Department of Industrial Engineering & Maintenance

Beyond Traditional Methods: Data Mining for Next-Generation Reliability Assessment

HIRECHE Zoulikha, SOUABNI Chaima

Master's thesis
in Industrial Maintenance Management and Engineering

HIRECHE Zoulikha
SOUABNI Chaima

Advisor:
GHOMARI Leila
REZGUI Wail

Academic year:
2023-2024

Abstract: The advent of Industry 4.0, marked by intricate machinery and systems, has brought to the fore an urgent need for robust reliability assessment methods. These methods are crucial to ensure the uninterrupted performance of industrial systems. While traditional techniques have been the go-to for reliability assessment, they often fall short of capturing the wealth of data that modern systems generate. This study explores the potential of data mining to enhance reliability assessment in industrial settings. Data mining offers powerful tools to discover hidden patterns and insights within this data. Unlike traditional methods, data mining can uncover these patterns without preconceived assumptions, leading to a more comprehensive understanding of system behavior. By leveraging these techniques, the goal is to enhance the reliability of industrial systems by uncovering hidden insights and patterns.

Key-Words: Data Mining, Reliability Assessment, Maintenance Management, Predictive Analytics, Machine Learning.

Contents

1. Introduction

The industrial landscape is witnessing various transformations characterized by faster innovation, personalized goods, greater flexibility, reduced hierarchies, and enhanced resource efficiency. To address these changes, industries have embraced a new level of value chain organization and control called Industry 4.0 [1]. As a result, industries have exhibited a heightened dependency on complex machinery and systems, ranging from software applications to mechanical engineering. This increased complexity requires a strong focus on reliability.

Reliability assessment is a systematic process that evaluates the performance of a product, system, or service under specific conditions. Its goal is to determine the ability of the entity to perform its intended function without failure over a defined period [1]. Since its inception in the 1950s, reliability theory has rapidly evolved, playing a pivotal role in advancing critical sectors such as aviation, aerospace, and nuclear energy. It has also significantly enhanced the quality of everyday items like computers, appliances, and vehicles. The ability to manufacture high-end equipment with high dependability and extended life has emerged as a key strategic indicator of a country's global power and competitiveness [2]. Therefore, ensuring and enhancing system reliability is of paramount importance.

Traditionally, reliability assessment has relied on established techniques varying from statistical to probabilistic methods to understand and quantify a system's reliability, availability, and maintainability. However, in today's data-rich environment, the growing volume and complexity of data generated by modern systems present challenges and opportunities for reliability assessment. While traditional methods provide a foundation, they may only partially capture the rich information hidden within this data.

Therefore, data mining emerges as a powerful tool to address these limitations of traditional methods. Data mining is a multidisciplinary field designed to handle various data's vast and complex characteristics. In the reliability assessment context, data mining is a vital facilitator. By leveraging data mining algorithms, organizations can discover hidden patterns (or rules) and information from the existing data [3]. These insights can then be used to assess and improve system reliability. Furthermore, data mining techniques offer distinct advantages over traditional methods. Unlike traditional approaches that rely on pre-conceived assumptions or specific data distribution, data mining can uncover hidden patterns without strict preconditions [4] [5].

This study aims to explore the diverse data mining techniques for reliability assessment. This will involve identifying a set of data mining algorithms suitable for analyzing the data generated by industrial machinery. We will also include a description of significant works addressing the application of these techniques, aiming to improve the reliability of industrial systems by uncovering hidden patterns and insights.

The remaining part of the paper is organized as follows: section 2 is about the fundamentals of reliability assessment, highlighting traditional methods and their associated challenges. Section 3 delves into data mining and the exploration of its diverse techniques, emphasizing their relevance to reliability assessment. This is followed by the 4th section, which discusses related works in data mining algorithms for reliability assessment. The 5th section is a discussion section, in which we present our findings of data mining analysis in the context of reliability assessment. Lastly, we conclude this study in section 6 by highlighting its findings and limitations.

2. Reliability

Reliability, a fundamental engineering and statistical analysis concept, is essential in industry and technology. It has always been a critical aspect in assessing industrial products and/or equipment [6], ensuring the ability of those numerous equipment, processes, and systems to perform their required functions under required conditions and over specific durations. This ability is commonly measured using probabilities. Reliability is, therefore, the probability that the complementary event will occur to failure, resulting in:

$$Reliability = 1 - FailureProbability[7] \quad (1)$$

Reliability is vital for maintaining operational continuity, ensuring safety, and upholding quality standards. Inadequate reliability considerations can result in severe consequences, leading to catastrophic events such as civil aviation disasters, nuclear power plant accidents, spacecraft launch failures, power system shutdowns, and other significant accidents [2]. These incidents highlight the crucial role of robust reliability assessments and maintenance protocols in mitigating risks and protecting against future breakdowns and potential failures.

Over the years, various reliability assessment methods, including qualitative and quantitative methods, have been developed to mitigate these risks and ensure the smooth operation of complex systems.

Qualitative methods employ models, diagrams, or other visual representations to analyze and understand the reliability of a manufacturing system and its components [1]. They provide a structured framework for identifying probable failure modes, evaluating their consequences, and prioritizing mitigation activities based on severity and likelihood of occurrence. These models are typically based on the experience of experts in the field. They may involve the use of Fault Tree Analysis (FTA), Failure Mode and Effects Analysis (FMEA), its extension, Failure Mode and Effects and Criticality Analysis (FMECA), and Hazard and Operability Study (HAZOP), or other similar techniques to evaluate the likelihood of different failure scenarios [1].

On the other hand, **quantitative methods** involve mathematical models, statistical analysis, and probabilistic techniques to assess the reliability of a system and its components [1]. Examples of these quantitative methods include Reliability Block Diagram (RBD), Markov Analysis, and Weibull Analysis. These methods are more data-driven since they rely heavily on the availability and quality of historical data. This allows for the calculation of specific metrics like Mean Time Between Failures (MTBF), Mean Time to Failure (MTTF), failure rate, and availability, providing numerical values that, in turn, offer insights into the reliability performance of a system, allowing engineers to make informed decisions regarding maintenance schedules, design improvements, and overall system optimization.

One significant advantage of qualitative approaches is their capacity to detect probable failure modes and evaluate their influence on system performance. In contrast, quantitative approaches give a more thorough and exact assessment of system dependability based on actual data and statistical analysis [1]. However, combining qualitative and quantitative methodologies can give the most complete and valuable insights into a system's reliability in many cases and practices.

