Fault Detection in Induction Motors Through Statistical Descriptors Using Vibration Signals

Detección de Fallas en Motores de Inducción a Través de Descriptores Estadísticos Utilizando Señales de Vibración

Diego Jesus Vasquez-Calderon[1], Juan Vazquez-Paniagua[1], Uriel Calderon-Uribe[1]
[1] Laboratorio de visión por computadora, Tecnológico Nacional de México / ITS Sur de Guanajuato, Av. Educación Superior 2000, Col. Juárez, Uriangato, Gto, CP. 38982, México.
diegojesusvc2003@gmail.com, juanvazquezpaniag@gmail.com, urielcal92@gmail.com

Abstract

Operational reliability in modern industry largely depends on the continuous operation of induction motors, as they power most essential production processes. A method is proposed for early fault diagnosis that extracts statistical descriptors from vibration signals, second to fourth-order centered moments, RMS, and crest/shape factors, and classifies motor states with a shallow decision tree. Tri-axial vibration data from a three-phase induction motor (1.1 kW, 220/380 V, 5 A, 2800 RPM, 50 Hz) were recorded under different operating conditions, segmented for uniformity, and transformed into the described features. Several supervised models were evaluated on a stratified 90/10 train–test split; the decision tree achieved the best accuracy while remaining easy to interpret. The findings indicate that simple statistical features coupled with an interpretable classifier can deliver accurate diagnosis with modest computational cost, supporting predictive-maintenance adoption in resource-constrained environments.

Keywords: Induction motor, Fault detection, Vibration signal, Statistical descriptor, Decision tree

Introduction

Operational reliability in modern industry largely depends on the continuous operation of induction motors, as they power most essential production processes. A significant portion of industrial loads is driven by this type of motor; therefore, faults can strongly impact efficiency by forcing unplanned stops, wasting resources, and increasing maintenance costs.

Over the last decade, digitalization and the Industry 4.0 paradigm have encouraged data-based diagnostic techniques. Low-cost current and vibration sensors now produce large volumes of information that, when properly analyzed, enable the detection of incipient faults before they escalate to critical downtime. However, many current solutions rely on deep neural networks that demand considerable computation, and large datasets are rarely available in real-world environments, particularly in small and medium-sized enterprises. This context motivates lighter, more transparent methods that combine basic motor physics with accessible statistics and decision algorithms that practitioners can readily interpret. This project addresses that need by proposing a diagnosis scheme based on statistical descriptors of vibration signals specifically second to fourth-order centered moments, RMS, and shape/crest factors together with a shallow decision tree. The aim is to balance accuracy, simplicity, and implementation cost so that the system can run on standard industrial hardware and remain understandable to maintenance personnel.

To this end, the project aims to develop and validate an early fault-diagnosis method for induction motors that leverages vibration-based statistical descriptors and an interpretable decision tree. Concretely, we will characterize frequent faults under different load conditions by acquiring vibration signals; extract centered moments (second to fourth order), RMS, crest and shape factors, and other low-cost computational metrics; design and train an optimized decision tree to classify motor states; and evaluate the method's accuracy and practical feasibility against more complex baselines. Taken together, this approach seeks to bring predictive maintenance within reach of companies with limited resources, thereby enhancing competitiveness and operational safety.

Induction motors are critical components in industry due to their efficiency and robustness, but their continuous operation exposes them to failures such as bearing damage, rotor defects, or stator short circuits, which can

lead to unexpected shutdowns and major economic losses (Wildi, 2014). Traditionally, the diagnosis of these failures was carried out through visual inspection and vibration analysis, but advances in electronics and digitalization have enabled new strategies based on analyzing data collected by current and vibration sensors (Chang *et al.*, 2022).

Over the last decade, the literature has documented the use of advanced machine-learning methods for early fault diagnosis. For example, Chang *et al.* (2022) explores the use of deep neural networks and synthetic data-generation techniques to address imbalance problems in industrial datasets. However, these methods usually require large volumes of data and high computational capacity, which may be unfeasible in small and medium-sized plants. Therefore, the development of lighter and more transparent approaches has been promoted—approaches that can be implemented on standard hardware and interpreted by maintenance personnel (Toma *et al.*, 2020).

