Machine Learning Approach for Risk-Based Inspection Screening Assessment

Andika Rachman, R.M. Chandima Ratnayake
Department of Mechanical and Structural Engineering and Material Science, University of Stavanger
N-4036 Stavanger, Norway
(andika.r.yahya@uis.no, chandima.ratnayake@uis.no)

Abstract

Risk-based inspection (RBI) screening assessment is used to identify equipment that makes a significant contribution to the system's total risk of failure (RoF), so that the RBI detailed assessment can focus on analyzing higher-risk equipment. However, due to its qualitative nature and high dependency on sound engineering judgment, screening assessment is vulnerable to human biases and errors, and thus subject to output variability and threatens the integrity of the assets. Moreover, screening assessment is deemed time-consuming and makes great demands on effort. This paper attempts to tackle these challenges by utilizing a machine learning approach to conduct screening assessment. A case study using a dataset of RBI assessment for three onshore and four offshore oil and gas production and processing units is provided, to illustrate the development of an intelligent system, based on a machine learning model for performing RBI screening assessment. The best performing model can achieve overall accuracy and precision of 92.33% and 84.58%, respectively. Additionally, a comparative analysis between the performance of the intelligent system and the conventional RBI screening assessment is performed to examine the benefits of applying machine learning approach in the RBI screening assessment. The result shows that the application of machine learning approach potentially improves the quality of the conventional RBI screening assessment output by reducing output variability and increasing accuracy and precision. The machine learning model should be complemented by human intelligence, so that it can refine the insights provided by the machine intelligence.

Keywords: Risk-based inspection, machine learning, screening assessment, knowledge transfer and reuse, risk assessment

1 Introduction

Equipment failure is one of the major causes of unexpected and undesirable events (e.g. hydrocarbon discharges, fire, and explosions) in the oil and gas industry [1]. Technical malfunction is commonly caused by inadequate assessment and control of the assets' technical integrity, which leads to the inability to control degradation rate and, ultimately, the release of hazardous substances from the pressure-containing envelopes [2]. To prevent the occurrences of equipment breakdown and to maintain the integrity of the assets, inspection has been used to ensure all equipment in the production and processing facility is fit-for-service. However, it is not cost-effective to apply fully comprehensive inspection to all equipment, considering the complexity and the quantity of equipment in a production and processing facility. Risk-based inspection (RBI) is typically performed to find the optimum inspection plan that can improve the safety and reliability of the production and processing facility in an effective and efficient manner.

RBI is an element of technical integrity management (TIM), which is associated with efforts to sustain the ability of physical assets to perform the required functions effectively and efficiently without harming personnel and the environment [3]. RBI uses equipment risk of failure (RoF) to prioritize inspection, thus allowing the organization to focus inspection resources on equipment with a high RoF. There are two types of assessment in the RBI methodology: screening and detailed assessment. Screening assessment is performed prior to detailed assessment, in order to identify equipment in the system that makes a considerable contribution to the system's total RoF [4]. This allows detailed assessment efforts to focus on higher-risk items; thus, resources (e.g., time and labors) can be utilized in a more effective and efficient manner [5].

Despite providing considerable benefits to the overall RBI process, screening assessment has a qualitative nature and high dependency on sound engineering judgment, which makes it vulnerable to human biases and errors, and thus subject to appraiser-to-appraiser output variation [6]. The research conducted by Geary [7] shows that qualitative assessment, based on subjective judgment and a limited amount of data and information, causes considerable variation in the assessment outputs. This variation can be threatening to the integrity of the assets, causing under-inspection of higher-risk equipment and over-inspection of lower-risk equipment [7]. Additionally, qualitative analysis is deemed time-consuming and makes great demands on effort [8]. While performing a screening assessment is not mandatory in RBI methodology, omitting it may generate more expensive and laborious RBI detailed assessments.

This paper attempts to address the aforementioned challenges by transferring and reusing the information and knowledge generated from past RBI detailed assessments as the references for conducting RBI screening assessment. RBI detailed assessment uses a more quantitative approach that has less dependency on subjective judgment and is less prone to human biases and appraiser-to-appraiser output variation. The premise is that, by transferring and reusing knowledge and information from past RBI detailed assessments, the subjectivity inherent in RBI screening assessment can be reduced. Knowledge transfer between activities/projects is deemed essential because it can reduce uncertainties and eliminate past failures and errors [9]. Furthermore, knowledge transfer and reuse can drive a leaner overall RBI assessment process by creating an uninterrupted flow of value-added information that alleviates the challenges involved in eliminating knowledge regeneration (i.e., the inability to acquire and reuse knowledge from past endeavors that generate non-value-added activities in the form of 'reinventing the wheel') and thus reducing the lead time of the corresponding activity [10, 11].

Hoppmann, et al. [12] suggest the utilization of explicit documentation such as a database and checklists to facilitate the transfer and reuse of legacy knowledge. While firms have applied the practice of recording and documenting past projects files and lessons learned in repositories, knowledge assets are still disregarded and abandoned [13]. Storing data and information without transmission and dissemination throughout the organization will not leverage knowledge assets, as it only creates worthless islands of isolated knowledge [14]. Ward and Sobek II [13] argue that data, information, and knowledge of past projects/activities are discarded due to lack of learning culture in the organization, insufficient time to reflect on and synthesize the generated knowledge, and lack of understanding of how to convert data and information into actionable knowledge. Additionally, the complexities inherent in data and information [15] create high barriers to the entry, retrieval, updating, and reuse of the knowledge and lead to the tendency to neglect knowledge transfer [11].

In this study, machine learning is proposed as the mechanism to implement the transfer and reuse of knowledge in the RBI screening assessment. The study of machine learning focuses on the development of computational techniques to mechanize the acquisition of knowledge from experience [16]. Langley and Simon [16] state that “Machine learning aims to provide increasing levels of automation in the knowledge engineering process, replacing much time-consuming human activity with automatic techniques that improve accuracy or efficiency by discovering and exploiting regularities in data”. Bose and Mahapatra [17] argue that the evolution of the machine learning approach has attempted to reduce/eliminate the costly and laborious knowledge-engineering process involved in the development of knowledge-based systems. Accordingly, machine learning systems are capable of converting data and information into knowledge and enable cost-effective exploitation of knowledge resources [18]. Consequently, machine learning facilitates knowledge transfer and reuse by lowering the barriers to the entry, retrieval, and updating of legacy knowledge.

Based on the argument above, the key objectives of this study are to demonstrate the application of machine learning approach for performing RBI screening assessment. A case study, using a dataset of RBI detailed assessment for three onshore and four offshore oil and gas production and processing units, is provided to illustrate the development of an intelligent system, based on a machine learning model for performing RBI screening assessment. Additionally, the benefits of applying machine learning approach in the RBI screening assessment compared to the conventional RBI screening assessment are examined through a comparative analysis.

The remainder of the paper is structured as follows. The second section provides a brief literature review on the RBI methodology and the current application of machine learning in inspection and diagnostic subjects. The third section elaborates the research methodology. The fourth section provides the results and analysis. Finally, the fifth section concludes the paper.

2 Related Work

2.1 Risk-based inspection

RBI assessment is a methodology for achieving a cost-effective inspection plan and ensuring regulatory and corporate compliance. RBI has profound applications in the chemical and oil and gas industries. For instance, RBI methodology has been applied to optimize in-service inspection of H_2S-based process plants [19]. Seo, et al. [20] and Kamsu-Foguem [21] utilize RBI assessment to determine the current RoF of subsea pipelines and to estimate the remaining life of the pipelines. The RBI methodology has also been used to develop an optimal inspection and repair strategy for structural systems [22].

The notion of using RoF as a measure to optimize inspection resources was proposed in the late 1980s [19]. In fact, the RBI methodology emerged from the Risk-Informed In-Service Inspection (RI-ISI) methodology [23], which has been prominent in the nuclear industries and in power plants [24, 25]. Generally, the objective of RI-ISI methodology is analogous with that of the RBI methodology: to prevent unnecessary inspections by prioritizing the inspection activities based on RoF [26]. In RI-ISI methodology, the degradation potential and the consequence of failure for each component in the system are assessed [27]. Subsequently, the components are ranked, based on the evaluated RoF, to determine which components the inspection resources should be concentrated on [26].

Despite its various applications, the main purpose of the RBI methodology is the same across industries and types of equipment: to focus the available inspection resources on the higher-risk components, such that the optimum inspection plan can be achieved (i.e., preventing under-inspection of higher-risk components and over-inspection of lower-risk components). A generic RBI methodology is shown in Figure 1. RBI starts with the identification of data requirements. The typical data needs may include, but are not limited to: design, fabrication, and construction documents; process and operating condition records; historical inspection and maintenance records; hazard analysis, material selection, and corrosion engineering records; cost and project engineering documents [5].

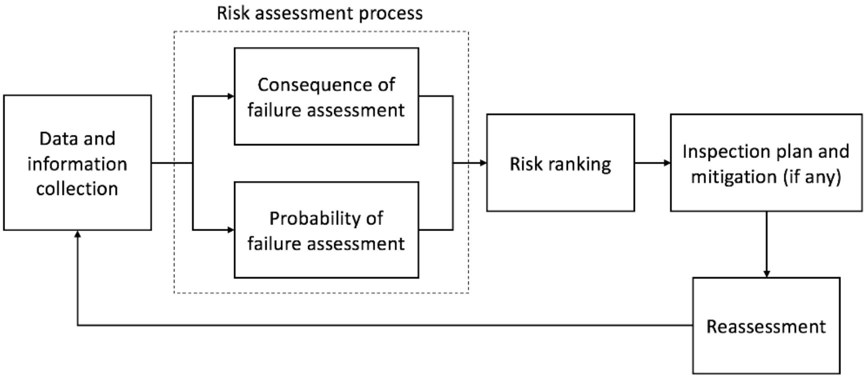

Risk assessment process
Data and information collection
Consequence of failure assessment
Probability of failure assessment
Risk ranking
Inspection plan and mitigation (if any)
Reassessment

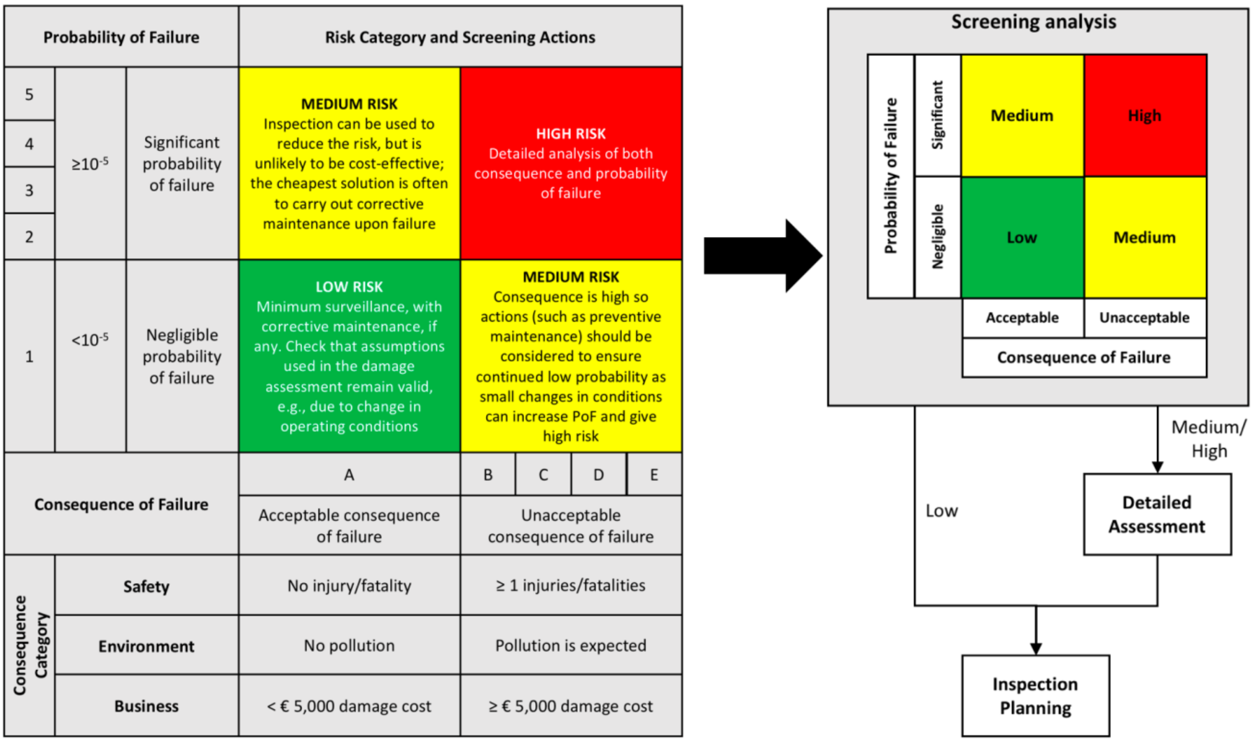
Figure 1 Generic RBI methodology [5]

