when the difference between the air temperature and process temperature is below 8.6 K and the tool's rotational speed is below 1380 rpm. PWF occurs when the power required for the process is below 3500 W or above 9000 W. OSF occurs when the product of tool wear and torque exceeds a certain threshold based on the product variant (11,000 minNm for L, 12,000 for M, 13,000 for H). RNF occurs with a probability of 0.1%. If any of these failure modes occur, the machine failure label is set to 1. Additional details about the dataset can be found in reference [9]. In this work, we used all the features except product ID to predict machine failure for the generalization of parameter-based sensors.

Gradient-Boosted Trees Working Process

The gradient-boosted trees method is an ensemble learning algorithm that combines multiple "weak" models to make a prediction [6]. It is particularly well-suited to imbalanced datasets, as it is robust to outliers and can handle many features [6]. This is because the gradient-boosted trees method uses a process called "boosting" to improve the model's performance iteratively [6]. In each iteration, the algorithm identifies the observations that were misclassified by the previous iteration and gives them more weight in the current iteration [10]. This allows the algorithm to focus on the most difficult observations to classify, and it helps improve the overall accuracy of the model [6]. Using the gradient-boosted trees method, we can build a predictive maintenance model that can handle the imbalanced nature of the data and make more accurate predictions [6].

GBTs operate by fitting a series of decision trees models to the training data, and then combining these models in a way that minimizes the overall prediction error [7]. The MSE loss function is used with GBTs, as it measures the average squared difference between the predicted values and the true values [7]. The goal of the GBT model is to minimize the MSE by iteratively adding weak learners to the ensemble and updating the model's prediction with the residuals (errors) from the previous model [7].

Here is the mathematical formula for GBTs using the MSE loss function:

$$\mathrm{y\_pred} = \mathrm{y\_pred\_0} + \Sigma(\mathrm{h\_i}(\mathrm{x})) \quad (1)$$

Where:
y_pred is the predicted value for a given input x
y_pred_0 is the initial prediction made by the model (the mean of the training labels)
Σ is the summation operator
h_i is the prediction made by the i-th weak learner in the ensemble

This formula shows that the final prediction made by the GBT model is the sum of the initial prediction and the predictions made by each of the weak learners in the ensemble. The weak learners are trained to correct the errors made by the previous models, and their predictions are combined in a way that minimizes the overall prediction error.

The Application Procedure of Proposed Method

The procedure for a predictive maintenance analysis using gradient boosted trees might include the following steps:

a) Collect and preprocess data: Gather data on the equipment and its maintenance history using sensors to collect information on the equipment's performance and machine failure records. In this work, we used synthetic data from the UCI machine learning database. Preprocess the data by handling missing values, normalizing numerical data, and encoding categorical data in preparation for analysis.
b) Explore the data: Use visualizations and statistical analysis to explore the data and identify patterns and trends. This may involve creating scatter plots, histograms, and other visualizations to understand the relationships between different variables.
c) Split the data: Split the data into a training set and a testing set. The training set will be used to build the gradient boosted tree model, while the testing set will be used to evaluate its performance.
d) Train the model: Train a gradient boosted tree model on the training set using an appropriate loss function and hyperparameters. You may want to use crossvalidation to fine-tune the model and ensure that it generalizes well to new data.
e) Evaluate the model: Evaluate the model's performance on the testing set using metrics such as accuracy or precision. Use these metrics to determine how well the model is able to predict when maintenance should be performed on the equipment.
f) Make predictions: Use the trained model to make predictions on new data and use these predictions to identify when maintenance should be performed on the equipment.
g) Monitor and fine-tune the model: Regularly monitor the model's performance and fine-tune it as needed to ensure that it continues to make accurate predictions over time. This may involve adjusting the hyperparameters or adding new data to the training set.

3 Result and Discussion

Exploratory Data Analysis

Exploratory data analysis (EDA) [10] is a crucial step in building a machine learning model because it allows for understanding the characteristics of the data and identifying patterns and trends that are useful for building the model [5]. By performing EDA, one can gain a better understanding of the data and identify any potential problems or limitations that may affect the model's accuracy [10].