A practical alternative is to use statistical descriptors derived from the motor's electrical and vibration signals. These features, such as second to fourth-order centered moments, root-mean-square (RMS), and crest factors, summarize the essential information in the signals at low computational cost (James *et al.*, 2013). Statistical moments are measures that capture the shape and dispersion of a data distribution: the second moment corresponds to variance, the third to skewness, and the fourth to kurtosis, providing information on the presence of anomalies in the motor's dynamics (DeGroot & Schervish, 2012; Spiegel *et al.*, 2012).

The use of decision trees as a classification method has proved especially valuable in industrial applications due to their ability to create interpretable rules and their low computational requirements (scikit-learn developers, 2025b). A decision tree is a supervised learning algorithm that partitions the data based on selected features, generating decision nodes that can be easily understood and validated by domain experts (James *et al.*, 2013). According to Toma *et al.* (2020), the combination of statistical descriptors and decision trees makes it possible to achieve high levels of accuracy in detecting bearing faults in induction motors while facilitating knowledge transfer to plant technical personnel.

Other strategies, such as electrical phasor analysis combined with fuzzy logic (Reyes-Malanche *et al.*, 2023) or the integration of genetic algorithms and machine-learning classifiers (Toma *et al.*, 2020), have been proposed to improve the robustness and accuracy of diagnosis. However, these approaches tend to increase system complexity and may hinder their implementation in contexts where transparency and ease of maintenance are priorities.

Meanwhile, the availability of machine-learning libraries such as scikit-learn has simplified the experimentation and validation of various classifier models, including decision trees (scikit-learn developers, 2025b), Naive Bayes (GaussianNB) (scikit-learn developers, 2025c), k-nearest neighbors (KNeighborsClassifier) (scikit-learn developers, 2025d), logistic regression (scikit-learn developers, 2025e), and support vector machines (SVC) (scikit-learn developers, 2025f). This enables comparisons of their performance in terms of accuracy, response time, and ease of interpretation.

In this way, the approach based on statistical descriptors and decision trees facilitates early fault detection with limited resources and encourages the adoption of predictive maintenance in industry (Wildi, 2014; James *et al.*, 2013).

Materials and Methods

The proposed method for diagnosing faults in induction motors was developed in three main stages: signal acquisition, data processing, and machine-learning model design.

Dataset Description

Signals were obtained from a three-phase induction motor rated at 1.1 kW, 220/380 V, 5 A, 2800 RPM, 50 Hz, using a digital ADXL345 tri-axial accelerometer to record vibration. Data were collected under different operating conditions and stored in CSV files for later analysis.

Preprocessing

In order to guarantee dataset uniformity, the vibration signals were segmented into equal-length portions. Each signal file was divided into segments of fixed size, resulting in consistent and comparable subsets. For each segment, second, third, and fourth-order centered moments were computed using algebraic formulas derived from non-centered moments. These descriptors were used to extract the dynamic characteristics of motor

vibration and facilitate the identification of patterns associated with potential failures. The resulting metrics were organized into a structured file, including the class labels for each analyzed condition.

The statistical descriptors calculated are the following:

- First-order non-centered moment (α): represents the mean value of the signal segment.

$$\alpha = \frac{1}{N} \sum_{i=1}^{N} x_i \quad (1)$$

Where x_i is the value of the signal at time step *i*, and N is the number of samples in the segment.

- Second-order non-centered moment (α2): Used to compute the second centered moment (variance).

$$\alpha_2 = \frac{1}{N} \sum_{i=1}^{N} x_i^2 \quad (2)$$

- Third-order non-centered moment (α3): Used to compute the third centered moment (skewness).

$$\alpha_3 = \frac{1}{N} \sum_{i=1}^{N} x_i^3 \quad (3)$$

- Fourth-order non-centered moment (α4): Used to compute the fourth centered moment (kurtosis).

$$\alpha_4 = \frac{1}{N} \sum_{i=1}^{N} x_i^4 \quad (4)$$

Based on these non-centered values, the centered statistical moments are calculated as follows:

- Variance (second-order centered moment): This metric measures the dispersion of the values in the signal with respect to the mean.

$$\mu_2 = \alpha_2 - \alpha^2 \quad (5)$$

- Skewness (third-order centered moment): This metric quantifies the asymmetry of the signal distribution around its mean.

$$\mu_3 = \alpha_3 - 3\alpha\alpha_2 + 2\alpha^3 \quad (6)$$

- Kurtosis (fourth-order centered moment): This metric measures the concentration of signal values around the mean, indicating whether the distribution is flat or peaked.