Figure 1 provides a comprehensive illustration of various reliability assessment methods, highlighting their significance within the context of this study.

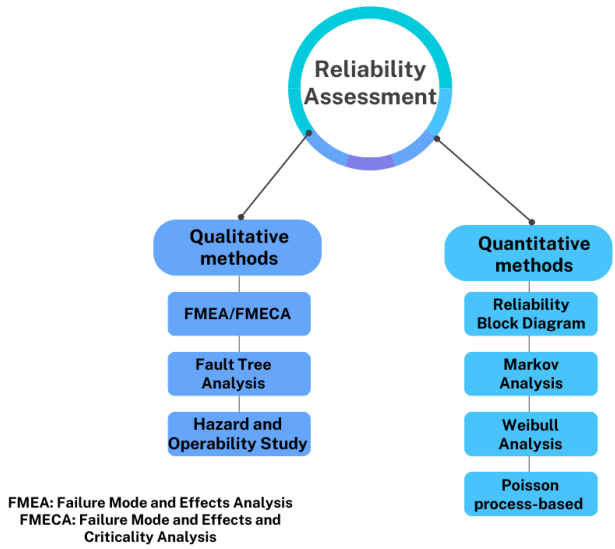

Figure 1: Reliability Assessment Methods.

The limitations of traditional reliability assessment methods, particularly their dependence on readily available and high-quality data, highlight the need for more robust approaches. In the reliability engineering realm, data's significance cannot be overstated. However, the quality of data collected in many industrial settings can be unreliable, hindering the effectiveness of traditional methods.

This is where **data mining** emerges as a powerful tool. Data mining techniques can address the challenge of uncertain and inconsistent data commonly found in CMMS (Computerized Maintenance Management System) systems, as highlighted by [4]. By leveraging sophisticated algorithms like K-means and association rules, data mining can extract valuable insights and hidden patterns even from unreliable datasets, as demonstrated by [8] and [9] in [4].

Data mining offers a pathway to transform unreliable data into a valuable asset for reliability assessment. The following section will explore data mining, its stages, and its techniques in more detail.

3. Data mining

Since the 80s, more data has been generated, driven by advancements in technology, the advent of the internet, and the digitization of various industries. Moreover, data alone is useless given that it consists of raw, unrefined, and commonly unfiltered information [10] [11] [12]. Thus, the transformation of data into useful and valuable information is needed to discover insights and create knowledge and information-based decision-making [10] [13] [14]. Knowledge discovery from data, described as data mining, is finding and extracting critical information from the data gathered [10] [15].

At a granular level, data mining represents a pivotal step in Knowledge Discovery in Databases (KDD), extracting previously unknown information and understandable hidden patterns in data [16]. While data mining and KDD are occasionally used interchangeably, they are commonly recognized as distinct entities. Data mining is a common catchphrase for data analysis. Since the evolution of data warehousing technology, the acceptance of data mining methods has quickly accelerated over the last ten years [3]. So, data mining is the process of sorting large data sets to identify patterns and relationships that can help solve business problems through data analysis. This entails exploring large batches of raw data to reveal concealed trends, employing advanced analytics techniques within the broader field of data science. Moreover, Data mining improves decision-making ability. It can target datasets and predict outcomes using machine-learning techniques [3]. Thus, this synergy between reliability and Data mining (DM) signifies a paradigm shift in industrial maintenance practices.

3.1. Data Mining Steps

The data mining process is broadly described through four steps that support the data analysis process [17]. The different stages of the data mining process are presented in Table 1.

Data gathering : The data-gathering stage is essential to initiate the process, as collecting relevant data is done in three steps. Data is generally stored in data warehouses or in big data environments in an unorganized way. Hence, the collection of relevant data is initiated in the data-gathering step [18]. It is impor-

Data Mining Steps	Description
Data gathering	Helps in the collection of relatable data from data warehouses or data lakes.
Data preparation	Prepare the data for mining through error-fixing.
Mining the data	Implement AI and automation to sort the data according to the analysis.
Analyzing the collected data and making interpretations	Helps to establish relations and find patterns for improved decision-making.

Table 1: Stages of Data Mining.

tant to collect relevant data to achieve a reliable result for analysis. Therefore, the significance of the step is related to the overall data extensive data analysis process. After completing the step, a data scientist moves a data set to a data pool with associates' access to further data processing [19].

Data Preparation : The data preparation steps become decisive when preparing the data for mining. The steps begin with exploring the collected data and making a precise data set for further processing [20]. Additionally, a consistent data set is achieved through data preparation. Moreover, a data set is filtered in the data preparation process for further analysis by removing errors and arranging it to support the expected outcome of the analysis [21] [19].

Mining the data : After the data set is prepared, it is handled with algorithmic operations to verify the collected data and prepare it for further analysis [19]. Machine learning and artificial intelligence are implemented in data mining to achieve results related to the analysis. Hence, data mining is significant for sorting a data set and extracting reliable data for analysis with the help of artificial intelligence [22] [19].

Analyzing the collected data and making interpretations : The last step of data mining is extracting valuable insights from the collected and processed data. Moreover, the decision-making process for the business depends on the steps of analysis and interpretation of the collected data. Finding relations and patterns is essential to make data-driven decisions for a business [23]. Therefore, achieving sustainability for the business patterns and data relationships is analyzed in this step. Besides that, the presentation of patterns and relations through business intelligence tools is conducted to understand the ladder and make decisions accordingly [24] [19].

Figure 2 offers an overview of the data mining and analytics process, as referenced in [25], showcasing its integral role in the research framework.

