Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Feature Engineering for Predictive Maintenance: Identifying Key Predictors of Machine Defects Using Machine Learning

Chinedu Sebastian Ani[1], Godwin Harold Chukwuemeka[2], Uchendu Onwusoronye Onwurah*[3]
Department of Industrial and production Engineering, Nnamdi Azikiwe University, Awka, Anambra State Nigeria

Article Info
Keywords:
Preventive Maintenance, Feature Engineering, Machine learning, Machine defects, Vibration signal, ANOVA, neural network.

ABSTRACT
In the modern industrial environments, the ability to predict equipment failure before it occurs is essential for minimizing downtime and maximizing operational efficiency. This research explores the use of feature engineering to identify key indicators of mechanical faults in a cement mill fan system. Vibration data were collected over 34 weeks from critical components of the fan and processed using several statistical techniques to extract relevant features. Various feature selection methods including Principal Component Analysis (PCA), Minimum Redundancy Maximum Relevance (mRMR), ReliefF, Chi-square, ANOVA, and Kruskal-Wallis were used to determine the most informative features. These features were then used to train and evaluate machine learning models, with neural networks demonstrating superior performance. Among all models, the neural network optimized with Chi-square-selected features achieved the highest classification accuracy, fastest prediction speed, and lowest misclassification cost. These results highlight the effectiveness of combining robust feature selection with deep learning methods for reliable fault detection and predictive maintenance in industrial systems.

Corresponding Author:
Uchendu Onwusoronye Onwurah
Department of Industrial and production Engineering, Nnamdi Azikiwe University, Awka, Anambra State Nigeria
ou.onwurah@unizik.edu.ng

INTRODUCTION

Predictive maintenance is essential in reducing equipment downtime and ensuring operational efficiency in industrial systems. The application of machine learning, particularly Gaussian process models, has become instrumental in forecasting equipment degradation by analysing large volumes of sensor data. This enables real-time condition monitoring and supports a shift from traditional reactive maintenance strategies to data-driven approaches. Bezerra et al. (2024) emphasized the benefits of integrating predictive maintenance into industrial processes, and also drew attention to the difficulties associated with managing complex datasets and ensuring fast data processing. Within this context, feature selection becomes a key step in model development. It involves identifying the most relevant input variables to reduce data dimensionality and enhance computational efficiency (Debal & Sitote, 2022). A feature refers to a measurable characteristic of the system being analysed, and selecting the right features improves model accuracy and reduces training time. Bharti et al. (2021) further described feature selection as a method for discarding less useful data, while

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Ileberi et al. (2022) highlighted its role in solving optimization problems with reduced computational costs.

Sharma and Mishra (2022) demonstrated the importance of feature selection techniques such as Correlation based feature selection (CFS), sequential feature selection (SFS) and information gain (IG) in achieving higher accuracy comparatively with other techniques. They further emphasized that highly correlated feature gives redundant information that does not contribute to performance improvement. Bezerra et al. (2024) demonstrated that applying techniques like Principal Component Analysis (PCA), Minimum Redundancy Maximum Relevance (mRMR), and Denoising Autoencoders (DAE) significantly improves the accuracy, precision, and recall of ML models in classifying machine failures, with Random Forest achieving up to 98% accuracy. Similarly, Ahmad et al. (2022) emphasized that integrating sequential feature selection enhances diagnostic precision in health applications, while Buchaiah and Shakya (2022) successfully used hybrid feature selection methods for bearing fault diagnosis. A comprehensive review by Theng and Bhoyar (2024) highlights the transformative impact of feature selection across domains, improving model generalization and reducing overfitting. Ghosh et al. (2021) further validated the effectiveness of Relief and LASSO methods in cardiovascular risk prediction, reinforcing the value of statistical feature selection. Moreover, Khattach et al. (2024) noted that identifying key features also guides sensor deployment, reducing redundancy and data collection costs. These studies collectively underline the importance of aligning data preprocessing techniques with predictive maintenance goals in industrial applications.

Feature selection techniques in machine learning are generally classified into three main categories: filter, wrapper, and embedded methods. Among these, filter and wrapper approaches are most frequently employed to identify the most relevant features. Filter techniques evaluate feature relevance based on statistical measures—such as correlation, distance metrics, and information gain—independently of any predictive model. They work directly with the training dataset and do not involve model training, which makes them computationally efficient. On the other hand, wrapper methods incorporate a learning algorithm to assess feature subsets based on model performance. This allows them to identify feature combinations that are specifically optimized for a particular algorithm. While wrappers often yield more accurate results, they require significantly more computational resources. A key limitation is that the optimal feature subset found may not perform as well with other algorithms. Embedded methods offer a hybrid approach, integrating feature selection within the model training process itself, effectively combining the strengths of both filter and wrapper strategies (Suruliandi et al., 2021).

Despite these advances, there remains a need to systematically identify and validate the most informative features for industrial machine defect prediction. This study addresses that gap by exploring feature engineering techniques to isolate key predictors that improve failure detection accuracy and optimize data-driven maintenance strategies.

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

MATERIAL AND METHODS

Figure 1 shows the steps taken in achieving the purpose of this study. The process involves data collection, feature selection, machine learning models generation, models evaluation, and comparison of the models and selection of the best model.

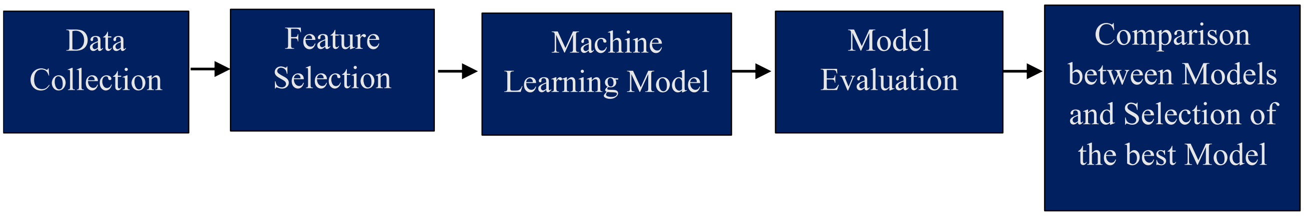

Figure 1. Process Steps

Data collection and Description

The data used in this study were collected from induced fan line 3 at Dangote Cement PLC, located in Obajana, Kogi State. The fan operates at a rated power of 5000 kW and plays a critical role in cement milling operations. Data collection was carried out using Pfrutechnik's VibXpert vibration data analyser in conjunction with a high-frequency accelerometer (VIB 6.140) and appropriate cabling. Sensor readings were obtained from both the drive-end and non-drive-end bearings of the cement mill fan, as well as from the fan's rotational speed (RPM). Manual techniques comprise collecting data from bearing housing three axes: namely, horizontal, vertical and axial direction. Measurements were taken over a period of 34 weeks to capture a wide range of operational conditions in 2024.

In total, more than seven thousand data points were gathered and subsequently used for exploring feature engineering techniques to isolate key predictors that improve failure detection accuracy and optimize data-driven maintenance strategies in MATLAB. Table 1 shows the descriptive statistics of the vibration data collected. A closer look at the sample data in Table 1 shows that sample time 1 recorded the highest mean vibration level, pointing to a period of elevated vibration in the fan system. From sample time 5 onward, the minimum values drop significantly, which could indicate the presence of sudden impacts or irregular behaviour. The variability in the data becomes most pronounced at sample time 9, where the standard deviation hits its highest point. This level of fluctuation suggests that the system may have been operating under unstable conditions. Between samples time 7 and 10, both the wide range of values and the sharp negative shifts in the signal point toward possible mechanical issues or early signs of system failure.