$$\mu_4 = \alpha_4 - 4\alpha\alpha_3 + 6\alpha^2\alpha_2 - 3\alpha^4 \quad (7)$$

Model Evaluation

From the extracted features, a dataset was built to train and test different supervised classification models. Several algorithms were evaluated, including logistic regression, support vector machines (SVM), Naive Bayes, and k-nearest neighbors (KNN), implemented with the scikit-learn library. The data were split into training (90%) and test (10%) sets while preserving class proportions (stratified split). After comparing classifier performance in terms of accuracy and interpretability, the decision tree was selected as the optimal model. The final model was validated by visualizing class separation in feature space.

Results

In order to evaluate the results of the final model, various metrics were used to measure its performance. These metrics are:

- Accuracy: which is a measure used to measure the correct predictions of the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP} \quad (8)$$

Where TP represents the positive instances assigned to the positive class. TN are the negative instances classified as negative. FP are positive instances assigned as negative instances. Finally, FN represents the negative instances classified as positive instances.

- Precision: the number of true positives divided by the total number of positive predictions.

$$Precision = \frac{TP}{TP + F} \quad (9)$$

- Recall: is a measure of how often a model correctly identifies positive instances (true positives) from all the actual positive samples in the dataset.

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

- F1-score: is the harmonic mean of precision and recall, capturing a balance between them. It is especially useful when classes are imbalanced or when both false positives and false negatives matter.

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (11)$$

- Specificity: is a measure of how often a model correctly identifies negative instances (true negatives) out of all actual negative samples in the dataset. It complements recall by quantifying how well the model avoids false alarms.

$$Specificity = \frac{TN}{TN + FP} \quad (12)$$

The performance of several classification models was evaluated using statistical features extracted from vibration signals. The decision-tree model stood out for both its accuracy and interpretability. Table 1 summarizes the performance metrics obtained for each class in the test set.

Table 1. Decision-tree performance metrics.

Class	Precision	Recall	F1-score	Support
Full Load Normal Bearing	1.00	1.00	1.00	1
Half Load Normal Bearing	1.00	1.00	1.00	1
Full Load Faulty Bearing	1.00	1.00	1.00	1
No Load Normal Bearing	1.00	1.00	1.00	1
Accuracy			1.00	4
Macro avg	1.00	1.00	1.00	4
Weighted avg	1.00	1.00	1.00	4

The model achieved an overall accuracy of 100%, correctly classifying all samples in the test set. Decision boundaries and class separation in representative feature spaces (e.g., using mc2 and mc3) show clear differentiation among motor states, as illustrated in Figure 1.

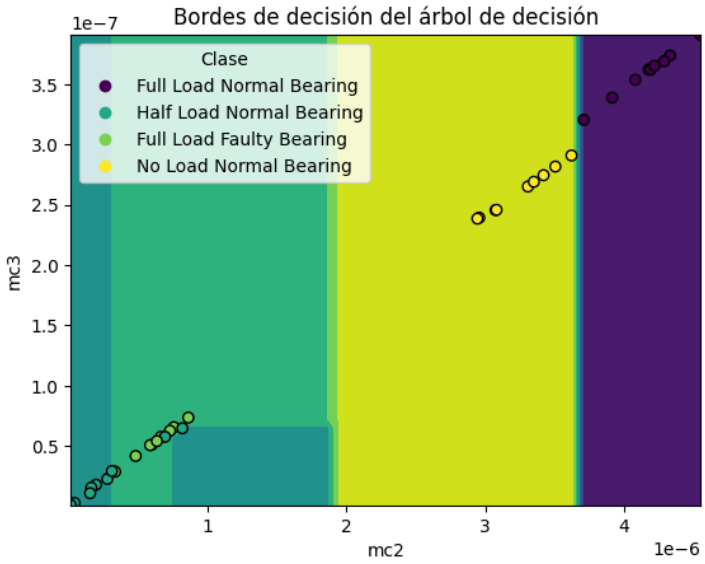

Figure 1. Decision-tree decision boundaries and class distribution in the mc2 and mc3. Source: Own authorship.

As observed, the decision tree model, trained with centered moments extracted from vibration signals, correctly classified all the evaluated conditions. This outstanding performance reflects the ability of the selected statistical features to distinguish among the different motor states. It is likely that the clarity of the class separation and the size of the dataset favored this outcome; therefore, validation with more samples and more complex scenarios is recommended.