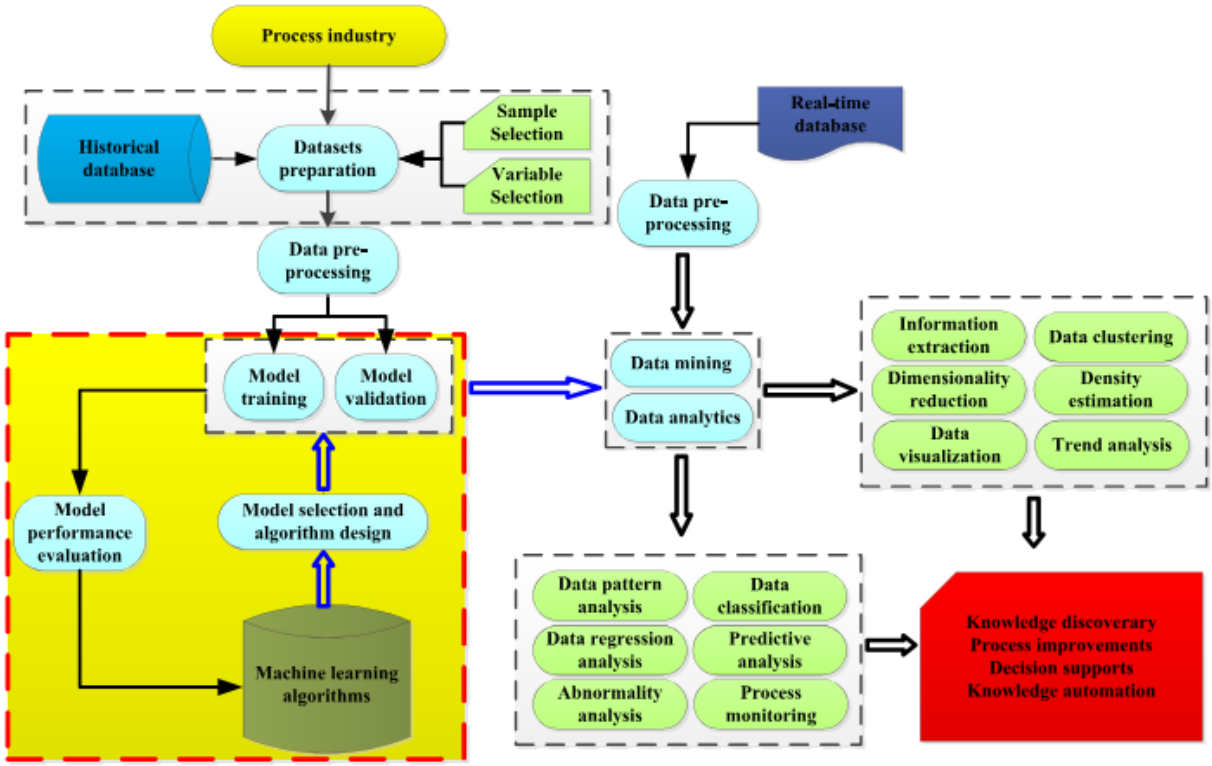

Figure 2: Overview of the data mining and analytics methodology [25].

3.2. Data Mining Techniques

This subsection will explore the principal data mining techniques used in reliability assessment. Each technique offers unique strengths in uncovering hidden patterns and insights. Those techniques can be classified into two main categories based on their learning approach: supervised and unsupervised. This classification is based on the nature of the data used for training and the learning goals of the algorithms employed. In the following, we will explore some primary data mining techniques used in reliability analysis.

Figure 3 below illustrates a hierarchy of data mining techniques mentioned in the section.

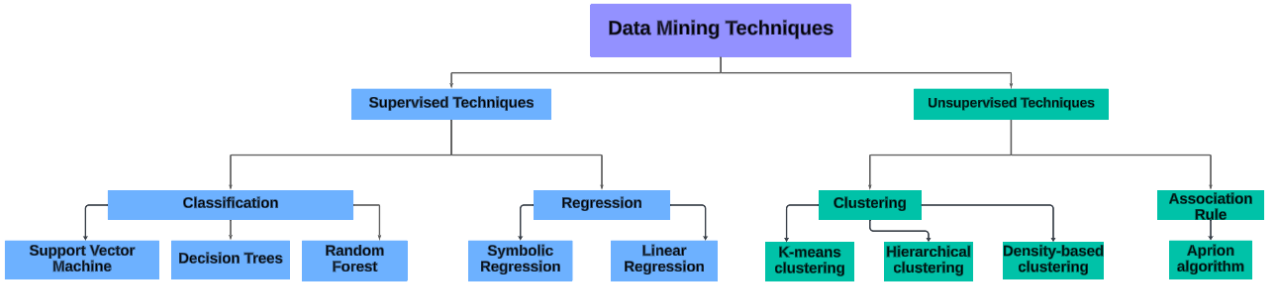

Figure 3: Data Mining Techniques.

3.2.1 Supervised Methods :

Supervised Learning Methods involve only labeled data (training patterns with known category labels) [26]. According to [27], supervised algorithms are performed for classification tasks when the goal is to predict the categorical class label of new instances based on past observations, on the other hand, they are used for regression tasks to predict a continuous numerical value based on input features.

- Classification: It aims to categorize each item in a set of input data into one of a predefined set of classes or groups. The data analysis task classification is where a model or classifier is constructed to predict categorical labels (the class label attributes) [28]. Within the context of reliability analysis, it is commonly used for generic failure prediction [29], prediction of gradual states of degradation [30], and simultaneous failure prediction. Notable algorithms within this domain include Decision Trees, Random Forest, and Support Vector Machines (SVM).

- **Decision Trees:** Decision trees consist of nodes that specify a particular attribute of the data, branches that represent a test of each attribute value, and leaves that correspond to the terminal decision of class assignment, for instance, in the dataset [31]. It facilitates understanding how individuals make decisions by adhering to the tree's structure. In the DT, the inner nodes represent the property, the branch represents the decision rule, and the leaf represents the outcomes [29].
- **Random Forest:** At its core, the RF method combines an ensemble of unpruned decision trees to achieve more efficient results with more than one decision maker, as in other methods [29] [32].
- **Support Vector Machine:** Shortened as SVM, is widely known to optimize the expected solution. Decision functions are determined directly from the training data using SVM so that the existing separation (margin) between the decision borders is maximized in a highly dimensional space called the feature space. This classification strategy minimizes the classification errors of the training data and obtains a better generalization ability, i.e., the classification skills of SVMs and other techniques differ significantly, especially when the number of input data is small [33].

- Regression: Regression in data mining is a predictive modeling technique that compares past successes and failures and then uses those formulas to predict future outcomes [34]. This is achieved by determining the relationship between one dependent variable and a series of other mutable (independent variables) [35]. Among the many algorithms that fall within the regression category, two stand out: Linear Regression (LiR) and Symbolic Regression (SR).

