JMES The International Journal of
Mechanical Engineering and Sciences
https://iptek.its.ac.id/index.php/jmes
e-ISSN 2580-7471

Machine Learning-Based Prediction of Diesel Engine Health Using Operational Parameters: Comparison of SVM and BPNN Models

Fadli Nurdin[1], Mohammad Khoirul Effendi[1*], D. Mohakul[3]

[1]Department of Mechanical Engineering, ITS, Sukolilo Surabaya 60111, Indonesia
[2]Computer Science, DIAT, Pune, India.

Received: 19 March 2025, Revised: 16 September 2025, Accepted: 2 October 2025

Abstract

Condition-based maintenance (CBM) is crucial for enhancing the reliability of diesel engines. This study evaluates the effectiveness of support vector machine (SVM) and backpropagation neural network (BPNN) in predicting engine faults using operational parameters, such as engine RPM, lubricating oil pressure, fuel pressure, coolant pressure, oil temperature, and coolant temperature. Unlike previous research, this study validates engine condition labels based on standardized operational parameter thresholds, ensuring a more reliable and realistic representation of data. Statistical analyses using Spearman correlation and ANOVA deviance reveal that engine RPM and coolant temperature are significant predictors of engine health ($p < 0.05$). The findings show a notable difference in performance between the two classification models assessed. The Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel achieved an accuracy of 85.46%. However, the BPNN, configured with a [2-3-3-2] layer architecture and utilizing the tansig activation function, significantly outperformed the SVM, achieving an accuracy of 97.16%. These results suggest that the BPNN is more adept at capturing nonlinear patterns and providing more accurate predictions. Overall, this study underscores the importance of integrating domain-based data validation with machine learning techniques to develop reliable predictive maintenance systems.

Keywords: Diesel Engine, Condition-Based Maintenance, Support Vector Machine, Backpropagation Neural Network, Engine Health Prediction, Machine Learning, Operational Parameters

1. Introduction

Machine health monitoring has become a critical aspect in various sectors, including industrial systems, automotive applications, and marine transportation. Unexpected machine failures can lead to significant operational losses, increased maintenance costs, and safety risks [1,2]. Therefore, effective maintenance strategies are required to ensure optimal machine performance and reliability. One of the most widely adopted approaches is Condition-Based Maintenance (CBM), which utilizes real-time operational data to assess machine condition and predict potential failures before they occur [3]. Compared to traditional maintenance strategies such as time-based or corrective maintenance, CBM offers several advantages, including reduced downtime, improved maintenance efficiency, and extended machine lifespan [4]. However, the main challenge in CBM lies in processing and interpreting complex sensor data accurately to generate reliable predictions [5]. To address this issue, machine learning techniques have been increasingly applied for predictive maintenance tasks. Among various machine learning methods, Support Vector Machine (SVM) and Backpropagation Neural Network (BPNN) are widely used for machine health prediction. SVM is known for its strong generalization capability, especially in high-dimensional data spaces, while BPNN is effective in modeling complex nonlinear relationships within operational data [6, 7]. Previous studies have demonstrated the effectiveness of these methods in predicting machine conditions. For example, SVM-based models have achieved satisfactory classification performance in predictive maintenance applications, while optimized BPNN models have shown high accuracy in detecting engine degradation patterns [8,9]. Despite these advancements, several limitations remain. Previous studies often rely on conventional SVM kernels such as linear and polynomial, which may not be optimal for handling nonlinear characteristics of engine data [10]. In addition, the reliability of labeled datasets is often overlooked, where labels are directly adopted without validation against actual engineering standards. This may reduce the accuracy and practical applicability of predictive models. To address these gaps, this study proposes a comparative analysis between modified SVM models and BPNN for diesel engine health prediction. The SVM model is enhanced using non-linear kernels, specifically Radial Basis Function (RB-

*Corresponding author. Email: khoirul_effendi@me.its.ac.id

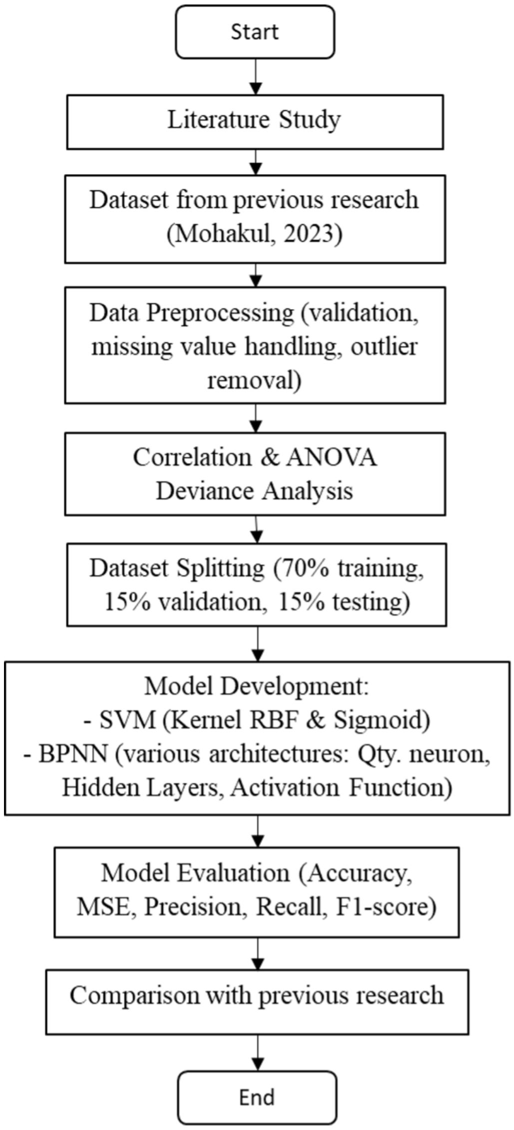


Figure 1. Research workflow.

F) and Sigmoid kernels, to better capture complex data patterns. Furthermore, instead of applying data relabeling techniques, this study validates engine condition labels based on standardized operational parameter thresholds derived from engineering references. This approach ensures that the dataset reflects realistic machine conditions and improves model reliability. In addition, statistical analyses, including Spearman correlation and ANOVA Deviance, are employed to identify significant operational parameters influencing engine health. The validated dataset is then used to develop and evaluate predictive models using SVM and BPNN, followed by a performance comparison with previous studies. The main contributions of this study can be summarized as follows: introducing an engineering-based validation approach for engine condition labels, improving SVM performance through the use of non-linear kernels, and providing a comprehensive comparison between SVM and BPNN for diesel engine health prediction.