The risk assessment process aims to prioritize equipment based on equipment RoF, thus allowing an organization to focus its inspection efforts on high-risk equipment and preventing over-inspection of low-risk equipment [28]. RoF in RBI is defined as the product of probability of failure (PoF) and consequence of failure (CoF). PoF is evaluated based on the variables that influence the failure rate of the equipment, such as type of degradation mechanism, degradation rate, operational conditions, equipment design, previous inspection effectiveness and results, and equipment age. In general, degradation mechanisms in RBI assessment include general/localized thinning, stress corrosion cracking, corrosion under insulation, brittle fracture, embrittlement, and fatigue [8]. CoF is assessed based on the factors that impact the magnitude of hazards in the event of hydrocarbon release, such as type of substance contained in the equipment and process conditions. RBI considers four categories of consequence effect: personnel safety and health impact, environmental impact, production losses, and facility repair costs [9].

The results of PoF and CoF assessment are then combined and represented in a risk matrix to communicate the risk assessment outputs. Higher-risk equipment shall be given higher inspection priority than lower-risk equipment; i.e., higher-risk equipment needs more frequent inspection and/or more rigorous inspection techniques, to reduce the RoF [29]. The risk acceptance level is normally used to determine whether the inspection planning can reduce the equipment RoF to an acceptable level [5]. Inspection does not reduce RoF by mitigating the degradation mechanisms but by reducing the uncertainties regarding equipment condition, under the assumption that the organization will act properly based on the inspection results [5]. Risk mitigation activities are required if inspection activities do not effectively reduce the equipment RoF. This is normally the case where the equipment has low PoF and high CoF. As inspection can only diminish PoF level, mitigation (e.g., implementing integrity operating window, upgrading hazard detection system, etc.) is required to reduce the RoF.

In most processing systems, a significant portion of the system's total risk is accumulated in a relatively small percentage of the total equipment [5]. Therefore, screening assessment is conducted prior to detailed assessment, to identify equipment that makes a considerable contribution to the system's total RoF [4], such that the detailed assessment focuses only on higher-risk items [5]. This allows engineering resources (e.g., time and labor) to be utilized in a more effective and efficient manner [5]. A simplified qualitative analysis is typically performed for the screening assessment. Qualitative analysis relies more on a higher level of engineering judgment and requires less data than a more quantitative approach. It provides a simplistic approach that can generate quick, but less detailed, results. However, it is more dependent on skilled personnel with a high level of knowledge of the facility being analyzed.

Based on the knowledge of installation history and degradation mechanisms, along with sound engineering judgment, the significance of PoF and consequence CoF are determined, typically by using the risk matrix shown in Figure 2a. Generally, low-risk equipment (i.e., equipment with negligible PoF and insignificant CoF) do not to undergo a detailed assessment, as minimum surveillance and corrective maintenance are deemed sufficient to sustain the operations [4]. Meanwhile, medium-risk and high-risk equipment shall be assessed further by the detailed assessment method. Detailed assessment has a more quantitative approach that utilizes a more rigorous and systematic risk calculation model and requires more detailed data input. Furthermore, quantitative analysis typically utilizes sophisticated software that generates quantitative insights regarding the equipment RoF, which are less subjective, as they depend less on engineering judgment and intuition.



Probability of Failure			Risk Category and Screening Actions				
5	$\geq 10^{-5}$	Significant probability of failure	**MEDIUM RISK** Inspection can be used to reduce the risk, but is unlikely to be cost-effective; the cheapest solution is often to carry out corrective maintenance upon failure	**HIGH RISK** Detailed analysis of both consequence and probability of failure			
4							
3							
2							
1	$<10^{-5}$	Negligible probability of failure	**LOW RISK** Minimum surveillance, with corrective maintenance, if any. Check that assumptions used in the damage assessment remain valid, e.g., due to change in operating conditions	**MEDIUM RISK** Consequence is high so actions (such as preventive maintenance) should be considered to ensure continued low probability as small changes in conditions can increase PoF and give high risk			
Consequence of Failure			A	B	C	D	E
			Acceptable consequence of failure	Unacceptable consequence of failure			
Consequence Category	Safety		No injury/fatality	≥ 1 injuries/fatalities			
	Environment		No pollution	Pollution is expected			
	Business		< € 5,000 damage cost	≥ € 5,000 damage cost			

(a)

(b)

Figure 2 (a) Risk matrix for RBI screening assessment and (b) an overview of inspection planning process in RBI [4]

2.2 Application of machine learning in inspection and diagnostic

Numerous researches have been carried out to examine the application of machine learning techniques in the area of inspection and diagnostic in the oil and gas industry. El-Abbasy, et al. [30] developed models based on regression analysis to help oil and gas pipeline operators in evaluating the condition of the pipelines and planning the inspection, maintenance, and refurbishment of the pipelines, based on regression analysis. Several general and operational features of the pipelines, such as pipeline age, diameter, coating condition, metal loss, crossings, free spans, operating pressure, and cathodic protection, are included. The output of the model is the overall pipeline condition, scaled from 0 to 10, with 0 being the worst and critical condition and 10 the perfect and excellent condition. The models achieved an average validity score above 96%. El-Abbasy, et al. [31] built the improvement of the models developed in [30] by using neural network (NN), instead of regression analysis, as the learning algorithm. The average validity score of the models was improved to 97.40%. Similar models were constructed by El-Abbasy, et al. [32], based on regression analysis, NN, and decision tree methods to assess the condition of unpiggable oil and gas pipelines. The unpiggable pipeline models have a slightly lower performance than the piggable pipeline models, with an average validity score of between 87% and 96%.

Layouni, et al. [33] used the combination of pattern-adapted wavelets and NN to detect, locate, and estimate the size of metal loss in oil and gas pipelines, based on the magnetic flux leakage (MFL) technique. This approach achieves a depth-prediction accuracy of 80%, with certainty of less than ±1% for noiseless signals and ±10% to ±13% for noisy signals [33]. A similar system was developed by Mohamed, et al. [34], who applied wavelet transform technique and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) to identify and estimate metal loss in oil and gas pipelines, based on the MFL signals. At the error tolerance of ±1%, ±5%, and ±10%, the ANFIS model achieved 11%, 48%, and 71% defect-depth prediction accuracy [34]. Carvalho, et al. [35] utilized NN to classify pipeline weld defects, based on the MFL signals, with the classification accuracy of the proposed model reaching 71.70%. Abdulla and Herzallah [36] used multiple model NN to determine the occurrence of leakage in oil and gas pipelines, based on inputs of inlet pressure, outlet pressure, and outlet flow signals. The multiple model NN has better performance than the single model NN, shown by the reduction in the validation error from 6.06% to 0% and the reduction in testing error from 3.5% to 1.1%[36]. Zadkarami, et al. [37] utilized NN to detect, locate, and evaluate the severity of leakage occurrence in oil and gas pipelines, based on the negative pressure wave (NPW) signals. The proposed model was able to attain 91.89% of the correct classification rate [37].

Wenhui, et al. [38] developed a deep neural network (DNN) model to automatically detect the existence of defects in the weld region from x-ray images. A maximum classification accuracy of 91.84% was obtained from the proposed model [38]. Vilar, et al. [39] constructed an automatic detection system to detect and discern five types of defects in the weld region (i.e., no defect, slag inclusion, porosity, longitudinal crack, and transversal crack) from radiographic images of welded joints. The system uses image processing techniques (i.e., noise reduction, contrast enhancement, thresholding, and labelling), before applying feature extraction and ANN for weld classification [39]. A similar system was developed by Kumar, et al. [40], who also used image processing techniques (i.e., RGB to grey conversion, region of interest (ROI) selection, noise reduction, and contrast enhancement), feature extraction (i.e., grey level co-occurrence matrix) and ANN, to classify nine type of weld flaws (i.e., gas cavity, lack of penetration, porosity, slag inclusion, crack, lack of fusion, worm hole, undercut, and non-defect) from digitized radiographic images. The system achieved classification accuracy of 86.10% [40]. Meanwhile, Mirapeix, et al. [41] built an arc-welding defect detection and classification system, based on principal component analysis (PCA) and ANN, using the plasma spectra captured from the welding process as the input.

Besides the oil and gas sector, the healthcare sector has significantly used machine learning to help in diagnosing various diseases. Chen, et al. [42] developed convolutional neural network (CNN) models for risk prediction of cerebral infarction disease, with the prediction accuracy of the proposed models reaching 94.8%. Onan [43] utilized the fuzzy-rough nearest neighbor algorithm, consistency-based feature selection, and fuzzy-rough instance selection for medical diagnosis of breast cancer. The proposed models attained a promising classification accuracy of 99.71% [43]. Similar models were developed by Asri, et al. [44], who used support vector machine (SVM), decision tree, Naïve Bayes (NB), and *k*-nearest neighbors (*k*-NN) to predict and diagnose breast cancer risk. Their models obtained prediction accuracy of 97.13% [44]. Polat and Güneş [45] developed a system that combines PCA and ANFIS to diagnose diabetes disease, with the classification accuracy of the developed system achieving 89.47%. Das, et al. [46] proposed a neural network ensemble method for diagnosing heart disease, which achieves 89.01% prediction accuracy.

Despite significant efforts to apply machine learning techniques in the area of inspection and diagnostic, there is no existing study that discusses the application of machine learning techniques in the subject of RBI assessment. Hence, this paper tries to cover that gap by presenting the utilization of the machine learning approach for classifying equipment RoF in RBI screening assessment. The aforementioned research show that the developed machine learning models can yield satisfactory predictive performance for inspection- and diagnostic-related tasks, and RBI assessment is not an exception.

3 Research Methodology

As mentioned in Section 1, one of the main objectives of this study is to develop an intelligent system for performing RBI screening assessment based on a machine learning approach, to facilitate knowledge transfer and the reuse of past RBI detailed assessments. The intelligent system task is to determine whether the equipment has low or high/medium RoF (as described in Section 2.1), based on several features related to design and operational conditions, in order to justify whether the equipment is required to undergo detailed assessment. In a machine learning domain, this type of task is called a supervised classification problem. Given a set of observations with a set of independent variables (i.e., input features) and a dependent variable (i.e., output feature), supervised learning aims to create a set of rules that can be used to assign and predict the dependent variable of new observations [47]. The steps required to develop the intelligent system are shown in Figure 3. Each step is discussed individually in the following subsections.

