Journal of Applied Sciences, Engineering and Technology for Development. Volume 2, Issue 1

Research Article **ISSN 2309-0936**

A DECISION TREE-BASED CLASSIFICATION FRAMEWORK FOR USED OIL ANALYSIS APPLYING RANDOM FOREST FEATURE SELECTION

WAKIRU*[1], J., PINTELON[1], L., CHEMWENO[1], P., & MUCHIRI[2], P.N.

[1]Centre for Industrial Management, Traffic and Infrastructure, KU Leuven, Celestijnenlaan 300A, Heverlee 3001, Belgium.
[b]Dedan Kimathi University of Technology. P. O. Box 657-10100 Nyeri, Kenya.

Received January 2018; Revised March 2018; Accepted April 2018

ABSTRACT

Lubricant condition monitoring (LCM), part of condition monitoring techniques under Condition Based Maintenance, monitors the condition and state of the lubricant which reveal the condition and state of the equipment. LCM has proved and evidenced to represent a key concept driving maintenance decision making involving sizeable number of parameter (variables) tests requiring classification and interpretation based on the lubricant's condition. Reduction of the variables to a manageable and admissible level and utilization for prediction is key to ensuring optimization of equipment performance and lubricant condition. This study advances a methodology on feature selection and predictive modelling of in-service oil analysis data to assist in maintenance decision making of critical equipment.

Proposed methodology includes data pre-processing involving cleaning, expert assessment and standardization due to the different measurement scales. Limits provided by the Original Equipment Manufacturers (OEM) are used by the analysts to manually classify and indicate samples with significant lubricant deterioration. In the last part of the methodology, Random Forest (RF) is used as a feature selection tool and a Decision Tree-based (DT) classification of the in-service oil samples. A case study of a thermal power plant is advanced, to which the framework is applied.

The selection of admissible variables using Random Forest exposes critical used oil analysis (UOA) variables indicative of lubricant/machine degradation, while DT model, besides predicting the classification of samples, offers visual interpretability of parametric impact to the classification outcome. The model evaluation returned acceptable predictive, while the framework renders speedy classification with insights for maintenance decision making, thus ensuring timely interventions. Moreover, the framework highlights critical and relevant oil analysis parameters that are indicative of lubricant degradation; hence, by addressing such critical parameters, organizations can better enhance the reliability of their critical operable equipment.

Keywords: Random forest, Decision trees, oil analysis, Maintenance decision support, Lubricant condition monitoring

*Correspondence to: James Wakiru, Centre for Industrial Management, Traffic and Infrastructure, KU Leuven, Celestijnenlaan, 300A, Heverlee 3001, Belgium. Email: jamesmutuota.wakiru@kuleuven.be

DOI: 10.33803/JASETD.2017.3-1.7

1. INTRODUCTION

In a lubricant condition monitoring or used oil analysis (UOA) program, four main areas are monitored and highlighted, that is, changes in the physical and chemical properties, contamination, component wear through ingression of wear particles and additive analysis which would indicate depletion of the crucial components in the additive. A single sample of oil may have over twenty parameters tested whilst the analyst should manually review each parameter and possibly derive some trend graphs to confirm if the sample is okay or necessitates some relevant action like top-up or change. As Wakiru *et.al*., (2017) alludes, the exercise takes considerable time and introduces errors due to the high dimensionality of the parameters which would to a more considerable extent and require reduction without compromising importance. Moreover, use of historical performance is not considered while classifying the samples manually, rendering the results insufficient. This study was motivated by the need to develop a moderate method of select admissible variables, and further developing a classification model which could provide parametric associative and interactive insights in a fast and accurate manner.

The random forest (RF) also known as random decision forest is a machine learning classification tool that uses a group of classification or regression tress to rank explanatory variables or predictors through its inbuilt measures of variable importance (Janitza et.al., 2016). RF commences with a standard decision tree as a weaker learner, it ensembles the frail learners jointly developing a “strong learner” offering a more altruistic classification to the underlying data. Important predictor variables in a dataset can be selected using variable of importance measure, an important feature of random forests (Hapfelmeier, *et.al.,* 2014).

RF offers high prediction accuracy and is able to identify or rate a variable’s influence to the outcome or prediction which possesses no missing values (Hapfelmeier, *et al.*, 2014). This is the fundamentalreason for the increased use of random forest (RF) techniques in variable selection. However, methods like multiple imputation and complete case analysiscould be employed when the data embodies missing values.In theirstudy(Jotheeswaran & Koteeswaran, 2016) compared RF with Principal component analysis (PCA) and Decision Trees (DT) in variable selection and concluded that using RF, the precision of classifiers improved than the others. In his study to downscale temperatures on the land surface, (Hutengs & Vohland, 2016), used RF to select the critical variables suggesting the number of variables included influenced the importance score, while a change in the importance score could also be attributed to predictors changing or replaced (Aldrich & Auret, 2010), while investigating fault conditions, employed RF to identify variables in the process that had high contribution to faultiness. RF was used to reduce 33 variables measured in a diagnosis exercise using breast images to 15 important variables (Nguyen, et.al., 2013). RF was proven useful for extracting the important variables with missing data (Hapfelmeier, *et al.*, 2014). On the contrast, RF has a limitation in variable selection in that it does not discriminate non-correlated attributes (Cerrada, *et al*., 2016). From the review, the researcher did not find any application of RF variable selection in LCM.