2. Experimental/theoretical method

2.1. System Design

The system developed in this study leverages machine learning techniques to accurately assess the health status of diesel engines, enabling proactive maintenance decisions within a condition-based maintenance (CBM) framework. By continuously monitoring engine parameters and analyzing operational data, the system can detect early signs of degradation or potential faults, thereby reducing unexpected breakdowns and optimizing maintenance schedules. This predictive capability enhances engine reliability, lowers maintenance costs, and extends the service life of critical components. Integrating machine learning models into the CBM approach allows for adaptive learning from historical and real-time data, improving diagnostic precision over time. The system is designed to extract meaningful features from sensor inputs and employ advanced algorithms to classify engine conditions. The overall research workflow is illustrated in Figure 1. The process begins with a literature study to establish a theoretical foundation and identify relevant methods for engine health prediction. The dataset used in this study is obtained from previous research conducted by Mohakul (2023), which contains operational parameters of a marine diesel engine, including engine RPM, lubricating oil pressure, fuel pressure, coolant pressure, oil temperature, and coolant temperature. In contrast to previous approaches, this study applies a data validation process based on standardized operational parameter thresholds to ensure the reliability of engine condition labels. This step includes data cleaning, such as handling missing values and removing outliers, followed by validation of the engine condition labels to better represent actual operating conditions. Subsequently, statistical analysis is performed to understand the relationship between operational parameters and engine condition. Spearman correlation is used to evaluate the strength and direction of relationships between variables. Furthermore, ANOVA Deviance analysis based on logistic regression is conducted to identify significant parameters affecting engine health, including significance testing, goodness-of-fit evaluation, and main effect analysis. After the statistical analysis stage, the dataset is normalized using min-max normalization to ensure consistency in the scale of input variables. The dataset is then divided into training (70%), validation (15%), and testing (15%) sets to support model development and evaluation. In the modeling stage, two machine learning approaches are implemented, namely Support Vector Machine (SVM) and Backpropagation Neural Network (BPNN). The SVM model is developed using Radial Basis Function (RBF) and Sigmoid kernels with various parameter configurations. Meanwhile, the BPNN model is constructed using different network architectures, including variations in the number of hidden layers, neurons, activation functions, and the Levenberg-Marquardt training algorithm. Finally, model performance is evaluated using several metrics, including accuracy, mean squared error (MSE), precision, recall, and

Table 1. Dataset.

No	Engine rpm	Lub oil pressure	Fuel pressure	Coolant pressure	Lub oil temp	Coolant temp	Engine Condition
1	520	2.96	6.55	1.06	77.75	79.65	1
2	1221	3.99	6.68	2.21	76.40	75.67	0
3	729	3.85	10.19	2.36	77.92	71.67	1
4	845	4.88	3.64	3.53	76.30	70.50	0
5	824	3.74	7.63	1.30	77.07	85.14	0
6	1230	3.43	10.84	1.83	77.41	85.92	0
7	538	4.26	7.69	2.08	80.18	81.18	1
8	1187	2.59	6.89	1.83	78.10	84.97	1
9	609	3.75	10.09	3.00	77.28	75.58	1
10	606	2.27	5.49	1.91	75.17	77.73	1
...	...	...	...	...	...	...	...
165	1286	5.12	3.83	3.25	77.37	71.86	1
166	524	3.22	8.19	1.86	79.28	68.36	1
167	980	3.53	9.05	1.02	76.80	80.01	0
168	571	3.56	7.63	2.68	76.32	69.89	1
169	541	3.11	6.36	1.72	76.65	86.09	1
170	422	2.80	9.51	1.31	77.18	71.52	1
171	430	2.20	5.47	3.32	77.59	77.07	1
172	801	4.84	5.80	1.12	80.36	84.06	1
173	588	2.28	6.52	1.87	75.68	73.38	0
174	709	2.04	5.20	2.55	75.93	80.22	1

F1-score. The results of SVM and BPNN models are then compared with previous studies to assess performance improvements and determine the most effective method for diesel engine health prediction.

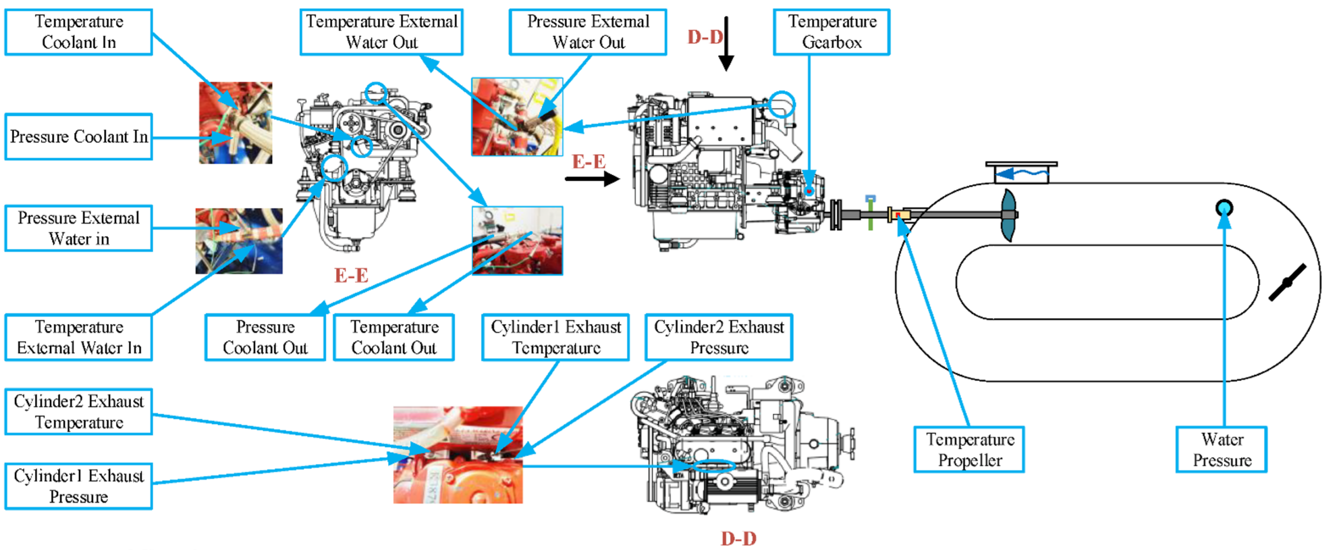
2.2. Dataset and Engine Data Acquisition

This study employs a dataset consisting of diesel engine operational parameters, including pressure and temperature-related variables. The dataset is sourced from previous studies conducted by D. Mohakul (2023) [10,11], which focus on the MTU Series 1400 marine diesel engine. The dataset includes several key operational parameters, engine speed (Engine RPM), lubricating oil pressure (Lub Oil Pressure) in bar, fuel pressure (Fuel Pressure) in bar, coolant pressure (Coolant Pressure) in bar, lubricating oil temperature (Lub Oil Temperature) in °C, and coolant temperature (Coolant Temperature) in °C. These parameters are used as input variables for predicting engine health condition.

The output variable is defined as engine condition, which is categorized into two classes: healthy (1) and unhealthy (0). In contrast to previous approaches that directly adopt existing labels, this study performs a validation process on the engine condition labels based on standardized operational parameter thresholds. This validation refers to established engineering standards, ensuring that the labeled data accurately represent actual engine conditions. The validated dataset is subsequently used for training and testing predictive models to evaluate the performance of Support Vector Machine (SVM) and Backpropagation Neural Network (BPNN) in predicting diesel engine health. The dataset is shown as in Table 1 and the data acquisition scheme is depicted as in Figure 2.