Moreover, to examine the benefits of applying machine learning approach in the RBI screening assessment compared to the conventional RBI screening assessment, a comparative analysis between the performance of the developed machine learning classifier with human appraisers is performed.

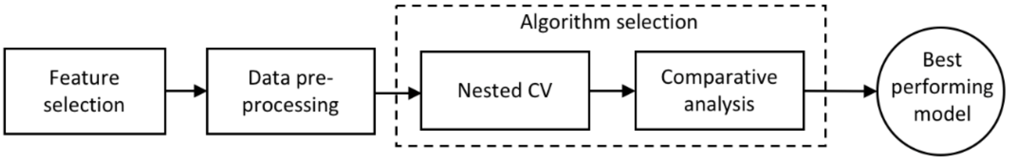


Figure 3 The steps for developing the intelligent system

In this study, standard MacBook Pro (8GB 1867 MHz LPDDR3 memory, 2 GHz Intel Core i5 processor, 250 GB of flash storage, and Mac OS X as operating system) with Python 3.6 (with the necessary packages such as *NumPy*, *Pandas*, and *Scikit-learn*) and Jupyter Notebook are used.

3.1 Dataset descriptions

The dataset was acquired from an RBI detailed assessment project conducted for three onshore and four offshore oil and gas production and processing units, comprising over 200 pressure vessels and 3000 piping lines. A total of 1581 instances are contained in the dataset. The number of instances in the dataset is fewer than the original number of pressure vessels and piping tags because each instance represents a group of piping and pressure vessels with similar characteristics (e.g., similar operating pressure/temperature, fluid containment, dimensions, etc.). Henceforward, the term ‘instance’ is also used to refer to the piping and pressure vessels assessed in the dataset.

The anatomy of the dataset is shown in Table 1. The listed features in the dataset can be divided into two groups: input and output/target features. Input features are variables that are entered by users and are sourced from the basic data of equipment as well as operational history and characteristics. Output features are variables that are influenced by the input features and generated as the results of risk assessment steps. The output features in this dataset were computed by RBI assessment software, based on API 581 RBI detailed assessment methodology [48].

Table 1 Some features contained in the dataset

Type		Feature examples
Input features	Design characteristics	Component type, equipment diameter, material of construction, furnished thickness, design code, corrosion allowance, commissioning date, post-weld heat treatment, weld joint efficiency, insulation type, etc.
	Operational characteristics	Operating pressure, operating temperature, dead leg, fluid containment, steam out, fluid phase, type of external environment, heat tracing, etc.
	Damage mechanism characteristics	Damage type, damage susceptibility, external damage type, thinning type, corrosion rate, etc.
	Inspection characteristics	Inspection date, type of inspection, external damage inspection effectiveness, internal damage inspection effectiveness, etc.
Output features	PoF assessment results	Damage factors, PoF value, PoF category, etc.
	CoF assessment results	CoF value, CoF category, etc.
	RoF results	Risk category, financial risk, etc.

3.2 Feature selection

Feature selection involves the process of choosing a subset of pertinent features, required for use in the model construction [49]. Feature selection aims to eliminate irrelevant and redundant features that degrade the speed and accuracy of the predictive model and to gain an understanding regarding how the data are generated [50].

There are three categories of methodologies that are commonly utilized in the feature selection process: filters, wrappers, and embedded solutions [51]. Filter methods are performed based on a proxy (e.g., Pearson correlation, significance test, etc.) rather than a machine learning algorithm to evaluate a set of features [51]. Due to their effectiveness in computation time, filter methods are suitable for use when the dimensionality of the data is high [52]. Wrapper methods use a machine learning algorithm to select the best subset of features [53]. Each subset in the space of all possible subsets of features is used to train the model under consideration, which is then tested on a validation dataset [53]. The subset that yields the best predictive power is selected [54]. Wrapper methods typically generate better predictive models than filter methods, but they can be computationally expensive, especially when the data dimensionality is high [52]. Embedded methods integrate a feature selection step as a part of the training process. Embedded methods generally have a more efficient process than wrapper methods because they eliminate the need to retrain every single subset of features being examined [54].

The three types of feature selection mentioned above are data-driven approaches and implemented when the domain knowledge is insufficient or unavailable to select the appropriate features [51, 54]. When the nature of a problem is well-understood, domain knowledge can be used as a feature selection method to describe the relationship among features and to avoid data overfitting [55-

57]. Guyon and Elisseeff [54] suggest constructing a set of features in an *ad hoc* manner if sufficient domain knowledge is available. As the nature of problems in RBI assessment is well established, domain knowledge can be used. Expert knowledge and engineering standards and codes related to RBI assessment (e.g., API 580 [5], API 581 [48], and DNV-RP-G101[4]) can be utilized as the guidance for selecting the features.

In this study, a systematic feature selection is performed by the combination of domain knowledge and data-driven approach. Filter method is selected as the data-driven approach because it is independent of the learning algorithm [58]. The complete feature selection process used in this study is shown in

Figure 4.

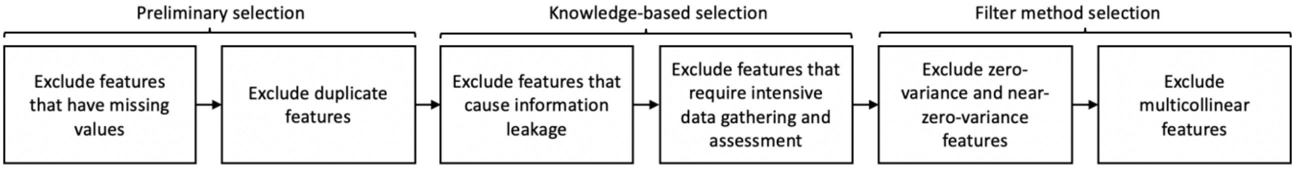

Preliminary selection
Knowledge-based selection
Filter method selection
Exclude features that have missing values
Exclude duplicate features
Exclude features that cause information leakage
Exclude features that require intensive data gathering and assessment
Exclude zero-variance and near-zero-variance features
Exclude multicollinear features

Figure 4 Feature selection process used in the study

The preliminary selection attempts to remove duplicate features and features with missing values. A feature is considered a duplicate if it contains the same information as another feature. Missing values can occur for a number of reasons (e.g., unrecorded data, data corruption, etc.). In this study, features that contain missing values are excluded, to prevent errors in the learning algorithms.

Subsequently, knowledge-based feature selection is performed. Features that cause information leakage and require intensive data gathering and assessment are removed. Information leakage happens when the training dataset contains features that the model is trying to predict. To prevent information leakage, features that are generated as part of the RBI assessment outputs (i.e., output features) are not selected as the input features. Considering the simplistic nature of RBI screening assessment, features that require intensive data gathering and assessment are removed. For instance, some features contained in the dataset (i.e., damage mechanism- and inspection-related features) are generated as the results of RBI detailed assessment, making them unsuitable as input features, due to heavy pre-assessment and pre-computation.

Then, filter methods are used for removing zero-variance/near-zero-variance features and multicollinear features. Zero-variance/near-zero-variance features are variables that are nearly identical and constant among instances in the dataset, making them uninformative, with little effect on the computations [59]. In this study, a zero-variance/near-zero-variance feature is defined as a feature with a frequency ratio (i.e., the frequency of the most prevalent value to the frequency of the second most prevalent value [59]) greater than 10. Additionally, multicollinearity among features is eliminated, due to its detrimental effect on the model generalizability [60]. To identify multicollinearity, a correlation matrix is constructed to compute the Pearson correlation coefficient between two features. Pairwise features that have a correlation coefficient larger than 0.5 are identified. Then, the feature within the pair that has a higher average correlation coefficient with the other features is removed [59].

For output feature, *risk category* is selected, based on the aim of RBI screening assessment: to determine whether particular equipment has a low or high/medium RoF. *Risk category* contains information regarding the qualitative risk level of the equipment (i.e., low, medium, medium-high, or high), which is adequate to represent the system's objective. Table 2 shows the complete list of input and output features to be included.

Table 2 Input and output features used in the intelligent system

Input features		Output feature
• Material of construction	• Insulation type	Risk category
• Component type	• Operating pressure	
• Equipment diameter	• Operation temperature	
• Commissioning date	• Fluid containment	
• Post-weld heat treatment	• Corrosion allowance	

3.3 Data pre-processing

Data pre-processing is performed to extract and integrate data from different sources, to transform raw data into a readable and understandable format, and to cleanse cluttered and noisy data [61]. Raw data are generally incomplete, inconsistent, and contain many errors; data pre-processing ensures that the data are prepared for further processing. Typical data pre-processing procedures are [62]:

- data cleaning: includes filling in empty values, correcting data format, removing unnecessary features, resolving inconsistencies, and identifying and removing outliers;
- data integration: merging data from multiple databases;
- data reduction: reduces the volume of data without sacrificing the quality of the analytical results;
- data transformation: includes smoothing, aggregation, generalization, normalization, and attribute construction.

In general, minimum data cleaning is required for the given dataset. Minor data format correction is performed to transform numerical features' values (e.g., *operating pressure* and *operating temperature*) from string-type to float-type, to enable the manipulation of numerical data. Data integration is needed, as the original dataset is separated into seven workbooks. Feature selection (Section 3.2) is a part of data reduction, which shrinks the dimensionality of the dataset, while attempting to minimize the loss of information [52].

A data transformation process that is performed on the dataset is generalization, which reduces the number of distinct values in a certain feature by replacing low-level feature values with higher-level values [63]. Generalization is performed to avoid undergeneralization, which causes a certain feature to be uninformative and unhelpful in refining the predictive power of the model [62]. The decision regarding how high a feature should be generalized is a subjective matter [62]. Attribute generalization threshold control, which limits the number of distinct values in a feature, is commonly applied to prevent undergeneralization [62, 64]. The threshold for the number of distinct values in a feature normally ranges from 2 to 8, and the determination of the threshold value should be based on expert judgment [62]. In this study, the threshold value, which is set to 7, is applied to categorical features.

Four features are generalized: *fluid containment*, *material of construction*, *insulation type*, and *risk category*. The *fluid containment* feature is generalized and transformed into three values (i.e., *non-hydrocarbon*, *light hydrocarbon*, and *heavy hydrocarbon*) that represent three types of fluid that give different magnitudes of CoF and RoF. The *material of construction* feature is generalized and transformed into four distinct values (i.e., *carbon steel*, *SS316L*, *22Cr*, and *aluminum*) that represent higher level of abstraction in material type concept hierarchy. These four values represent material types that have different levels of resistance to various degradation mechanisms and, thus, have a distinctive impact on the PoF and RoF levels. The *insulation type* feature is generalized into *none* and *insulated*, which indicates whether equipment is insulated or not. The presence of insulation influences the type of damage mechanism that occurs on the equipment and thus impacts the PoF and RoF levels.

For the *risk category* feature, generalization is conducted in accordance with the objective of RBI screening assessment. The *risk category* from the original dataset contains the PoF and CoF classification of each item of equipment (see Figure 2a). As mentioned in Section 3, the goal of the model is to classify whether particular equipment is in the 1A risk category (i.e., low risk) or not (i.e., medium/high risk). Thus, the *risk category* feature value is transformed to low class (i.e., 1A risk level) or medium/high class (i.e., higher than 1A risk level). In other words, the RBI screening assessment is essentially a binary classification task. From the dataset, 375 of the instances (23% of the total instances) are in the low-risk category, while the remaining 1206 instances (77% of the total instances) are in the medium/high class. The complete list and details of feature generalization are shown in Table 3.