- **Linear Regression (LiR):** Refers to a multivariate linear combination of regression coefficients (i.e., constants and weights of input variables). The generalized least square technique estimates the coefficients. Given that linear regression is deterministic and parameterless, there is no need to configure anything other than a data split for model training and testing [36].
- **Symbolic Regression (SR):** Refers to models as a syntax tree consisting of arbitrary mathematical symbols (terminals: constants and variables, nonterminals: mathematical functions), which can be seamlessly translated to plain mathematical functions. For target estimation, syntax trees are evaluated top-down. Syntax trees are developed using evolutionary algorithms' stochastic genetic programming technique [36].

3.2.2 Unsupervised Learning :

Unsupervised Learning methods, unlike Supervised Learning, involve only unlabeled data [26], or we can say that in unsupervised learning, the desired output is not given [37]. The primary goal is to uncover in-

herent patterns within the unlabeled and then assign a label to each data value without the guidance of known outcomes. That means the techniques are appropriate for the clustering and association rule mining tasks. They are suitable for creating the labels in the data that are subsequently used to implement supervised learning tasks [38].

- Clustering: In data mining, clustering is more challenging than classification. An operational definition of clustering can be stated as follows: Given a representation of n objects, find K groups based on a similarity measure such that the similarities between objects in the same group are high. In contrast, the similarities between objects in different groups are low [26]. It is mainly used to detect anomalies by identifying patterns in the provided data.

The most common algorithms used for clustering tasks are K-means, Hierarchical clustering, and Density-based clustering, with Hierarchical clustering being the oldest among the three.

- **K-means clustering :** It is by far the most popular and used clustering technique. K-means clustering is a method that partitions n data points into k clusters. Each data point is assigned to the cluster with the nearest mean, the cluster center. K-means clustering aims to minimize the within-cluster variance, measured by the squared Euclidean distance [4]. The iterative process of K-means ends when the centroids no longer move, indicating convergence.

- **Hierarchical clustering :** As the name indicates, provides a hierarchical decomposition of the data, with a typical representation being a dendrogram. Hierarchical clustering techniques recursively find nested clusters in an agglomerative or divisive manner. Agglomerative clustering is one where each data point starts in its cluster and, after that, merges a similar pair of clusters successively, resulting in a hierarchy. Alternatively, divisive clustering starts with all the data points in one cluster and repeatedly divides each cluster into smaller ones. Once divisions or fusions are made, they cannot be reversed; thus, re-adjustment is impossible with hierarchical clustering [39].

- **Density-based clustering :** Unlike the K-means explained previously, does not divide the data into a predefined number of clusters. Instead, it focuses on identifying clusters based on the density of data points and then grouping densely packed points. The most popular implementation of Density-based clustering is the applications with Noise (DBSCAN).

- Association Rule : Association analysis, also known as association rule learning, aims to identify the association between seemingly uncorrelated data by measuring the degree of association between two variables [4]. It is generally ideal for extensive data handling and can model complex multi-way relationships given an adequate data set. Among the prominent association rule mining algorithms, the Aprion algorithm is one of the most well-known and utilized.

4. Related Works

This section addresses a literature review of existing research on applying data mining techniques for reliability assessment. The methodology employed for this review adhered to a systematic approach, starting with a comprehensive literature search carried out across well-known electronic scientific databases, including Google Scholar, ScienceDirect, IEEE Xplore, and Scopus.
The choice of keywords was based on terms that are common in the literature and related to this review, such as (“data mining” OR “data mining techniques” OR ‘’data analysis”) AND (“reliability” OR ‘’ maintenance reliability ‘’ OR “reliability assessment” OR ‘’ reliability analysis ‘’).
The quality assessment criteria that guide the search were:

- Inclusion: Papers written in English.
- Inclusion: Recent Works within the last 15 years.
- Inclusion: Papers that include the application of data mining techniques.
- Exclusion: theses (ongoing works) and books (mature works).
- Exclusion: Papers about reliability but not from a data mining perspective.

The survey results were collected in scientific databases, using keywords and applying the research criteria, and included 59 research papers.
In what follows, we will highlight the contributions of ten selected works that dealt with applying data mining techniques for reliability assessment (the most relevant ones).

- [4] investigates the causes of low-quality maintenance data in Computerized Maintenance Management Systems (CMMS) within the petrochemical industry. They argue that despite the potential of CMMS data for reliability analysis and equipment health assessments, poor data quality can significantly hinder these efforts. The main focus of this study is to apply data mining technology, including quality metrics, the association rule, and clustering, to identify the root causes of low-quality maintenance data. First, they establish a reference standard for CMMS function modules and data columns specific to the petrochemical industry. Second, they employ data quality metrics to assess the completeness and accuracy of maintenance records. Finally, data mining techniques like K-means clustering and association rule analysis explore operational and management practices contributing to poor data quality. The authors find that a combination of ineffective maintenance policies, the low integrity of key system columns, nonadherence to the policy, and misunderstanding of column definitions were the primary reasons behind the low-quality data in their case study. They conclude by proposing that their data mining-based method offers a systematic approach to identifying areas for improvement in CMMS data quality, ultimately leading to more reliable equipment health assessments and informed maintenance decisions.

- The research in [40] delves into reliability modeling for CNC systems, addressing a common challenge: incorporating failure correlations between different failure modes and causes. They argue that traditional methods relying solely on lifetime failure data miss these dependencies, leading to inaccurate models. Their work proposes a data mining-based approach for CNC systems that leverages association rule mining to identify these correlations. Using the Apriori algorithm, the study first establishes the relationship between failure positions and causes. This information is then used to define a failure correlation factor integrated into a Weibull distribution model for reliability estimation. The results demonstrate that this approach accurately represents the system's reliability more than models that neglect failure correlations. This work highlights the potential of data mining techniques like association rule mining to improve the accuracy of reliability models, particularly in scenarios with complex failure dependencies.

- [41] contributes to the reliability analysis in agricultural machinery, explicitly focusing on grain combine harvesters, by proposing an innovative approach based on Failure Mode, Effects, and Criticality Analysis (FMECA) coupled with Data Mining Technology. This study responds to the limitations of traditional FMECA methods, which can be subjective and expensive due to reliance on expert judgment and limited test data. The authors' approach leverages data mining techniques implemented in Python to collect and analyze fault data from the grain harvester header. Specifically, they use the Apriori algorithm to identify frequently occurring failure modes and the Analytic Hierarchy Process (AHP) to prioritize critical failure causes. Their results indicate that the cutter blade is the most critical component within the header assembly, highlighting the importance of prioritizing inspections and maintenance for this component. This study demonstrates the potential of data mining to improve the objectivity, efficiency, and cost-effectiveness of FMECA for agricultural machinery reliability analysis.