Comparing these findings with the referenced literature, Toma *et al.* (2020) implemented a combination of statistical descriptors and genetic algorithms with various machine-learning classifiers, achieving high precision but with greater complexity and computational demand. Chang *et al.* (2022) employed deep neural networks and synthetic data-generation techniques to address class imbalance, which required large volumes of data and processing resources. In contrast, the method presented here achieved comparable results using a much simpler and more transparent approach, suitable for industrial contexts with limited resources. Reyes-Malanche *et al.* (2023) proposed the use of electric phasor analysis together with fuzzy logic, which increases robustness but also system complexity.

The implications of these findings suggest that a scheme based on simple statistical features, combined with interpretable models such as decision trees, can be sufficient for early fault diagnosis in induction motors, provided the classes are well defined. This represents an opportunity for small and medium-sized enterprises to access affordable, low-cost, and easy-to-implement solutions for predictive maintenance.

The limitations of this study include the reduced size of the dataset, the exclusive use of vibration signals, and the consideration of a limited number of fault conditions. Future work should expand the database, include different types of faults, account for operational variability and noise, and conduct tests in real industrial environments. It would also be of interest to broaden the analysis to include other data sources and to explore the applicability of more complex techniques in more diverse industrial scenarios.

Conclusions

This work presented and validated a method for the early diagnosis of faults in induction motors, using statistical descriptors extracted from vibration signals together with a decision tree model. The results showed that this approach accurately classified the different motor states evaluated, achieving outstanding performance on the test set.

The proposed procedure met the stated objectives by characterizing frequent faults through measurement campaigns under various load conditions and by extracting key metrics such as centered moments and shape factors at low computational cost. Using a decision tree as the classifier facilitated interpretation of the results

and demonstrated the feasibility of implementing accessible solutions in industrial environments that do not have extensive technological resources.

When compared with the reviewed literature, it was observed that precision levels comparable to those reported by more complex methods can be achieved using a simpler, more efficient strategy. This highlights the potential of applying statistical techniques and interpretable models to bring predictive maintenance closer to small and medium-sized enterprises.

Finally, although the findings are encouraging, the study has limitations related to the size of the dataset and the number of scenarios analyzed. Therefore, future research could focus on expanding the database, incorporating additional electrical signals, and validating the system under real industrial conditions, in order to strengthen the robustness and generalization of the developed method.

References

Chang, H., Wang, Y., Shih, Y., & Kuo, C. (2022). Fault Diagnosis of Induction Motors with Imbalanced Data Using Deep Convolutional Generative Adversarial Network. *Applied Sciences,* 12(8), 4080. https://doi.org/10.3390/app12084080

DeGroot, M. H., & Schervish, M. J. (2012). *Probability and statistics* (4a ed., pp. 234‑236). Pearson Education.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R* (pp. 303‑316). Springer Nature.

Reyes-Malanche, J. A., Villalobos-Pina, F. J., Ramırez-Velasco, E., Cabal-Yepez, E., Hernandez-Gomez, G., & Lopez-Ramirez, M. (2023). Short-Circuit Fault Diagnosis on Induction Motors through Electric Current Phasor Analysis and Fuzzy Logic. *Energies,* 16(1), 516. https://doi.org/10.3390/en16010516

scikit-learn developers. (2025a). DecisionBoundaryDisplay. https://scikit-learn.org/stable/modules/generated/sklearn.inspection.DecisionBoundaryDisplay.html

scikit-learn developers. (2025b). DecisionTreeClassifier. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

scikit-learn developers. (2025c). GaussianNB. https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

scikit-learn developers. (2025d). KNeighborsClassifier. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

scikit-learn developers. (2025e). LogisticRegression. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

scikit-learn developers. (2025f). SVC. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Spiegel, M. R., Schiller, J. J., & Srinivasan, R. A. (2012). Schaum’s outline of probability and statistics (4a ed., pp. 75‑79). McGraw Hill Education.

Toma, R. N., Prosvirin, A. E., & Kim, J.-M. (2020). Bearing Fault Diagnosis of Induction Motors Using a Genetic Algorithm and Machine Learning Classifiers. *Sensors,* 20(7), 1884. https://doi.org/10.3390/s20071884

Wildi, T. (2014). Electrical machines, drives, and power systems (6a ed., pp. 271‑314). Pearson Education Limited.