Table 3 Generalization of features

Features	Number of distinct values	Original values	Transformed values
Fluid containment	160	Water; heating medium; closed drain liquid; fuel gas; diesel; re-compressor gas; cold separator liquid; etc.	Non-hydrocarbon; light hydrocarbon; heavy hydrocarbon
Material of construction	42	SA106; SA516; SA333; SS316L; API5L; SA179; SA283; SA36; A790; SA789; SA182; SB209; etc.	Carbon steel; SS316L; 22Cr; aluminum
Insulation type	8	None; calcium silicate; mineral wool; asbestos; fiberglass; foam glass	None; insulated
Risk category	25	1A; 2A; 3B; 4C; 5D; 2C; 5E; 5B; 3C; etc.	Low; medium/high

The other data transformation that is performed on the dataset is attribute construction, which is the addition of new features, constructed using the dataset original set of features. In this study, the *equipment age* feature is added and computed by subtracting the RBI assessment date with the *commissioning date* feature. The *commissioning date* feature itself is not included as one of the model input features, replaced by the *equipment age* feature.

All non-ordinal categorical features (i.e., *material of construction*, *equipment type*, *post-weld heat treatment*, *insulation type*, and *fluid containment*) with N values are converted into $N-1$ dummy variables, in order to enable machine learning algorithms to process these features [65]. Moreover, the values of all input features are standardized such that they will have the properties of standard normal distribution [66]. Standardization is essential, as the input features have values with different ranges and units [67]. Standardization makes all input features contribute proportionally to the final outputs and causes the optimization algorithm to converge faster than without standardization [65].

3.4 Algorithm selection

As mentioned in Section 3.3, RBI screening assessment is essentially a supervised classification problem with binary output feature values (i.e., low risk and high/medium risk). There is a wide range of machine learning techniques that can solve binary classification problems. A set of binary classification techniques is selected and each of them is compared to determine the algorithm that yields the best results. Six established classifiers are selected: logistic regression (LR), support vector machines (SVM), *k*-nearest neighbors (*k*-NN), gradient boosting decision trees (GBDT), AdaBoost (AB), and random forests (RF). LR, SVM, and *k*-NN are single classifiers, while GBDT, AB, and RF are ensemble classifiers. A brief explanation of each classifier is presented in the following subsections.

3.4.1 Logistic regression

LR uses the logistic regression function, which measures the relationship between input and output features, in order to estimate the probability of instances belonging to certain classes [68]. In the case of RBI screening assessment, let us assume that $y = 1$ if the instance has low risk and $y = 0$ if the instance has high/medium risk. If $p = \text{Probability}(y = 1|\mathbf{x})$ and $\mathbf{x}$ is a column matrix of n input features, then the logistic regression function can be defined as follows:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta^T\mathbf{x} \quad (1)$$

where α is the intercept from the linear regression function and β^T is a vector of regression coefficients [69]. A complete description regarding the LR algorithm can be seen in [70].

3.4.2 SVM

In a two-class classification problem, the basic principle of SVM is to separate a given set of binary labeled instances, which are represented as points in a feature space, with an optimal hyperplane (i.e., a subspace that gives the greatest margin of separation between the two classes) [71]. SVM can perform non-linear classification by applying the kernel technique, which implicitly transforms the input into a higher-dimensional feature space in order to find a non-linear decision boundary in the original feature space [71]. Consider a set of training instances $\mathbf{x}_i$ $(i = 1, 2, \ldots, N)$ that are vectors in a space $\mathbf{x}_i \in R^n$ and belong to two separate classes $y_i \in \{0, 1\}$; the hyperplane that separates these two classes is:

$$f(\mathbf{x}) = w^T \cdot \Phi(\mathbf{x}) + b = 0 \quad (2)$$

where w^T is a vector normal to the hyperplane, b is the bias, and $\Phi: R^n \to R^m$ is the feature map that transforms the input feature space into a higher-dimensional space, such that the instances can be linearly separable [72]. The values of w and b are optimized such that classification error is minimized. A comprehensive explanation about SVM is provided in [73].

3.4.3 *k*-NN

k-NN performs classification by measuring the similarity of two or more instances [74]. The decision rule of *k*-NN is quite simple: the membership of an unclassified sample point is the same as the membership of a set of k nearest classified points. Euclidean distance is the most common metric for measuring the similarity between two instances and is used in this study. The Euclidean distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as follows [69]:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \cdot (\mathbf{x}_i - \mathbf{x}_j)} \quad (3)$$

k-NN is also called a 'lazy learner' because the training data are simply stored to be compared with unclassified data points, instead of being used to construct a generic function for classification [75]. Cover and Hart [76] provide a comprehensive description of *k*-NN.

3.4.4 Gradient boosting decision trees

GBDT creates an ensemble of "weak" decision trees (i.e., learning algorithms that have better accuracy than a simple guess), by sequentially constructing a decision tree based on the residuals of the entire ensemble generated so far [77]. GBDT improves the predictive power of the ensemble through incremental minimization of errors in each iteration of new decision tree construction [69]. GBDT model has the following generic form [78]:

$$F(\mathbf{x}) = \sum_{j=1}^{M} \beta_j \, h_j(\mathbf{x}) \quad (4)$$

where $h_j(\mathbf{x})$ is the individual decision tree generated in each sequence and β_j is the coefficient calculated by the gradient boosting algorithm. A complete description of GBDT is given by [77] and [78].

3.4.5 AdaBoost

The AB approach starts with finding a weak classifier and fitting it to a subset of training data in order to generate a new classifier [77]. These steps are performed iteratively with an emphasis on the subsets of training data that are not well modeled (i.e., misclassified) by the previously generated classifiers [79]. This ensemble of classifiers is then combined into a model that generally has better accuracy than a single classifier. Let $\mathbf{x}$ be a column matrix of n input features and $\{h_t(\mathbf{x})|t = 1, 2, \ldots, T\}$ be a set of weak classifiers that are generated in T rounds; the combined classifier is [77]:

$$H(\mathbf{x}) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})\right) \quad (5)$$

where α_t is the importance measure of weak classifier $h_t(\mathbf{x})$. α_t is calculated by the following function:

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-r_t}{r_t}\right) \quad (6)$$

where r_t is the error of weak classifier $h_t(\mathbf{x})$. As the error becomes smaller, the importance of a weak classifier becomes greater. A comprehensive description regarding AB can be seen in [80].

3.4.6 Random forests

RF is a type of ensemble learning that works, based on the premise that a multitude of classifiers performs better than a single classifier [81]. RF comprises a collection of decision trees that are constructed using a random subset taken independently and with replacement from the original dataset [82]. Classification is made by casting a vote from each tree and aggregating these votes to determine each instance's class through majority voting. The utilization of an ensemble of decision trees reduces the risk of overfitting to the training dataset [68]. A comprehensive explanation of RF is provided in [83].

3.5 Nested cross-validation

Stratified nested cross-validation (CV) is a technique used for model selection and evaluation. Model selection is fundamentally a means to optimize the classifier by tuning its corresponding hyperparameters. Meanwhile, model evaluation is comprised of a procedure to approximate the performance of the classifier, based on the selected hyperparameters. Varma and Simon [84] report that nested CV is able to produce an unbiased estimate of the classifier's error.

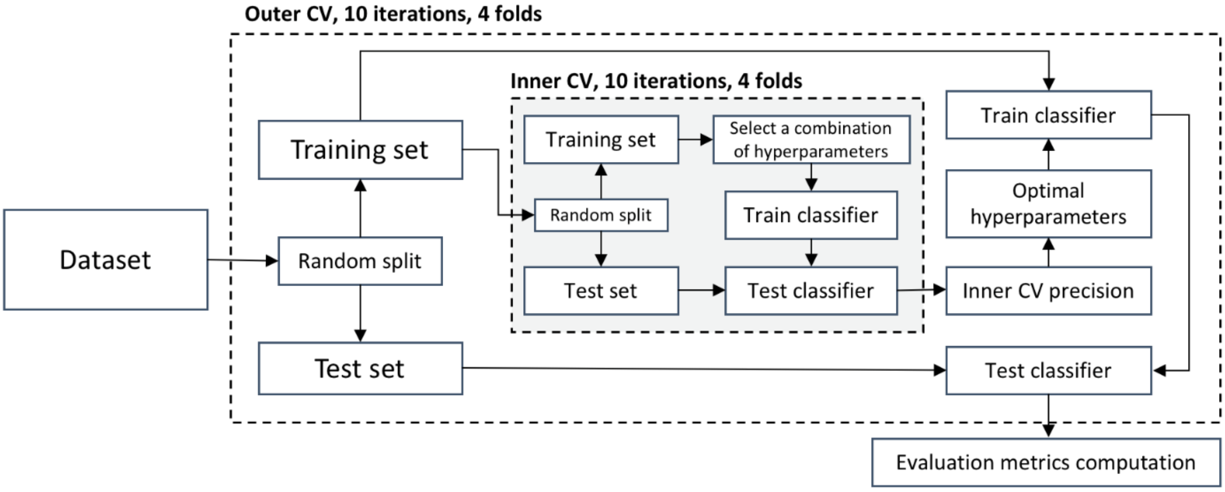


Figure 5 Model selection and evaluation using nested CV (adapted from [85])

The workflow of nested CV is given in Figure 5. Nested CV starts by randomly splitting the dataset into training and test folds. In this paper, 4-fold outer CV is used for model evaluation purposes. A training set from each outer CV fold is split again into training and test folds to conduct inner CV for model selection purpose. In other words, a training set from each outer CV fold has its own inner CV. The inner CV is essentially a grid search, which is a method of hyperparameter optimization through exhaustive searching of a manually specified subset of hyperparameter values. Four-fold inner CV is utilized to assess the performance of the classifier for all possible combinations of hyperparameter values. The combination of hyperparameters that gives the best precision is retained. The logic for selecting precision as the metric for choosing the best combination of hyperparameter values is given in Section 3.6.

A limited number of values for each hyperparameter is selected, as it is unfeasible to try the entire range of possible values. The hyperparameter value range for each classifier is discussed as follows:

- For *LR*, default hyperparameters are used.
- *SVM* – Three hyperparameters are considered for optimization: (1) kernel type, (2) C, which controls the trade-off between simplicity of the decision boundary and correct classification of training instances, and (3) gamma, which defines how far the influence of a single training instance reaches [68]. Four kernel types are tested: linear, sigmoid, polynomial, and RBF (Radial Basis Function). Value range with exponentially growing sequence of [10^{-5}, 10^{-3}, 10^{-1}, 10, 100] and [10^{-7}, 10^{-5}, …, 10^{4}] are assessed for C and gamma, respectively [86].
- *k-NN* – The number of neighbors is the hyperparameter to be tuned in *k*-NN [68]. The value range of [1, 10], with a spacing between values, is assessed for the number of neighbors.
- *GBDT* – There are two key hyperparameters that should be tuned in GBDT: (1) learning rate, which decreases the correction contribution of each tree, and (2) maximum depth of the individual regression estimators, which limits the number of nodes in

the tree [68]. A value range of [0.1, 1], with 0.1 as the distance between two adjacent values, is assessed for learning rate. For maximum depth, a range of [1, 5] with a spacing between values is examined.

- *AB* – Two hyperparameters are required to be tuned in AB: (1) learning rate, which shrinks the correction contribution of each classifier, and (2) the number of classifiers, which limits the construction of classifiers during boosting. For learning rate, the value range of [0.1, 1], with 0.1 as the distance between two adjacent values, is examined. The value range of [10, 50, 100, 250, 500, 1000] is evaluated for the number of estimators.
- *RF* – Two hyperparameters need to be examined: (1) the number of trees in the forest and (2) the number of features used to grow each tree [69]. For the number of trees, the value range of [10, 50, 100, 250, 500, 1000] is evaluated, while the range of [2, 32], with 2 as the spacing between values, is examined for the number of features.