2.3. Data Preprocessing

Data preprocessing is conducted to ensure that the dataset used in this study is clean, consistent, and suitable for further analysis and model development [12]. This process begins with data cleaning, which includes handling missing values using mean imputation and detecting outliers using the interquartile range (IQR) method. The IQR approach is applied to identify and remove extreme values that may negatively affect model performance. In contrast to previous approaches that rely on data relabeling using clustering techniques, this study performs a validation process on the existing engine condition labels based on standardized operational parameter thresholds and actual engine conditions during data acquisition. Each data instance is evaluated by comparing its operational parameters with established standard ranges. If one or more classified as unhealthy (0); otherwise, it is classified as healthy (1). In addition, the validation process is suppor-



Sensor Specification:

Physical Channel	Sensor Model	Sensitivity	Physical Channel	Sensor Model	Sensitivity
Pressure Intake Air	XMLPM01GC71F		Temperature Coolant Out	YMC 8604	[0,200]°C = [1,5]v
Pressure Coolant In	Druck PMP4010	[0,5]bar = [0,10]v	Temperature Cyl Exhaust	YMC 8604	[0,200]°C = [1,5]v
Pressure Cylinder1 Exhaust	XMLPM05GC71F	[−1,5]bar = [0,10]v	Temperature ExWater In	YMC 8604	[0,200]°C = [1,5]v
Pressure Cylinder2 Exhaust	XMLPM05GC71F	[−1,5]bar = [0,10]v	Pressure ExWater Out		[0,10]bar = [0,5]v
Pressure Fuel Supply	XMLPM01GC71F		Pressure ExWater In	Druck PMP4010	[0,5]bar = [0,10]v
Temperature Gearbox Oil	YMC 8604	[0,200]°C = [1,5]v	Pressure Coolant Out		[0,10]bar = [0,5]v
Temperature External Water Out	YMC 8604	[0,200]°C = [1,5]v	Diesel Lub Oil Temp		[0,150]°C = [0,5]v
Temperature Cylinder2 Exhaust	YMC 8604	[0,200]°C = [1,5]v	Diesel Lub Oil Press		[0,10]bar = [0,5]v
Temperature Bushing	YMC 8604	[0,200]°C = [1,5]v	Engine RPM		
Temperature Coolant in	YMC 8604	[0,200]°C = [1,5]v	Water Pressure		[−1,1]bar = [0,10]v

Figure 2. Diesel engine operation data acquisition scheme [13].

ted by actual engine operating conditions observed during data collection. Indicators such as abnormal vibration, engine misfiring, unstable operation, or unexpected engine shutdown are considered to confirm the classification of unhealthy conditions. This approach ensures that the dataset reflects realistic engine behavior and improves the reliability of the classification labels. After the initial validation process, min-max normalization is applied to the dataset to standardize the range of feature values across all variables. This normalization technique rescales each feature to a fixed range, typically between -1 and 1 using mapminmax function in Matlab, which is essential for ensuring that the input data is compatible and balanced for machine learning algorithms, such as support vector machines (SVMs) and backpropagation neural networks (BPNNs). By transforming the features to a uniform scale, the models are prevented from being biased toward variables with larger magnitude values, thereby improving the convergence speed and overall predictive performance. Subsequently, the normalized dataset is partitioned into three distinct subsets to facilitate effective model training and evaluation. Seventy percent of the data is allocated for training, enabling the models to learn underlying patterns and relationships within the data. Fifteen percent is designated as validation data, which is used to fine-tune model parameters and prevent overfitting by monitoring performance on unseen samples during training. The remaining 15% is reserved as testing data, serving as an independent set for assessing the final model's generalization capability and robustness on new, unseen instances. This stratified division ensures a balanced approach to model development, validation, and unbiased performance assessment.

2.4. Analysis of Engine Operational Parameter Relationships

Spearman's correlation analysis was employed to investigate the relationship between diesel engine operational parameters and engine health condition, with the results visualized in the form of a heatmap [14]. This method was selected due to its capability to measure monotonic relationships without requiring normally distributed data, making it suitable for datasets with non-linear characteristics and binary output variables. The correlation coefficient was calculated to assess the strength and direction of the relationship between each operational parameter and the engine condition, as well as among the input variables themselves. The results are presented in a correlation matrix, where the magnitude of the coefficient indicates the level of association. In the heatmap representation, color intensity reflects the strength of the relationship: values approaching +1 indicate a strong positive correlation, values approaching -1 indicate a strong negative correlation, and values close to 0 represent weak

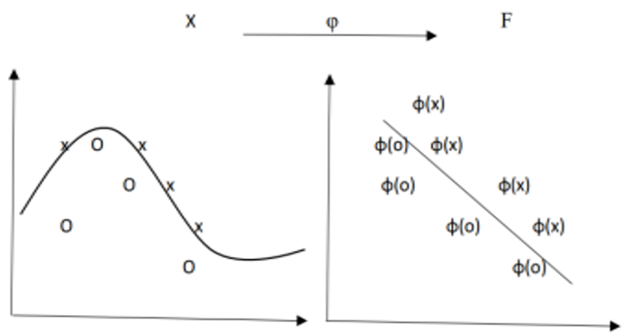


Figure 3. Kernel transforms non-linear problems into linear problems in a new space.

or negligible relationships. This analysis plays an important role in identifying the most influential parameters affecting engine health. By understanding these relationships, the study provides a foundation for improving predictive model performance and supporting more effective condition-based maintenance strategies.

2.5. ANOVA Deviance

Deviance-based analysis of variance (ANOVA) within the logistic regression framework is employed to evaluate the significance of operational parameters in relation to diesel engine health, which is represented as a binary outcome variable (healthy = 1, unhealthy = 0) [15]. This method is particularly suitable for classification problems involving categorical dependent variables, as it assesses the contribution of predictor variables based on likelihood estimation. The analysis is conducted by comparing a full model, which includes all predictor variables, with a reduced model using the likelihood ratio test (LRT). The statistical significance of each parameter is determined based on the p-value, where a value less than 0.05 indicates that the variable has a significant effect on engine health prediction. The ANOVA deviance procedure in this study consists of three main stages. First, the deviance test is performed to identify significant operational parameters influencing engine condition. Second, post-analysis is conducted using odds ratio estimation to quantify the effect of each parameter on the likelihood of engine health status. Third, model validation is carried out through goodness-of-fit testing to ensure that the logistic regression model adequately represents the observed data. This analytical approach provides a comprehensive understanding of the relationship between operational parameters and engine condition. The results are subsequently used to support feature relevance analysis and improve the reliability of predictive modeling using SVM and BPNN.