Table 1. Descriptive Statistics of the Vibration Data

Sample time (microsecond)	Mean	Minimum	Maximum	Standard deviation
0	0.7395825	-2.726749564	9.13700166	2.186673734
1	2.419397836	-2.713179879	9.091520978	2.45979591
2	2.24436162	-2.245224048	6.571475824	2.161266262
3	1.084355854	-2.234058204	5.383715608	1.228454925

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

4	1.156204408	-1.768482176	3.657721331	1.229654094
5	0.277710863	-5.649532405	3.639524251	1.532745775
6	1.632737768	-5.621407459	4.542298957	2.072760548
7	1.423688884	-11.09478636	4.519690828	2.719978701
8	1.299932285	-11.03957012	4.497214909	2.588471569
9	0.978850055	-14.64422665	4.450614518	3.078500988
10	0.336813726	-14.57139025	3.536279257	2.753700919

Methods of data analysis

Data cleaning

Data cleaning was done in order to identify and handle missing values, eliminate outliers, and reduce noise within the signal. The cleaning procedures help improve the reliability and consistency of the data, thereby enhancing the effectiveness of feature extraction.

Feature selection

In order to effectively assess the condition of the fan system, key statistical features were extracted from the time-domain vibration signals. These features helped to reduce the dimensionality of the data while retaining essential diagnostic information. By focusing on these condensed metrics, the analysis becomes more efficient without sacrificing accuracy. In this study, a filter-based feature selection method was employed for its effectiveness and robustness in identifying features that significantly contributed to the model performance.
The following statistical features from the time-domain vibration signals – peak value, impulse factor, crest factor, clearance factor, signal-to-noise ratio, total harmonic distortion, signal to noise and distortion ratio, shape factor, and kurtosis, were examined in this study.

1. Peak value
   The peak value represents the maximum absolute amplitude observed in the time-domain signal. It is essential for identifying extreme vibrations and serves as the basis for computing several impulse-related features. The peak is calculated using equation (1).
   $$Peak = max(|x(t)|) \tag{1}$$
   Where x = amplitude, and t = time.
2. Impulse Factor
   The impulse factor quantifies the ratio between the highest peak and the average signal level, providing insight into sudden impulses. The impulse factor is calculated using equation (2).
   $$Impulse\ Factor = \frac{max(|x(t)|)}{mean(|x(t)|)} \tag{2}$$
3. Crest Factor
   Crest factor is the ratio of the peak value to the root mean square (RMS) of the signal as shown in equation (3). It is highly sensitive to fault onset, as incipient faults typically cause spikes in peak values before affecting overall energy.

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

$$Crest\ Factor = \frac{max(|x(t)|)}{\sqrt{\frac{1}{N}\sum_{i=1} N\, x_i^2}} \tag{3}$$

Where N is the number of samples.

4. Clearance Factor

   Clearance factor measures the sharpness of the signal peaks by comparing the peak amplitude to the square of the mean of the square roots of the absolute amplitudes as shown in equation (4). It is especially useful for diagnosing bearing faults.

$$\text{Clearance } Factor\ (CF) = \frac{max(|x(t)|)}{(\sqrt{\frac{1}{N}\sum_{i=1} N \sqrt{|x_i|}})^2} \tag{4}$$

5. Signal-to-Noise Ratio (SNR)

   SNR evaluates the proportion of the true signal's power to the background noise power. A higher SNR indicates a clearer and more reliable signal. The SNR is calculated using equation (5).

$$SNR = 10 * log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) \tag{5}$$

Where: P_{signal} is the signal power, and P_{noise} is the noise power.

6. Total Harmonic Distortion (THD)

   THD assesses the degree of harmonic distortion in the signal by comparing the power of all the harmonic components to that of the fundamental frequency.

$$THD = \frac{\sqrt{\sum_{n=2}^{\infty} V_n^2}}{V_1} \tag{6}$$

Where: V_1 is the root mean square of the fundamental frequency, and V_n is the root mean square of the harmonic frequency.

7. Signal to Noise and Distortion Ratio (SINAD)

   SINAD measures the ratio of total signal power to the combined power of noise and distortion. It reflects overall signal quality.

$$SINAD\ (DB) = 10 * log_{10}\left(\frac{P_{signal}}{P_{noise} + P_{distortion}}\right) \tag{7}$$

8. Shape Factor

   Shape factor is the RMS of the signal divided by the mean of its absolute values. It reflects waveform structure while being dimension independent.

$$Shape\ Factor = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_1^2}}{\frac{1}{N}\sqrt{\sum_{i=1}|x_i|}} \tag{8}$$

9. Kurtosis

   Kurtosis quantifies the "tailed ness" of the signal distribution, indicating the presence of extreme values or outliers. It is especially useful for detecting developing faults, as these often increase outlier frequency.

$$Kurtosis = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2)^2} \quad (9)$$

Where μ is the mean of the signal. A Gaussian distribution yields a kurtosis of 3.

Feature Selection Techniques and Comparison

In order to determine the important predictors (features) to include in the model, feature selection techniques were used to compare, rank, and select the best features related to machine failures. The techniques include: principal component analysis (PCA), minimum redundancy maximum relevance (mRMR), ReliefF algorithm, chi-square technique, ANOVA, and Kruskal Wallis.

1. Chi-Square Technique

   Chi-square feature selection has gained prominence as an effective technique for identifying the most relevant features in diverse machine learning applications, particularly in classification tasks. Chi-square (χ^2) is calculated using equation (10).

   $$\chi^2 = \sum \frac{(O-E)^2}{E} \quad (10)$$

   Where: O is the observed frequency, and E is the expected frequency.

2. Minimum Redundancy Maximum Relevance Algorithm

   The mRMR algorithm, introduced by Xie et al., (2023), is designed to select a subset of features that are both highly relevant to the target variable and minimally redundant with one another. Relevance is quantified through mutual information between each feature and the target variable, while redundancy is measured via mutual information among features themselves. The objective is to identify a feature subset S that maximizes the overall relevance V_s to the response variable y while minimizing the internal redundancy W_s .These metrics are computed using mutual information as shown below (Xie et al., 2023).

   $$V_s = \frac{1}{|S|}\sum_{x\in S} I(x,y), \quad (11)$$

   $$W_s = \frac{1}{|S|2}\sum_{x,z\in S} I(x,z), \quad (12)$$

   $|S|$ represents the number of features in S

   Finding an optimal set S requires considering all 2|Ω| combinations, where Ω is the entire feature set. Instead, the mRMR algorithm ranks features through the forward addition scheme, which requires O(|Ω|·|S|) computations, by using the mutual information quotient (MIQ) value.

   $$\mathrm{MIQx} = \mathrm{Vx}\frac{V_x}{W_x} \quad (13)$$

   Where V_x and W_x are the relevance and redundancy of a feature, respectively:

   $$V_x = I(x,y) \quad (14)$$

   $$W_x = \frac{1}{|S|}\sum_{z\in S} I(x,z) \quad (15)$$

   The function ranks all features in Ω and returns idx (the indices of features ordered by feature importance) using the mRMR algorithm. Therefore, the computation cost becomes O(|Ω|2). The function quantifies the importance of a feature using a heuristic

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

algorithm and returns a score (scores). A large score value indicates that the corresponding predictor is important. Also, a drop in the feature importance score represents the confidence of feature selection.