After optimal hyperparameters are found, the classifier is trained by using the training set of the outer CV folds. In machine learning, training can be understood as the process of finding and extracting patterns in the data to generate rules for generalizing the data [16]. It can also be defined as the process of learning the relationship between input features and output features to make a predictive model, based on the inferred relationship [74]. Thus, it is necessary for the training set for having known output and input features to be fed to the learning algorithm. Then, the test set of the outer CV folds is used to evaluate the generalizability of the trained model to unseen data. The evaluation is performed based on certain metrics, which are discussed in Section 3.6. The nested CV is repeated 10 times, to obtain a more reliable performance estimation and comparison [87]. The data is reshuffled in each iteration; thus, the model is trained and tested on different splits of the dataset.

Because each classifier will undergo 4-fold nested CV with 10 iterations, there will be 40 values of evaluation metric per classifier. To summarize the performance of the classifier for each evaluation metric, the following equation is defined:

$$\overline{EM} = \frac{1}{10}\sum_{j=1}^{10}\left(\frac{1}{4}\sum_{i=1}^{4} EM_j^{(i)}\right) \quad (7)$$

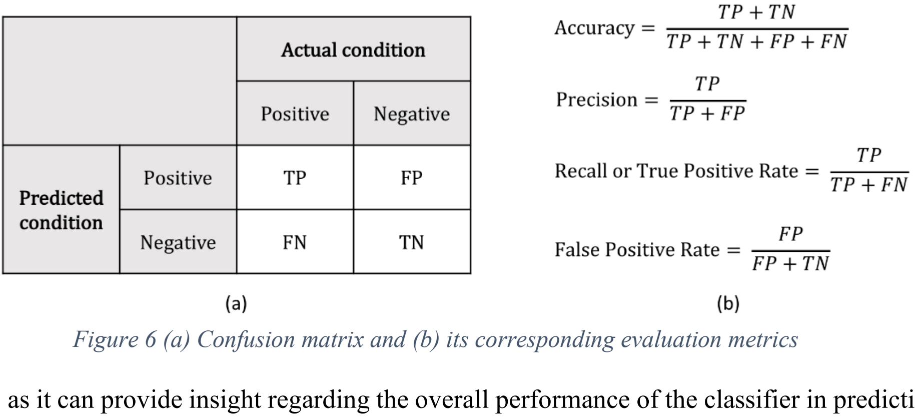
where $EM_j^{(i)}$ is the evaluation metric value of the i-th fold and the j-th iteration for a classifier, $\overline{EM}$ is the average of classifier evaluation metric values across all folds and iterations (i.e., the average value of an evaluation metric).

Besides assessing the average value of evaluation metrics, it is important to observe the stability of classifier performance on different cuts of the dataset. A box-and-whisker plot is used to examine the statistical dispersion of evaluation metrics' value for all classifiers on all folds and iterations.

3.6 Evaluation metrics

It should be noted that different evaluation metrics emphasize different aspects of a classifier's performance. Therefore, the objectives of the evaluation should be clearly defined to enable appropriate selection of metrics that correctly represent the aspects to be examined.

Before discussing evaluation metrics, it is necessary to clarify that, in a binary classification problem, the output of the model is either positive or negative. In this study, a low-risk class is defined as positive, while a medium/high-risk class is defined as negative. A confusion matrix can be used as the representation of the decision made in a binary classification problem (see Figure 6a). A confusion matrix has four elements: True positives (TP) are instances that are correctly classified as positive; false negatives (FN) are positive instances that are wrongly classified as negative; false positives (FP) are negative instances that are incorrectly labeled as positive; and true negatives (TN) are instances correctly labeled as negative [88]. From these elements, commonly used evaluation metrics for binary classification models can be defined, as shown in Figure 6b.



		Actual condition	
		Positive	Negative
Predicted condition	Positive	TP	FP
	Negative	FN	TN

(a)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall or True Positive Rate} = \frac{TP}{TP + FN}$$

$$\text{False Positive Rate} = \frac{FP}{FP + TN}$$

(b)

Figure 6 (a) Confusion matrix and (b) its corresponding evaluation metrics

Accuracy is selected, as it can provide insight regarding the overall performance of the classifier in predicting unseen instances. However, the dataset used in this paper consists of instances with unbalanced class distribution (see Section 3.3). In addition, the cost of misclassifying medium/high-risk items as low risk (i.e., false positive) is higher than incorrectly classifying low-risk items as medium/high risk (i.e., false negative). Misclassification of medium/high-risk items into a low-risk class may cause under-inspection (i.e., the allocation of inspection below the minimum requirement), which can be hazardous to the overall system integrity. Incorrectly classifying low-risk items into a medium/high-risk class will not do any harm with regard to the technical integrity of the system, except that the cost of inspection may rise. While an accuracy score is appropriate for observing the general performance of the classifier, it has severe limitations with respect to skewed class distribution [89] and different misclassification cost [90].

Precision and recall can be used to tackle the inherent limitations of accuracy. Precision and recall are metrics with single-class focus, which enable performance measurement with an emphasis on an individual class, while addressing class imbalance [90]. Due to the high cost of making a false positive decision in RBI screening assessment, precision has significant importance, as it measures the ability of the classifier to make a correct positive prediction. Precision addresses the question: "Among all positive predictions, what is the probability of them being correct?" However, a classifier with a high precision score does not necessarily imply a good prediction model, as it may indicate that the classifier misses numerous positive instances in the dataset. Thus, recall is used to complement precision by measuring the ability of the classifier to discover instances with a positive class. Recall addresses the question: "Of all items that belong to a positive class, what fractions does the classifier correctly detect as positive?"

Besides accuracy, precision, and recall, the area under the curve (AUC) of the receiving operating characteristics (ROC) is also utilized for evaluating the performance of the classifiers. ROC is a graphical plot that provides an illustration of the trade-off between the true positive rate and the false positive rate at various decision boundaries [69]. The true positive rate is synonymous with recall. The basic notion of ROC is that the behavior of a classifier throughout the decision boundary can be observed, given that it generates a score (typically in the interval of 0 to 1) on every instance, rather than binary labels [90]. The decision boundary is used to determine whether a specific instance belongs to a positive or negative class. AUC is basically a single scalar value to quantify the ROC curve for more convenient comparative analysis [91]. An advantage of using ROC curve and AUC is that they are insensitive to class distributions and misclassification cost [90]. Additionally, ROC and AUC provide a way to assess a classifier, without relying on the selection of a decision boundary.

As precision decreases, recall increases, and vice versa. Therefore, it is important to apprehend the precision of a particular classifier at all possible values of recall. To acknowledge this, average precision (AP) is computed as a metric that indicates the average precision score at all recall values [92]. Similar to ROC and AUC, AP can provide information regarding the precision of a particular classifier, without relying on the selection of a decision boundary.

3.7 Comparative analysis of classifiers

In order to demonstrate the difference in performance between classifiers and to select the best performing model, pairwise comparisons are performed for all evaluation metrics mentioned in Section 3.6. To determine whether the observed difference in each evaluation metric is statistically significant or simply due to chance, randomization tests are performed. A randomization test is a type of non-parametric test because it is independent from any assumptions regarding the underlying distribution of the test statistics [93]. The advantage of this test is that any metrics can be used as the test statistic, regardless of the underlying distributions of the observations [93, 94]. Under the null hypothesis, the two classifiers are not really different, and the observed empirical difference occurs purely by chance.

In this paper, a randomization test is performed based on Good [93]. The test statistic is defined as the difference in average value of an evaluation metric between two classifiers. First, the values of the metric produced by the two classifiers are concatenated. Each classifier will have 10 values for each evaluation metric, as the results of nested CV with 10 iterations (see Section 3.5). These values are then randomly rearranged, and the test statistic based on the new arrangement is computed. This procedure is repeated 10,000 times. Then, a comparison is made between the test statistics calculated from the randomized and non-randomized observations. The p-value is calculated as the proportion of times that the test statistic of the randomized observations is more extreme than the one generated by the non-randomized observations. A significance level of 0.01 ($\alpha = 0.01$) is used for this test. If the p-value is less than α, this indicates that the difference in a metric average value between two classifiers is unlikely to be due to chance, and, thus, the null hypothesis can be rejected. Mlxtend [95], a python library for data science tasks, is used to conduct the randomization test.

The applicability and generalizability of the result of this comparative analysis to the other RBI assessment datasets depends on how the datasets are generated. The dataset used in this study is generated from API 581 RBI detailed assessment methodology. If the datasets are produced by the same methodology, the result of the comparative analysis in this study can be applicable and generalized to select the optimum classifier for applying machine learning approaches to RBI screening assessment.

3.8 Comparative analysis of the best performing classifier and human appraisers

It is stated in Section 1 that the conventional RBI screening assessment is vulnerable to human biases and appraiser-to-appraiser output variation due to its qualitative nature and high dependency on sound engineering judgment. RBI assessment is performed by a team of engineers that have experience in materials science, risk assessment, mechanical engineering, and process engineering [5]. Each engineer normally takes a portion of the facilities to be assessed individually. However, the conventional RBI screening assessment is vulnerable to human biases and appraiser-to-appraiser output variation due to its qualitative nature and high dependency on sound engineering judgment. This variability can be detrimental to the quality of the output.

It is argued that the machine learning approach of RBI screening assessment can reduce the subjectivity inherent in the assessment. To demonstrate this, a comparative analysis between the performance of a machine learning classifier with human appraisers is performed. Three human appraisers who have related experience in performing RBI and integrity assessment are selected. The number of years of related experience for each appraiser is shown in Table 4.

Table 4 Human appraisers' number of years of related experience

	Number of years of experience
1st Appraiser	10
2nd Appraiser	5
3rd Appraiser	7

In practice, there is only one classifier needed to assess all instances in the facility under consideration because of the scalability of the machine learning classifier. However, to create a fair comparison with the human appraisers, three machine learning classifiers are used. These classifiers are created based on the best performing model as described in Section 3.7. Basically, these classifiers are identical to each other because they are trained by using the same features, algorithm, and dataset.

Each appraiser and classifier assess the same instances and perform the assessment individually. Not all instances in the dataset are included in the comparative analysis because of the time and intellectual effort required to perform manual RBI screening assessment. Therefore, twenty instances from the dataset are selected randomly for the purpose of the comparative analysis. The same set of features used by the machine learning algorithm (see Section 3.2) is given to the appraisers as the basis of the assessment. In addition, process flow diagrams of the oil and gas production and processing units being assessed are given to the appraisers to provide references regarding the production and processing flow.

The appraiser-to-appraiser output variation is evaluated by using the following equation [96]:

$$\text{Appraiser} - \text{to} - \text{appraiser variation} = \frac{\text{Total number of times classification for all not concur}}{\text{Total number of classification}} \quad (8)$$

The importance of evaluating appraiser-to-appraiser variation is to discover a bias due to appraisers and to learn if different appraisers generate inconsistent results [96]. Additionally, accuracy, precision, and recall metric are used to compare the performance of the machine learning classifier and human appraisers. The description about accuracy, precision, and recall metric is given in Section 3.6.

4 Results and Analysis

4.1 Comparative analysis of classifiers

Table 5 shows the average value of evaluation metrics for all classification models. LR shows reasonable performance in terms of accuracy (88.30%), recall (75.71%), and AUC (91.21%), but it has a relatively low score in precision (75.34%) and AP (77.32%), compared to the other classifiers. SVM has the lowest accuracy (84.89%) and recall (49.26%) of all the classifiers. Although SVM has better precision (79.64%) than LR, its AP score (74.75%) is worse than that of LR. *k*-NN produces a comparable score of accuracy (87.47%) with LR and SVM. *k*-NN performs better than SVM in terms of precision (82.36%) and recall (60.11%), but it has the worst AUC (88.62%) and AP (70.38%) amongst all the other classifiers.