2.6. Support Vector Machine

The Support Vector Machine (SVM) modeling approach in this study utilizes two distinct kernel functions: the Radial Basis Function (RBF) and the Sigmoid kernel [13], each offering unique advantages in capturing non-linear patterns in diesel engine health data. The RBF kernel excels at mapping input features into a higher-dimensional space, thereby allowing the SVM to establish more flexible decision boundaries that effectively separate data points with complex patterns. This capability makes it particularly suitable for scenarios in which the relationship between variables is not linearly separable, which is a common occurrence in engine condition monitoring. Conversely, the sigmoid kernel functions similarly to neural networks by introducing a nonlinear transformation that models complex interactions between features. Its application in SVM enables the model to approximate nonlinear decision surfaces, thereby enhancing the classification performance when the underlying data distribution exhibits sigmoid-like characteristics. By employing both the RBF and sigmoid kernels, this study harnesses complementary strengths to improve the robustness and accuracy of the SVM model in diagnosing and predicting diesel engine health status, thereby effectively addressing various forms of nonlinearity and data complexity. The kernel transform scheme is shown in Figure 3.

2.6.1. Radial Basis Function (RBF) Kernel

The radial basis function (RBF) kernel is a widely used kernel function in machine learning, particularly in support vector machines (SVMs), for addressing nonlinear classification and regression challenges. By implicitly mapping input data into a higher-dimensional feature space, the RBF kernel enables linear algorithms to separate data that are not linearly separable in the original input space. This transformation occurs without explicitly calculating the coordinates in the high-dimensional space, thereby allowing for efficient computation, even with complex datasets. Mathematically, the RBF kernel assesses the similarity between two data points based on their Euclidean distance, with the kernel value decreasing exponentially as the distance increases as shown in the following formulation.

$$K(x_i, x_j) = \exp\left(-\lambda \cdot \|x_i - x_j\|^2\right) \tag{1}$$

The parameter gamma (γ) plays a crucial role in controlling the sensitivity of the model to variations in the input data. In particular, a high γ value causes the model to focus intensely on individual data points, which can lead to overfitting, in which the model captures noise and specific patterns that do not generalize well to unseen data. Conversely, a low γ value results in a smoother decision boundary, making the model less flexible and potentially underfitting the data by failing to capture important complexities or patterns. To achieve optimal performance, both the regularization parameter C and γ are systematically tuned using a grid search. This method involves exhaustively searching through a predefined set of parameter values to identify the combination that yields the best balance between bias and variance, thereby maximizing the model's predictive accuracy. By carefully optimizing these parameters, the model can generalize better to new

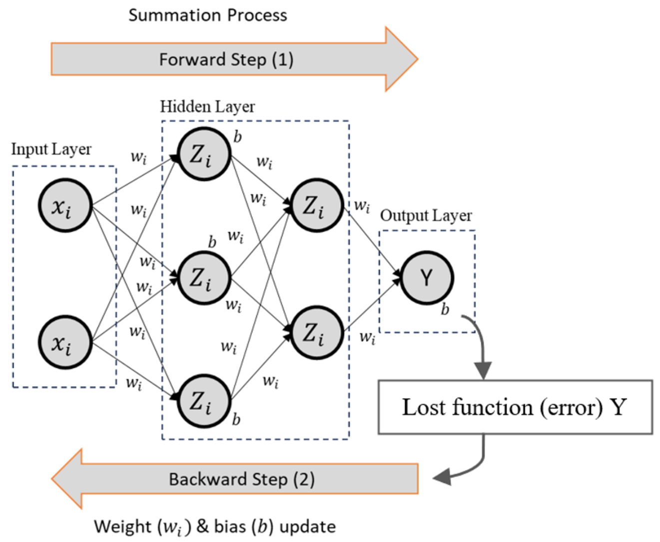


Figure 4. Backpropagation algorithmic scheme.

data, avoiding the pitfalls of both overfitting and underfitting.

2.6.2. Sigmoid Kernel

The sigmoid kernel is a favored option in machine learning, especially within support vector machines (SVMs), because of its capability to model complex nonlinear relationships between data points. This similarity enables the sigmoid kernel to transform input features into a higher-dimensional space where linear separation becomes feasible, thus allowing the algorithm to capture intricate patterns that linear kernels cannot. It is mathematically represented as a hyperbolic tangent function, which closely resembles the activation function used in artificial neural networks, as demonstrated in the following formulation.

$$K(x_i, x_j) = \tanh\left(\gamma \cdot (x_i, x_j) + c\right) \tag{2}$$

The kernel function described uses the parameter gamma to control the slope of the curve, thereby effectively determining the sharpness of the function transition. This slope adjustment enables the kernel to model complex nonlinear relationships within data, making it particularly useful for datasets that exhibit patterns similar to those found in artificial neural networks (ANNs). The parameter c serves as a horizontal shift, enabling fine-tuning of the decision boundary to better separate data points belonging to different classes. Together, these parameters provide flexibility in shaping the kernel to fit the underlying data distribution more accurately.

2.7. Backpropagation Neural Network (BPNN)

A backpropagation neural network (BPNN) model is utilized to predict the health status of machines [9] by examining operational parameters, such as engine speed, oil pressure, engine temperature, coolant pressure, coolant temperature, and fuel pressure. This network comprises three main layers: the input layer, hidden layer, and output layer. The input layer receives operational parameters as data, whereas the hidden layer processes this data to discern nonlinear relationships. The architecture of the neural network involves a variable number of hidden layers, typically ranging from one to three. Each hidden layer comprises one to three neurons, thereby allowing the model to capture different levels of complexity and abstraction in the input data. This flexible design enables the network to balance computational efficiency with the capacity to learn meaningful patterns relevant to the machine's operational status. Following the hidden layers, the output layer functions as the decision-making component and produces predictions about the condition of the machine. It classifies the machine as either healthy or unhealthy based on the processed information from the preceding layers. The backpropagation neural network scheme is shown in Figure 4.

In the BPNN modeling, various activation functions play a crucial role in determining each neuron's output based on its input. Activation functions play a crucial role in introducing nonlinearity into neural networks, thereby enabling them to model complex relationships within data that linear models cannot capture. By transforming the summed input signals of a neuron into an output signal, these functions determine whether a neuron should be activated, thereby influencing the network's ability to learn intricate patterns. For instance, the sigmoid function maps input values into a smooth range between 0 and 1, making it particularly suitable for binary classification tasks, wherein the outputs represent probabilities. However, its tendency to saturate at extreme values can slow down learning owing to vanishing gradients. To address such limitations, alternative activation functions like ReLU (Rectified Linear Unit) and Tansig have been developed. ReLU accelerates the training process by outputting zero for negative inputs and a linear relationship for positive inputs, which helps mitigate the vanishing gradient problem common in deep networks. This property allows gradients to propagate more effectively during backpropagation, thereby improving convergence speed and performance. Meanwhile, the Tansig function, which outputs values between -1 and 1, offers greater stability than the sigmoid function by centering the data around zero, which can facilitate faster learning and better gradient flow.

