JOURNAL OF ADVANCED INDUSTRIAL TECHNOLOGY AND APPLICATION
e-ISSN: 2716-7097

Vol. 4 No. 2 (2023) 48-65
https://publisher.uthm.edu.my/ojs/index.php/jaita

JAITA

A Study on Comparison of Classification Algorithms for Pump Failure Prediction

Ruwaida Aliyu[1]*, Ainul Akmar Mokhtar[1], Hilmi Hussin[1]

[1] *Mechanical Engineering Department,*
Universiti Teknologi PETRONAS, Bandar Seri Iskandar, 32610 Perak Darul Ridzuan, MALAYSIA

*Corresponding Author: ruwaidaaliyu@gmail.com
DOI: https://doi.org/10.30880/jaita.2023.04.02.005

Article Info

Received: 17 September 2023
Accepted: 20 November 2023
Available online: 12 December 2023

Keywords

Fault prediction, pump failure, machine learning, classification algorithms

Abstract

The reliability of pumps can be compromised by faults, impacting their functionality. Detecting these faults is crucial, and many studies have utilized motor current signals for this purpose. However, as pumps are rotational equipped, vibrations also play a vital role in fault identification. Rising pump failures have led to increased maintenance costs and unavailability, emphasizing the need for cost-effective and dependable machinery operation. This study addresses the imperative challenge of defect classification through the lens of predictive modeling. With a problem statement centered on achieving accurate and efficient identification of defects, this study's objective is to evaluate the performance of five distinct algorithms: Fine Decision Tree, Medium Decision Tree, Bagged Trees (Ensemble), RUS-Boosted Trees, and Boosted Trees. Leveraging a comprehensive dataset, the study meticulously trained and tested each model, analyzing training accuracy, test accuracy, and Area Under the Curve (AUC) metrics. The results showcase the supremacy of the Fine Decision Tree (91.2% training accuracy, 74% test accuracy, AUC 0.80), the robustness of the Ensemble approach (Bagged Trees with 94.9% training accuracy, 99.9% test accuracy, and AUC 1.00), and the competitiveness of Boosted Trees (89.4% training accuracy, 72.2% test accuracy, AUC 0.79) in defect classification. Notably, Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbors (KNN) exhibited comparatively lower performance. Our study contributes valuable insights into the efficacy of these algorithms, guiding practitioners toward optimal model selection for defect classification scenarios. This research lays a foundation for enhanced decision-making in quality control and predictive maintenance, fostering advancements in the realm of defect prediction and classification.

1. Introduction

The process of locating and resolving issues with mechanical machinery is given great priority in today's advanced industries [1]. The hydraulic turbomachines of choice for a range of industries, including those that deal with food processing and oil refineries, are centrifugal pumps, also referred to as CPs. They feature a very strong design that can handle a wide range of flow needs. To run a conventional chemical factory, it is believed that each person uses an average of one CP [2]. Moreover, it is estimated that around 20% of the overall energy generated on a global scale is allocated for the functioning of pumps [3]. As a result, they are critical components in keeping the plant's process course running. Prediction of defects in mechanical systems can be performed in several methods, including qualitative and statistical analysis of obtained failure data, development of system mathematical models, and application of machine learning techniques. The specific fault frequency variations are examined and studied during a qualitative analysis to establish which faults are present [4].

On the other hand, there is a substantial margin for error due to the involvement of humans. CP flaws have been identified by several studies through the use of modeling tools [5]–[8]. Conversely, precisely modeling the interrelated defects and locating them can be a very difficult challenge. Online machine learning and condition monitoring are becoming increasingly popular as the capabilities of current computers continue to improve [9]. Artificial neural networks (ANNs), fuzzy logic, empirical mode decomposition (EMD), decision trees, support vector machines (SVM), and deep learning are some of the most common and widely used machine learning approaches. The machine learning approaches discussed above have been employed by researchers to identify and diagnose defects in various mechanical systems, such as bearings, [10]–[18] gears [19]–[25], induction motors [26]–[32], and pumps [33], [34], [43], [44], [35]–[42].

It is crucial to choose the right signal to use for CP condition monitoring. CP vibration [2], [32], [45]–[48], motor line current [2], [45], [49]–[51] and auditory emission signals [45], [52], [53] are common examples of signals. Failures also alter how the CP system operates, changing the motor's load and, consequently, the motor's line-current. As a result, the utilization of motor current has also been employed for the detection of CP defects. The most frequently advised signals for the diagnosis of CP defects are vibration signatures due to their immense utility, sensitivity to fault conditions, and ease of acquisition. Given the advantages vibration signals provide, this study would use the data from these signals to identify CP faults. Analysis of the collected signal is possible in the time, frequency, and time-frequency domains. Many studies use the time domain to identify faults since it provides a physical comprehension of the signal and is sensitive. However, compared to time-domain analysis, spectral/frequency analysis has three key advantages. The first is that it makes the waveform easier to grasp. Second, the physical characteristics of the signal typically depend on its frequency, and third, it is a tool used in mathematics to solve equations [54].

The analysis of fault data encompasses various data forms, including temperature, pressure, and vibration fluctuations [47]. Managing vibration data, in particular, is challenging due to its rapid expansion, leading to storage and analysis complexities. Given the irregularity and insufficiency of these data forms, effective methodologies for analysis are elusive. To address this, researchers are increasingly drawn to studying this data type to improve vibration data interpretation [54]. This study aims to evaluate classifier performance based on true positive rate (TPR), false negative rate (FNR), and area under the curve (AUC), seeking both an optimal classification algorithm and enhanced insights into pump failure. To achieve this, the vibration dataset from the Kaggle Machine Learning Repository serves as the data mining foundation for this study's objectives. Classification, a cornerstone of supervised learning, is central. The study applies five distinct classification techniques—Decision Tree (DT), Ensemble, Support Vector Machine (SVM), Neural Networks (ANN), and k-Nearest Neighbor (KNN)—to categorize the dataset into NORMAL, BROKEN, and RECOVERING classes. In Section 2, the study delves into the functional categorization characteristics of each algorithm. In summary, this research addresses the challenges of managing and interpreting fault data, focusing on classifier evaluation to identify an effective algorithm for understanding pump failure dynamics.

The major contributions of this study are;

- Enhanced Methodologies for Vibration Data Analysis: The study acknowledges the challenges associated with managing and interpreting vibration data due to its rapid growth and irregularity. By applying a range of classification techniques such as Decision Tree, Ensemble, Support Vector Machine, Neural Networks, and k-Nearest Neighbor, the study contributes to advancing effective methodologies for analyzing complex vibration data. This can lead to improved accuracy in identifying and understanding patterns related to pump failure.
- Classifier Performance Evaluation for Pump Failure Prediction: The study's focus on evaluating classifier performance using metrics like true positive rate, false negative rate, and area under the curve represents a significant contribution. By systematically comparing the performance of different classification algorithms on the vibration dataset, the study provides valuable insights into the strengths and weaknesses of each algorithm for predicting pump failure. This can guide practitioners in selecting the most suitable algorithm for real-world applications.
- Insights into Pump Failure Dynamics: Through the classification of vibration data into categories of NORMAL, BROKEN, and RECOVERING, the study contributes to a deeper understanding of pump failure dynamics. By leveraging various classification techniques, the study sheds light on distinguishing patterns associated with different failure modes. This understanding can potentially lead to proactive maintenance strategies and improved decision-making in industrial settings, ultimately contributing to reduced downtime and enhanced operational efficiency.

In summary, the study's major contributions lie in advancing vibration data analysis methodologies, evaluating classifier performance for pump failure prediction, and providing insights into the dynamics of pump

J. of Advanced Industrial Technology and Application Vol. 4 No. 2 (2023) p. 48-65

failures through effective classification techniques. The remaining parts of this article are structured as follows: in section 2, a discussion on literature review of the algorithms chosen for this study as well as relevant works, and Section 3 explains the methodology of the study. The results that were obtained are discussed in Section 4, while the conclusions and suggestions for the future are presented in Section 5.

2. Literature Review on Classification Algorithms

This section provides a detailed explanation of the classification algorithms employed in this study. The section further gives an overview of the related studies in the literature concerning the application of the classification algorithms including their cons and pros. Several researchers have focused on the condition motoring of centrifugal pumps so far. McKee et al., [55] introduced centrifugal pump fault types and analyzed known diagnostic and prognostic methods for centrifugal pumps. In comparison to operating conditions such as current, voltage, vacuum gauge reading, and pressure gauge reading, vibration signals are commonly used in centrifugal pump condition motoring. Because vibration signals provided dynamic information about the machine condition, Wang and Chen, [37] demonstrated that they were useful in defect diagnostics of centrifugal pumps. Xue et al., [56] further stated that the vibration signature is the most telling indicator of the condition of spinning machinery. Effective feature extraction of vibration signals is a critical step in ensuring problem diagnosis accuracy.