Ensemble classification techniques (i.e., GBDT, AB, and RF) generally produce better results than single classifiers (i.e., LR, SVM, and *k*-NN). GBDT, RF, and AB have the top three accuracies (92.33%, 92.18%, and 89.65%, respectively), AUC (97.07%, 96.33%, and 94.12%, respectively), and AP (91.05%, 88.56%, 81.37%, respectively) among all the classifiers. GBDT and RF are also the top two performers with regard to precision (84.58% and 85.21%, respectively) and recall (82.96% and 81.33%, respectively).

Table 5 The average value of evaluation metrics for all classification models

Classifier	Accuracy (%)	Precision (%)	Recall (%)	AUC (%)	AP (%)
Single Classifier					
LR	88.30	75.34	75.71	91.21	77.32
SVM	84.89	79.64	49.26	89.60	74.75
k-NN	87.47	82.36	60.11	88.62	70.38
Ensemble Classifier					
GBDT	92.33	84.58	82.96	97.07	91.05
AB	89.65	80.66	74.31	94.12	81.37
RF	92.18	85.21	81.33	96.33	88.56

Box-and-whisker plots (Figure 7) are utilized to assess the statistical dispersion of the classifiers' performance on all folds and repetitions. SVM has the highest accuracy score dispersion, while RF has the lowest accuracy score dispersion. The other classifiers have comparable accuracy score dispersion.

In general, the value of precision has more spread than the value of accuracy for all classifiers. SVM has the least stable precision score among all classifiers, as it has the tallest box plot appearance. The other classifiers have a comparable spread of precision score. SVM clearly has the highest recall score dispersion, compared to that of the other classifiers. The remaining classifiers have similar recall score dispersion.

Despite some outlier points, the AUC score of GBDT and RF has less variance than that of the other classifiers, due to the compactness of both its interquartile range and whiskers. SVM and *k*-NN are the models with the greatest AUC score dispersion. GBDT and RF have the least AP score dispersion while SVM and AB have the greatest AP score spread.

Overall, SVM is the least stable model because, comparatively, it has high variance in all metrics. All other classifiers have comparable variance on accuracy, precision, and recall metrics. GBDT and RF have the lowest dispersion in AUC and AP metrics.

Table 6 summarizes the paired difference and the corresponding *p*-value between two classifiers for all evaluation metrics. It can be inferred that most of the differences between pairs of classifiers are statistically significant for all evaluation metrics. However, a few exceptions are found, where the difference between classifiers is not robust (denoted by bold typeface in Table 6). LR - AB, LR - *k*-NN, SVM - AB, and AB - *k*-NN pairs are statistically insignificant in only one evaluation metric, while SVM - *k*-

NN pair is statistically insignificant in two evaluation metrics. The most apparent one is the GBDT - RF pair, whose paired difference in accuracy, precision, and recall is not statistically significant.

In general, the ensemble techniques consistently outperform the non-ensemble techniques in all evaluation metrics. The research on empirical performance of various classifiers conducted by Zhang, et al. [68] produces similar results, of which RF and GBDT yield the best classification accuracy on several datasets. This is not surprising, given the inherent characteristic of ensemble learning that improves the generalizability and robustness of the classifier [77]. The superiority of ensemble techniques may also be due to the characteristic of the input features: five out of ten input features have a discrete/categorical nature. All ensemble classifiers examined in this paper are constructed based on information-based learning (i.e., decision trees), which is better in terms of handling discrete/categorical features [97]. Among the ensemble techniques, AB is the least preferred because it performs the worst in all metric categories and has the highest performance dispersion. It is difficult to select the preferred classifier between GBDT and RF, due to the marginal performance disparity between them, the lack of robustness in their performance difference, and their analogous performance dispersion. GBDT is slightly preferred because it outperforms RF in all evaluation metrics, except precision. Moreover, GBDT is faster in both training and testing running time than RF [68].

Despite the excellent prediction performance of ensemble methods, they have three inherent drawbacks [97]. First, ensemble classifiers have higher storage requirement than single classifiers. The storage size increases as the number of classifiers in a single ensemble expands. Second, ensemble classifiers are computationally expensive, due to the need to process multiple classifiers. Consequently, ensemble classifiers tend to have slow training time efficiency [68]. Third, the involvement of multiple classifiers reduces the comprehensibility of the ensemble classifiers [97]. These disadvantages are more prominent as the size of the dataset becomes larger.

It should be noticed that there is no single classifier that works best on all datasets [98]. Despite using the same dataset, if the data pre-processing and setting are different (e.g., different selection of input features and different features' transformation), the performance of each classifier may change. In the case of RBI, the type of assessment used (e.g., qualitative, semi-quantitative, or quantitative), the type of system being analyzed (e.g., offshore/onshore, downstream/upstream), and the type of methodology used (e.g., API, DNV, etc.) may influence the selection, as well as the performance, of the classifiers.

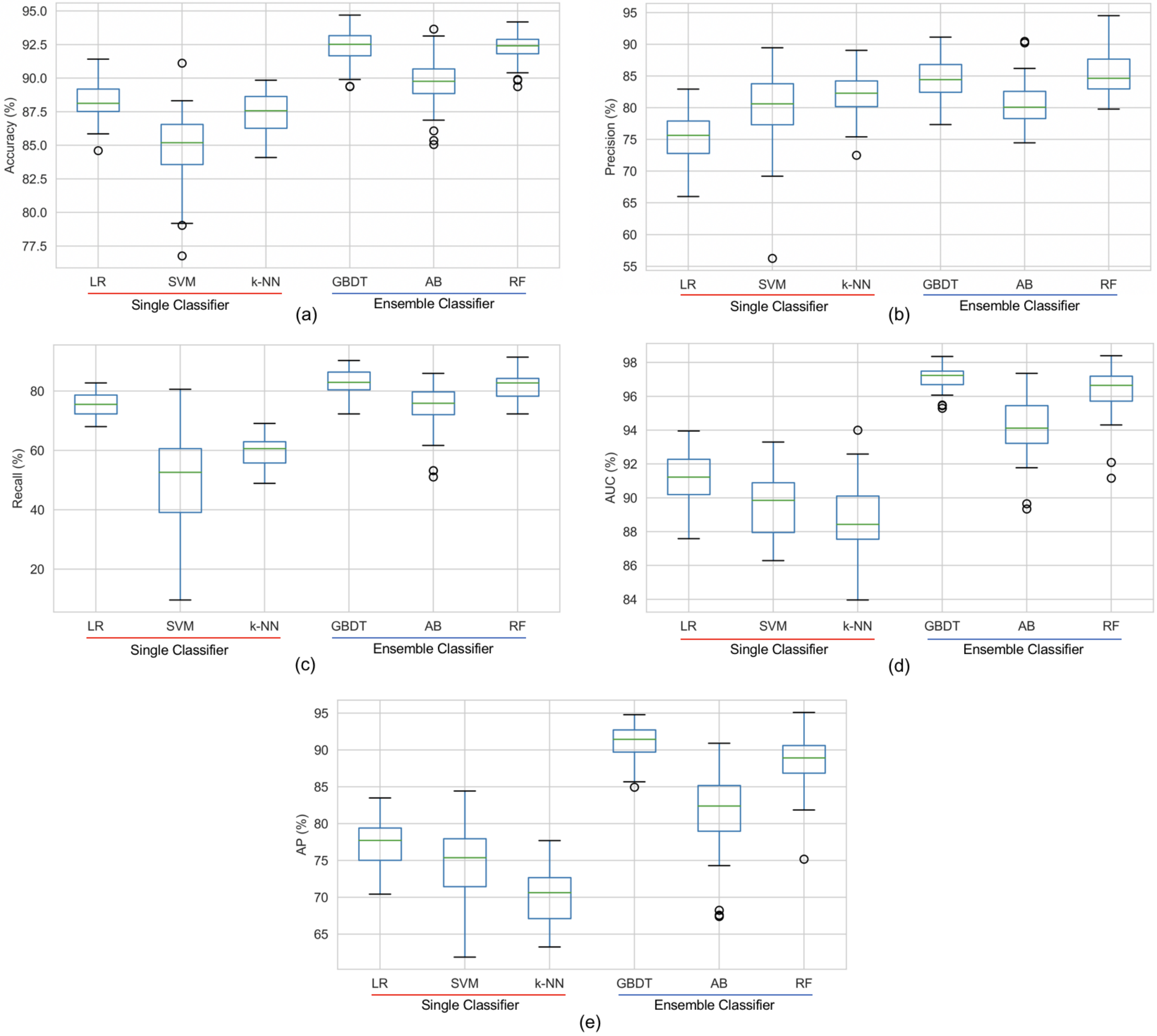


Figure 7 Box plots for (a) accuracy, (b) precision, (c) recall, (d) AUC and (e) AP from 10 iterations of nested CV

Table 6 Paired difference and p-value between two classifiers for all evaluation metrics

Model pair	Accuracy (%)		Precision (%)		Recall (%)		AUC (%)		AP (%)	
	Paired difference	*p*-value	Paired difference	*p*-value	Paired difference	*p*-value	Paired difference	*p*-value	Paired difference	*p*-value
LR - SVM	3.41	< 0.001	-4.30	< 0.001	26.45	< 0.001	1.61	< 0.001	2.57	0.005
LR - GBDT	-4.03	< 0.001	-9.24	< 0.001	-7.25	< 0.001	-5.86	< 0.001	-13.73	< 0.001
LR - AB	-1.35	< 0.001	-5.32	< 0.001	1.40	**0.348**	-2.91	< 0.001	-4.05	< 0.001
LR - *k*-NN	0.83	**0.014**	-7.02	< 0.001	15.59	< 0.001	2.59	< 0.001	6.94	< 0.001
LR – RF	-3.88	< 0.001	-9.87	< 0.001	-5.62	< 0.001	-5.12	< 0.001	-11.24	< 0.001
SVM - GBDT	-7.44	< 0.001	-4.94	< 0.001	-33.7	< 0.001	-7.47	< 0.001	-16.3	< 0.001
SVM - AB	-4.76	< 0.001	-1.02	**0.378**	-25.05	< 0.001	-4.52	< 0.001	-6.62	< 0.001
SVM - *k*-NN	-2.58	< 0.001	-2.72	**0.019**	-10.85	< 0.001	0.98	**0.043**	4.37	< 0.001
SVM - RF	-7.29	< 0.001	-5.57	< 0.001	-32.07	< 0.001	-6.73	< 0.001	-13.81	< 0.001
GBDT - AB	2.68	< 0.001	3.92	< 0.001	8.65	< 0.001	2.95	< 0.001	9.68	< 0.001
GBDT - *k*-NN	4.86	< 0.001	2.22	0.007	22.85	< 0.001	8.45	< 0.001	20.67	< 0.001
GBDT - RF	0.15	**0.568**	-0.63	**0.401**	1.63	**0.122**	0.74	0.002	2.49	< 0.001
AB - *k*-NN	2.18	< 0.001	-1.70	**0.041**	14.20	< 0.001	5.50	< 0.001	10.99	< 0.001
AB - RF	-2.53	< 0.001	-4.55	< 0.001	-7.02	< 0.001	-2.21	< 0.001	-7.19	< 0.001
k-NN - RF	-4.71	< 0.001	-2.85	< 0.001	-21.22	< 0.001	-7.71	< 0.001	-18.18	< 0.001

4.2 Comparative analysis of the best performing classifier and human appraisers

Table 7 shows the assessment results generated by the machine learning classifier and the human appraisers. It can be seen that there are some variations of output from one human appraiser to the other. The appraiser-to-appraiser variation score shown in Table 8 asserts that the human appraisers generate inconsistent outputs due to the qualitative and subjective nature of the assessment [6]. Meanwhile, the machine learning classifiers have zero appraiser-to-appraiser variation score because they produce identical outputs. Hence, the utilization of machine learning classifier potentially eliminates appraiser-to-appraiser variation by removing the influence of subjective judgment from the assessment. The classifier is trained based on the results of RBI detailed assessment, which use a quantitative approach and has less dependency on subjective judgment [5].