3. ReliefF Algorithm
   ReliefF finds the weights of predictors in the case where y is a multiclass categorical variable. The algorithm penalizes the predictors that give different values to neighbours of the same class, and rewards predictors that give different values to neighbours of different classes (Yang et al., 2025).
   ReliefF first sets all predictor weights W_j to 0. Then, the algorithm iteratively selects a random observation x_f, finds the k-nearest observations to x_r for each class, and updates, for each nearest neighbour x_q, all the weights for the predictors F_j as follows:
   If x_r and x_q are in the same class,
$$W_j^i = W_j^{i-0} - \frac{\Delta_j(x_r, x_q)}{m} . d_{rq} \tag{16}$$
   If x_r and x_q are in the different class,
$$W_j^i = W_j^{i-1} + \frac{Py_q}{1 - Py_r} \frac{\Delta_j(x_r, x_q)}{m} . d_{rq} \tag{17}$$
   Where: W_j^i is the weight of the predictor F_j at the ith iteration step.
   Py_r is the prior probability of the class to which x_r belongs, and Py_q is the prior probability of the class to which x_q belongs.
   m is the number of iterations specified by 'updates'.
   $\Delta_j(x_r, x_q)$ is the difference in the value of the predictor F_j between observations x_r and x_q. Let x_{rj} denote the value of the jth predictor for observation x_r and let x_{qj} denote the value of the jth predictor for observation x_q.
   For discrete F_j,
$$\Delta_j(x_r, x_q) = \begin{cases} 0, x_{rj} = x_{qj} \\ \quad x_{rj} \neq x_{qj} \end{cases} \tag{18}$$
$$\Delta_j(x_r, x_q) = \frac{|x_{rj} - x_{qj}|}{\max(F_j) - \min(F_j)} \tag{19}$$
   d_{rq} is a distance function of the form.
$$d_{rq} = \frac{d_{rq}}{\sum_{i=1}^{k} d_{ri}} \tag{20}$$
   The distance is subject to the scaling,
$$d_{rq} = e^{-(\frac{rank(r,q)}{sigma})^2} \tag{21}$$
   Where: rank (r, q) is the position of the qth observation among the nearest neighbours of the rth observation, sorted by distance; k is the number of nearest neighbours, specified by k. The scaling can be changed by specifying 'sigma'.
4. Principal Component Analysis (PCA)
   Principal Component Analysis (PCA) is a widely adopted multivariate statistical technique used to reduce high-dimensional data into a smaller set of uncorrelated variables, known as principal components, while retaining as much of the original variance as possible. As Greenacre et al. (2022) described, PCA constructs linear

combinations of the original features such that the first component captures the maximum possible variance, with subsequent components accounting for the largest share of the remaining variability under the constraint of orthogonality. This transformation not only reduces dimensionality but also enhances the interpretability of complex datasets by eliminating redundancy. In this vibration signal analysis for rotating machinery, PCA assisted in isolating the most informative features while mitigating the influence of noise and irrelevant variables. PCA serves as a crucial feature selection tool in industrial applications, supporting accurate and timely fault detection when integrated with machine learning algorithms. In this work, PCA was applied to the data as indicated by Bezerra et al., (2024) to verify variable's percentage of total variance using the main components for optimize neural network.

5. Analysis of Variance (ANOVA)
   Analysis of Variance (ANOVA) is a fundamental statistical technique employed to evaluate whether significant differences exist between the means of multiple groups under varying experimental conditions. In this predictive maintenance research, ANOVA was particularly valuable for identifying the influence of operational factors and fault conditions on system performance indicators. The method not only provided a parsimonious representation of the data but also validated the stability of sensor responses under steady-state operations, ensuring that the derived models met the assumptions of independence and homoscedasticity. Importantly, ANOVA served as the foundation for subsequent model selection and machine learning integration, thereby enhancing both the interpretability and predictive accuracy of the maintenance framework.
6. Kruskal-Walli's test
   The Kruskal-Walli's test is a nonparametric version of classical one-way ANOVA, and an extension of the Wilcoxon rank sum test to more than two groups. The Kruskal-Wallis test is valid for data that has two or more groups. It compares the medians of the groups of data in x to determine if the samples come from the same population (or, equivalently, from different populations with the same distribution). The Kruskal-Wallis test uses ranks of the data, rather than numeric values, to compute the test statistics. It finds ranks by ordering the data from smallest to largest across all groups and taking the numeric index of this ordering. The rank for a tied observation is equal to the average rank of all observations tied with it.

Fault Classification Using Machine Learning Techniques

The classification of the features was done using machine learning techniques – neural networks and decision tree. The classification was done with the help of MATLAB's Classification Learner Application, which allows training and evaluating multiple models in a single session.

1. Neural Networks
   In this study, neural network models were employed and optimized to enhance the classification and feature selection process. A dense architecture comprising 39 input nodes was implemented, utilizing the Tanh activation function in the encoding layer,

with a suitable activation function applied during decoding. Bayesian optimization was conducted over 1,000 epochs to fine-tune hyperparameters and improve model performance. These methods were applied to extract relevant features from the dataset and to reduce its dimensionality, thereby enhancing computational efficiency and interpretability. The significance of individual features was also examined based on their influence on the model's output. Following feature extraction, a comparative analysis was conducted to evaluate the most impactful features identified by each selection technique. Notably, the Chi-square method, when integrated within a neural network framework, facilitates feature selection by assessing the contribution of each input variable to the model's predictive capability. The model training was conducted with 90% of the data for training and 10% for testing. The accuracy, precision, and recall metrics, as indicated in Mahmood (2021), were employed to evaluate the model.

2. Decision Tree
   In this study, decision tree was employed in the classification of the features. Budcci et al. (2018) described the Decision Tree as a predictive model that works by testing conditions at each tree level, and moving down the tree where different decisions are identified.

Validation and Performance Metrics

The models were assessed based on multiple performance metrics, including validation accuracy, test accuracy, total misclassification cost, prediction speed, and training time. The model training was conducted with 90% of the data for training and 10% for testing.

RESULTS AND DISCUSSIONS

This study evaluated multiple feature selection approaches namely PCA, mRMR, ReliefF, Chi-square, ANOVA, and Kruskal-Wallis, to pinpoint the most influential predictors for classifying machine defects. The classification methods (decision tree and neural networks) were used with the selection techniques in order to select the most suitable feature technique and machine learning technique for the vibration signals prediction in the company.

Analysis of the Time-Domain Vibration Signals Features

The statistical features from the time-domain vibration signals – peak value, impulse factor, crest factor, clearance factor, signal-to-noise ratio, total harmonic distortion, signal to noise and distortion ratio, shape factor, and kurtosis, were analysed to examine various machine conditions. The machine fault conditions examined were misalignment, normal vibration, unbalance fault, and bearing damage.