In the feature analysis of a centrifugal pump, various methodologies were used. Samanipour et al., [57] used pressure time domain features to detect centrifugal pump cavitations. Kumar and Kumar, [58] demonstrated an automatic centrifugal pump detection approach based on time-frequency properties. Muralidharan and Sugumaran, [39], [46], [59] proposed wavelet-based fault diagnostic algorithms for monoblock centrifugal pumps and also presented a comparative study between support vector machine (SVM) and extreme learning machine (ELM) for fault detection in pumps. Azizi et al., [60] presented a hybrid feature selection technique for centrifugal pump cavitation severity. Yu et al., [61] suggested a sewage source heat pump system fault detection methodology based on principal component analysis (PCA). Zhang et al., [62] presented a variational mode decomposition method for detecting rolling bearing faults in a multistage centrifugal pump. Sakthivel et al., [48] investigated dimensionality reduction strategies for failure diagnosis of a single-block centrifugal pump using vibration signals. Liu et al.[63] used locally linear embedding, wavelet modification, and singular value decomposition approaches to reduce the dimensionality of submersible plunger pump failure diagnostics.

Several artificial intelligence algorithms have been employed in centrifugal pumps in terms of comparative investigation of identification performance. Muralidharan and colleagues [64], [65] used naive Bayes, Bayes net, and support vector machine (SVM) to diagnose faults in a single-block centrifugal pump. Sakthivel et al.[48], [52] used rough-set, fuzzy set, and gene expression programming to diagnose faults in a monoblock centrifugal pump. Deep belief networks and lowest entropy deconvolution were introduced by Wang et al. [66], [67] to detect problems in axial piston pumps. Panda et al. [68] used SVM to investigate flow obstructions in the inlet pipe and approaching bubble production in the centrifugal pump. Kang et al. [69] used the status coupling relationship to improve fault detection sensitivity and reduce false alarm rates in the pipeline and pump unit systems fault diagnostics. Buono et al.[70] studied the possibility of cavitation in gerotor pumps and developed a defect diagnosis method based on the auto-regressive-moving-average methodology. Al-Tobi et al. [71] used a genetic algorithm, multilayer feedforward perceptron, SVM, and continuous wavelet transform to diagnose a centrifugal pump failure. Vibration analysis was used by Al-Obaidi et al. [72] to detect and diagnose the cavitation phenomena within centrifugal pumps. Bordoloi and Tiwari [72] talked about the best SVM method for detecting pressure flow obstructions and cavitations in casing of pumps. Azadeh et al. [47] suggested a flexible approach for centrifugal pump defect diagnostics using ANN and SVM. Nevertheless, one of the limitations of ANN is that it is impossible to determine the size of the hidden layer or the learning rate. The precision of the classifier is always a problem for fuzzy logic. SVM has higher classification accuracy, but the negative is the high algorithmic complexity and the need to properly set its hyperparameters in advance.

WSNs are able to communicate with a wide variety of fields, including but not limited to Household Robotics Systems, research, ship monitoring, underwater acoustics, and medical diagnosis. The principles of AI serve as the foundation for ML techniques, which are implemented in computing problems in instances where conventional methods are unable to be utilized. To compare their efficacy, the author set linear regression, Stochastic Gradient Decent and Naive Bayes (NB), to the test for classification (SGD) using multiple machine learning algorithms on several different datasets [10]. (SGD). Classifiers were utilized for the purpose of classification, and among the classifiers that were utilized, Gaussian naive Bayes performed best. Preprocessing of Electrocardiogram Signals and support vector machine coupling were used in order to classify the collected data. We used the adaptive filter in order to cut down on the latency as well as the computational overhead. For the purpose of classification utilizing parameters like ECG and HRV, [11] utilized techniques such as SVM, Principal Component Analysis (PCA), knowledge-based system, KNN, and ANN. In this study, four machine learning classifiers were employed: DT, KNN, ANN, SVM, and ENSEMBLE. The simulation results demonstrate their detection accuracies. Fig. 1 depicts the types of classification algorithms employed in the study.

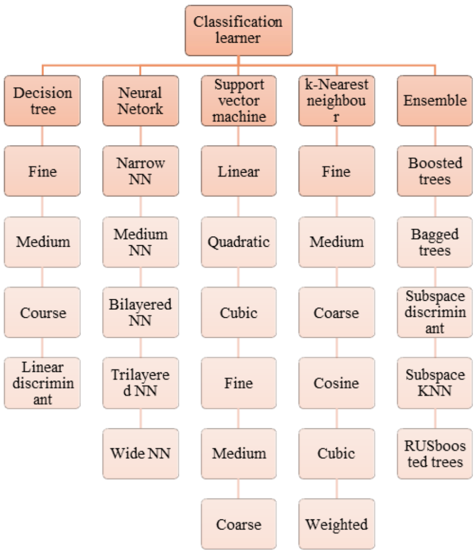

Fig. 1 *Classification learner algorithms*

2.1 Decision Tree

In machine learning, Decision Trees (DT algorithms are used for both prediction and classification. Using the decision tree and a given set of inputs, one can map the many outcomes resulting from the decisions or consequences. DT classifies occurrences by sorting them according to the feature values they possess. A decision tree is a diagram in which each node represents a property of an instance that can be classified, and each subdivision signifies a probable value for the corresponding node [73]. Instances are classified and ordered according to their feature values, beginning with the root node. Decision tree learning is a technique used in data mining and ML in which a DT is used as a prediction model to map observations to a target value. Classification trees and regression trees are common names for these tree-based models. Post-pruning processes in decision tree classifiers typically use a validation set to assess how well the trimmed trees perform. If a node has a high frequency of a certain class among the sorted training samples, it can be removed [74]. Fig. 2 is an illustration of the decision trees.

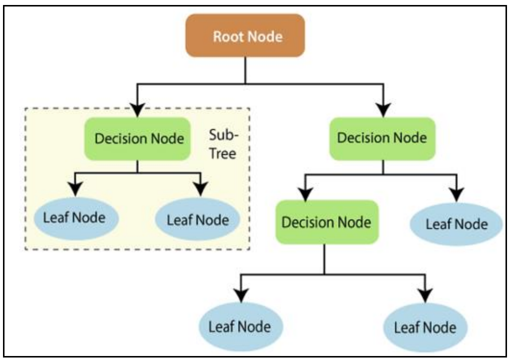

Fig. 2 *Decision tree*

J. of Advanced Industrial Technology and Application Vol. 4 No. 2 (2023) p. 48-65

2.2 Ensemble

The ensemble learning methodology is a close emulation of the human socio-cultural tendency of asking the opinions of multiple individuals before making any significant choice. This behavior is common among humans. An ensemble algorithm is a generic meta-approach to machine learning that combines the predictions from several different models to achieve improved predictive performance [10]. Even though there are an infinite number of ensembles that you can construct for your predictive modeling challenge, the subject of ensemble learning is mostly dominated by three distinct methodologies [75]. To such an extent that, rather than algorithms per se, each is now considered to be a field of study that has given rise to a lot of additional approaches that are more specialized. The ensemble Kalman filter EnKf is a collection of reservoir simulations with varying initial circumstances that attempt to capture and update a probability distribution. Each model is an advanced one-time step using previous data, and the Kalman filter provides a revised ensemble of models. This approach has the advantage of always being available as an estimate of the probability density function of such true reservoir state, as well as the estimate is continuously updated as new data becomes available. The ensemble constructs new models in the revision process by modifying the parameters of the original models or by mixing parameters from more than one earlier model as shown in Fig. 3.

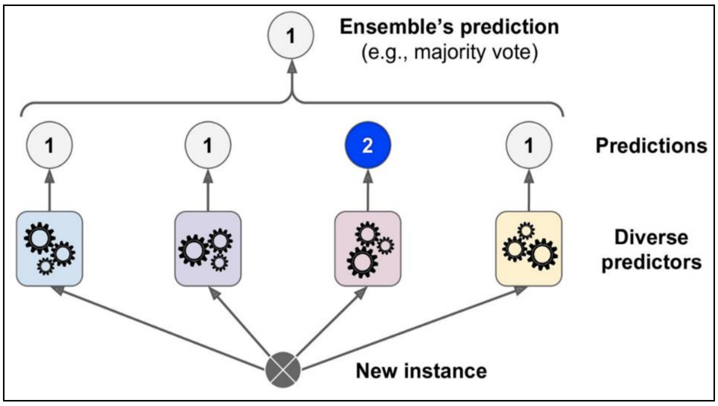

Fig. 3 *Ensemble method [75]*

2.3 Artificial Neural Network (ANN)

A neural network is described by Haykin [76] as a fundamentally distributed massively parallel processor made up of fundamental processing units that have a natural preference for accumulating and exchanging experience data. The network makes use of synaptic weights, which are also known as the strengths of interneurons' connections to one another to store the knowledge it has learned from its environment. Layers of connected neurons make up an ANN. Numerous tiny nodes or neurons, which are the small neuron processing units that make up each layer, communicate with one another through weighted numerical connections. There are n layers of neurons in it, with two serving as input and output layers. Only the first layer is capable of receiving and sending outside signals, while the last layer is in charge of sending calculations' results. When incoming signals are processed by relays, the n-2 hidden layers that are the deepest take out any relevant characteristics. After that, the yield layer stays adjusted so that it conforms to the necessary qualities. To maximize the network's capacity for spotting significant data or patterns, complex neural networks may include some hidden layers, feedback loops, and time-delay elements. Using a single layer for input, hidden, and output, Fig. 4 depicts the simple architecture of a typical ANN [77].