Table 7 The assessment results based on the human appraisers and the machine learning classifiers

Instance Number	Prediction						Actual output
	Human			Machine classifier			
	1st Appraiser	2nd Appraiser	3rd Appraiser	1st Appraiser	2nd Appraiser	3rd Appraiser	
#1	M/H	M/H	L	M/H	M/H	M/H	M/H
#2	M/H	M/H	M/H	M/H	M/H	M/H	M/H
#3	M/H	M/H	M/H	M/H	M/H	M/H	M/H
#4	L	L	L	M/H	M/H	M/H	M/H
#5	M/H	M/H	M/H	M/H	M/H	M/H	M/H
#6	L	M/H	L	M/H	M/H	M/H	M/H
#7	M/H	M/H	M/H	M/H	M/H	M/H	M/H
#8	M/H	M/H	M/H	M/H	M/H	M/H	M/H
#9	L	M/H	L	L	L	L	L
#10	M/H	M/H	L	M/H	M/H	M/H	M/H
#11	M/H	L	M/H	M/H	M/H	M/H	M/H
#12	L	L	L	L	L	L	M/H
#13	M/H	M/H	M/H	M/H	M/H	M/H	M/H
#14	M/H	L	M/H	M/H	M/H	M/H	M/H
#15	M/H	M/H	L	M/H	M/H	M/H	M/H
#16	L	M/H	L	L	L	L	L
#17	M/H	M/H	L	M/H	M/H	M/H	M/H
#18	L	L	M/H	L	L	L	L
#19	M/H	M/H	M/H	M/H	M/H	M/H	M/H
#20	L	M/H	L	M/H	M/H	M/H	M/H

Note: M/H = Medium/high risk, L = Low risk

Table 8 Appraiser-to-appraiser variation score

	Appraiser-to-appraiser variation
Human appraisers	0.55
Machine learning classifiers	0.0

Table 9 summarizes the accuracy, precision, and recall score for the human appraisers and the machine learning classifiers. The scores for machine learning classifiers are aggregated into a column because they generate identical scores. It can be seen that the machine learning classifiers outscore human appraisers in all evaluation metrics. The machine learning classifiers only misclassify one instance and are able to achieve 95% accuracy, 75% precision, and 100% recall. Meanwhile, the best human appraiser achieves 80% accuracy, 42.8% precision, and 100% recall. Based on these results, the application of machine learning approach potentially improves the quality of the RBI screening assessment outputs.

Table 9 The accuracy, precision and recall metric score for the human appraisers and the machine learning classifiers

Evaluation metric	1st Appraiser	2nd Appraiser	3rd Appraiser	Machine learning classifiers
Accuracy	80.0%	70.0%	55.0%	95.0%
Precision	42.8%	20.0%	20.0%	75.0%
Recall	100.0%	33.3%	66.6%	100.0%

However, even the best machine learning model generated in this study cannot achieve perfectly accurate and precise predictions. Thus, pure dependency on machine learning models in RBI screening assessment is not recommended, considering the high cost of misclassification. Similarly, relying only on human judgment is also not advised as it has its own set of biases and errors [99]. A mixture of machine and human intelligence is needed, such that the inductive learning can be complemented by the tacit knowledge of human workers. It is believed the inclusion of human intelligence can refine the insights provided by the machine intelligence [100]. This is in line with the concept of *jidoka* (or autonomation), derived from the Toyota Production System (TPS), which can be translated as automation with a “human touch” [101].

The proposed scheme to combine machine and human intelligence in the RBI screening assessment is presented in Figure 8. First, all instances in the dataset are assessed by the machine learning classifier. Then, the subsequent action depends on the classification results that the classifier makes. As explained in Section 3.6, the cost of misclassifying medium/high-risk instances as low risk (i.e., false positive) is greater than the cost of misclassifying low-risk instances as medium/high risk (i.e., false negative). The instances that are discerned as high/medium risk by the machine learning classifier are directly included in the RBI detailed assessment. While the possibility of misclassifying these instances exists, this will not do any harm with regard to the technical integrity of the system. In other words, the classification made by the classifier for the high/medium risk instances are generally agreed upon. Meanwhile, the instances that are classified as low risk are intervened by re-assessing them using sound engineering judgement and additional data (if necessary). The consideration to re-assess these low risk instances comes from the significance consequence of misclassifying medium/high-risk instances into low risk class, which may cause under-inspection (i.e., the allocation of inspection below the minimum requirement) and lead to the deterioration of the overall system integrity. The re-assessment provides an opportunity for engineers to use their experience and tacit knowledge to find irregularities in the results generated by the machine learning classifier.

The proposed scheme is expected to reduce the volume of instances that are needed to be evaluated by the engineers. In this case, the machine learning approach has not completely automated the RBI screening assessment process, but it provides the foundation to reduce the effort and time in evaluating instances and becomes the basis for the subsequent engineering evaluation.

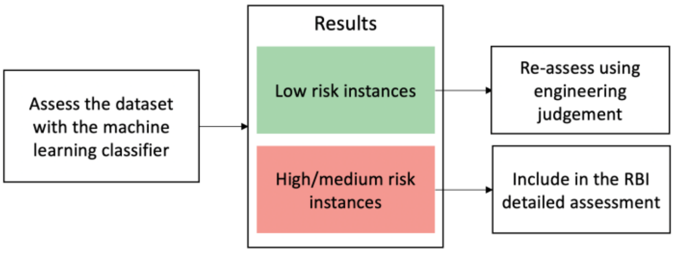

Assess the dataset with the machine learning classifier
Results
Low risk instances
High/medium risk instances
Re-assess using engineering judgement
Include in the RBI detailed assessment

Figure 8 A scheme to combine machine and human intelligence in the RBI screening assessment

5 Conclusion

This paper attempts to develop an intelligent system, based on a machine learning approach, to facilitate knowledge transfer and reuse and to reduce the human biases and errors and appraiser-to-appraiser output variation inherent in the conventional RBI screening assessment. A machine learning approach enables the reduction of the costly and laborious knowledge-engineering process involved in the development of knowledge-based systems. A case of the screening phase of RBI assessment is used to demonstrate the knowledge transfer and reuse mechanism by the machine learning approach. The dataset used in this study was acquired from an RBI assessment conducted on three onshore and four offshore oil and gas production and processing units, comprising over 200 pressure vessels and 3000 piping lines.

Six machine learning algorithms are selected and compared, to determine the best performing model. Ten iterations of 4-fold nested CV are performed for model selection and evaluation. The result of the study shows that ensemble techniques (i.e., GBDT, RF, and AB) perform better than single classifiers (i.e., LR, SVM, and *k*-NN). Among the ensemble classifiers, AB is the less preferred, as it performs the worst of all ensemble techniques. It is challenging to choose the preferred classifier between GBDT and RF, due to marginal performance disparity between them, lack of robustness in their performance difference, and their analogous performance variability. GBDT is slightly preferred because it outperforms RF in all evaluation metrics, except precision. Despite their excellent prediction performance, ensemble techniques are computationally more expensive and less comprehensible than single classifiers.

Additionally, a comparative analysis between the performance of the best performing machine learning model and the conventional assessment (i.e., manual assessment) is performed to examine the benefits of applying machine learning approach in the RBI screening assessment. It is shown that the application of machine learning approach eliminates appraiser-to-appraiser variation by removing the influence of subjective judgment from the assessment. The machine learning model also has better performance than the conventional assessment in terms of accuracy, precision, and recall score. It is recommended that the intelligent system based on machine learning should be complemented by human intelligence, so that it can refine the insights provided by the machine intelligence.

Further research should be conducted to incorporate other classification techniques (e.g., artificial neural networks and XGBoost, voting classifier, etc.) that have the potential to improve the predictive strength in the case of RBI screening assessment. A future study should be performed on the utilization of a machine learning approach, to tackle multiclass classification in the RBI detailed assessment.

Acknowledgement

This work has been carried out as a part of a Ph.D. research project at the University of Stavanger, funded by the Norwegian Ministry of Education and Research.

References

[1] Gordon RP. The contribution of human factors to accidents in the offshore oil industry. Reliability Engineering & System Safety. 1998;61:95-108.
[2] das Chagas Moura M, Lins ID, Droguett EL, Soares RF, Pascual R. A multi-objective genetic algorithm for determining efficient risk-based inspection programs. Reliability Engineering & System Safety. 2015;133:253-65.
[3] Ratnayake RMC, Markeset T. Technical integrity management: measuring HSE awareness using AHP in selecting a maintenance strategy. Journal of Quality in Maintenance Engineering. 2010;16:44-63.
[4] DNV GL. DNV GL-RP-G101: Risk Based Inspection of Offshore Topsides Static Mechanical Equipment. DNV-GL; 2017.
[5] API. Risk-Based Inspection: API Recommended Practice 580 (3rd ed.). Washington, D.C.: API; 2016.
[6] Drucker P. Knowledge-worker productivity: the biggest challenge. California Management Review. 1999;41:79-94.
[7] Geary W. Risk Based Inspection: A Case Study Evaluation of Offshore Process Plant. Sheffield: Health and Safety Laboratory; 2002.
[8] Ratnayake RMC. Application of a fuzzy inference system for functional failure risk rank estimation: RBM of rotating equipment and instrumentation. Journal of Loss Prevention in the Process Industries. 2014;29:216-24.
[9] Oppenheim BW. Lean product development flow. Systems Engineering. 2004;7:352-76.
[10] Morgan JM, Liker JK. The Toyota Product Development System: Integrating People, Process, and Technology. New York: Productivity Press; 2006.
[11] Thomke S, Fujimoto T. The effect of "front-loading" problem-solving on product development performance. Journal of Product Innovation Management. 2000;17:128-42.
[12] Hoppmann J, Rebentisch E, Dombrowski U, Zahn T. A framework for organizing lean product development. Engineering Management Journal. 2011;23:3-15.
[13] Ward AC, Sobek II DK. Lean Product and Process Development. Cambridge: Lean Enterprise Institute; 2014.
[14] Dutta S. Strategies for implementing knowledge-based systems. IEEE Transactions on Engineering Management. 1997;44:79-90.
[15] Huber GP. A theory of the effects of advanced information technologies on organizational design, intelligence, and decision making. Academy of Management Review. 1990;15:47-71.
[16] Langley P, Simon HA. Applications of machine learning and rule induction. Communications of the ACM. 1995;38:54-64.
[17] Bose I, Mahapatra RK. Business data mining—a machine learning perspective. Information & Management. 2001;39:211-25.
[18] Ford N. From information-to knowledge-management: the role of rule induction and neural net machine learning techniques in knowledge generation. Journal of Information Science. 1989;15:299-304.
[19] Vinod G, Sharma PK, Santosh TV, Hari Prasad M, Vaze KK. New approach for risk based inspection of H2S based process plants. Annals of Nuclear Energy. 2014;66:13-9.
[20] Seo JK, Cui Y, Mohd MH, Ha YC, Kim BJ, Paik JK. A risk-based inspection planning method for corroded subsea pipelines. Ocean Engineering. 2015;109:539-52.
[21] Kamsu-Foguem B. Information structuring and risk-based inspection for the marine oil pipelines. Applied Ocean Research. 2016;56:132-42.
[22] Luque J, Straub D. Risk-based optimal inspection strategies for structural systems using dynamic Bayesian networks. Structural Safety. 2019;76:68-80.