Peak Value

Figure 2 shows the peak value analysis of the time-domain vibration signals. From the figure, the peak value plots show a clear variation among different machine fault conditions. Low peak values (0–10 mm/s) are common in both normal operation and misalignment. In contrast, bearing faults are mostly found in the 10–20 mm/s range, indicating higher intensity vibrations. Unbalance issues appear across a broader span, including higher peak ranges, which may reflect differences in fault severity. While peak value helps in identifying bearing

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

damage and some unbalance cases, distinguishing misalignment from normal states may require additional signal features.

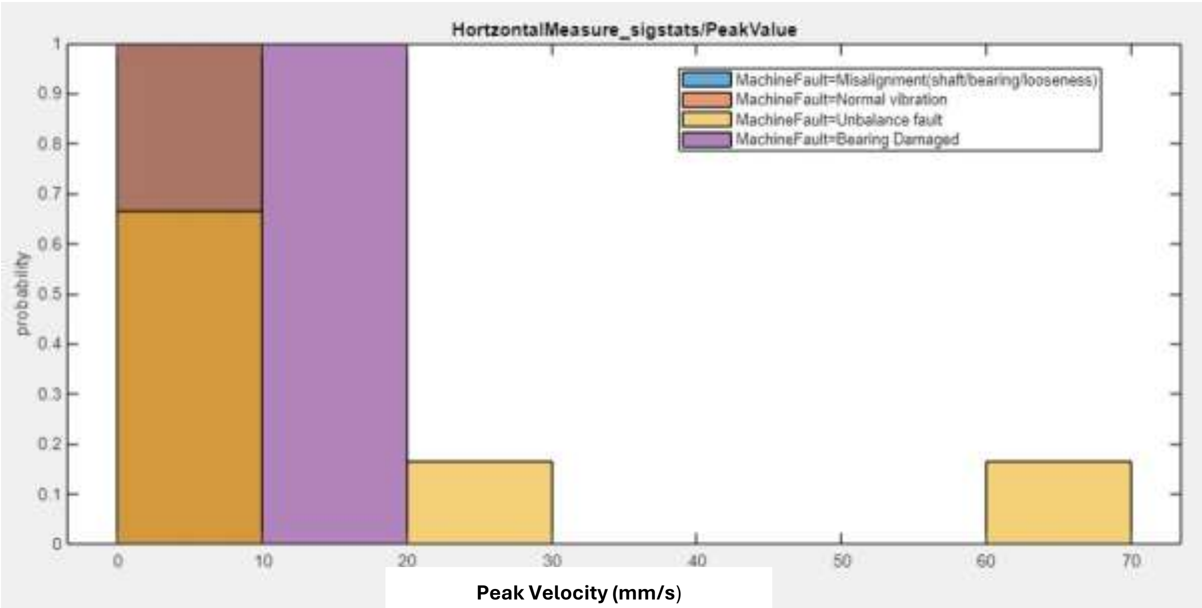

Figure 2. Bar Chart for Peak Value for the Probe Horizontal Measure

Impulse Factor

Figure 3 illustrates the probability distribution of the impulse factor extracted from the horizontal vibration signals under various machine fault conditions. The results show that most fault types, particularly bearing damage and unbalance fault, are concentrated within a low impulse factor range (approximately 3–6), indicating moderate transient behaviour. In contrast, normal vibration appears across both low and high impulse factor values, suggesting variability that may include occasional spikes or noise. The absence of data for misalignment suggests it either did not occur in the measured samples or its impulse factor was negligible. Overall, the impulse factor effectively distinguishes between fault conditions, especially in detecting bearing-related issues.

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

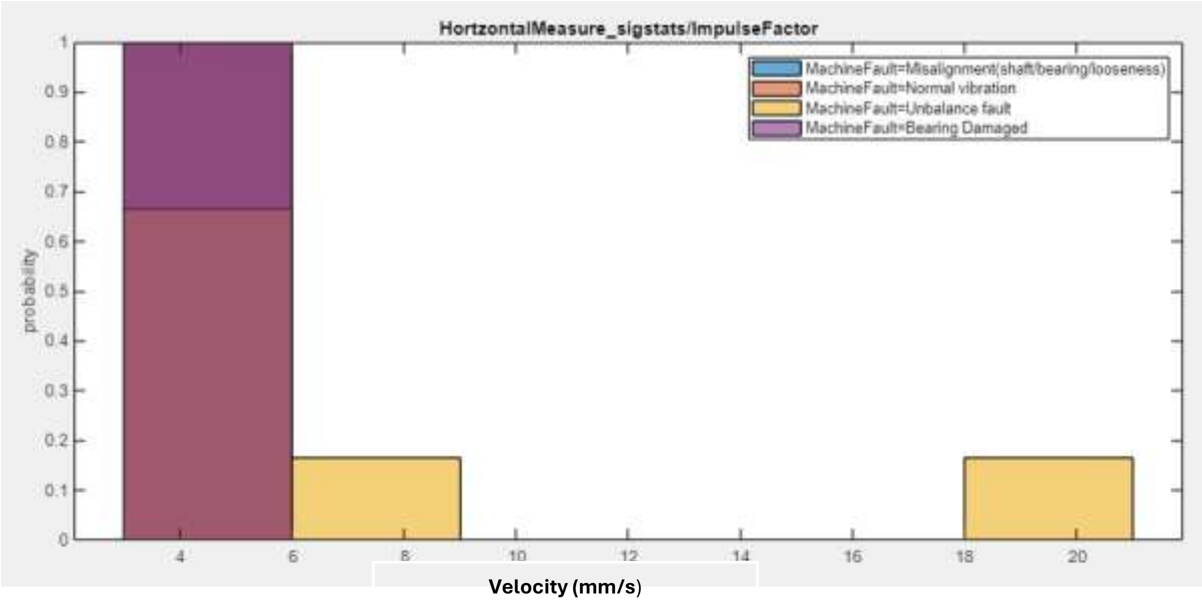

Figure 3. Bar Chart for Impulse Factor in Probe Horizontal Measure

Crest Factor

Figure 4 illustrates the probability distribution of crest factor values corresponding to various machine conditions. Notably, conditions such as unbalance and bearing damage exhibit higher crest factor probabilities, signifying sudden increases in peak amplitudes compared to normal vibration. This demonstrates the crest factor's usefulness in identifying incipient mechanical issues like shaft misalignment, bearing looseness, and early bearing faults.