According to Bello et al. [78], ANNs enable the examination and diagnosis of nonlinear behaviors in compound systems that may be investigated, and operators and decision-makers can use them as an effective performance evaluation tool. ANNs can be taught to learn from prior examples and uncover intricate practical correlations among the data supplied, even if the underlying links are difficult to articulate or unknown. This is made feasible by their education. These methods enable the representation of complex physical processes, including those with nonlinear, high-order, and time-varying dynamics, as well as those for which analytic models are not yet widely available.

A basic neural network consists of three parts, as shown in Fig. 4: the input layer (layer L1), the hidden layers (layers L2, L3), and the output layer (layer L4). Each layer has its weight and the weight changes as the model go through the hidden layers until it reaches the output layer. The components of each layer are completely independent of each other. As Fig. 4 shows, more hidden layers would result in different outputs. The performance of the obtained ANN is based on its ability to generalize from the training to the validation data set. Neural networks with more than two hidden layers are known as deep neural networks (DNN) [79].

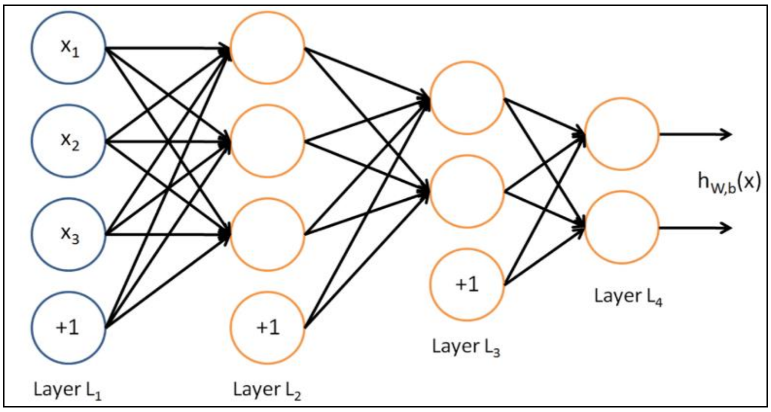

Fig. 4 *Structure of ANN [80]*

2.4 Support Vector Machines (SVM)

Using the concept of statistical learning, Vapnik developed support vector machines (SVMs) in the late 1960s. SVMs are a method for supervised machine learning that allows for both classification and regression data analysis, but their main use is for classification. As processing power became more widely available in the middle of the 1990s, they started to emerge, opening the door for several useful applications. It bases its operations on the reduction of structural risk. A hyperplane and a group of support vectors are used to divide the data into two classes in a basic SVM model [81]. To be thorough, a basic overview of SVM is provided below. Between two data sets, the SVM can be viewed as creating a classification line or hyperplane. In a two-dimensional context, the SVM's process can be basically explicated without sacrificing simplification. Class A is represented by a series of circles, whereas class B is represented by a series of squares (class B). The SVM searches for and places a straight boundary (solid line) between the categories to increase the margin (shown by dashed lines). The SVM aims to find the border in each class as close to the nearest point of data as possible. After that, the boundary is established in the center of this region. The boundary is then placed in the center of this area [73]. The closest data points are used as support vectors for establishing the limits (SV, represented by a gray circle and square).

The remaining features in the feature set can be discarded once the support vectors have been chosen because they provide all of the classification data required. By utilizing structural risk reduction to solve a constrained quadratic optimization problem, SVM splits data transversely the choice borderline of the hyperplane f(x) = 0. The objects with distinct labels that correspond to positive and negative classifications are included in the provided data input (xi I = 1, 2, N). The separation hyperplane is defined by the vector W and scalar b from Fig. 4. The separating hyperplane that produces the largest margin or separation between the plane and the data that is nearest to it is the best one. When the kernel function is used, SVM can be applied to non-linear classification tasks [82]. Due to the management of non-linearly separable features, working in a high-dimensional feature space presents challenges that can be overcome with the kernel function. Because it specifies the feature space that will be used to classify the training dataset, selecting the right kernel function is essential [83]. An illustration of how SVM classifies data is shown in Fig. 5.

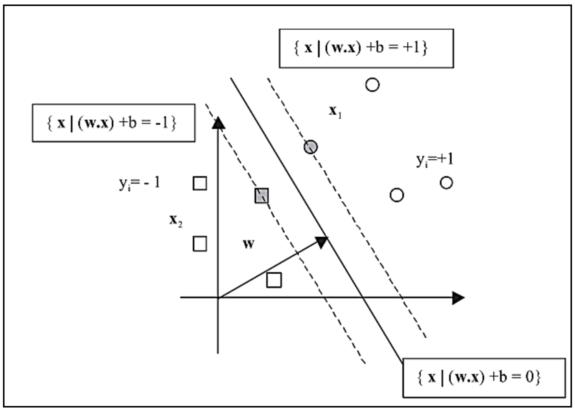

Fig. 5 *Classification of data with SVM*

2.5 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a distance-based classification technique developed by Fukunaga et al. [84]. It uses labeled samples stored in a training phase and labeled samples from that class to determine whether a sample belongs to a particular class. In effective methods of classification and regression, the contributions of neighbors are given weights. This means that neighbors who are closer to one another contribute more to the average than neighbors who are further apart. For instance, one common weighing technique uses a distance measure (d) to give each neighbor a weight of 1/d. For both K-NN classification and K-NN regression, the neighbors are drawn from a pool of examples whose classes or property values are already known. We could think of this as the algorithm's training set, though it is not strictly necessary. In Fig. 6, we see a KNN example in action.

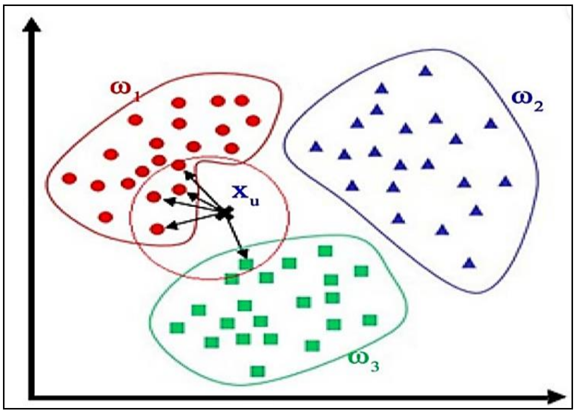

Fig. 6 *Illustration of KNN [80]*

Table 1 *Comparison of strength and weakness of algorithms*

Algorithm	Strength	Weakness
Fine Decision Tree	Interpretability: Decision trees are inherently interpretable, making it easier to understand the decision-making process. Handling Non-Linearity: Effective in capturing non-linear	Overfitting: Fine decision trees may be prone to overfitting, capturing noise in the data and reducing generalization to unseen data.

	relationships within the data.	
Medium Decision Tree	Interpretability: Similar to the fine decision tree, it provides a clear and interpretable decision structure	Reduced Complexity: The medium tree might sacrifice some complexity, potentially missing finer patterns in the data.
Bagged Trees (Ensemble):	Improved Accuracy: Ensemble methods, like bagging, reduce overfitting and enhance model accuracy by aggregating multiple models. Robustness: Less sensitive to noise and outliers due to averaging predictions across multiple trees.	Lack of Interpretability: Ensemble methods can be less interpretable compared to individual decision trees.
RUS-Boosted Trees:	Boosted Accuracy: Boosting focuses on correcting errors made by previous models, leading to improved accuracy. Handles Class Imbalance: Particularly effective when dealing with imbalanced datasets.	Sensitivity to Noisy Data: RUS-Boosted trees may be sensitive to noisy data, potentially resulting in overfitting.
Boosted Trees:	Accuracy Improvement: Boosting helps in improving the accuracy of the model by giving more weight to misclassified instances. Versatility: Can be applied to various types of data	Computational Complexity: Training boosted trees can be computationally expensive, especially with large datasets.

The research findings, elucidating the strengths and weaknesses of various classification algorithms for pump failure prediction based on vibration data, carry significant implications for industrial maintenance and safety. The research findings, particularly the success of the Ensemble method, pave the way for optimized predictive maintenance strategies. Implementing the Ensemble method can enhance the accuracy of pump failure predictions, allowing for proactive maintenance interventions. This, in turn, minimizes unexpected breakdowns, reduces downtime, and extends the lifespan of industrial equipment. Maintenance schedules can be optimized based on more accurate predictions, leading to cost savings and improved operational efficiency.

The detailed insights into pump failure dynamics, obtained through the classification of vibration data into categories like NORMAL, BROKEN, and RECOVERING, enable early fault detection. The Ensemble method's ability to distinguish patterns associated with different failure modes empowers maintenance teams to identify issues at their nascent stages. Early intervention can prevent catastrophic failures, ensuring the safety of personnel and minimizing the risk of accidents or damage to equipment. Understanding the strengths and weaknesses of each algorithm allows industrial practitioners to make informed decisions based on the specific requirements of their maintenance scenarios. For instance, the interpretable nature of Decision Trees could be preferred in situations where transparency in decision-making is crucial. On the other hand, the high accuracy of Ensemble methods may be prioritized in scenarios where predictive accuracy is paramount. By selecting the most appropriate algorithm for pump failure prediction, industrial maintenance teams can allocate resources more efficiently. This includes manpower, spare parts, and equipment. Proactively addressing potential failures reduces the need for reactive, emergency responses, leading to cost savings and an overall more efficient allocation of resources.