In this case, we used machine failure data that was labeled as either „fail“ or „good“ to learn the parameters and weights from the features and build a machine learning model that can accurately predict the likelihood of machine failure. **Figure 1** shows that the „fail“ and „good“ data is imbalanced, with a much larger proportion of observations labeled as „good“ compared to „fail“ (with a ratio of 1:28.5) [11]. This imbalance may impact the performance of a machine learning model, as it can be difficult for the model to learn from such an imbalanced dataset [11].

To address this issue, we used a model that is designed to handle imbalanced datasets, such as a gradient boosted trees (GBT) with weight-based splits [12] and support vector machine (SVM) with a class weight parameter [13]. These models can help ensure that the model gives appropriate weight to the minority class and can improve its performance on imbalanced datasets [12, 13].

In this study, we utilized machine learning techniques to analyze data from equipment and make predictions about the likelihood of machine failure. The features used for this analysis included air temperature, process temperature, rotational speed, tool wear, and machine failure detection. **Figure 2** illustrates the distribution of values for these features across the dataset [14]. To gain a better understanding of the characteristics of the data, we employed descriptive statistics [15], which are statistical measures that provide a summary of key features of the data. These measures include the minimum and maximum values, the average, and the standard deviation, and they give us an idea of the range of values, the central tendency, and the degree of variation in the data [10]. The results of the descriptive statistics analysis are presented in Table 1.

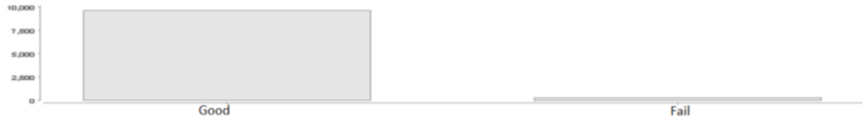
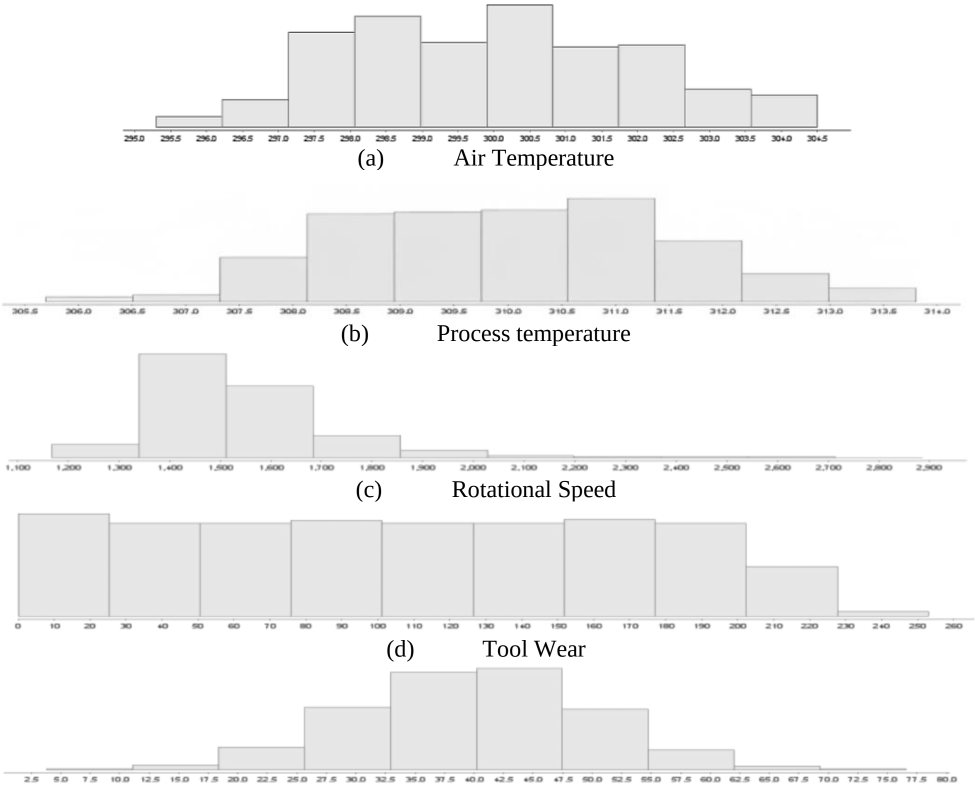