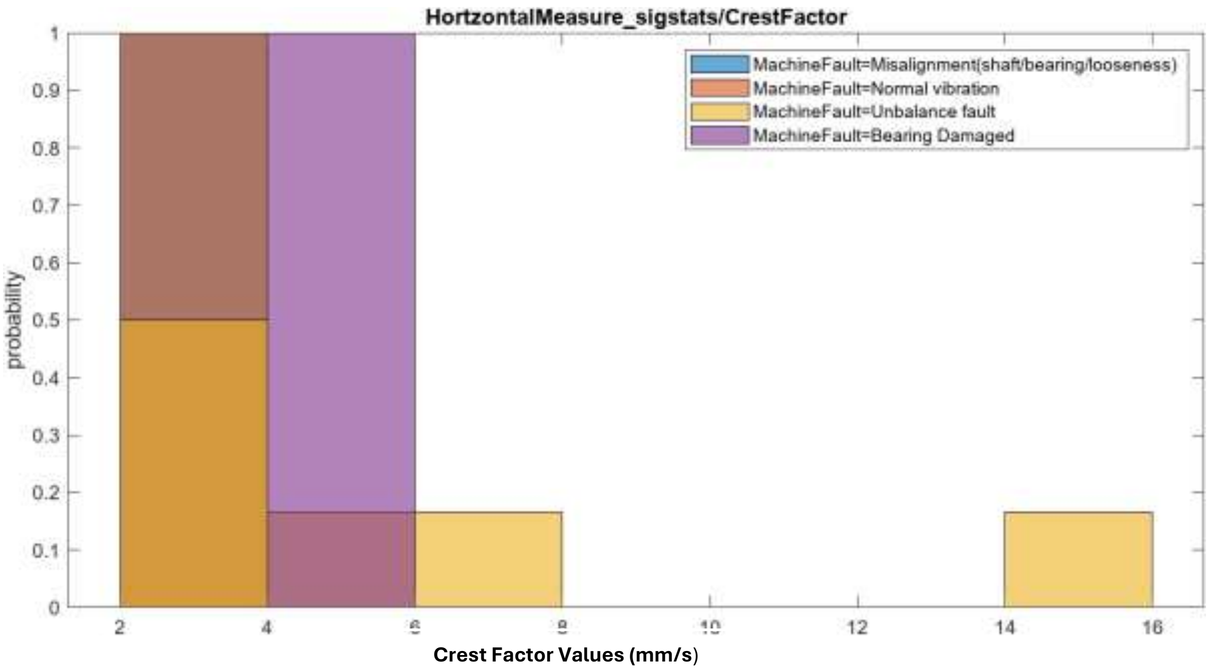

Figure 4. Bar Chart for Crest Factor in Probe Horizontal Measure

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Clearance Factor

The plotted distribution in figure 5 shows that conditions such as unbalance and bearing damage correspond to higher clearance factor values compared to normal vibration, reflecting the presence of more pronounced peak amplitudes. The elevated probabilities for bearing damage further highlight the sensitivity of this feature to localized mechanical faults, making it a valuable diagnostic tool for early fault detection.

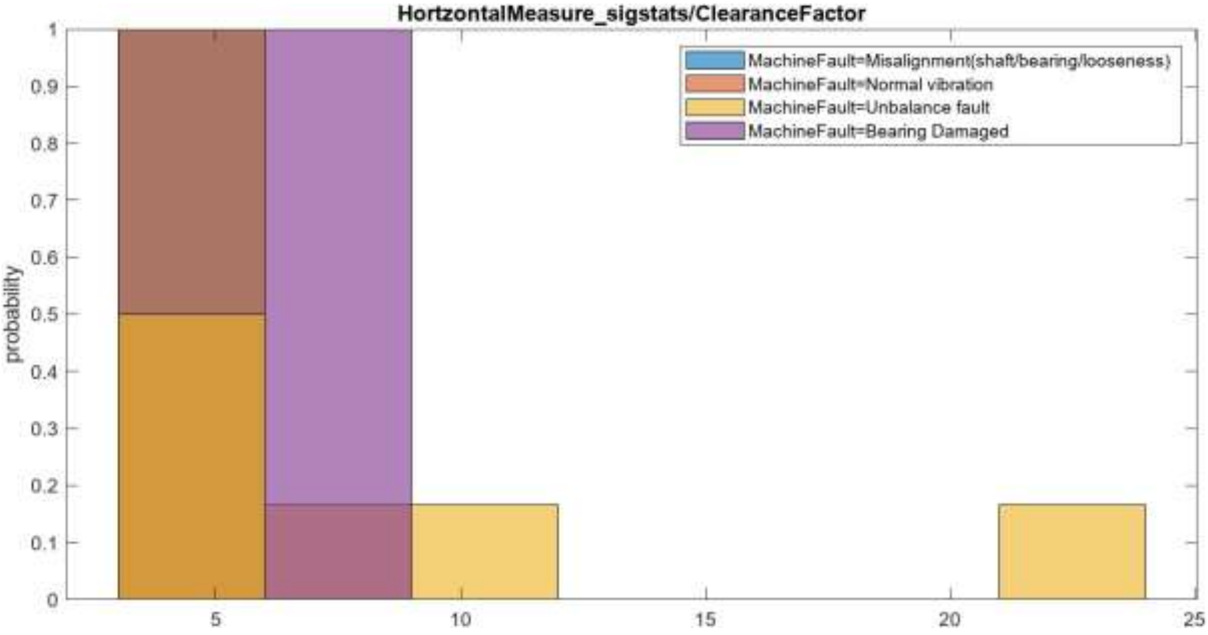

Figure 5. Bar Chart for Clearance Factor in Probe Horizontal Measure

Signal to Noise Ratio

Figure 6 shows the results of the analysis of the signal to noise ratio of the vibration signals data. As shown in figure 6, unbalance faults exhibit relatively higher SNR values (0-15), indicating that the vibration patterns are clearer and more pronounced compared to the noise background. Conversely, bearing damage and normal vibration are associated with lower or even negative SNR values, suggesting that the signal is heavily masked by noise or weak in amplitude. This demonstrates the diagnostic importance of SNR in distinguishing between strong, fault-related signals and those obscured by noise.

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

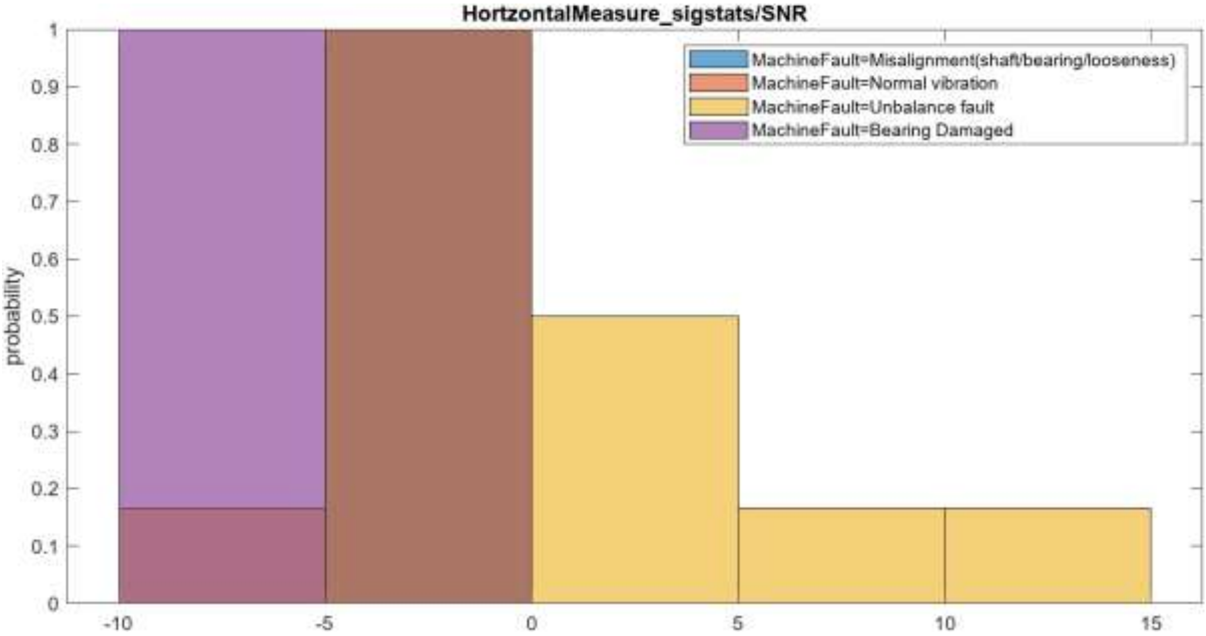

Figure 6. Bar Chart for SNR in Probe Horizontal Measure

Total Harmonic Distortion (THD)