The implementation of advanced algorithms requires a skilled workforce. The research findings highlight the importance of understanding algorithmic strengths and weaknesses. Organizations can invest in training programs to equip their personnel with the skills needed to leverage these algorithms effectively, ensuring the successful implementation of predictive maintenance strategies. The research findings underscore the dynamic nature of predictive maintenance. Regularly reassessing and adapting algorithms based on evolving datasets and technological advancements is essential. Continuous improvement efforts can be guided by a nuanced understanding of each algorithm's performance in different contexts, ensuring that industrial maintenance practices remain at the forefront of safety and efficiency. In summary, the research findings offer a roadmap for implementing advanced predictive maintenance strategies in industrial settings. By leveraging the strengths of specific algorithms and addressing their limitations, organizations can enhance safety, optimize maintenance practices, and foster a culture of continuous improvement in the realm of industrial maintenance and safety.

3. Methodology

The proposed methodology consists of the following steps as shown in Fig. 7 below.

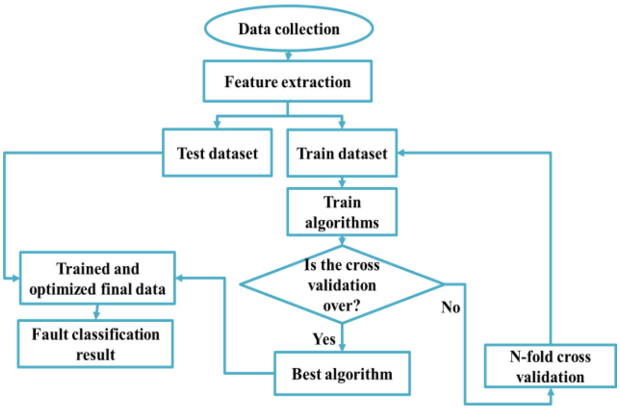

Fig. 7 *Methodology flow of classification*

An example of algorithm model for the classification is shown in Fig .8.

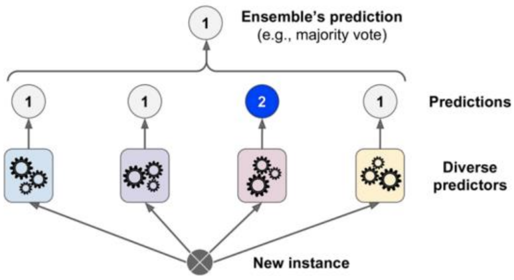

Fig. 8 *Ensemble algorithm model*

3.1 Data Collection

The dataset consisting of pump-related sensor data was provided by Kaggle and was utilized in this study. The data collection contains a timestamp, 52 sensor data, and machine status. In this section, the data from the sensors is recorded once every minute, and the state of the machine is also presented. There are 222,0320 data points and 55 characteristics included in the dataset. Within these 55 features, the timestamp and the machine status both belong to the object datatype, all sensor data belongs to the float datatype, and the Unnamed: 0 feature is an int datatype. Our investigation revealed that there are no overlapping records. The label data includes three different values for the machine state. The BROKEN status indicates that the machine has failed. The RECOVERING status indicates that the machine is attempting to recover from a failed state. The machine is operating normally, as indicated by the status indicator NORMAL.

3.2 Data Cleaning

The objective of machine learning (ML) and data mining (DM) is to extract actionable insights from sets of real-world data. Real-world data often comprises extraneous or inconsequential information known as "noise," which can have a substantial impact on a variety of data analyses Association analysis, classification, and clustering are machine learning tasks. It is evident that this the confusion must be addressed, as it negatively impacts virtually every type of data analysis. In machine learning datasets, two distinct forms of noise may be present: attribute noise, which affects the predictive attributes, and class noise, which affects the target attribute. Noise introduced into a dataset has the potential to augment the intricacy of models and lengthen the learning process, thereby impairing the efficacy of learning algorithms [2]. Consequently, it is essential to detect and manage this disturbance in data sets.

Since the given data is highly imbalanced, further preprocessing procedures are required. This is since unbalanced data has a major effect on the ML model and this will highly affect the accuracy of the prediction [216]. Hence, to fill in the missing gaps, and rectify the unbalanced nature of the data, the forward fill method is adopted in this study for the feature engineering as presented in section 4.1.3.

After studying the pattern of the missing data, the forward fill data imputation was used to fill the missing data instances. The reason for adopting the forward fill method for processing the data is that the pattern of the missing data was moving in a forward direction. Therefore, the forward fill will be the best option to employ for filling the missing gaps before the data is being trained in the ML model.

3.3 Normalize Value

The objective of normalization is to scale numerical data from various columns to a common scale. Standard Scaler normalization is employed. A standard Scaler normalizes a characteristic by removing the mean and then scaling to unit variance. Unit variance is calculated by dividing every value by the standard deviation.

3.4 Encode Label

The category value is changed to a numerical value in this step. The machine status values of NORMAL are mapped to 1, RECOVERING, and BROKEN are each mapped to the number 0, respectively.

3.5 Feature Selection

Feature selection is another method for reducing the number of features in a dataset. The feature selection method seeks to rank the value of existing characteristics in the dataset and eliminate less significant ones (no new features are created). Lowering the number of features used in a statistical analysis may result in numerous benefits including increased accuracy, risk of overfitting is reduced, increased training speed, improved data visualization, and improved model's explainability [217]. In this study, the principal component analysis (PCA) feature selection technique is used, and after that, the Synthetic Minority Over-Sampling Technique (SMOTE) technique is employed to smoothen the data. The primary difficulty encountered when utilising Principal Component Analysis (PCA) on a dataset is the determination of the optimal number of principal components. The process of optimising hyperparameters, despite its apparent complexity, can be effectively accomplished by utilising the GridSearchCV function within the sklearn module. The process of determining the optimal number of principal components is commonly referred to as a hyperparameter tuning procedure, wherein the hyperparameter n-components [218] is selected to achieve optimal performance. The present study utilises Principal Component Analysis (PCA) to derive both nonlinear and linear representations of the initial dataset within a reduced-dimensional space.

After the application of PCA for the feature selection, and the Synthetic Minority Over-Sampling Technique (SMOTE) is carried out to fine tune the data, the model is trained and tested again with the algorithms considered in this study. SMOTE is a method used in classification tasks to address the issue of imbalanced data [219]. In imbalanced data, the number of instances belonging to one class (the minority class) is significantly lower than the number of instances belonging to another class (the majority class). This can lead to biased models, where the majority class dominates the prediction results. SMOTE addresses this issue by oversampling the minority class by creating synthetic examples that are similar to the existing minority class examples. The synthetic examples are generated by interpolating between minority class examples, creating new samples that are located on the line segment connecting two existing minority class samples. The steps of the SMOTE method are as follows:

1. Select a minority class example to oversample.
2. Select one of its k nearest minority class neighbours.
3. Generate a synthetic example by interpolating between the selected example and its neighbour.
4. Repeat steps 1-3 until the desired number of synthetic examples have been generated.

After applying SMOTE, the dataset will be more balanced, and the classifier can be trained on the oversampled dataset. This approach can help to improve the classifier's performance on the minority class. Afterwards, the datasets are trained and tested using the intended algorithms for the study.

3.6 Selection and Validation of Models

Using a web exploration and the k-fold cross-validation method, the hyperparameters of the machine learning model were optimized. The accuracy of the model's performance in real-world scenarios was assessed during the results validation process using K-fold cross-validation, which also served to prevent overfitting. The loop repeats k times, using a different training set and validation set for each iteration. By averaging the results across all iterations, the scoring criteria for the model are established. The value of the k parameter was selected in such a way as to ensure that each fold of the k equal-sized subsets contains at least one failure event. As a result, data samples that occurred prior to the occurrence of a failure event were not divided across the folds. Due to the data's non-uniform distribution of failure events, an ideal value for k has been determined. A more thorough illustration of the system model for fault detection and classification is provided in the diagram below in Fig. 9.

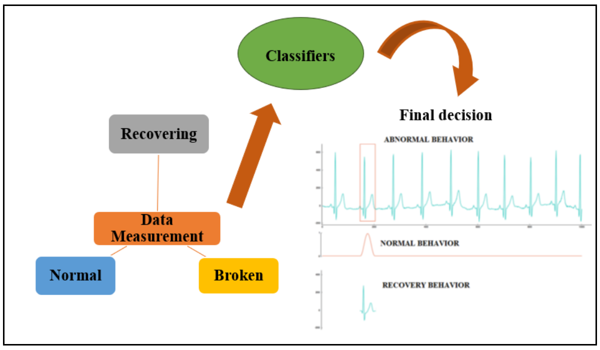

Fig. 9 *Fault detection and classification model*

3.7 Fault Classification

Faults are separated into their appropriate groups, and the classification accuracy percentage indicates the performance of the classifier. The flow of the classification process is shown in Fig. 10 below.