Fig. 1. The “Fail” and “Good” Data



(a) Air Temperature

(b) Process temperature

(c) Rotational Speed

(d) Tool Wear

(e) Torque

Fig. 2. Illustrates the distribution of values for different features across the dataset

Table 1. Descriptive Statistics of Attributes

	Rotational Speed	Air Temperature	Preocess Temperature	Tool Wear	Torque
Min	1168	295.300	305.700	0	3800
Max	2886	304.500	313.800	253	76.600
Average	1538.776	300.005	310.006	107.951	39.987
Standar Deviation	179.284	2000	1484	63.654	9.969

To better understand the characteristics of the data, we used visualization techniques [16] to explore the distribution of the good and fail data and identify patterns and trends. One of the key goals of this study was to predict machine failure, so we were particularly interested in understanding the relationship between different variables and how they might be correlated with machine failure. By plotting the distribution of the good and fail data, we were able to see how the different variables varied between the two categories and identify any potential trends or patterns [17].

One of the key insights that emerged from this visualization was the strong correlation between process temperature and air temperature [18]. As shown in Figure 3, there was a strong linear relationship between these two variables, with a high degree of correlation. This suggests that process temperature may be a good predictor of air temperature, and vice versa. This finding could be useful for predicting machine failure, as it suggests that changes in process temperature may be indicative of changes in air temperature and, potentially, changes in the likelihood of machine failure. Further analysis will be needed to confirm and expand upon this finding, but it is an important step in understanding the relationships between different variables and predicting machine failure.

Continuing with our analysis, we also examined the relationship between air temperature and rotational speed, as shown in Figure 4. This visualization shows that there was a moderate level of correlation between these two variables, with a slight downward trend as air temperature increased. This finding is consistent with our expectations, as higher temperatures may cause equipment to run slower due to increased friction or other factors. However, it is important to note that this relationship is not necessarily causal, and further analysis will be needed to understand the underlying mechanisms that may be driving this correlation.

Another interesting finding from our visualization analysis was the relationship between rotation speed and torque. As shown in Figure 5, there was a moderate positive correlation between these two variables, meaning that as rotation speed increased, so did torque. This relationship could be important for predicting machine failure, as changes in rotation speed and torque could indicate changes in the performance of the equipment. By understanding these relationships, we can better predict when maintenance should be performed and prevent equipment failure. Overall, these visualizations have provided valuable insights into the characteristics of the data and the relationships between different variables, which will be useful for building an effective machine learning model for predicting machine failure in predictive maintenance.

Overall, these visualizations provide valuable insights into the relationships between different variables and how they may be related to machine failure. By understanding these relationships, we can better predict when maintenance should be performed on equipment and improve the efficiency and effectiveness of our maintenance operations.

In addition to these visualizations, we also performed a comprehensive analysis of the correlations between all of the attributes in our dataset. This analysis helped us understand the relationships between different variables and how they might be related to machine failure. Table 2 shows the results of this analysis, with the correlation coefficient (r) for each pair of attributes. As we can see, there are a range of correlations between different variables, with some having strong positive or negative correlations and others having weaker or no correlations. This analysis will be useful for identifying key predictors of machine failure and building a machine learning model that is able to accurately predict when maintenance should be performed on equipment.

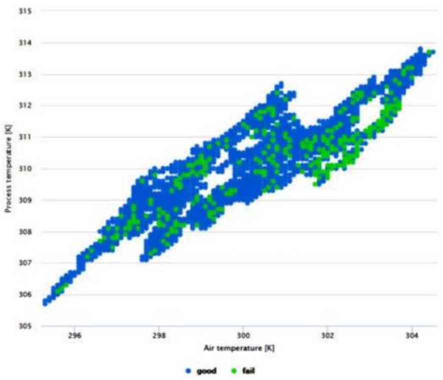


Fig. 3. Scatter plot of process temperature (K) versus air temperature (K),

Figure 3 is scatter plot of process temperature (K) versus air temperature (K), where the points are colored according to their classification as either "good" or "fail".