Figure 7 shows the total harmonic distortion for the vibration signals data collected. A higher THD value indicates a greater presence of harmonic content, which typically arises from nonlinearities or mechanical irregularities in the system such as shaft misalignment, looseness, or bearing defects. Since harmonic distortion directly affects signal clarity and reliability, THD serves as an important indicator for diagnosing machine faults. Elevated THD values may suggest severe mechanical issues, while lower values are generally associated with stable and balanced operating conditions.

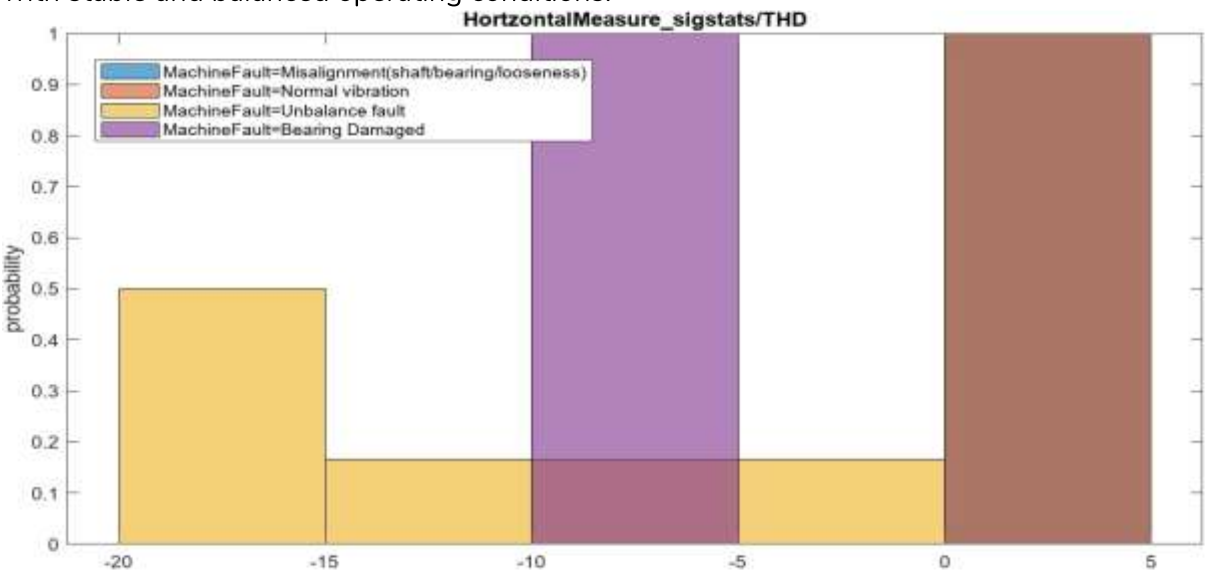

Figure 7. Bar Chart for THD in Probe Horizontal Measure

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Signal to Noise and Distortion Ratio (SINAD)

Figure 8 shows the signal to noise and distortion ratio of the vibration data analysed in order to detect fault. A higher SINAD value in figure 8 indicates a cleaner and more reliable signal, essential for accurate fault detection. Conversely, lower SINAD values such as from -12 to 0 suggest that the signal is significantly affected by unwanted noise and distortion components, potentially masking fault-related characteristics. This makes SINAD particularly useful for assessing the overall fidelity of vibration signals and distinguishing between healthy machine states and those with incipient or advanced faults.

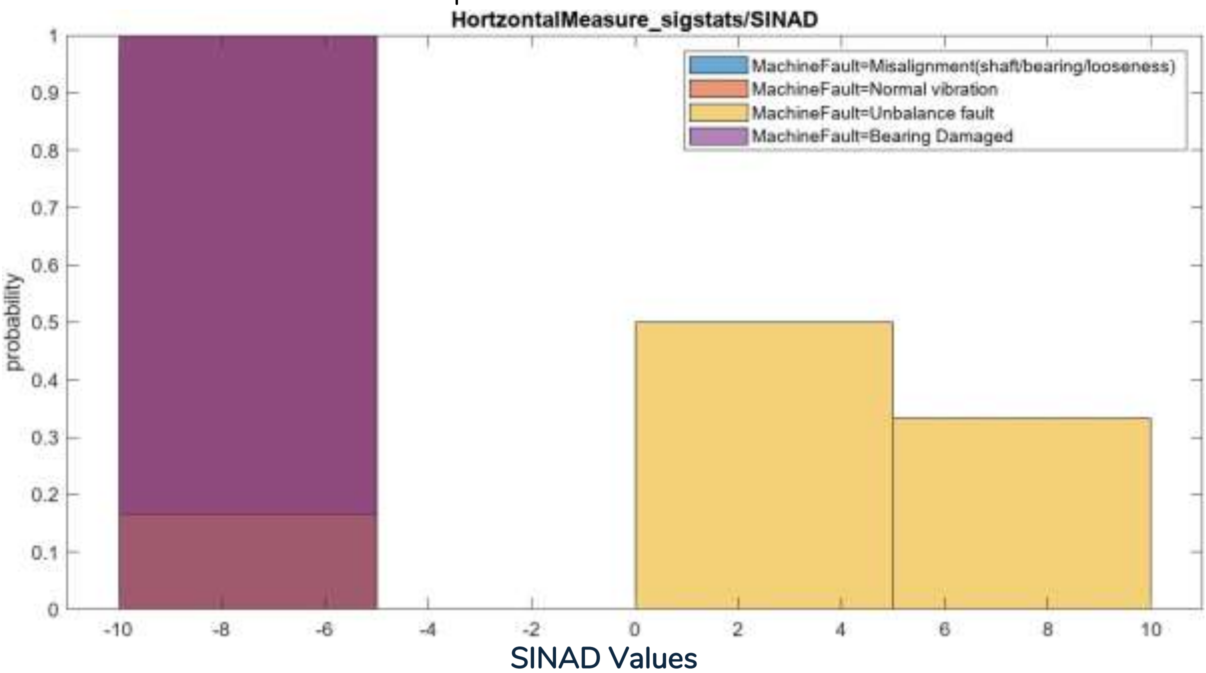

Figure 8. Bar Chart for SINAD in Probe Horizontal Measure

Shape Factor

The shape factor of the vibration signals data analysed to detect the possible fault in the system is shown in Figure 9. From the figure, a Shape Factor of 1.25 is strongly indicative of the bearing damage, while 1.2 suggests unbalance issues.