- In [42], a novel system identification methodology that integrates data mining techniques to enhance the reliability of identification processes was introduced. Traditionally, system identification involves comparing predicted and observed responses to determine the state of a system and its parameters. However, the reliability of such identification methods has yet to be studied in prior research. The proposed methodology addresses this gap by generating a population of candidate models and assessing their characteristics to gauge identification reliability. By leveraging data mining techniques, the methodology extracts features from the candidate models, providing insights into identification quality and facilitating improvements. This approach extends beyond structural engineering applications, offering potential utility across various domains. The study demonstrates the effectiveness of correlation measurements, principal component analysis (PCA), and decision trees in identifying key variables and separating good and bad models. Overall, this methodology is a valuable tool for engineers involved in monitoring and maintaining engineering systems, showcasing the potential of data mining in enhancing reliability analysis across diverse fields.

- While data mining is commonly used to build software fault prediction models, [43] explores its potential for broadly improving software reliability. They argue that traditional development practices rely on poorly documented APIs, leading to misuse and potential security vulnerabilities. Their work highlights the applicability of data mining techniques beyond just fault prediction. The authors propose leveraging data mining to identify usage patterns and relationships within Application Programming Interfaces (APIs). Specifically, they suggest techniques like association rule mining and clustering to uncover patterns that indicate correct API usage. This information can guide developers and improve software reliability by reducing API misuse and associated errors.

- Investigations by [44] introduce a novel approach to prognostics in the Aeronautics sector, based on classification algorithms. This approach focuses on binary classification tasks rather than the traditional ones to estimate the probability of failure in the upcoming timeframe, all this using advanced data mining algorithms such as Support Vector Machine (SVM) and Decision Trees (DT).

- In [45], random forest (RF) was employed as a classification algorithm, accompanied by two feature selection methods: a wrapper approach based on the beam search algorithm, as well as a new filter method based on the Kolmogorov–Smirnov test, the results of which are compared to those of a human expert. The work develops a data-driven method for predicting future failures of air compression in commercial vehicles. Furthermore, according to the authors, the techniques developed and tested to handle feature selection with inconsistent data sets, imbalanced and noisy class labels, and multiple examples per vehicle.

- [46] employed a Support Vector Machine (SVM) to develop a predictive maintenance module to predict integral-type faults. The model proposed in this research is specifically targeting ion-implantation tools. Furthermore, this method utilizes process iteration data to achieve real-time prediction. Although there are no comparisons between SVMs and other data mining techniques compare the cost. The proposed approach with classical preventive maintenance approaches. Hence, the authors state the potential of the recommended module to minimize overall maintenance costs.

- The authors in [47] employed another type of SVM, but in this case, it was for regression purposes called

Support Vector Regression (SVR). The study proposes a modification to the SVR kernel to address challenges in prognostic applications, aiming to improve the prediction accuracy of remaining useful life for industrial systems and equipment. The proposed model was tested on a simplified simulated Time-series data set and it showed improvement over the traditional SVR Formulation.

- [48] employed the K-means clustering method to automatically group the dissolved gas data from the insulating oil of a power transformer. The aim was to characterize each identified cluster, thereby indicating a specific fault or signaling a potential maintenance action. Utilizing the Euclidean distance as the criterion for similarity within the k-means algorithm, the researchers successfully delineated four distinct clusters: the first class revealed the presence of electric arcing with high energy, while the second cluster indicated an abnormal temperature rise of the oil: The third cluster highlighted a period characterized by an accelerated increase in the production of all gases, signaling potential operational issues or anomalies and the fourth cluster was associated with post-treatment periods of the oil, indicating maintenance or remedial actions to restore the oil properties and the transformer's performance.

Table 2 below presents a summary of studied papers, detailing the techniques/algorithms used, contributions, and limitations.

Additionally, the reviewed papers exhibit a variety of data mining techniques employed for improving system reliability, ranging from classification algorithms like random forests and support vector machines (SVMs) for fault prediction to association rule mining that uncovers relationships between variables to understand failure correlations. These studies, which draw on disciplines like data science, highlight the multidisciplinary character of reliability engineering. This fusion of knowledge allows researchers to identify potential failures and understand the underlying causes, leading to more targeted and effective preventative measures.
Despite the wide range of application domains—from software fault prediction and maintenance management to system identification—common contributions emerge, highlighting the integration of data-driven methodologies to improve reliability, spot system flaws, and develop new analytical strategies. Furthermore, the focus on specific application domains like software engineering, semiconductor manufacturing, and power transformers showcases the versatility of data mining techniques. This adaptability demonstrates the potential for broader application across various industries where system reliability is paramount.
These developments are accompanied, nonetheless, by a recognition of persistent limitations and challenges. Notably, the struggle to capture non-linear relationships between variables, data quality issues, and the integration hurdles of predictive maintenance systems highlight areas ripe for future exploration. Researchers are actively exploring solutions, with advanced techniques showing promise for complex data. Integrating domain expertise with data mining can also lead to more robust models. Addressing these challenges offers exciting avenues for advancement, necessitating the development of stronger algorithms, improved data preprocessing, and seamless integration strategies. Therefore, this synthesis not only demonstrates the breadth of contemporary reliability research but also offers insights into the ongoing push to expand the boundaries of this field.