The Levenberg-Marquardt algorithm (trainlm) was employed for the training process. This optimization technique combines gradient descent with Newton's method and was chosen for its ability to enhance convergence speed and efficiency compared to traditional optimization methods when dealing with nonlinear models, such as BPNN. This approach enables the model to adapt more quickly to the data, thereby improving the accuracy of predicting machine health conditions.

Table 2. Model evaluation matrices.

	Actual (+)	Actual (–)	Precision
Predicted (+)	True + (TP)	False + (FP)	TP / (TP + FP)
Predicted (–)	False – (FN)	True – (TN)	TN / (TN + FN)
Recall	TP / (TP + FN)	TN / (TN + FP)	–
F-Measure	(2 × Precision × Sensitivity) / (Precision + Sensitivity)		
Accuracy	(TP + TN) / (TP + TN + FP + FN)		

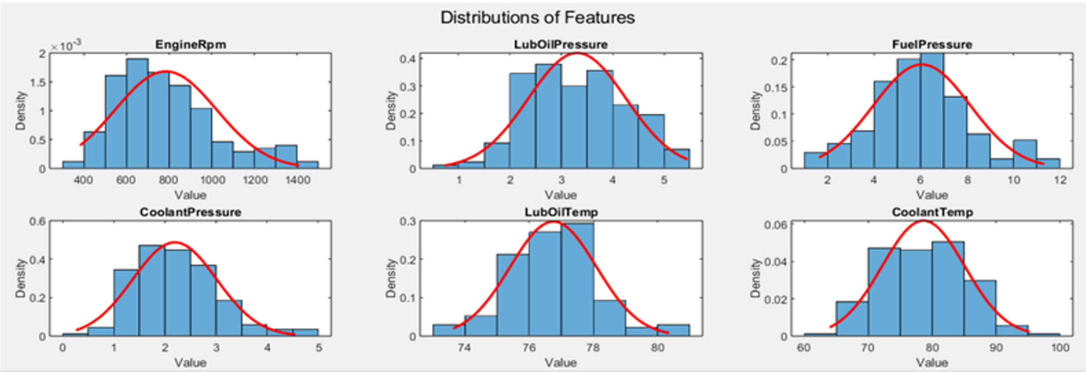
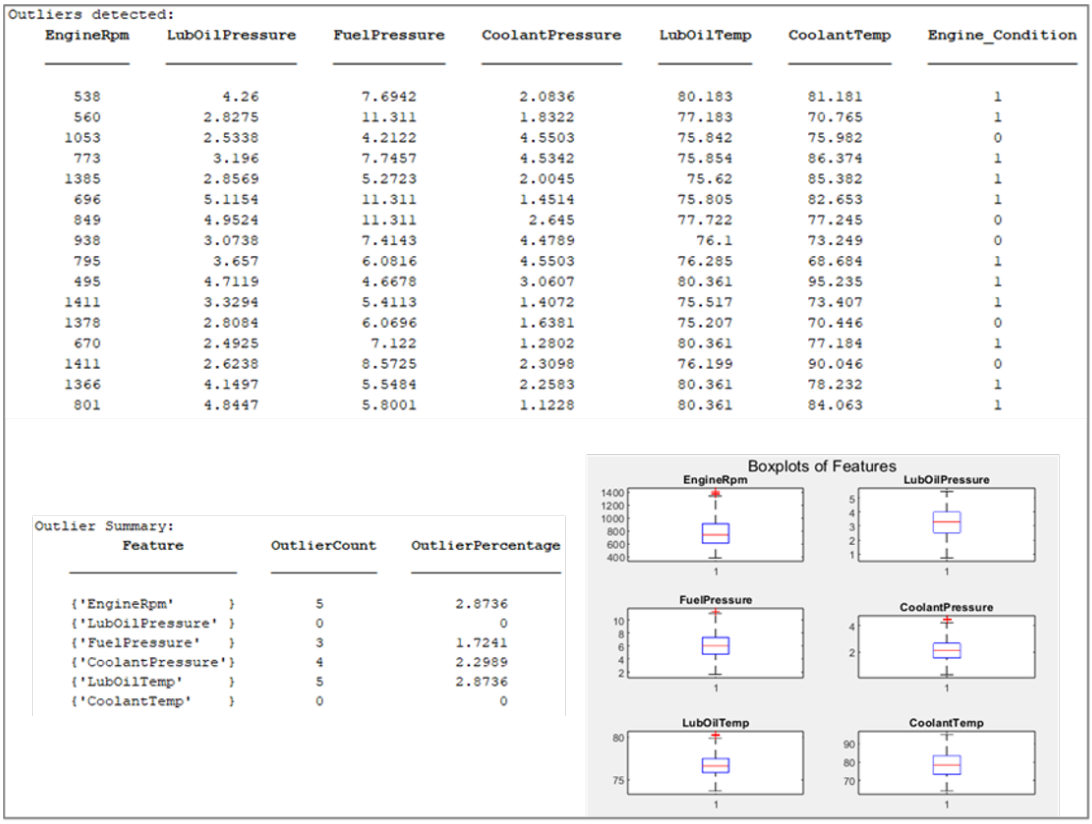


Figure 5. Engine parameter data distributed.



Outliers detected:

EngineRpm	LubOilPressure	FuelPressure	CoolantPressure	LubOilTemp	CoolantTemp	Engine_Condition
538	4.26	7.6942	2.0836	80.183	81.181	1
560	2.8275	11.311	1.8322	77.183	70.765	1
1053	2.5338	4.2122	4.5503	75.842	75.982	0
773	3.196	7.7457	4.5342	75.854	86.374	1
1385	2.8569	5.2723	2.0045	75.62	85.382	1
696	5.1154	11.311	1.4514	75.805	82.653	1
849	4.9524	11.311	2.645	77.722	77.245	0
938	3.0738	7.4143	4.4789	76.1	73.249	0
795	3.657	6.0816	4.5503	76.285	68.684	1
495	4.7119	4.6678	3.0607	80.361	95.235	1
1411	3.3294	5.4113	1.4072	75.517	73.407	1
1378	2.8084	6.0696	1.6381	75.207	70.446	0
670	2.4925	7.122	1.2802	80.361	77.184	1
1411	2.6238	8.5725	2.3098	76.199	90.046	0
1366	4.1497	5.5484	2.2583	80.361	78.232	1
801	4.8447	5.8001	1.1228	80.361	84.063	1

Outlier Summary:

Feature	OutlierCount	OutlierPercentage
{'EngineRpm' }	5	2.8736
{'LubOilPressure' }	0	0
{'FuelPressure' }	3	1.7241
{'CoolantPressure'}	4	2.2989
{'LubOilTemp' }	5	2.8736
{'CoolantTemp' }	0	0

Figure 6. Summary outlier detected and box plot data outlier.