Decision trees (DT), is a machine learning tool that utilizes tree-like structures or model of decision and their possible consequence or outcome building an associated decision tree incrementally (Nasridinov, *et.al*., 2013). In his outline on DT, (Rokach & Maimom, 2007) outlines some benefits of the method, such as flexibility for a wide variety of data mining tasks, performing variable screening, requires relatively less effort in data preparation and self-explanatory and easy to interpret results. Other advantages include robustness in performance even with nonlinear parametric relations, and versatility in handling a variety of input data. DT classification was used in LCM to generate the model predicting wear conditions of the equipment's using UOA wear particles data and failure events data (Ide, *et.al.*, 2015). In this study selected variables were used with focus on the wear metals. The basis and procedure of selecting the variables was not clearly outlined. In other fields, DT was used to classify Power Quality disturbances (Ray, *et.al.*, 2014), in banking application (Chitra Devi, 2014), crime prediction (Nasridinov, *et al.*, 2013), classifying mobile LiDAR data(Guan, *et.al.*, 2015), automatic classification of patients in the Hashimoto's disease diagnosis (Omiotek, *et.al.*, 2013), classification in the prediction antibody incompatible kidney transplantation (Shaikhina, *et al.*, 2015).

Despite the primary advantage of ease of interpretability, DT has several limitations such as, a modest change in one of more variables generating a different tree, could lead to over fitted results and it is insensitive to missing data as well as inclusion of irrelevant predictors as outlined by (Rokach & Maimom, 2007).

2. METHODOLOGY

In-service oil data

The data used for this study involved 1103 in-service oil analysis samples from a thermal power plant running on heavy fuel oil for the period between 2010 and 2015. The power plant carries out scheduled in-service oil analysis, where the oil samples are tested in an independent laboratory and the analysts manually classify the sample results indicating the health condition of the oil either okay or fail, where maintenance intervention is needed.The variables tested included Nickel, Calcium, Viscosity at 40^{o}C, Aluminium, Sodium, TBN, Silicon, Iron, Lead, Zinc, Viscosity 100^{o}C, Pentane insolubles,Flash point, Water, and Chromium. Data was split and varied in different ratios such as 80:20 to 70:30, where the 70:30 depicts that70% of the samples, randomly sampled, were used for model training, while 30% were used for testing independently.

Variable selection by Random forest

While using Random Forest, the predictor variables are ranked according to their contribution in predicting the output, response or prediction. The random forest algorithm introduces random sampling by building several decision trees using bagging (collection of random sample of observations into a bag or bootstrap aggregation). From the training data set, randomly selected observations are obtained to create multiple. The importance or influence of the variables while being employed to build the multiple decision trees is considered while selecting the important variables (De Rivas, *et.al.*, 2017).

While establishing the important variables, two measures are computed. The first measure is the mean decrease in accuracy that reports the model's accuracy decline and is based on influence of each variable to the prediction error. The second measure of importance is the mean decrease, Gini, which is employed to select the splitting criterion or decision tree node's impurity decrease while the variable is being split. A generalization splitting criterion of a binomial classification employing 10-fold (default) cross-validation is termed as Gini index. Gini index has an advantage that it can be used with missing data but with a shortcoming like undesired variable ranks with categorical frequencies (Hapfelmeier, *et al.*, 2014). Mean Decrease Gini index is then a measure of node purity since Gini Index is taken as an impurity measure (Freyhult, Gustafsson, & Strömbergsson, 2015; Rutkowski, Jaworski, Pietruczuk, & Duda, 2014).

Classification using Decision Tree

Decision trees using a decision tree as a predictive model, are widely used as they offer information about a variable to the conclusion on its classification considering the target value or outcome. The variables are split repetitively until the classification is attained signifying the termination(Ray, *et al.*, 2014).

Model building

While forming the decision tree model, the occurrences reflected from the decision and internal nodes, form a hierarchy of branches. Considering a tree, the path from the root (root node) to the leaf (leaf node) indicates a classification decision rule.The DT algorithm can use the following definitions.

$$X = \{X_1, X_2, \dots \dots .. X_m\}^T \qquad (1)$$

$$X_i = \{x_1, x_2, x_{ij} \; \dots . x_{in}\} \qquad (2)$$

$$S = \{S_1, S_2, \dots S_i \; \dots . S_m\} \qquad (3)$$

Where *m* is the available observations number in this study we have two observations of “FAIL” or “PASS”, *n* is the independent variables number (twenty UOA parameters), S is the *m*-dimension vector from X, while X_i, is the *i*th component vector of n-dimension variables. The autonomous variables x_{i1}, $x_{i2}, \dots x_{ij} \dots, x_{in}$form the pattern vector$X_i$, while T is the transpose notation vector.