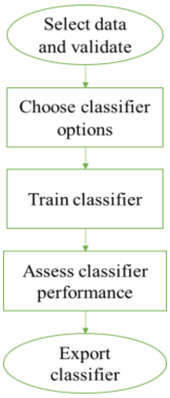

Fig. 10 *Workflow for training classification models*

3.8 Classification Accuracy

As a performance measure, the classification accuracy of the algorithms selected for the study is calculated using the equation 3.1 below. The accuracies of the trained and tested algorithms are presented in section 4, Table 1.

$$Accuracy = \frac{Number\ of\ correctly\ predicted\ data}{Total\ number\ of\ testing\ data} \times 100\% \quad (1)$$

4. Results and Discussion

This section provides a discussion of the obtained results from the study. Five different algorithms were employed to determine the fault of the pump and the results were assessed based on the classification accuracies (TPR/FNR, ROC AND AUC) and the classification accuracies were obtained using the formular provided in equation 3.1. The following sub sections explain the results obtained for the failure prediction for each algorithm.

4.1 Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic (ROC) Curve is an additional popular tool for depicting the accuracy of a classification algorithm. This study is the outcome of comparing the true negative rate to the false positive rate while changing the decision threshold. However, the precision and recall values for all five classifiers are not significantly differentiated from each other, and also there is no dominating relation between ROC curves in the entire range. In this situation, AUC (Area Under Curve) provides a good summary for comparing the classifiers. Ling et al. [85] also compared the accuracy and the AUC with different classifiers in various datasets. They conclude that the best tool for classifier comparison is AUC which helps users to better understand the performance of the classifiers. Table 2 depicts the comparison of the results obtained from the classification algorithms used in this study.

Table 2 *Comparison of the classification methods from classification learner*

Algorithm	Class	Training Accuracy	Test Accuracy	Area Under Curve (AUC)
Decision tree	Fine tree	91.2%	74%	0.80
	Medium tree	87.7%	70.4%	0.74
	Course tree	84.7%	53.5%	0.69
	Linear discriminant	65%	67.4%	0.75
Ensemble	Bagged trees	94.9%	99.9%	1.00
	Subspace discriminant	78%	68.3%	0.78
	RUS boosted trees	89%	72.6%	0.78
	Boosted trees	89.4%	72.2%	0.79
Neural network	Narrow NN	56.4%	77.4%	0.85
	Medium NN	71.4%	81.9%	0.83
	Trilayered NN	60.4%	79.1%	0.86
	Wide NN	72%	89.8%	0.97
	Bilayered NN	63.9%	78.2%	0.86
SVM	Cubic SVM	70.2%	84.8%	0.92
	Coarse gaussian SVM	66.3%	72%	0.81
	Fine gaussian SVM	72.2%	94.6%	0.99
	Medium gaussian SVM	69.4%	81.8%	0.89
	Quadratic SVM	69.3%	80.3%	0.87
	Linear SVM	61.1%	68.1%	0.77
KNN	Coarse KNN	71.7%	81.5%	0.90
	Cosine KNN	74.4%	93.3%	0.90
	Fine KNN	74.4%	99.9%	0.94
	Medium KNN	74.4%	93.2%	0.99
	Weighted KNN	74.3%	99.9%	1.00

The table is a summarized form of the results obtained for comparison of the performance of the algorithms used in the study. The five columns represent the algorithms employed in this study, the class of each algorithm used, the training accuracy, test accuracy, and the AUC. From the table, the Decision tree (Fine tree with training accuracy of 91.2%, test accuracy of 74%, and AUC 0.80. Medium tree with training accuracy of 87.7%, test accuracy of 70.4% and AUC 0.74). Ensemble (Bagged trees with training accuracy 94.9%, test accuracy 99.9%, and AUC 1.00. RUS-boosted trees had a training accuracy of 89%, test accuracy of 72.6%, and AUC of 0.78. while Boosted trees had a training accuracy of 89.4%, test accuracy of 72.2%, and AUC of 0.79. According to the results presented in the table, amongst the five algorithms implemented for this particular study, in terms of correctly classifying the defects, the DT and En algorithms were the most accurate. The rest of the algorithms (SVM, ANN, and KNN) showed lower performance.

Decision Trees (Fine and Medium) are known for their interpretability and ease of understanding. The fine tree, with higher accuracy, performed better on the test set than the medium tree. However, there is a trade-off between complexity and performance. Meanwhile, bagging (Bootstrap Aggregating) typically improves the stability and accuracy of a model. In this case, the ensemble of decision trees achieved high accuracy on both the training and test sets, with an AUC of 1.00, indicating excellent performance. Furthermore, both RUS-boosted and Boosted trees performed similarly, with accuracies and AUC values falling between those of the Decision Trees and the Ensemble. The implication on the obtained result is Decision trees, especially the fine one, and the ensemble of bagged trees stand out in terms of accuracy and AUC. These models seem well-suited for defect classification in this particular study. In comparison with Other Algorithms (SVM, ANN, KNN), the results indicate that SVM, ANN, and KNN showed lower performance compared to Decision Trees and the Ensemble. Table 3 on the other hand, shows the comparison of the true positive and false negative rate accuracies of the compared algorithms employed for the study.

Table 3 *TPR, FNR comparison*

Algorithm	Class	TPR			FNR		
		Broken	Normal	Recovering	Broken	Normal	Recovering
Decision tree	Fine tree	71.4%	79.5%	37.5%	28.6%	20.5%	62.5%
	Medium tree	63.6%	80.1%	22.7%	36.4%	19.9%	73.3%
	Course tree	0.1%	100.%	0%	99.9%	0.0%	100%
	Linear discriminant	51.7%	83.7%	24.4%	48.3%	16.3%	75.6%
	Quadratic discriminant	77.1%	62.2%	99.4%	22.9%	37.8%	0.6%
Ensemble	Bagged trees	99.8%	99.9%	100%	0.2%	0.1%	0%
	Subspace discriminant	43.5%	91.2%	29.8%	56.5%	8.8%	70.2%
	RUS boosted trees	71.4%	71.1%	99.7%	28.6%	28.9%	0.3%
	Boosted trees	58.1%	88.2%	18.4%	41.9%	11.8%	81.6%
Neural network	Narrow NN	74.1%	78.2%	97.3%	25.9%	21.8%	2.7%
	Medium NN	78.4%	83.0%	99.5%	21.6%	17.0%	0.5%
	Trilayered NN	72.1%	82.8%	99.1%	27.9%	17.2%	0.9%
	Wide NN	88.7%	89.8%	99.7%	11.3%	10.2%	0.3%
	Bilayered NN	72.3%	81.0%	98.4%	27.7%	19.0%	1.6%
SVM	Cubic SVM	77.2%	89.3%	99.8%	22.8%	10.7%	0.2%
	Coarse gaussian SVM	61.5%	84.3%	27.4%	38.5%	15.7%	72.6%
	Fine gaussian SVM	93.4%	95.1%	99.9%	6.6%	4.9%	0.1%
	Medium gaussian SVM	74.4%	85.9%	99.3%	25.6%	14.1%	0.7%
	Quadratic SVM	74.2%	83.3%	99.7%	25.8%	16.7%	0.3%
	Linear SVM	52.9%	83.4%	31.4%	47.1%	16.6%	68.6%
KNN	Coarse KNN	72.8%	87.2%	94.4%	27.2%	12.8%	5.6%
	Cosine KNN	93.2%	92.9%	99.7%	6.8%	7.1%	0.3%
	Fine KNN	99.9%	99.9%	100%	0.1%	0.1%	0%

Medium KNN	92.8%	93.0%	99.7%	7.2%	7.0%	0.3%
Weighted KNN	100%	99.8%	100%	0%	0.2%	0%

The comparison results affirm the study's success in advancing methodologies for vibration data analysis. The utilization of Decision Tree, Ensemble, Support Vector Machine, Neural Networks, and k-Nearest Neighbor techniques showcases a comprehensive approach to handling the challenges posed by the rapid growth and irregularity of vibration data. The models' varied performances suggest that the Decision Tree and Ensemble approaches offer particularly effective solutions. The achieved accuracy rates, as indicated in the results, signify a notable improvement in identifying and understanding patterns related to pump failure, supporting the notion that a diversified approach enhances the robustness of the analysis. The study's emphasis on evaluating classifier performance metrics, including true positive rate, false negative rate, and area under the curve, manifests as a significant contribution to the field. The results indicate that the Ensemble method outperformed others in terms of accuracy, with a near-perfect test accuracy of 99.9%. The meticulous comparison of different classification algorithms on the vibration dataset provides practitioners with valuable insights into algorithmic strengths and weaknesses. This empirical evaluation guides the selection of the most suitable algorithm for real-world applications, facilitating informed decision-making in predictive maintenance strategies. The study's classification of vibration data into categories of NORMAL, BROKEN, and RECOVERING yields rich insights into pump failure dynamics. The results highlight the effectiveness of the Ensemble method in achieving a flawless AUC of 1.00, emphasizing its capability to distinguish patterns associated with different failure modes. This understanding of pump failure dynamics is a crucial step towards proactive maintenance strategies. The fine-grained classification offered by the Ensemble method enables targeted interventions, potentially reducing downtime and enhancing operational efficiency. The study not only provides accurate predictions but also translates these predictions into actionable insights for effective decision-making in industrial settings.