5. Discussion

Incorporating data mining tools into reliability analysis significantly transforms industrial maintenance methods. Engineers can use data mining to enhance conventional reliability assessment techniques, providing more precise insights and facilitating proactive maintenance plans.
One of the primary challenges in reliability analysis lies in the quality of data [4] available for assessment. The reliability of equipment safety and the determination of inspection periods heavily rely on the quality of data extracted from Computerized Maintenance Management Systems (CMMS). However, it's widely acknowledged that CMMS maintenance records often need more quality, which can undermine reliability assessments. The causes of low data quality can be divided into two issues: one related to the system itself and the other related to operation and management [4]. This limitation has made engineers hesitant to use such data for decision-making processes, as discussed in [4]. Because of this, engineers should always look for methods and tools to make the most out of the data they use for their research.
Data mining offers a solution to the challenge of low-quality maintenance data, which in turn helps improve reliability by enabling engineers to extract valuable insights from large datasets, as mentioned in [19], even in the presence of noise and inconsistencies. Techniques such as K-means clustering, association rule analysis, classification methods, and others have proven invaluable in uncovering hidden patterns and relationships within the data, thereby facilitating the identification of root causes behind data quality issues. This may take the reliability assessment to another level and help gain time, effort, and precious information.
Although progress has been achieved in addressing problems related to data quality, several areas in the current research environment still offer exciting opportunities for further study because data mining for reliability has yet to be extensively addressed in prior research. While recent research has concentrated on creating methodologies for evaluating and enhancing data quality, more nuanced approaches tailored to particular industry contexts are still required. These approaches could create tools and methodologies for evaluating and enhancing data quality and other maintenance systems. Improving the accuracy and depend-

Reference	Techniques/algorithms used	Contribution	Limitations
[4]	Data mining technology, quality metrics, association rule, and clustering.	Integrates techniques to investigate low-quality maintenance records, showing promising results.	Additionally, the study identifies deficiencies in the CMMS system, such as issues related to policy, software, and operation dimensions, leading to conflicts, low-quality records, and erroneous data accumulation.
	Review of RAGAGEP, failure analysis standards, and data quality metrics.	Identifies deficiencies in software design, data quality trends, and policy shortcomings. Offers a systematic analysis of maintenance record quality that is more objective than expert-based methods. Helps improve maintenance record quality, reducing resistance to equipment diagnosis.	Data cleaning and conversion challenges can compromise the reliability and usefulness of the resulting data for failure analysis. The article also highlights the limitations of decision trees in handling combinations of variables that determine the classes of data points.
[40]	Association rule mining technique was used for reliability modeling.	Proposed reliability modeling based on degree of failure correlation.	Lack of objective basis for determining failure correlation factor.
	Apriori algorithm was employed for association rule mining.	Introduced failure correlation factor into parameter estimation for reliability modeling.	Reliability modeling mainly focused on failure time data processing.
	The maximum likelihood estimation method was used for parameter estimation.	Used association rule mining to study failure correlation in CNC system. Model with failure correlation factor suitable for multiple failure modes.	
[41]	FMECA analysis method based on Data Mining Technology.	Proposes FMECA analysis method based on Data Mining Technology.	Subjectivity, ambiguity, high test cost, and difficult data acquisition.
	Data visualization using Python for reliability research.	Identifies blade part as the most hazardous in grain harvester.	Problems in failure mode, influence, and hazard analysis.
	Data collection with Python Programming Language.	Enhances reliability analysis with Python for fault data processing. Integrates intelligent monitoring for equipment reliability and safety analysis.	Existing issues in reliability testing of agricultural machinery.
[42]	PGSL algorithm for stochastic global search in system identification.	Data mining techniques improve system identification reliability.	Cannot obtain relationships between more than two parameters simultaneously.
	PCA is used as a weighting method for model characteristics.	Identifying candidate models and reliability indications.	Linear data mining methods like PCA cannot reveal non-linear relationships.

			Decision trees struggle with combinations of linear or non-linear relationships. Techniques used cannot identify non-linear relationships between model variables.
[43]	Classification trees, association discovery, clustering, artificial neural networks, etc. Quinlan proposed the ID3 algorithm for classification tree induction.	Data mining techniques for software fault prediction and quality enhancement. Exploration of data mining algorithms for software reliability improvement.	Incomplete specifications can lead to perceived unreliability in software systems. Few works on clustering techniques applied to software engineering data.
[44]	Random forests, support vector machines, nearest neighbors, and deep learning techniques. Grid search, evolutionary search, signal processing, Principle Component Analysis (PCA). K-Nearest Neighbors (KNN), Gaussian Support Vector Machines (GSVM), Random Forests (RF).	Proposed novel classification models for prognostics without RUL estimates. Evaluated machine learning classifiers on real-world aeronautics case studies.	Traditional classifiers have low performance compared to deep learning methods. Dataset quality affects predictability, especially for challenging datasets.
[45]	Random forest classifier algorithm used for prediction models. Feature selection methods: Wear, Usage, Wear with age normalization, Usage with age normalization. LVD database analyzed for vehicle usage patterns and key parameters.	Predictive maintenance solutions for the automotive industry using machine learning. Comparison of machine learning features with human expert features. Challenges in predictive maintenance for heavy-duty vehicles discussed. Use of logged on-board data for predicting air compressor failures.	Data sources not designed for data mining. Loss of data over time due to lack of collection.
[46]	Classification methods for prediction of integral type faults. Support Vector Machines (SVMs) for dealing with classification problems. Linear SVMs and Radial Basis Function (RBF) SVMs for comparison.	Testing the PdM system on a real production dataset. Prediction of integral type failures in semiconductor manufacturing processes. Development of a PdM module for ion-source tungsten filament breaks. Application of Classification methods for predicting equipment failures.	Historical data-based PvM approach lacks current machine state utilization. Limited to predicting ion-source tungsten filament breaks in semiconductor manufacturing.
[47]	Support Vector Regression (SVR) with a modified regression kernel.	Proposed regression kernel for support vector regression in prognostics.	Data sets are unstructured with unique characteristics from multiple sources.

	Kernel-based approaches for time-series data sets.	Highlighted challenges in developing prognostic algorithms using machine learning. Reviewed traditional and modern data-driven approaches for prognostic applications. Suggested a regression kernel modification for support vector regression.	Time-series data poses challenges for traditional machine learning algorithms. Non-stationary characteristics of data are difficult to model. The precise definition of remaining useful life is unclear.
[48]	Principal Component Analysis (PCA) for variable identification. Clustering by k-means method for classification of dissolved gas data.	Analysis of power transformer operating periods through dissolved gas concentrations. Unsupervised classification of gas data to identify major events.	The challenges in data cleaning and conversion. The proposed methodologies, such as correlation measurement, PCA, and decision trees, are unable to capture non-linear relationships between model variables.

Table 2: Summary of Studied Papers.

ability of maintenance records involves investigating methods for data validation, anomaly detection, and error repair.