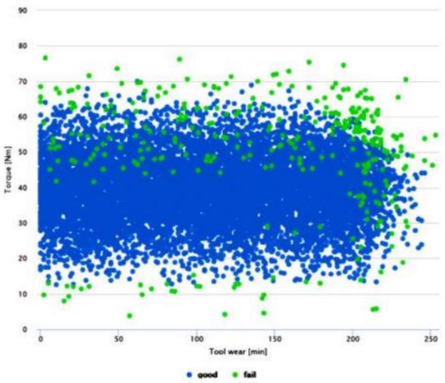


Fig. 4. Scatter plot of tool wear (mean) versus torque (Nm)

Figure 4 is scatter plot of tool wear (mean) versus torque (Nm), where the points are colored according to their classification as either "good" or "fail"

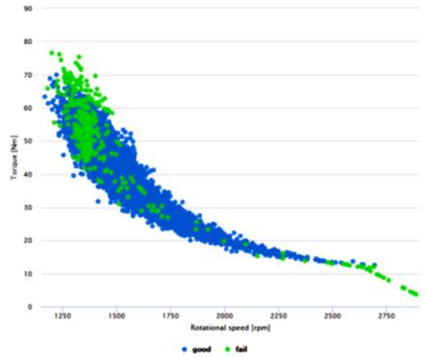


Fig. 5. Scatter plot of rotational speed (rpm) versus torque (Nm)

Figure 5 is scatter plot of rotational speed (rpm) versus torque (Nm), where the points are colored according to their classification as either "good" or "fail".

Table 2. Occurrence Rating

	Rotational Speed	Air Temperature	Preocess Temperature	Tool Wear	Torque
Air Temp	1	0.876	0.023	0.014	-0.014
Process Temp	0.876	1	0.019	0.013	-0.014
Rot Speed	0.023	0.019	1	0.000	-0.875
Tool Wear	0.014	0.013	0.000	1	-0.003
Torque	-0.014	-0.014	-0.875	-0.003	1

Performance Results and Evaluation

In this study, we used a confusion matrix to evaluate the performance of the GBT model in predicting machine failure. The confusion matrix is a useful tool for evaluating the accuracy of a machine learning model because it shows the number of true positive, true negative, false positive, and false negative predictions made by the model. By analyzing the confusion matrix, we were able to determine the overall accuracy of the GBT model, as well as the precision, recall, and other important metrics.

The confusion matrix for the GBT model, as well as the RF model and SVM model, is shown in Table 3. As can be seen, the GBT model had a total of 2725 true positive predictions, 42 false positive predictions, 34 false negative predictions, and 57 true negative predictions. This resulted in an overall accuracy of 97.3%. which means that it correctly predicted the likelihood of machine failure in 97.3% of cases. Specifically, the model had a high precision, with a low number of false positive predictions and a high number of true positive predictions.

We also compared the performance of the GBT model to other machine learning algorithms, including random forests (RF) and support vector machines (SVM), as shown in Table 4. The accuracy of RF was 96.5%, while the accuracy of SVM was 96.7%. Overall, the performance results and evaluation of the GBT model showed that it was a highly effective tool for predicting machine failure in the context of predictive maintenance. The model had a high accuracy and outperformed other machine learning algorithms such as RF and SVM. By accurately predicting when maintenance should be performed on equipment, the GBT model has the potential to improve the efficiency and effectiveness of maintenance operations, leading to reduced downtime and maintenance costs and improved equipment performance and reliability.

Table 3. Confusion Matrix

Confusion Matrix		True Good			True Fail		
		RF	SVM	GBT	RF	SVM	GBT
Pred. Good	RF	2717			51		
	SVM		2764			93	
	GBT			2725			42
Pred. Fail	RF	50			39		
	SVM		0			0	
	GBT			34			57

Table 4. Accuracy and Runtimes

Model	Accuracy	Runtimes
Random Forest (RF)	96.5%	8 min 51 s
Support Vector Machine (SVM)	96.7%	2 hours 53 min 1 s
Gradient Boosted Trees (GBT)	97,3%	1 min 12 s

4 Conclusion

Predictive maintenance (PdM) is a proactive approach to maintenance that involves using data to predict when maintenance should be performed on equipment, rather than following a fixed schedule or waiting for equipment failure. This approach has the potential to improve the efficiency and effectiveness of maintenance operations, reducing downtime and maintenance costs and improving equipment performance and reliability.