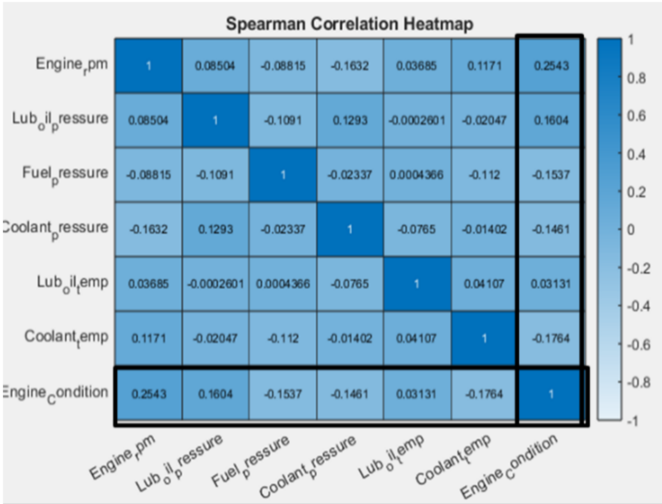


Figure 7. Heatmap correlation engine data parameter.

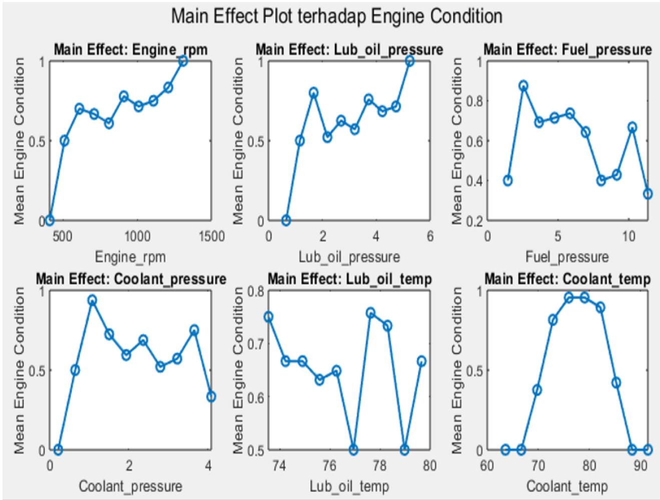


Figure 8. Main effect plot engine data parameter.

Table 3. ANOVA deviance model result.

Model	Deviance	df	p-value
Full Model	176.18	151	0.0085
Reduction Model	183.10	152	–

2.8. Model Evaluation

Model evaluation using metrics such as Accuracy, Precision, Recall, and F1-score, which are essential for assessing the model's ability to accurately classify machine conditions. Various equations can be formulated from the confusion matrix to further analyze the model performance. The model evaluation is presented in Table 2.

3. Results and Discussion

3.1. Data Distribution and Outlier Analysis

The initial stage of analysis focuses on understanding the distribution characteristics of the dataset. The histogram of feature distributions is presented in Figure 5, illustrating the spread of each operational parameter. The results show that several variables do not strictly follow a normal distribution, which justifies the use of non-parametric statistical methods such as Spearman correlation in subsequent analysis.

Outlier detection was performed using the Interquartile Range (IQR) method. The summary of detected outliers and the corresponding boxplot visualization are shown in Figure 6. The results reveal that a number of extreme values exist in parameters such as lubricating oil pressure and coolant temperature. These outliers were carefully handled to prevent distortion in model training while preserving meaningful variations in the dataset.

3.2. Correlation Analysis of Engine Parameters

The relationship between operational parameters and engine condition was analyzed using Spearman correlation. The heatmap of correlation is presented in Figure 7. The results show that engine RPM ($r = 0.2543$), lubricating oil pressure ($r = 0.1604$), and coolant temperature ($r = 0.1764$) exhibit the strongest correlation with engine condition. Although the correlation strength is relatively moderate, these parameters demonstrate consistent influence on engine health. This finding aligns with the physical behavior of diesel engines, where variations in speed, lubrication, and cooling significantly affect performance stability.

3.3. Main Effect Analysis of Engine Parameters

To further understand the influence of each parameter on engine condition, a main effect plot is presented in Figure 8. This analysis shows how changes in individual parameters impact the probability of engine health status.

The results show that engine RPM and coolant temperature exhibit more pronounced trends compared to other parameters. Increasing deviations from their optimal operating ranges correspond to a higher likelihood of unhealthy engine conditions. This supports the hypothesis that thermal and rotational dynamics play a critical role in engine performance.

3.4. ANOVA Deviance Analysis of Engine Parameters

The significance of operational parameters was evaluated using ANOVA Deviance based on logistic regression. The results are presented in Table 3 and supported by paramater analysis in Table 4. The analysis confirms that overall the p-value of the full model is $0.0023 < 0.05$, thus H_o is rejected, the full model is significant or there is at least one variable that has a significant impact on the health of the diesel engine. Engine RPM (p-value = 0.002) and coolant temperature (p-value = 0.01) are statistically significant predictors of engine condition ($p\text{-value} < 0.05$)

Table 4. Parameter results of the analysis of each engine parameter.

Parameter	Estimate	SE	t-stat	p-value
Intercept	5.3327	11.852	0.44993	0.65276
Engine_rpm	0.0029949	0.00097761	3.0635	0.0021874
Lub_oil_pressure	0.38583	0.20058	1.9235	0.054412
Fuel_pressure	-0.15916	0.093386	-1.7043	0.088328
Coolant_pressure	-0.40816	0.25783	-1.5831	0.11341
Lub_oil_temp	-0.0051003	0.15098	-0.033781	0.97305
Coolant_temp	-0.075417	0.029556	-2.5517	0.010721

```
Deviance Chi-Square = 176.182 | df = 151 | p-value = 0.0788
Pearson Chi-Square  = 155.82 | df = 151 | p-value = 0.3772
```

Figure 9. Goodness of fit results model.

The goodness-of-fit test results (p-value > 0.05), shown in Figure 9, indicate that the model adequately represents the observed data.

Additionally, Variance Inflation Factor (VIF) analysis, shown in Figure 10, demonstrates that multicollinearity among variables is within acceptable limits. These findings validate the relevance of selected parameters and provide a strong statistical foundation for subsequent machine learning modeling.

3.5. Prediction Results Using SVM

This study employs a Support Vector Machine (SVM) to predict the health condition of diesel engines using validated operational data. The dataset used in this study has undergone a validation process based on standardized operational parameter thresholds and actual engine operating conditions, ensuring that the labels accurately represent real engine behavior. The SVM model was selected due to its strong generalization capability in handling complex and nonlinear datasets. Two kernel types were implemented, namely Radial Basis Function (RBF) and Sigmoid, each evaluated using multiple parameter combinations to determine the optimal model performance. The prediction results are presented in Tables 5 and 6, which summarize the evaluation metrics, including average accuracy, mean squared error (MSE), precision, recall, and F1-score across the training, validation, and testing phases.