5. Conclusion and Future Work

Classification algorithms for predicting pump failure were compared in this research. In the pursuit of accurately classifying defects, our study employed five distinct algorithms: Fine Decision Tree, Medium Decision Tree, Bagged Trees (Ensemble), RUS-Boosted Trees, and Boosted Trees. The analysis of the results revealed nuanced performances across these models. The Fine Decision Tree exhibited a commendable training accuracy of 91.2%, with a test accuracy of 74% and an AUC of 0.80. Its medium counterpart, while slightly less accurate, still demonstrated a respectable performance with a training accuracy of 87.7%, a test accuracy of 70.4%, and an AUC of 0.74. These results emphasize the reliability and interpretability of decision trees, with the fine-grained version proving particularly effective in defect classification. The Ensemble approach, represented by Bagged Trees, emerged as a standout performer, boasting a remarkable training accuracy of 94.9%, an almost perfect test accuracy of 99.9%, and a flawless AUC of 1.00. This underscores the efficacy of aggregating decision trees to enhance predictive accuracy and model robustness. On the boosted front, both RUS-Boosted Trees and Boosted Trees demonstrated comparable performances, each achieving a training accuracy of around 89%, test accuracies of 72.6% and 72.2%, respectively, and AUC values of 0.78 and 0.79. These results affirm the boosting methodology as a valuable technique, albeit with slightly lower accuracy compared to the Ensemble. In comparison, Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbors (KNN) lagged behind in terms of performance, according to the provided information. However, specific details about their results were not provided, limiting a comprehensive understanding of their relative strengths and weaknesses.

In conclusion, our findings suggest that, for defect classification in our specific context, the Fine Decision Tree, the Ensemble of Bagged Trees, and Boosted Trees stand out as the most accurate models. The Ensemble approach, in particular, exhibited exceptional accuracy and robustness. Future work could involve delving deeper into misclassifications, assessing the interpretability of models, and evaluating their generalization to external datasets. Moreover, a more detailed comparison with other algorithms would provide a holistic perspective on the strengths of each approach. These results pave the way for informed model selection and optimization in defect classification applications, contributing to the advancement of predictive modeling in this domain.

Acknowledgment

The authors would like to express their gratitude to Universiti Teknologi PETRONAS for supporting these research activities.

Author Contributions

Each of the authors worked on every aspect of the manuscript. R.A. conceived of the project and drafted the manuscript, and A.A.M. oversaw the effort. The manuscript was edited and reviewed by H.H. Each author has reviewed the final manuscript and agrees with its contents.

Funding

This study did not receive any funding from outside sources.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Commented [MABZ1]: Use Acknowledgement Style

Reference

[1] T. Wuest, D. Weimer, C. Irgens, and K. D. Thoben, “Machine learning in manufacturing: Advantages, challenges, and applications,” *Prod. Manuf. Res.*, vol. 4, no. 1, pp. 23–45, 2016, doi: 10.1080/21693277.2016.1192517.

[2] J. S. Rapur and R. Tiwari, “Automation of multi-fault diagnosing of centrifugal pumps using multi-class support vector machine with vibration and motor current signals in frequency domain,” *J. Brazilian Soc. Mech. Sci. Eng.*, vol. 40, no. 6, 2018, doi: 10.1007/s40430-018-1202-9.

[3] A. Shankar, V. Kalaiselvan, U. Subramaniam, and P. Shanmugam, “A comprehensive review on energy efficiency enhancement initiatives in centrifugal pumping system,” *Appl. Energy*, vol. 181, pp. 495–513, 2016, doi: 10.1016/j.apenergy.2016.08.070.

[4] A. H. Ã, S. Zhang, A. C. C. Tan, and J. Mathew, “Rotating machinery prognostics : State of the art , challenges and opportunities,” vol. 23, pp. 724–739, 2009, doi: 10.1016/j.ymssp.2008.06.009.

[5] D. Zhou, H. Zhang, and S. Weng, “A novel prognostic model of performance degradation trend for power machinery maintenance,” *Energy*, vol. 78, pp. 740–746, 2014, doi: 10.1016/j.energy.2014.10.067.

[6] A. Kumar, R. B. Chinnam, and F. Tseng, “An HMM and polynomial regression based approach for remaining useful life and health state estimation of cutting tools,” *Comput. Ind. Eng.*, vol. 128, no. May 2018, pp. 1008–1014, 2019, doi: 10.1016/j.cie.2018.05.017.

[7] J. R. Ruiz-Sarmiento, J. Monroy, F. A. Moreno, C. Galindo, J. M. Bonelo, and J. Gonzalez-Jimenez, “A predictive model for the maintenance of industrial machinery in the context of industry 4.0,” *Eng. Appl. Artif. Intell.*, vol. 87, no. October 2019, p. 103289, 2020, doi: 10.1016/j.engappai.2019.103289.

[8] M. S. Khaled, S. Imtiaz, S. Ahmed, and S. Zendehboudi, “Dynamic simulation of offshore gas processing plant for normal and abnormal operations,” *Chem. Eng. Sci.*, vol. 230, p. 116159, 2021, doi: 10.1016/j.ces.2020.116159.

[9] A. A. M. Ruwaida Aliyu, “Research Advances in the Application of FlexSim: A Perspective on Machine Reliability, Availability, and Maintainability Optimization,” *J. Hunan Univ. Nat. Sci.*, vol. 48, no. 9, 2021.

[10] M. Irfan *et al.*, “Condition monitoring of water pump bearings using ensemble classifier,” *Adv. Mech. Eng.*, vol. 14, no. 3, pp. 1–14, 2022, doi: 10.1177/16878132221089170.

[11] J. Ben Ali, B. Chebel-Morello, L. Saidi, S. Malinowski, and F. Fnaiech, “Accurate bearing remaining useful life prediction based on Weibull distribution and artificial neural network,” *Mech. Syst. Signal Process.*, vol. 56, pp. 150–172, 2015, doi: 10.1016/j.ymssp.2014.10.014.

[12] E. Mendel *et al.*, “Automatic bearing fault pattern recognition using vibration signal analysis,” in *2008 IEEE International Symposium on Industrial Electronics*, Jun. 2008, pp. 955–960, doi: 10.1109/ISIE.2008.4677026.

[13] M. A. Abu-Zeid and S. M. Abdel-Rahman, “Bearing problems’ effects on the dynamic performance of pumping stations,” *Alexandria Eng. J.*, vol. 52, no. 3, pp. 241–248, 2013, doi: 10.1016/j.aej.2013.02.002.

[14] J. Lee *et al.*, “Remaining Useful Life Estimation for Ball Bearings Using Feature Engineering and Extreme Learning Machine,” *IFAC-PapersOnLine*, vol. 55, no. 7, pp. 198–203, 2022, doi: 10.1016/j.ifacol.2022.07.444.

[15] M. Qiu, W. Li, Z. Zhu, B. Wu, and G. Zhou, “Health monitoring of rolling element bearing using a spectrum searching strategy,” *J. Vibroengineering*, vol. 19, no. 6, pp. 4231–4246, 2017, doi: 10.21595/jve.2017.18331.

[16] B. Wu, W. Li, and M. Q. Qiu, “Remaining Useful Life Prediction of Bearing with Vibration Signals Based on a Novel Indicator,” *Shock Vib.*, vol. 2017, 2017, doi: 10.1155/2017/8927937.

[17] H. E. Kim, A. C. C. Tan, J. Mathew, and B. K. Choi, “Bearing fault prognosis based on health state probability estimation,” *Expert Syst. Appl.*, vol. 39, no. 5, pp. 5200–5213, 2012, doi: 10.1016/j.eswa.2011.11.019.

[18] B. Yang, Y. Lei, F. Jia, and S. Xing, “An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings,” *Mech. Syst. Signal Process.*, vol. 122, pp. 692–706, 2019, doi: 10.1016/j.ymssp.2018.12.051.

[19] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, “Prognostics and health management design for

rotary machinery systems - Reviews, methodology and applications," *Mech. Syst. Signal Process.*, vol. 42, no. 1–2, pp. 314–334, 2014, doi: 10.1016/j.ymssp.2013.06.004.

[20] B. Samanta, "Gear fault detection using artificial neural networks and support vector machines with genetic algorithms," *Mech. Syst. Signal Process.*, vol. 18, no. 3, pp. 625–644, 2004, doi: 10.1016/S0888-3270(03)00020-7.

[21] D. Wang, K. L. Tsui, and Q. Miao, "Prognostics and Health Management: A Review of Vibration Based Bearing and Gear Health Indicators," *IEEE Access*, vol. 6, pp. 665–676, 2017, doi: 10.1109/ACCESS.2017.2774261.