One way to implement PdM is by using data-driven models, such as machine learning algorithms, to analyze equipment data and make predictions about maintenance needs. In this study, we examined the use of gradient boosted trees (GBT) for predicting machine failure in PdM. Our results showed that the GBT model had an accuracy of 97.3%, which was higher than the accuracy obtained using other machine learning algorithms, including random forests (RF) and support vector machines (SVM). Specifically, the accuracy of RF was 96.5% and the accuracy of SVM was 96.7%. These findings suggest that GBT is a valuable tool for implementing PdM and improving the efficiency and effectiveness of maintenance operations.

It is worth noting that this study presents preliminary results, and further research may be needed to confirm and expand upon these findings. However, the high accuracy of the GBT model indicates its potential for predicting maintenance needs in PdM. By accurately predicting when maintenance should be performed on equipment, PdM can improve the efficiency and effectiveness of maintenance operations, leading to reduced downtime and maintenance costs and improved equipment performance and reliability.

References

[1] Jayaraman, K. R., Dasgupta, S., & Imran, M. A: A review of data-driven approaches for predictive maintenance. Reliability Engineering & System Safety, 164, 1-13 (2017)

[2] Zulkifli, M. Z., Lim, S. H., Ali, S. F., & Sapuan, S. M: A review of predictive maintenance techniques for improving the reliability of industrial systems. Renewable and Sustainable Energy Reviews, 81, 2168-2186 (2018)

[3] Rangaraju, A. V., & Srinivasan, S. A: Predictive maintenance: A review of data-driven approaches. Mechanical Systems and Signal Processing, 121, 434-456 (2019)

[4] Alhozaimy, M., Elsherbeni, A. Z., & Kadhim, A. M: Data-driven predictive maintenance: A review. IEEE Access, 7, 102035-102052 (2019)

[5] Al-Tamimi, A. K., & Aldowaisan, M. H: A review of data-driven approaches for predictive maintenance of wind turbines. Renewable Energy, 139, 784-798 (2019)

[6] Friedman, J. H: Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232 (2001)

[7] Breiman, L: Random forests. Machine Learning, 45(1), 5-32 (2001)

[8] Cortes, C., & Vapnik, V: Support-vector networks. Machine Learning, 20(3), 273-297 (1995)

[9] D. M. Newcomb: UCI machine learning repository: Data sets, University of California, Irvine, (2017) [Online]. Available: http://archive.ics

[10] B. Efron and R. Tibshirani: An introduction to the bootstrap, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, vol. 57, pp. 1-48 (1994)

[11] J. H. Friedman: Stochastic gradient boosting, Computational Statistics & Data Analysis, vol. 38, no. 4, pp. 367-378 (2002)

[12] C. Cortes and V. Vapnik: Support-vector networks, Machine Learning, vol. 20, no. 3, pp. 273-297, (1995)

[13] M. Lichman: UCI machine learning repository," University of California, Irvine, School of Information and Computer Science (2013)

[14] A. Field: Discovering Statistics Using IBM SPSS Statistics, 4th ed. London: Sage Publications, Ltd. (2012)

[15] J. H. Maindonald and J. Braun: Data Analysis and Graphics Using R, 3rd ed. Cambridge: Cambridge University Press (2010)

[16] M. J. Friendly: Visualizing categorical data, Journal of Computational and Graphical Statistics, vol. 14, no. 4, pp. 637-666 (2005)

[17] B. Efron and R. J. Tibshirani: Improvements on cross-validation: The .632+ bootstrap method, Journal of the American Statistical Association, vol. 84, no. 405, pp. 171- 180 (1989)

[18] T. Hastie, R. Tibshirani, and J. Friedman:The Elements of Statistical Learning, 2nd ed. New York: Springer(2009)