In the SVM model with the Sigmoid kernel, the parameters varied include the box constraint (C), constant (c), and gamma (γ), with combinations of C = 0.1, 1, 10, c = -1, 0, 1, and γ = 0.01, 0.1, 1. The optimal performance was achieved at C = 1, c = 0, and γ = 0.1, resulting in an average accuracy of 94.07%, MSE of 0.06, recall of 0.88, precision of 0.99, and an F1-score of 0.93. For the RBF kernel, the parameters varied include C = 0.1, 1, 10 and γ = 0.01, 0.1, 1. The best-performing model was obtained at C = 10 and γ = 0.1, achieving an average accuracy of 85.46%, MSE of 0.14, recall of 0.77, precision of 0.92, and an F1-score of 0.83. The results show that both kernels are capable of modeling nonlinear relationships in diesel engine operational data. However, the RBF kernel demonstrates more stable and consistent performance in capturing complex data patterns, making it more suitable for this dataset compared to the Sigmoid kernel. The improved performance of the RBF kernel can be attributed to its ability to map data into a higher-dimensional feature space, enabling better separation of classes in nonlinear conditions. In contrast, although the Sigmoid kernel achieves relatively high accuracy in certain configurations, its performance tends to be less stable and more sensitive to parameter selection. These findings suggest that the selection of kernel functions and their corresponding parameters plays a critical role in SVM performance. The RBF kernel, with its flexibility in handling nonlinear data distributions, provides a more reliable model for diesel engine health prediction within the context of this study.

```
=== VIF (Variance Inflation Factor) Tiap Variabel: ===
'Engine_rpm'          : 1.0533
'Lub_oil_pressure'    : 1.0458
'Fuel_pressure'       : 1.0180
'Coolant_pressure'    : 1.0595
'Lub_oil_temp'        : 1.0194
'Coolant_temp'        : 1.0171
```

Figure 10. Variance Inflation Factor results each engine paramater.

3.6. Prediction Results Using BPNN

The Backpropagation Neural Network (BPNN) model was implemented to predict diesel engine health using various network configurations. Several parameters were systematically varied, including the number of hidden layers (1-5), the number of neurons in each hidden layer (1-5 neuron), and activation functions such as logsig, tansig, purelin, satlin, and satlins. The training process utilized the Levenberg-Marquardt (trainlm) algorithm, which is well-known for its fast convergence in nonlinear optimization problems. The performance results of selected BPNN configurations are presented in Table 7, highlighting

Table 5. Evaluation metrics for SVM model prediction performance with sigmoid kernel.

Kernel	Parameter			Model Performance				
	C	c	γ	Avg Accuracy (%)	Avg MSE	Avg Recall	Avg Precision	Avg F1-Score
Sigmoid 1	0.1	-1	0.01	65.58	0.34	1.00	0.66	0.79
Sigmoid 2	1	0	0.01	65.58	0.34	1.00	0.66	0.79
Sigmoid 3	10	1	0.01	65.58	0.34	1.00	0.66	0.79
Sigmoid 4	0.1	-1	0.1	65.58	0.34	1.00	0.66	0.79
Sigmoid 5	1	0	0.1	65.58	0.34	1.00	0.66	0.79
Sigmoid 6	10	1	0.1	65.58	0.34	1.00	0.66	0.79
Sigmoid 7	0.1	-1	1	65.58	0.34	1.00	0.66	0.79
Sigmoid 8	1	0	1	65.28	0.35	0.97	0.66	0.79
Sigmoid 9	10	1	1	52.67	0.47	0.69	0.62	0.66

Table 6. Evaluation metrics for SVM model prediction performance with RBF kernel.

Kernel	Parameter			Model Performance				
	C	c	γ	Avg Accuracy (%)	Avg MSE	Avg Recall	Avg Precision	Avg F1-Score
RBF 1	0.1	–	0.01	65.58	0.34	1.00	0.66	0.79
RBF 2	0.1	–	0.1	65.58	0.34	1.00	0.66	0.79
RBF 3	0.1	–	1	65.58	0.34	1.00	0.66	0.79
RBF 4	1	–	0.01	77.29	0.23	1.00	0.77	0.86
RBF 5	1	–	0.1	77.29	0.23	1.00	0.77	0.86
RBF 6	1	–	1	77.39	0.23	0.93	0.78	0.85
RBF 7	10	–	0.01	77.29	0.23	1.00	0.77	0.86
RBF 8	10	–	0.1	77.29	0.23	1.00	0.77	0.86
RBF 9	10	–	1	85.46	0.15	0.94	0.86	0.90

the top-performing models based on evaluation metrics, including average accuracy, mean squared error (MSE), recall, precision, and F1-score across training, validation, and testing datasets. As shown in Table 7, the best-performing configuration is achieved by a network with three hidden layers consisting of [2-3-3-2] neurons, using the tansig activation function in all hidden and output layers. This model achieves an average accuracy of 97.16%, MSE of 0.03, recall of 0.98, precision of 0.98, and an F1-score of 0.98, indicating excellent classification performance and balanced prediction capability.

The best performance of this configuration demonstrates the effectiveness of deeper network architectures in capturing complex nonlinear relationships among engine operational parameters. The use of multiple hidden layers allows the model to learn hierarchical feature representations, which improves its ability to distinguish between healthy and unhealthy engine conditions. In addition, the tansig activation function contributes to stable and consistent performance by providing a bounded linear response, which helps reduce sensitivity to extreme values and improves generalization. Compared to other configurations, this combination of architecture and activation function offers a better balance between model complexity and predictive accuracy.

3.7. Comparison of SVM and BPNN Prediction Results with Related Studies.

A comparative analysis between the proposed models and related studies is presented in 8, which summarizes the performance of different machine learning approaches based on accuracy, precision, recall, and F1-score. The results show that the BPNN model developed in this study achieves the highest performance, with an accuracy of 97.16%, precision of 0.98, recall of 0.98, and F1-score of 0.98. This performance surpasses both SVM models implemented in this study as well as models reported in previous research.

In comparison, earlier studies such as those conducted by Mohakul et al. reported lower performance when using conventional SVM kernels. Although SVM demonstrates good generalization capability, its performance is limited in handling highly complex nonlinear relationships present in diesel engine operational data. The results of this study show that even with improved kernel selection, SVM still performs below the BPNN model. The superior performance of BPNN can be attributed to its a-

Table 7. Evaluation metrics of BPNN model prediction.