The stages of the data mining process—data gathering, data preparation, mining, and analyzing the collected data—provide a systematic framework for extracting meaningful insights from raw data. Each stage contributes to the overall reliability analysis by ensuring the derived insights' relevance, accuracy, and reliability.

Furthermore, data mining facilitates predictive analytics, which improves decision-making skills. Engineers can anticipate maintenance requirements and equipment breakdowns using machine learning techniques. This enables proactive interventions to prevent costly downtime and safety hazards. This predictive capacity turns maintenance procedures from reactive to proactive by eliminating operational risks and maximizing resource allocation.

Integrating data mining techniques into reliability assessment offers a transformative approach to maintenance practices. By overcoming the limitations of traditional methods and harnessing the power of big data, engineers can make more informed decisions, optimize maintenance schedules, and ensure the continued reliability and safety of industrial systems and equipment.

6. Conclusions

This research explores integrating data mining techniques into reliability assessment for industrial systems. The main finding highlights the transformative potential of data mining in enhancing maintenance practices, enabling engineers to make informed decisions, optimize schedules, and ensure system reliability and safety.

The study delves into various data mining techniques, showcasing their effectiveness in uncovering hidden patterns and relationships within large datasets. These methods, such as K-means clustering, association rule analysis, and classification methods, can identify root causes behind data quality issues and elevate reliability assessment to a new level. The predictive analytics enabled by data mining allows for proactive interventions, minimizing downtime and maximizing resource allocation.

This research demonstrates how data mining can revolutionize traditional methods, particularly in Industry 4.0, by overcoming the limitations of conventional techniques and leveraging the power of big data. Engineers can optimize maintenance procedures and ensure the continued reliability of industrial systems and equipment.

This paper's implications are significant for industries reliant on complex machinery and systems. Adopting data mining techniques can improve efficiency, lower operational risks, and improve maintenance practices. In the long run, switching from reactive to proactive maintenance techniques can result in lower costs and more output.

However, it is essential to acknowledge the limitations of this study. While data mining shows promise in enhancing reliability assessment, there are still challenges related to data quality and implementation. Future research should focus on developing more nuanced ap-

proaches tailored to specific industry contexts. Additionally, exploring the integration of data mining with other emerging technologies could further enhance the reliability and safety of industrial systems.

Acknowledgements

We would like to express our profound and sincere gratitude to both Mrs. GHOMARI Leila and Mr. REZGUI Wail for their guidance and unwavering support throughout our thesis. We consider ourselves very fortunate to be able to work with very considerate and encouraging professors like them and we will always cherish the countless hours they invested in reviewing our work, offering constructive feedback, and pushing us to strive for excellence. Their wealth of knowledge, dedication, and encouragement have been invaluable assets, providing us with the guidance and inspiration needed to navigate through the challenges of our research journey. Their mentorship has not only given us the chance to grow academically but has also ingrained in us a deeper appreciation for the pursuit of knowledge. We will always be thankful for their encouragement in each step of this journey.

References

[1] Jonas Friederich and Sanja Lazarova-Molnar. Reliability assessment of manufacturing systems: A comprehensive overview, challenges and opportunities. *Journal of Manufacturing Systems*, 72:38–58, 2024.

[2] Mingjian Zuo. System reliability and system resilience. *Frontiers of Engineering Management*, 8(4):615–619, 2021.

[3] Amitava Bondyopadhyay and Dr. A.C Mandal. Improvement of software reliability using data mining technique. *International Journal of Scientific and Research Publications (IJSRP)*, 12(06):183–187, June 2022.

[4] Yen-Ju Lu, Wei-Chen Lee, and Chen-Hua Wang. Using data mining technology to explore causes of inaccurate reliability data and suggestions for maintenance management. *Journal of Loss Prevention in the Process Industries*, 83:105063, 2023.

[5] Aurora Esteban, Amelia Zafra, and Sebastian Ventura. Data mining in predictive maintenance systems: A taxonomy and systematic review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(5):e1471, 2022.

[6] Andrew KS Jardine, Daming Lin, and Dragan Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, 20(7):1483–1510, 2006.

[7] Tran Van Ta, Doan Minh Thien, and Vo Trong Cang. Marine propulsion system reliability assesment by fault tree analysis. *International Journal of Mechanical Engineering and Applications*, 8(1):1–7, 2016.

[8] Adam Coates and Andrew Y. Ng. *Learning Feature Representations with K-Means*, pages 561–580. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[9] Krzysztof J. Cios, Roman W. Swiniarski, Witold Pedrycz, and Lukasz A. Kurgan. *Unsupervised Learning: Association Rules*, pages 289–306. Springer US, Boston, MA, 2007.

[10] Jovanne C Alejandrino, Jovito Bolacoy Jr, John Vianne B Murcia, et al. Supervised and unsupervised data mining approaches in loan default prediction. *International Journal of Electrical & Computer Engineering (2088-8708)*, 13(2), 2023.

[11] RA Evans. Information vs data, 1981.

[12] Jeanne Harris. Data is useless without the skills to analyze it. *Harvard Business Review*, 13, 2012.

[13] Varun Grover, Roger HL Chiang, Ting-Peng Liang, and Dongsong Zhang. Creating strategic business value from big data analytics: A research framework. *Journal of management information systems*, 35(2):388–423, 2018.

[14] Benjamin CM Fung, Ke Wang, Rui Chen, and Philip S Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (Csur)*, 42(4):1–53, 2010.

[15] Camilo Ernesto López Guarín, Elizabeth León Guzmán, and Fabio A González. A model to predict low academic performance at a specific enrollment using data mining. *IEEE Revista Iberoamericana de tecnologias del Aprendizaje*, 10(3):119–125, 2015.

[16] Toon Calders and Bart Custers. What is data mining and how does it work? In *Discrimination and privacy in the information society: Data mining and profiling in large databases*, pages 27–42. Springer, 2013.

[17] Wei Chen, Haoyuan Hong, Mahdi Panahi, Himan Shahabi, Yi Wang, Ataollah Shirzadi, Saied Pirasteh, Ali Asghar Alesheikh, Khabat Khosravi, Somayeh Panahi, Fatemeh Rezaie, Shaojun Li, Abolfazl Jaafari, Dieu Tien Bui, and Baharin Bin Ahmad. Spatial prediction of landslide susceptibility using gis-based data mining techniques of anfis with whale optimization algorithm (woa) and grey wolf optimizer (gwo). *Applied Sciences*, 9(18), 2019.