[22] Z. Y. Wang, C. Lu, and B. Zhou, "Fault diagnosis for rotary machinery with selective ensemble neural networks," *Mech. Syst. Signal Process.*, vol. 113, pp. 112–130, 2018, doi: 10.1016/j.ymssp.2017.03.051.

[23] M. Issues, *Maintenance Issues 15.1*. 2019.

[24] D. Kumar and C. P. Garg, "Evaluating sustainable supply chain indicators using fuzzy AHP: Case of Indian automotive industry," *Benchmarking*, vol. 24, no. 6, pp. 1742–1766, 2017, doi: 10.1108/BIJ-11-2015-0111.

[25] S. S. Jolly and B. J. Singh, "An approach to enhance availability of repairable systems: A case study of spms," *Int. J. Qual. Reliab. Manag.*, vol. 31, no. 9, pp. 1031–1051, 2014, doi: 10.1108/IJQRM-02-2014-0016.

[26] M. A. Jirdehi and A. Rezaei, "Parameters estimation of squirrel-cage induction motors using ANN and ANFIS," *Alexandria Eng. J.*, vol. 55, no. 1, pp. 357–368, 2016, doi: 10.1016/j.aej.2016.01.026.

[27] D. Andrés and M. Salinas, "Predictive Maintenance of Induction Motors using Deep Learning Anomaly Detection using an Autoencoder Neural Network and Fault Classification using a Convolutional Neural Network," *Degree Proj. Technol.*, 2022.

[28] M. A. Hmida and A. Braham, "An On-Line Condition Monitoring System for Incipient Fault Detection in Double-Cage Induction Motor," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 8, pp. 1850–1858, 2018, doi: 10.1109/TIM.2018.2806009.

[29] A. S. Guedes, S. M. Silva, B. de J. Cardoso Filho, and C. A. Conceição, "Evaluation of electrical insulation in three-phase induction motors and classification of failures using neural networks," *Electr. Power Syst. Res.*, vol. 140, pp. 263–273, 2016, doi: 10.1016/j.epsr.2016.06.016.

[30] P. P. and A. G., "Fault Diagnosis of Centrifugal Pumps Using Motor Electrical Signals," in *Centrifugal Pumps*, InTech, 2012.

[31] A. Giantomassi, F. Ferracuti, S. Iarlori, G. Ippoliti, and S. Longhi, "Electric motor fault detection and diagnosis by kernel density estimation and kullback-leibler divergence based on stator current measurements," *IEEE Trans. Ind. Electron.*, vol. 62, no. 3, pp. 1770–1780, 2015, doi: 10.1109/TIE.2014.2370936.

[32] M. Irfan *et al.*, "A novel feature extraction and fault detection technique for the intelligent fault identification of water pump bearings," *Sensors*, vol. 21, no. 12, 2021, doi: 10.3390/s21124225.

[33] A. de la Fuente, C. A. Márquez, E. Candón, J. F. Gómez, and J. Serra, "A comparison of machine learning techniques for LNG pumps fault prediction in regasification plants," *IFAC-PapersOnLine*, vol. 53, no. 3, pp. 125–130, 2020, doi: 10.1016/j.ifacol.2020.11.020.

[34] X. Zhou, P. H. A. J. M. van Gelder, Y. Liang, and H. Zhang, "An integrated methodology for the supply reliability analysis of multi-product pipeline systems under pumps failure," *Reliab. Eng. Syst. Saf.*, vol. 204, no. March, p. 107185, 2020, doi: 10.1016/j.ress.2020.107185.

[35] T. LI, S. WANG, J. SHI, and Z. MA, "An adaptive-order particle filter for remaining useful life prediction of aviation piston pumps," *Chinese J. Aeronaut.*, vol. 31, no. 5, pp. 941–948, 2018, doi: 10.1016/j.cja.2017.09.002.

[36] M. Muhammad, M. A. A. Majid, and N. A. Ibrahim, "A case study of reliability assessment for centrifugal pumps in a petrochemical plant," *Eng. Asset Lifecycle Manag. - Proc. 4th World Congr. Eng. Asset Manag. WCEAM 2009*, no. September, pp. 398–404, 2009, doi: 10.1007/978-0-85729-320-6_44.

[37] L. Chen, L. Wei, Y. Wang, J. Wang, and W. Li, "Monitoring and Predictive Maintenance of Centrifugal Pumps Based on Smart Sensors," *Sensors*, vol. 22, no. 6, 2022, doi: 10.3390/s22062106.

[38] N. Hashim, A. Hassan, and M. F. A. Hamid, "Predictive maintenance model for centrifugal pumps under improper maintenance conditions," *AIP Conf. Proc.*, vol. 2217, no. April, 2020, doi: 10.1063/5.0000876.

[39] V. Muralidharan and V. Sugumaran, "A Comparative study between Support Vector Machine (SVM) and Extreme Learning Machine (ELM) for Fault Detection in Pumps," *Indian J. Sci. Technol.*, vol. 9, no. 48, 2017, doi: 10.17485/ijst/2016/v9i48/107915.

[40] W. G. Li, "A method for analyzing the performance of centrifugal oil pumps," *J. Fluids Eng. Trans. ASME*, vol. 126, no. 3, pp. 482–485, 2004, doi: 10.1115/1.1760544.

[41] C. O'Brien, J. Monteagudo, C. Schad, E. Cheung, and W. Middlesworth, "Centrifugal pumps and hemolysis in pediatric extracorporeal membrane oxygenation (ECMO) patients: An analysis of Extracorporeal Life Support Organization (ELSO) registry data," *J. Pediatr. Surg.*, vol. 52, no. 6, pp. 975–978, 2017, doi:

10.1016/j.jpedsurg.2017.03.022.
[42] J. Ma, C. Lu, W. Zhang, and Y. Tang, "Health assessment and fault diagnosis for centrifugal pumps using softmax regression," *J. Vibroengineering*, vol. 16, no. 3, pp. 1464–1474, 2014.
[43] W. Zhao, Z. Wang, C. Lu, J. Ma, and L. Li, "Fault diagnosis for centrifugal pumps using deep learning and softmax regression," *Proc. World Congr. Intell. Control Autom.*, vol. 2016-Septe, pp. 165–169, 2016, doi: 10.1109/WCICA.2016.7578673.
[44] D. Kimera and F. N. Nangolo, "Predictive maintenance for ballast pumps on ship repair yards via machine learning," *Transp. Eng.*, vol. 2, no. September, p. 100020, 2020, doi: 10.1016/j.treng.2020.100020.
[45] Y. Chen, J. Yuan, Y. Luo, and W. Zhang, "Fault Prediction of Centrifugal Pump Based on Improved KNN," *Shock Vib.*, vol. 2021, 2021, doi: 10.1155/2021/7306131.
[46] V. Muralidharan and V. Sugumaran, "Feature extraction using wavelets and classification through decision tree algorithm for fault diagnosis of mono-block centrifugal pump," *Meas. J. Int. Meas. Confed.*, vol. 46, no. 1, pp. 353–359, 2013, doi: 10.1016/j.measurement.2012.07.007.
[47] A. Azadeh, M. Saberi, A. Kazem, V. Ebrahimipour, A. Nourmohammadzadeh, and Z. Saberi, "A flexible algorithm for fault diagnosis in a centrifugal pump with corrupted data and noise based on ANN and support vector machine with hyper-parameters optimization," *Appl. Soft Comput. J.*, vol. 13, no. 3, pp. 1478–1485, 2013, doi: 10.1016/j.asoc.2012.06.020.
[48] N. R. Sakthivel, V. Sugumaran, and B. B. Nair, "Comparison of decision tree-fuzzy and rough set-fuzzy methods for fault categorization of mono-block centrifugal pump," *Mech. Syst. Signal Process.*, vol. 24, no. 6, pp. 1887–1906, 2010, doi: 10.1016/j.ymssp.2010.01.008.
[49] M. S. Lee, T. A. Shifat, and J. W. Hur, "Kalman Filter Assisted Deep Feature Learning for RUL Prediction of Hydraulic Gear Pump," *IEEE Sens. J.*, vol. 22, no. 11, pp. 11088–11097, 2022, doi: 10.1109/JSEN.2022.3167926.
[50] S. Perovic, P. J. Unsworth, and E. H. Higham, "Fuzzy logic system to detect pump faults from motor current spectra," *Conf. Rec. - IAS Annu. Meet. (IEEE Ind. Appl. Soc.*, vol. 1, no. C, pp. 274–280, 2001, doi: 10.1109/ias.2001.955423.
[51] Y. Luo, X. Zhixiang, H. Sun, S. Yuan, and J. Yuan, "Research on the induction motor current signature for centrifugal pump at cavitation condition," *Adv. Mech. Eng.*, vol. 7, no. 11, pp. 1–13, 2015, doi: 10.1177/1687814015617134.
[52] N. R. Sakthivel, B. B. Nair, and V. Sugumaran, "Soft computing approach to fault diagnosis of centrifugal pump," *Appl. Soft Comput. J.*, vol. 12, no. 5, pp. 1574–1581, 2012, doi: 10.1016/j.asoc.2011.12.009.
[53] S. Farokhzad, H. Ahmadi, and A. Jafary, "Fault Classification of Centrifugal Water Pump Based on Decision Tree and Regression Model," *J. Sci. today's world*, vol. 2, no. 2, pp. 170–176, 2013, [Online]. Available: www.journalsci.com.
[54] R. Aliyu, A. A. Mokhtar, and H. Hussin, "Prognostic Health Management of Pumps Using Artificial Intelligence in the Oil and Gas Sector: A Review," *Appl. Sci.*, vol. 12, no. 22, p. 11691, 2022.
[55] Kristoffer K. McKee, G. Forbes, I. Mazhar, R. Entwistle, and I. Howard, "A review of major centrifugal pump failure modes with application to the water supply and sewerage industries," *ICOMS Asset Manag. Conf. , Gold Coast, QLD, Aust.*, pp. 1–12, 2011.
[56] W. Han, L. Nan, M. Su, Y. Chen, R. Li, and X. Zhang, "Research on the prediction method of centrifugal pump performance based on a double hidden layer BP neural network," *Energies*, vol. 12, no. 14, 2019, doi: 10.3390/en12142710.
[57] P. Samanipour, J. Poshtan, and H. Sadeghi, "Cavitation detection in centrifugal pumps using pressure time-domain features," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 25, no. 5, pp. 4287–4298, 2017, doi: 10.3906/elk-1701-2.
[58] A. Kumar and R. Kumar, "Time-frequency analysis and support vector machine in automatic detection of defect from vibration signal of centrifugal pump," *Meas. J. Int. Meas. Confed.*, vol. 108, pp. 119–133, 2017, doi: 10.1016/j.measurement.2017.04.041.
[59] V. Muralidharan and V. Sugumaran, "Rough set based rule learning and fuzzy classification of wavelet features for fault diagnosis of monoblock centrifugal pump," *Meas. J. Int. Meas. Confed.*, vol. 46, no. 9, pp. 3057–3063, 2013, doi: 10.1016/j.measurement.2013.06.002.
[60] R. Azizi, B. Attaran, A. Hajnayeb, A. Ghanbarzadeh, and M. Changizian, "Improving accuracy of cavitation severity detection in centrifugal pumps using a hybrid feature selection technique," *Meas. J. Int. Meas. Confed.*, vol. 108, pp. 9–17, 2017, doi: 10.1016/j.measurement.2017.05.020.
[61] H. Li, H. Yu, N. Cao, H. Tian, and S. Cheng, "Applications of Artificial Intelligence in Oil and Gas Development," *Arch. Comput. Methods Eng.*, vol. 28, no. 3, pp. 937–949, 2021, doi: 10.1007/s11831-020-09402-8.
[62] W. Zhang, X. Li, and Q. Ding, "Deep residual learning-based fault diagnosis method for rotating machinery," *ISA Trans.*, vol. 95, pp. 295–305, 2019, doi: 10.1016/j.isatra.2018.12.025.
[63] Y. Liu, Z. Yu, M. Zeng, and Y. Zhang, "LLE for submersible plunger pump fault diagnosis via joint wavelet