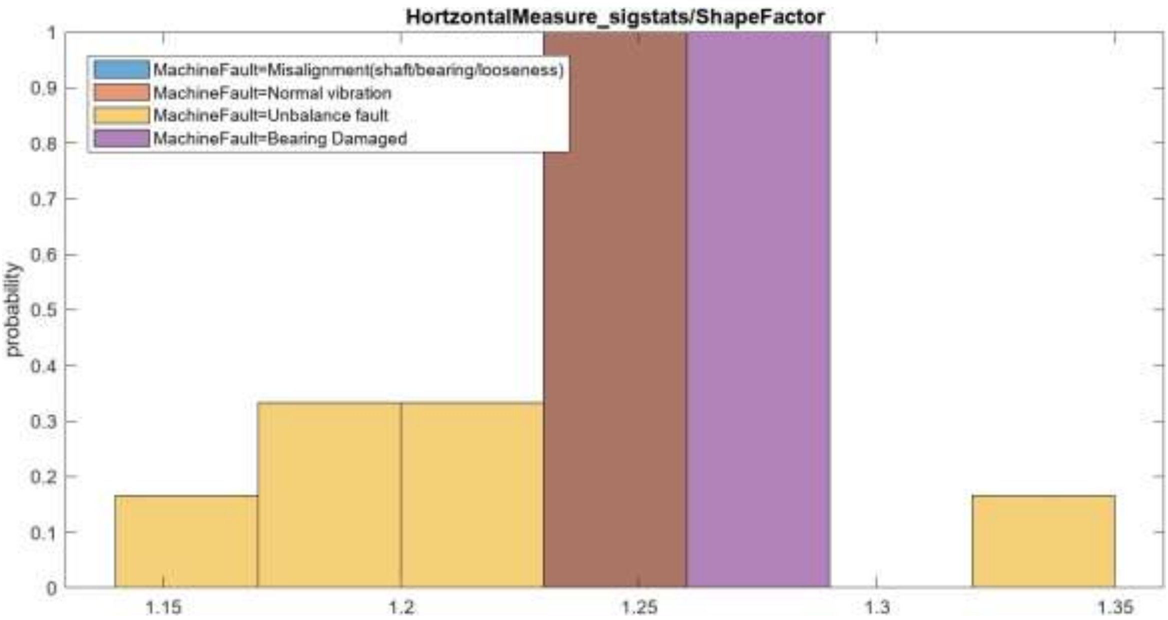

Figure 9. Bar Chart for Impulse Factor in Probe Horizontal Measure

Kurtosis

Figure 10 shows the vibration signals' kurtosis. In the figure, a low Kurtosis (near 0) strongly correlates with multiple fault types, especially bearing damage, while higher values suggest reduced fault likelihood.

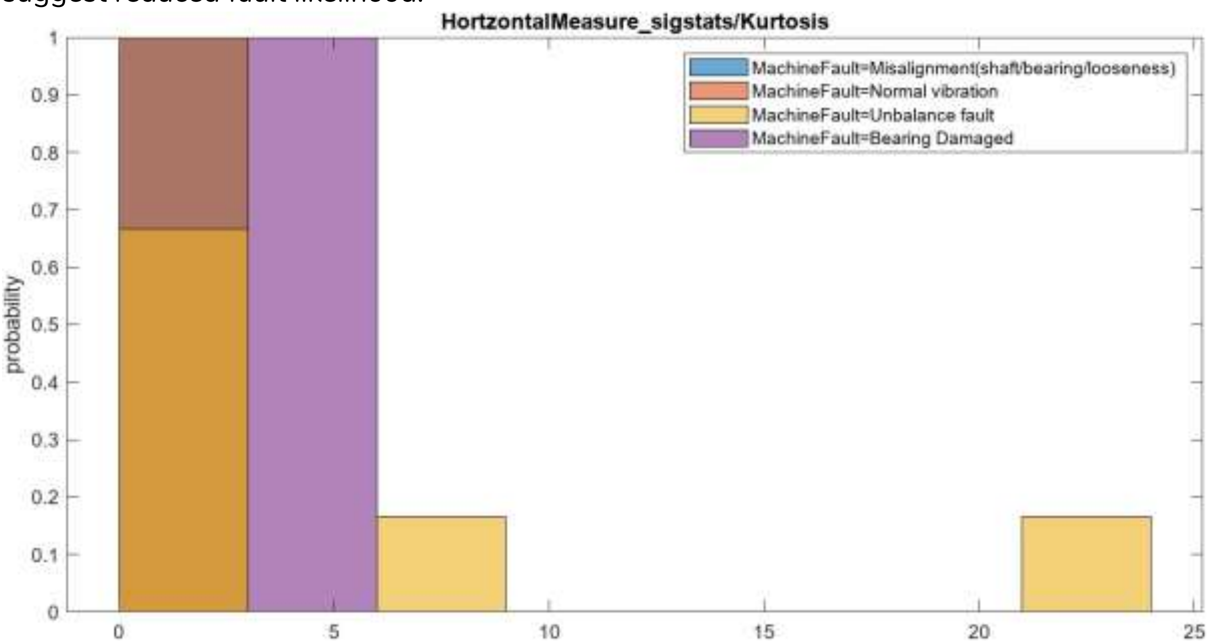

Figure 10. Bar Chart for Kurtosis in Probe Horizontal Measure

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Performance Comparison of Feature Selection with Classifier Based on Time-Domain Vibration Signals

Table 2 shows a performance evaluation of feature selection techniques with classification techniques, based only on the vibration signals characteristics such as the peak value, impulse factor, crest factor, clearance factor, signal-to-noise ratio, total harmonic distortion, signal to noise and distortion ratio, shape factor, and kurtosis.

As presented in Table 2, model performance varied significantly depending on the feature selection strategy applied. The neural network model enhanced by Chi-square feature selection delivered the most impressive results. It recorded a validation accuracy of 96.55% and a perfect test accuracy of 100%, with the lowest validation cost of 1. Furthermore, it achieved the highest prediction speed, processing 873 observations per second, and completed training in approximately 96.4 seconds.

This outcome suggests that the Chi-square method was particularly effective in isolating key input variables, contributing to both model precision and efficiency. Its combination with the neural network architecture allowed for robust generalization across new data samples, reinforcing its potential for real-time machine condition monitoring. This study is in agreement with previous studies that had highlighted the wide applicability and efficiency of chi-square in filtering out irrelevant features, thereby enhancing the performance of the predictive models (Alshaer et al., 2021; Çalışkan, 2023; Korial et al., 2024).

The mRMR-based model also showed strong results, posting a validation accuracy of 93.10% and matching the chi-square model's 100% test accuracy. However, it lagged slightly in speed and incurred a higher misclassification cost, suggesting it may be better suited for scenarios where accuracy is prioritized over processing speed. Other models using ReliefF, ANOVA, and Kruskal-Wallis also achieved full test accuracy, but with slightly reduced validation performance. These approaches, while still viable, might be preferable in cases where computational simplicity or specific statistical characteristics are more critical. The Fine Tree model, paired with PCA, proved less capable of managing the dataset's complexity. With a validation accuracy of 82.76% and a relatively low-test accuracy of 60%, it lacked the predictive strength of the neural networks, likely due to its limited ability to capture complex data patterns.