[18] S Iwin Thanakumar Joseph and Iwin Thanakumar. Survey of data mining algorithm's for intelligent computing system. *Journal of trends*

in Computer Science and Smart technology (TCSST), 1(01):14–24, 2019.

[19] Manish Sharma and Richa Gupta. The significance of using data extraction methods for an effective big data mining process. In *2023 2nd International Conference for Innovation in Technology (INOCON)*, pages 1–4. IEEE, 2023.

[20] Pooya Tabesh, Elham Mousavidin, and Sona Hasani. Implementing big data strategies: A managerial perspective. *Business Horizons*, 62:347–358, 03 2019.

[21] T Senthil Kumar. Data mining based marketing decision support system using hybrid machine learning algorithm. *Journal of Artificial Intelligence*, 2(03):185–193, 2020.

[22] Anurag Kumar Verma, Saurabh Pal, and Surjeet Kumar. Classification of skin disease using ensemble data mining techniques. *Asian Pacific journal of cancer prevention: APJCP*, 20(6):1887, 2019.

[23] Özerk Yavuz. A classification and clustering approach using data mining techniques in analysing gastrointestinal tract. *International Scientific and Vocational Studies Journal*, 5(2):254–265, 2022.

[24] Manish Sharma, Bhasker Pant, and Vijay Singh. Demographic profile building for cold start in recommender system: A social media fusion approach. *Materials Today: Proceedings*, 46:11208–11212, 2021.

[25] Zhiqiang Ge, Zhihuan Song, Steven X Ding, and Biao Huang. Data mining and analytics in the process industry: The role of machine learning. *Ieee Access*, 5:20590–20616, 2017.

[26] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[27] David R Musicant, Janara M Christensen, and Jamie F Olson. Supervised learning by training on aggregate outputs. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 252–261. IEEE, 2007.

[28] Gopalan Kesavaraj and Sreekumar Sukumaran. A study on classification techniques in data mining. In *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*, pages 1–7. IEEE, 2013.

[29] Mustafa Cakir, Mehmet Ali Guvenc, and Selcuk Mistikoglu. The experimental application of popular machine learning algorithms on predictive maintenance and the design of iiot based condition monitoring system. *Computers & Industrial Engineering*, 151:106948, 2021.

[30] Pablo Aqueveque, Luciano Radrigan, Francisco Pastene, Anibal S Morales, and Ernesto Guerra. Data-driven condition monitoring of mining mobile machinery in non-stationary operations using wireless accelerometer sensor modules. *IEEE Access*, 9:17365–17381, 2021.

[31] Michael D Twa, Srinivasan Parthasarathy, Cynthia Roberts, Ashraf M Mahmoud, Thomas W Raasch, and Mark A Bullimore. Automated decision tree classification of corneal shape. *Optometry and Vision Science*, 82(12):1038–1046, 2005.

[32] Yongheng Zhao and Yanxia Zhang. Comparison of decision tree methods for finding active objects. *Advances in Space Research*, 41(12):1955–1959, 2008.

[33] Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, and Asdrubal Lopez. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215, 2020.

[34] Prajak Chertchom. A comparison study between data mining tools over regression methods: Recommendation for smes. In *2018 5th International Conference on Business and Industrial Research (ICBIR)*, pages 46–50. IEEE, 2018.

[35] Salim Jibrin Danbatta and Asaf Varol. Predicting student's final graduation cgpa using data mining and regression methods: a case study of kano informatics institute. In *2020 8th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–7. IEEE, 2020.

[36] Jan Zenisek, Florian Holzinger, and Michael Affenzeller. Machine learning based concept drift detection for predictive maintenance. *Computers & Industrial Engineering*, 137:106031, 2019.

[37] Aized Soofi and Arshad Awan. Classification techniques in machine learning: Applications and issues. *Journal of Basic Applied Sciences*, 13:459–465, 08 2017.

[38] Michael W Berry, Azlinah Mohamed, and Bee Wah Yap. *Supervised and unsupervised learning for data science*. Springer, 2019.

[39] P. Govender and V. Sivakumar. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research*, 11(1):40–56, 2020.

[40] Guangpeng Liu and Chong Peng. Research on reliability modeling of cnc system based on association rule mining. *Procedia Manufacturing*, 11:1162–1169, 2017.

[41] YANG Xiaohui, Guohai ZHANG, Yao Jia, LIAN Jitan, WANG Xin, LV Danyang, DENG Yujie, and Aoqi ZHANG. Reliability analysis of grain

combine harvesters based on data mining technology. *INMATEH-Agricultural Engineering*, 67(2), 2022.

[42] Sandro Saitta, Benny Raphael, and Ian FC Smith. Data mining techniques for improving the reliability of system identification. *Advanced Engineering Informatics*, 19(4):289–298, 2005.

[43] Nadhem Sultan Ali and VP Pawar. The use of data mining techniques for improving software reliability. *International Journal of Advanced Research in Computer Science*, 4(2), 2013.

[44] Marcia L Baptista, Elsa MP Henriques, and Helmut Prendinger. Classification prognostics approaches in aviation. *Measurement*, 182:109756, 2021.

[45] Rune Prytz, Sławomir Nowaczyk, Thorsteinn Rögnvaldsson, and Stefan Byttner. Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data. *Engineering applications of artificial intelligence*, 41:139–150, 2015.

[46] Gian Antonio Susto, Sean McLoone, Daniele Pagano, Andrea Schirru, Simone Pampuri, and Alessandro Beghi. Prediction of integral type failures in semiconductor manufacturing through classification methods. In *2013 IEEE 18th Conference on Emerging Technologies & Factory Automation (ETFA)*, pages 1–4. IEEE, 2013.

[47] Josey Mathew, Ming Luo, and Chee Khiang Pang. Regression kernel for prognostics with support vector machines. In *2017 22nd IEEE international conference on emerging technologies and factory automation (ETFA)*, pages 1–5. IEEE, 2017.

[48] Samuel Eke, Thomas Aka-Ngnui, Guy Clerc, and Issouf Fofana. Characterization of the operating periods of a power transformer by clustering the dissolved gas data. In *2017 IEEE 11th International symposium on diagnostics for electrical machines, power electronics and drives (SDEMPED)*, pages 298–303. IEEE, 2017.