Configuration	Avg Accuracy (%)	Avg MSE	Avg Recall	Avg Precision	Avg F1-Score
Layer: 4 Neurons: [2 3 3 2] Activations: tansig, tansig, tansig, tansig OutputFcn: tansig	97.16	0.03	0.98	0.98	0.98
Layer: 5 Neurons: [2 1 1 3 1] Activations: tansig, tansig, tansig, tansig, tansig OutputFcn: tansig	96.86	0.03	0.97	0.98	0.98
Layer: 5 Neurons: [3 2 1 2 1] Activations: satlins, satlins, satlins, satlins, satlins OutputFcn: satlins	95.96	0.04	0.96	0.98	0.97

Table 8. Comparison of SVM and BPNN prediction results with related studies

Model	Accuracy	Precision	Recall	F1-Score	Source
SVM (Sigmoid)	65	66	100	79	Proposed Method
SVM (RBF)	85	86	94	90	
BPNN [2-3-3-2] (act. function: tansig)	97	98	98	98	
Logistic Regression	89	88	87	88	D. Mohakul, 2023
SVM (Linear)	89	88	87	88	
SVM (Polynomial)	88	87	84	86	
KNN	89	88	87	87	
Naïve Bayes	83	82	79	80	
Decision Tree	74	72	67	68	

bility to model intricate nonlinear interactions among multiple input parameters through deep network architectures. The use of multiple hidden layers combined with the tansig activation function enables the model to capture complex patterns more effectively and maintain stable prediction performance across different data subsets. In addition, the improved results in this study are also influenced by the data validation process applied prior to modeling. Unlike previous studies that rely on raw or directly labeled datasets, this study ensures label reliability through validation based on standardized operational thresholds and actual engine conditions. This approach enhances the quality of the training data and contributes to improved prediction accuracy.

4. Conclusions

This study evaluates the performance of Support Vector Machine (SVM) and Backpropagation Neural Network (BPNN) in predicting diesel engine health conditions using validated operational data. The results show that the SVM model with the RBF kernel achieved an accuracy of 85.46%, demonstrating its capability to handle nonlinear data, although its performance remains limited in capturing more complex patterns. In contrast, the BPNN model achieved superior performance, with the best configuration consisting of three hidden layers [2-3-3-2] and the tansig activation function. This model achieved an accuracy of 97.16%, along with high precision, recall, and F1-score values, indicating strong and consistent classification performance. These results confirm that BPNN is more effective in modeling complex nonlinear relationships in diesel engine operational data. Furthermore, the improved prediction performance in this study is strongly influenced by the data validation approach applied prior to modeling. By validating engine condition labels based on standardized operational parameter thresholds and actual engine conditions, the reliability of the dataset is significantly enhanced. This contributes to more accurate and robust prediction results compared to approaches that rely solely on raw labeled data. Overall, This study shows that integrating domain-based data validation with machine learning techniques provides a reliable framework for diesel engine health prediction within a Condition-Based Maintenance (CBM) system.

For future work, further improvements can be achieved by exploring other machine learning approaches such as Gradient Boosting, XGBoost, or advanced neural network architectures. Additionally, ensemble learning methods may be considered to combine the strengths of multiple models. Expanding the dataset with more diverse operational conditions and different types of diesel engines is also recommended to improve model generalization and applicability in real-world scenarios.

Acknowledgments

The author would like to express sincere gratitude to Institut Teknologi Sepuluh Nopember (ITS), Surabaya, particularly the Department of Mechanical Engineering, for providing academic support and research facilities throughout the completion of this study. The author is also deeply grateful to Mohammad Khoirul Effendi, Ph.D, for his continuous guidance, valuable insights, and constructive feedback during the research process. Appreciation is extended to colleagues and laboratory members for their assistance, technical support, and meaningful discussions that contributed to this work. In addition, the author would like to acknowledge all parties who contributed to the availability of the dataset and supporting references used in this research. Finally, the author sincerely thanks family and friends for their encouragement, understanding, and unwavering support throughout this study.

References

[1] K. S. H. Ong, D. Niyato, and C. Yuen, "Predictive maintenance for edge-based sensor networks: A deep reinforcement learning approach." arXiv preprint arXiv:2007.03313, July 2020. Accessed: Mar. 30, 2026. [Online]. Available: https://arxiv.org/abs/2007.03313v1.

[2] J. Zhu, Z. Zhu, B. Li, and X. Zhao, "Predictive maintenance strategy based on critical probabilistic cost," *Reliability Engineering & System Safety*, vol. 267, p. 111823, Mar. 2026.

[3] R. K. Ilmi, Y. Sukrawan, and T. Permana, "Condition based monitoring improves component durability in heavy equipment maintenance," *MOTOR: Journal of Automotive Engineering*, vol. 1, no. 2, 2024. [Online; accessed Mar. 30, 2026].

[4] S. B. Alsakinah, "Pemanfaatan teknologi big data pada Crane Health Management System (CHMS) dengan pendekatan integrative framework," *Proceeding of National Seminar on Maritime and Interdisciplinary Studies*, vol. 3, pp. 112–120, Dec. 2024.

[5] A. F. Yusri, "Optimalisasi perawatan pressure vacum valve dalam sistem perawatan kapal di MT. arzoyi," diploma thesis, Politeknik Ilmu Pelayaran Makassar, 2024. Accessed: Mar. 30, 2026. [Online]. Available: http://eprints.pipmakassar.ac.id/767/.

[6] M. O. Kuttan, J. Steinheimer, K. Zhou, A. Redelbach, and H. Stoecker, "Deep learning based impact parameter determination for the CBM experiment," *Particles*, vol. 4, no. 1, pp. 47–52, 2021.

[7] J. Sharma, M. L. Mittal, and G. Soni, "Condition-based maintenance using machine learning and role of interpretability: a review," *International Journal of System Assurance Engineering and Management*, vol. 15, pp. 1345–1360, Apr. 2024.

[8] S. C. R. H. Haliza and A. Qoiriah, "Predictive maintenance untuk kendaraan bermotor dengan menggunakan support vector machine (SVM)," *JINACS*, vol. 2, pp. 159–168, Jan. 2021.

[9] V. Mandala, T. Senthilnathan, S. Suganyadevi, S. Gobhinath, D. Selvaraj, and R. Dhanapal, "An optimized back propagation neural network for automated evaluation of health condition using sensor data," *Measurement: Sensors*, vol. 29, p. 100846, Oct. 2023.

[10] D. Mohakul, C. R. S. Kumar, S. Singh, and S. Katti, "Condition based predictive maintenance on ship's major equipment using AI," *International Journal of Scientific Research in Engineering and Management*, vol. 7, Feb. 2023. [Online]. Available: https://ijsrem.com/download/condition-based-predictive-maintenance-on-ships-major-equipment-using-ai/.

[11] D. Mohakul, C. Kumar, S. Singh, S. Katti, and S. Chougule, "Health monitoring of ship's engine with simulated data by using classifiers–preliminary result," in *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, pp. 1–5, Apr. 2023.

[12] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Analytics*, vol. 1, p. 9, Nov. 2016.

[13] C. Fu, X. Liang, Q. Li, K. Lu, F. Gu, A. D. Ball, and Z. Zheng, "Comparative study on health monitoring of a marine engine using multivariate physics-based models and unsupervised data-driven models," *Machines*, vol. 11, p. 557, May 2023.

[14] S. Zhao, Y. Guo, Q. Sheng, and Y. Shyr, "Advanced heat map and clustering analysis using heatmap3," *BioMed Research International*, vol. 2014, p. 986048, 2014.

[15] T. Wahyuni, A. Agoestanto, and E. Pujiastuti, "Analisis regresi logistik terhadap keputusan penerimaan beasiswa PPA di FMIPA unnes menggunakan software minitab," *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 1, pp. 755–764, Feb. 2018.