Table 2. Models Performance for Training and Testing

Model No.	Model Type	Feature Algorithm	Ranking	Accuracy % (Validation)	Accuracy % (Test)	Total cost (Validation)	Prediction Speed (obs/sec)	Training Time (sec)
1	Fine Tree	PCA		82.75862	60	5	190.01236	6.03636
2	Neural Network	None		89.65517	100	3	529.93299	228.97810
3	Neural Network	MRMR		93.10345	100	2	472.78164	161.53963

4	Neural Network	Chi2	96.55172	100	1	873.14945	96.41821
5	Neural Network	ReliefF	89.65517	100	3	486.75114	104.64299
6	Neural Network	ANOVA	89.65517	100	3	491.95985	159.32742
7	Neural Network	Kruskal-Wallis	86.20690	100	4	497.63879	145.05505

These findings highlight the crucial role of feature selection in building accurate and efficient predictive maintenance models. The Chi-square method emerged as the most effective, identifying features that significantly improved the model's accuracy and operational performance. The optimized neural network's superior results underline its capacity to learn complex fault patterns from high-dimensional data, especially when paired with a suitable feature selection method. The use of Bayesian optimization to tune model parameters also contributed to its high level of accuracy. On the other hand, while traditional models like decision trees offer quicker training and interpretability, their lower predictive performance makes them less ideal for applications involving complex sensor data.

In summary, the combination of careful feature selection and advanced classification techniques, particularly the use of neural networks with chi-square feature selection, can significantly enhance fault detection accuracy and support more proactive maintenance planning in industrial settings.

CONCLUSION

This study set out to improve the accuracy of predictive maintenance systems by identifying and selecting the most relevant features from vibration data collected from a cement mill fan. The comparison of several feature selection techniques revealed that the chi-square method provided the most impactful feature set, significantly improving model accuracy when combined with a neural network classifier. The resulting model achieved perfect test accuracy and outperformed other models in prediction speed and cost efficiency. While other feature selection methods like mRMR, ReliefF, ANOVA, and Kruskal-Wallis also showed strong results, chi-square stood out as the most effective overall. The findings confirm that thoughtful feature selection is critical to optimizing predictive models. The integration of statistical preprocessing with machine learning especially using neural networks, offers a practical and scalable approach for early fault detection in industrial machinery. These insights can guide future applications of artificial intelligence-driven maintenance solutions across various industrial domains. While the results are promising, the study was limited to a single machine type and offline analysis. Future research should focus on deploying the models in real-time industrial environments, expanding datasets across different equipment, and exploring deep learning approaches for automatic feature extraction. Incorporating explainable tools and integrating predictive outputs with maintenance scheduling systems would further enhance the practical utility of the model.

REFERENCES

Ahmad, G.N., Ullah, S., Algethami, A., Fatima, H., & Akhter, S.M.H. (2022). Comparative study of optimum medical diagnosis of human heart disease using machine learning technique with and without sequential feature selection. IEEE Access, 10, 23808–23828. https://ieeexplore.ieee.org/abstract/document/9718089/

Alshaer, H.N., Otair, M.A., Abualigah, L., Alshinwan, M., & Khasawneh, A.M. (2021). Feature selection method using improved Chi-square on Arabic text classifiers: Analysis and application. Multimedia Tools and Applications, 80(7), 10373–10390. https://doi.org/10.1007/s11042-020-10074-6

Balducci, F., Impedovo, D., & Pirlo, G. (2018). Machine learning applications on agricultural datasets for smart farm enhancement. Machines, 6, 38–59. https://doi.org/10.3390/machines6030038.

Bezerra, F. E., Oliveira Neto, G. C. de, Cervi, G. M., Francesconi Mazetto, R., Faria, A. M. de, Vido, M., Lima, G. A., Araújo, S. A. de, Sampaio, M., & Amorim, M. (2024). Impacts of feature selection on predicting machine failures by machine learning algorithms. Applied Sciences, 14(8), 3337. https://www.mdpi.com/2076-3417/14/8/3337.

Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. (2021). Prediction of heart disease using a combination of machine learning and deep learning. Computational Intelligence and Neuroscience, 2021(1), 8387680. https://doi.org/10.1155/2021/8387680

Buchaiah, S., & Shakya, P. (2022). Bearing fault diagnosis and prognosis using data fusion based feature extraction and feature selection. Measurement, 188, 110506. https://www.sciencedirect.com/science/article/pii/S0263224121013889.

Çalışkan, A. (2023). Diagnosis of malaria disease by integrating chi-square feature selection algorithm with convolutional neural networks and autoencoder network. Transactions of the Institute of Measurement and Control, 45(5), 975–985. https://doi.org/10.1177/01423312221147335.

Debal, D. A., & Sitote, T. M. (2022). Chronic kidney disease prediction using machine learning techniques. Journal of Big Data, 9(1), 109. https://doi.org/10.1186/s40537-022-00657-5.

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., Shultana, S., Beeravolu, A. R., & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access, 9, 19304–19326. https://ieeexplore.ieee.org/abstract/document/9333574/

Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. Journal of Big Data, 9(1), 24. https://doi.org/10.1186/s40537-022-00573-8

Khattach, O., Moussaoui, O., & Hassine, M. (2024). Feature selection strategies in failure prediction. In M. Serrhini & K. Ghoumid (Eds.), Advances in Smart Medical, IoT & Artificial Intelligence, 11, 185–192. https://doi.org/10.1007/978-3-031-66850-0_21

Korial, A. E., Gorial, I. I., & Humaidi, A. J. (2024). An improved ensemble-based cardiovascular disease detection system with chi-square feature selection. Computers, 13(6), 126. https://www.mdpi.com/2073-431X/13/6/126

Mahmood, M. R. (2021). Two feature selection methods comparison chi-square and relief-f for facial expression recognition. Journal of Physics: Conference Series, 1804 (1), 012056. https://iopscience.iop.org/article/10.1088/1742-6596/1804/1/012056/meta.

Rupapara, V., Rustam, F., Ishaq, A., Lee, E., & Ashraf, I. (2023). Chi-square and PCA based feature selection for diabetes detection with ensemble classifier. Intelligent Automation & Soft Computing, 36(2).

Sharma, A., & Mishra, P. K. (2022). Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. International Journal of Information Technology, 14(4), 1949–1960. https://doi.org/10.1007/s41870-021-00671-5

Suruliandi, A., Mariammal, G., & Raja, S. P. (2021). Crop prediction based on soil and environmental characteristics using feature selection techniques. Mathematical and Computer Modelling of Dynamical Systems, 27(1), 117–140. https://doi.org/10.1080/13873954.2021.1882505.

Theng, D., & Bhoyar, K. K. (2024). Feature selection techniques for machine learning: A survey of more than two decades of research. Knowledge and Information Systems, 66(3), 1575–1637. https://doi.org/10.1007/s10115-023-02010-5.

Xie, S., Zhang, Y., Lv, D., Chen, X., Lu, J., & Liu, J. (2023). A new improved maximal relevance and minimal redundancy method based on feature subset. The Journal of Supercomputing, 79(3), 3157–3180. https://doi.org/10.1007/s11227-022-04763-2

Yang, Y., Zhai, J., Wang, H., Xu, X., Hu, Y., & Wen, J. (2025). An improved fault diagnosis method for rolling bearing based on relief-F and optimized random forests algorithm. Machines, 13(3). https://doi.org/10.3390/machines13030183.