[23] Vinod G, Kushwaha HS, Verma AK, Srividya A. Optimisation of ISI interval using genetic algorithms for risk informed in-service inspection. Reliability Engineering & System Safety. 2004;86:307-16.
[24] Vinod G, Bidhar SK, Kushwaha HS, Verma AK, Srividya A. A comprehensive framework for evaluation of piping reliability due to erosion–corrosion for risk-informed inservice inspection. Reliability Engineering & System Safety. 2003;82:187-93.
[25] Fleming KN. Markov models for evaluating risk-informed in-service inspection strategies for nuclear power plant piping systems. Reliability Engineering & System Safety. 2004;83:27-45.
[26] Vinod G, Kushwaha HS, Verma AK, Srividya A. Importance measures in ranking piping components for risk informed in-service inspection. Reliability Engineering & System Safety. 2003;80:107-13.
[27] Simola K, Pulkkinen U, Talja H, Karjalainen-Roikonen P, Saarenheimo A. Comparison of approaches for estimating pipe rupture frequencies for risk-informed in-service inspections. Reliability Engineering & System Safety. 2004;84:65-74.
[28] Chang M-K, Chang R-R, Shu C-M, Lin K-N. Application of Risk Based Inspection in Refinery and Processing Piping. Journal of Loss Prevention in the Process Industries. 2005;18:397-402.
[29] Reynolds JT. Risk-based inspection - where are we today? Corrosion 2000. Houston: NACE International; 2000.
[30] El-Abbasy MS, Senouci A, Zayed T, Mirahadi F, Parvizsedghy L. Condition prediction models for oil and gas pipelines using regression analysis. Journal of Construction Engineering and Management. 2014;140:1-17.
[31] El-Abbasy MS, Senouci A, Zayed T, Mirahadi F, Parvizsedghy L. Artificial neural network models for predicting condition of offshore oil and gas pipelines. Automation in Construction. 2014;45:50-65.
[32] El-Abbasy MS, Senouci A, Zayed T, Parvizsedghy L, Mirahadi F. Unpiggable oil and gas pipeline condition forecasting models. Journal of Performance of Constructed Facilities. 2016;30:1-19.
[33] Layouni M, Hamdi MS, Tahar S. Detection and sizing of metal-loss defects in oil and gas pipelines using pattern-adapted wavelets and machine learning. Applied Soft Computing. 2017;52:247-61.
[34] Mohamed A, Hamdi MS, Tahar S. A Hybrid Intelligent Approach for Metal-Loss Defect Depth Prediction in Oil and Gas Pipelines. In: Bi Y, Kapoor S, Bhatia R, editors. Intelligent Systems and Applications: Extended and Selected Results from the SAI Intelligent Systems Conference (IntelliSys) 2015. Cham: Springer International Publishing; 2016. p. 1-18.
[35] Carvalho AA, Rebello JMA, Sagrilo LVS, Camerini CS, Miranda IVJ. MFL signals and artificial neural networks applied to detection and classification of pipe weld defects. NDT & E International. 2006;39:661-7.
[36] Abdulla MB, Herzallah R. Probabilistic multiple model neural network based leak detection system: Experimental study. Journal of Loss Prevention in the Process Industries. 2015;36:30-8.
[37] Zadkarami M, Shahbazian M, Salahshoor K. Pipeline leakage detection and isolation: an integrated approach of statistical and wavelet feature extraction with multi-layer perceptron neural network (MLPNN). Journal of Loss Prevention in the Process Industries. 2016;43:479-87.
[38] Wenhui H, Ye W, Jie G, Yi J, Chang'an Z. Automatic detection of welding defects using deep neural network. Journal of Physics: Conference Series. 2018;933:012006.
[39] Vilar R, Zapata J, Ruiz R. An automatic system of classification of weld defects in radiographic images. NDT & E International. 2009;42:467-76.
[40] Kumar J, Anand RS, Srivastava SP. Multi-class welding flaws classification using texture feature for radiographic images. International Conference on Advances in Electrical Engineering2014. p. 1-4.
[41] Mirapeix J, García-Allende PB, Cobo A, Conde OM, López-Higuera JM. Real-time arc-welding defect detection and classification with principal component analysis and artificial neural networks. NDT & E International. 2007;40:315-23.
[42] Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. IEEE Access. 2017;5:8869-79.
[43] Onan A. A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Systems with Applications. 2015;42:6844-52.
[44] Asri H, Mousannif H, Moatassime HA, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science. 2016;83:1064-9.
[45] Polat K, Güneş S. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing. 2007;17:702-10.
[46] Das R, Turkoglu I, Sengur A. Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications. 2009;36:7675-80.
[47] Abellán J, López G, De OñA J. Analysis of traffic accident severity using decision rules via decision trees. Expert Systems with Applications. 2013;40:6047-54.
[48] API. Risk-Based Inspection Methodology: API Recommended Practice 581 (3rd ed.). Washington, D.C.: API; 2016.
[49] Kira K, Rendell LA. A Practical Approach to Feature Selection. In: Sleeman D, Edwards P, editors. Machine Learning Proceedings 1992. San Francisco: Morgan Kaufmann; 1992. p. 249-56.
[50] Maldonado S, Weber R. A wrapper method for feature selection using Support Vector Machines. Information Sciences. 2009;179:2208-17.
[51] Stańczyk U. Feature Evaluation by Filter, Wrapper, and Embedded Approaches. In: Stańczyk U, Jain LC, editors. Feature Selection for Data and Pattern Recognition. Berlin: Springer Berlin Heidelberg; 2015. p. 29-44.
[52] Chizi B, Maimon O. Dimension Reduction and Feature Selection. In: Maimon O, Rokach L, editors. Data Mining and Knowledge Discovery Handbook. Boston: Springer US; 2010. p. 83-100.
[53] Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence. 1997;97:273-324.
[54] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research. 2003;3:1157-82.
[55] Tsang-Hsiang C, Chih-Ping W, Tseng VS. Feature selection for medical data mining: comparisons of expert judgment and automatic approaches. The 19th IEEE Symposium on Computer-Based Medical Systems2006. p. 165-70.
[56] Domingos P. The role of Occam's razor in knowledge discovery. Data Mining and Knowledge Discovery. 1999;3:409-25.
[57] Yoon S-C, Henschen LJ, Park EK, Makki S. Using domain knowledge in knowledge discovery. The 8th International Conference on Information and Knowledge Management. Kansas City: ACM; 1999. p. 243-50.
[58] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23:2507-17.
[59] Kuhn M, Johnson K. Applied Predictive Modeling. New York: Springer; 2013.
[60] Armitage DW, Ober HK. A comparison of supervised learning techniques in the classification of bat echolocation calls. Ecological Informatics. 2010;5:465-73.
[61] Hu X. DB-HReduction: a data preprocessing algorithm for data mining applications. Applied Mathematics Letters. 2003;16:889-95.
[62] Han J, Pei J, Kamber M. Data Mining: Concepts and Techniques. San Francisco: Elsevier; 2011.
[63] Al-Mamory SO, Hasson ST, Hammid MK. Enhancing attribute oriented induction of data mining. Journal of University of Babylon. 2013;21:2286-95.

[64] Han J, Cai Y, Cercone N, Huang Y. Discovery of data evolution regularities in large databases. Journal of Computer and Software Engineering. 1994.

[65] Zheng A, Casari A. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. Sebastopol: O'Reilly Media; 2018.

[66] Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann; 2016.

[67] Guyon I, Elisseeff A. An Introduction to Feature Extraction. In: Guyon I, Nikravesh M, Gunn S, Zadeh LA, editors. Feature Extraction: Foundations and Applications. Berlin: Springer; 2006. p. 1-25.

[68] Zhang C, Liu C, Zhang X, Almpanidis G. An up-to-date comparison of state-of-the-art classification algorithms. Expert Systems with Applications. 2017;82:128-50.

[69] Brown I, Mues C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications. 2012;39:3446-53.

[70] Cox DR. The regression analysis of binary sequences. Journal of the Royal Statistical Society Series B. 1958;20:215-42.

[71] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000;16:906-14.

[72] Huang W, Nakamori Y, Wang S-Y. Forecasting stock market movement direction with support vector machine. Computers & Operations Research. 2005;32:2513-22.

[73] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20:273-97.

[74] Kelleher JD, Mac Namee B, D'Arcy A. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. Cambridge: MIT Press; 2015.

[75] Aha DW. Editorial. In: Aha DW, editor. Lazy Learning. Dordrecht: Springer Netherlands; 1997. p. 7-10.

[76] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967;13:21-7.

[77] Schapire RE. The Boosting Approach to Machine Learning: An Overview. Nonlinear Estimation and Classification. New York: Springer; 2003. p. 149-71.

[78] Friedman JH. Greedy function approximation: a gradient boosting machine. The Annals of Statistics. 2001;29:1189-232.

[79] Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. Journal of Animal Ecology. 2008;77:802-13.

[80] Freund Y, Schapire RE. Experiments with a new boosting algorithm. International Conference on Machine Learning1996. p. 148-56.

[81] Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing. 2012;67:93-104.

[82] Pal M. Random forest classifier for remote sensing classification. International Journal of Remote Sensing. 2005;26:217-22.

[83] Breiman L. Random forests. Machine Learning. 2001;45:5-32.

[84] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91-.

[85] Kothari S, Phan JH, Young AN, Wang MD. Histological image classification using biologically interpretable shape-based features. BMC Medical Imaging. 2013;13:9.

[86] Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification. 2003.

[87] Refaeilzadeh P, Tang L, Liu H. Cross-Validation. In: Liu L, ÖZsu MT, editors. Encyclopedia of Database Systems. Boston: Springer US; 2009. p. 532-8.

[88] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. The 23rd International Conference on Machine Learning. Pittsburgh: ACM; 2006. p. 233-40.

[89] Provost F, Fawcett T. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. The 3rd International Conference on Knowledge Discovery and Data Mining. Newport Beach: AAAI Press; 1997. p. 43-8.

[90] Japkowicz N, Shah M. Evaluating Learning Algorithms: A Classification Perspective. New York: Cambridge University Press; 2011.

[91] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27:861-74.

[92] Parker C. An analysis of performance measures for binary classifiers. The 11th International Conference on Data Mining. Vancouver: IEEE; 2011. p. 517-26.

[93] Good P. Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. New York: Springer Science & Business Media; 1994.

[94] Yeh A. More accurate tests for the statistical significance of result differences. The 18th conference on Computational linguistics - Volume 2. Saarbrücken: Association for Computational Linguistics; 2000. p. 947-53.

[95] Raschkas S. Mlxtend. 2016.

[96] Pyzdek T, Keller PA. The Six Sigma Handbook. New York: McGraw-Hill; 2014.

[97] Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. Artificial Intelligence Review. 2006;26:159-90.

[98] Salzberg SL. On comparing classifiers: pitfalls to avoid and a recommended approach. Data Mining and Knowledge Discovery. 1997;1:317-28.

[99] Hodson J. How to Make Your Company Machine Learning Ready. HBR OnPoint: Harvard Business Review; 2017.

[100] Sharma R, Mithas S, Kankanhalli A. Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. European Journal of Information Systems. 2014;23:433-41.

[101] Hinckley CM. Combining mistake-proofing and Jidoka to achieve world class quality in clinical chemistry. Accreditation and Quality Assurance. 2007;12:223-30.