and SVD approach," *Neurocomputing*, vol. 185, pp. 202–211, 2016, doi: 10.1016/j.neucom.2015.12.054.
[64] V. Muralidharan and V. Sugumaran, "A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis," *Appl. Soft Comput. J.*, vol. 12, no. 8, pp. 2023–2029, 2012, doi: 10.1016/j.asoc.2012.03.021.
[65] V. Muralidharan, V. Sugumaran, and V. Indira, "Fault diagnosis of monoblock centrifugal pump using SVM," *Eng. Sci. Technol. an Int. J.*, vol. 17, no. 3, pp. 152–157, 2014, doi: 10.1016/j.jestch.2014.04.005.
[66] S. Wang, J. Xiang, Y. Zhong, and H. Tang, "A data indicator-based deep belief networks to detect multiple faults in axial piston pumps," *Mech. Syst. Signal Process.*, vol. 112, pp. 154–170, 2018, doi: 10.1016/j.ymssp.2018.04.038.
[67] S. Wang, J. Xiang, H. Tang, X. Liu, and Y. Zhong, "Minimum entropy deconvolution based on simulation-determined band pass filter to detect faults in axial piston pump bearings," *ISA Trans.*, vol. 88, pp. 186–198, 2019, doi: 10.1016/j.isatra.2018.11.040.
[68] A. K. Panda, J. S. Rapur, and R. Tiwari, "Prediction of flow blockages and impending cavitation in centrifugal pumps using Support Vector Machine (SVM) algorithms based on vibration measurements," *Meas. J. Int. Meas. Confed.*, vol. 130, pp. 44–56, 2018, doi: 10.1016/j.measurement.2018.07.092.
[69] J. Kang, L. Zhang, and W. Liang, "Fault diagnosis of pipeline and pump unit systems using status coupling analysis," *J. Loss Prev. Process Ind.*, vol. 33, pp. 70–76, 2015, doi: 10.1016/j.jlp.2014.11.015.
[70] D. Buono, D. Siano, E. Frosina, and A. Senatore, "Gerotor pump cavitation monitoring and fault diagnosis using vibration analysis through the employment of auto-regressive-moving-average technique," *Simul. Model. Pract. Theory*, vol. 71, pp. 61–82, 2017, doi: 10.1016/j.simpat.2016.11.005.
[71] M. A. S. ALTobi, G. Bevan, P. Wallace, D. Harrison, and K. P. Ramachandran, "Fault diagnosis of a centrifugal pump using MLP-GABP and SVM with CWT," *Eng. Sci. Technol. an Int. J.*, vol. 22, no. 3, pp. 854–861, 2019, doi: 10.1016/j.jestch.2019.01.005.
[72] A. R. Al-Obaidi, "Investigation of effect of pump rotational speed on performance and detection of cavitation within a centrifugal pump using vibration analysis," *Heliyon*, vol. 5, no. 6, p. e01910, 2019, doi: 10.1016/j.heliyon.2019.e01910.
[73] O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, "Supervised Machine Learning Algorithms: Classification and Comparison," *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128–138, Jun. 2017, doi: 10.14445/22312803/IJCTT-V48P126.
[74] M. Tarik and K. Jebari, "Maintenance prediction by machine learning: Study review of some supervised learning algorithms," *Proc. Int. Conf. Ind. Eng. Oper. Manag.*, vol. 59, pp. 2678–2686, 2020.
[75] D. Sui, R. Nybø, G. Gola, D. Roverso, and M. Hoffmann, "Ensemble methods for process monitoring in oil and gas industry operations," *J. Nat. Gas Sci. Eng.*, vol. 3, no. 6, pp. 748–753, 2011, doi: 10.1016/j.jngse.2011.05.004.
[76] M. Kubat, "Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7.," *Knowl. Eng. Rev.*, vol. 13, no. 4, pp. 409–412, 1999.
[77] L. Biggio and I. Kastanis, "Prognostics and Health Management of Industrial Assets: Current Progress and Road Ahead," *Front. Artif. Intell.*, vol. 3, no. November, pp. 1–24, 2020, doi: 10.3389/frai.2020.578613.
[78] P. C. Pumps, "West Indian," *Definitions*, vol. 42, no. 2, 2020, doi: 10.32388/vuwxv5.
[79] B. Sundui, O. A. Ramirez Calderon, O. M. Abdeldayem, J. Lázaro-Gil, E. R. Rene, and U. Sambuu, "Applications of machine learning algorithms for biological wastewater treatment: Updates and perspectives," *Clean Technol. Environ. Policy*, vol. 23, no. 1, pp. 127–143, 2021, doi: 10.1007/s10098-020-01993-x.
[80] W. Zhang, "Machine Learning Approaches to Predicting Company Bankruptcy," *J. Financ. Risk Manag.*, vol. 06, no. 04, pp. 364–374, 2017, doi: 10.4236/jfrm.2017.64026.
[81] E. Ebrahimi and M. Javidan, "Vibration-based classification of centrifugal pumps using support vector machine and discrete wavelet transform," *J. Vibroengineering*, vol. 19, no. 4, pp. 2586–2597, 2017, doi: 10.21595/jve.2017.18120.
[82] A. K. Panda, J. S. Rapur, and R. Tiwari, "Prediction of flow blockages and impending cavitation in centrifugal pumps using Support Vector Machine (SVM) algorithms based on vibration measurements," *Meas. J. Int. Meas. Confed.*, vol. 130, pp. 44–56, 2018, doi: 10.1016/j.measurement.2018.07.092.
[83] P. F. Orrù, A. Zoccheddu, L. Sassu, C. Mattia, R. Cozza, and S. Arena, "Machine learning approach using MLP and SVM algorithms for the fault prediction of a centrifugal pump in the oil and gas industry," *Sustain.*, vol. 12, no. 11, 2020, doi: 10.3390/su12114776.
[84] K. Fukunaga, "Introduction to Statistical Pattern Recognition. Elsevier Academic Press. San Diego, San Francisco, New York," 1990.
[85] C. Bento, A. Cardoso, and G. Dias, *Lecture Notes in Artificial Intelligence: Introduction*, vol. 3808 LNCS. 2005.