Model tuning

Pruning or tuning is a machine learning technique which eliminates over fitting, a common problem with decision tree models. This is performed by removing the nodes on the decision tree that demonstrate the least influence on the overall performance of the classification model (Breiman, *et.al.*, 1984). The minimum split was varied from 9 to the default 20 to eliminate nodes with a small number of observations, which generally have less contribution or influence in the developed classification model, where the optimal splitting was evaluated at value 10. The complexity parameter (cp) determines the significance of making or not making a split of the decision tree by quantifying a benefit conditionally to be gained before the process. This restricts the size of the decision tree and enhances selection of an optimal tree size. The data set was manipulated where the setting of the complexity parameter was varied from default 0.001 to 0.01, of which, optimal splitting was evaluated at value 0.0.

Model validation

The goodness of fit of the developed decision tree as a classification tool is depicted in terms of its performance and predictive power.Performance is indicated by the classification rate and misclassification rate, while the area under the receiver operator curve (ROC) demonstrates the model’s predictive power.The ROC chart is a graph plotting the sensitivity (true positive rate) against specificity (false positive rate). The lift chart depicts the effectiveness of the classification model by computing the ratio of the results obtained while using the model and while not employing the model.

3. RESULTS AND DISCUSSION

This section will illustrate the results and incorporate discussions with insights from the results.

Variable selection using RF

The visual plot of the the mean decrease in Gini index shown in Figure 1, illustrates the relative importance of the variables.

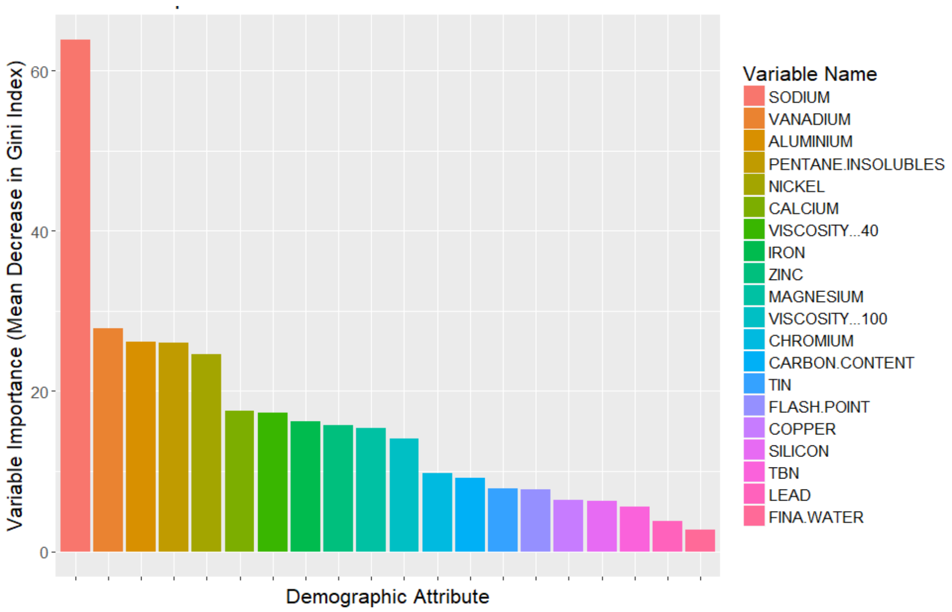

Figure 1: Graphical representation of variable importance using random forest

Sodium is the most significant admissable variable. Vanadium is listed next, then Aluminium, Pentane insolubles, Nickel, Calcium then Viscosity at 40^{o}C.The variables selected represent all fundamental categories of the UOA. Sodium assumes the most significant importance in this dataset which could infer a contamination problem. Sodium ingression can be from saline water or anti-freeze inhibitor found in the coolant, hence indicating high probability of high water or coolant leaks.

Classification using DT

DT classification model was built with the variables selected as important for the study. The model utilized the variables with more than 8% of mean decrease in Gini index which meant Water, Lead, TBN and Silicon were excluded in the model as the index made no significant change in the fourth and fifth variable.

Decision tree model tuning

The DT classification model was built using the default minimum split of 20, minimum bucket of 7, cp of 0.010 and provided an overall accuracy of 96.61%. The model was tuned and modified to a minimum split of 10, minimum bucket of 3, cp of 0.0001 while the surrogate values as 0 returning an overall accuracy of 97.53%. The same configuration was applied while varrying the data split ratios from 70:30 (70% training and 30% testing) to 90:10 which is illustrated in table 1.

From the comparison depicted in table 1, the data split ratio of 70:30 attained the highest area under the curve, meaning it offered the highest predictive power among the compared ratios. Likewise, the classification accuracy of 98.2% (0.982) was the highest, signifying high accuracy.

Table 1: Comparison of data split ratio variation with model performance

Data split ratio	AUC	Classification accuracy
70:30	0.9753	0.9820
75:25	0.9681	0.9782
80:20	0.9612	0.9740
85:15	0.9515	0.9639
90:10	0.9605	0.9640

In general, the variation of the AUC and and classification accuracy in the comparison is not significant and this shows the results whil not significantly changes using different data split ratios.

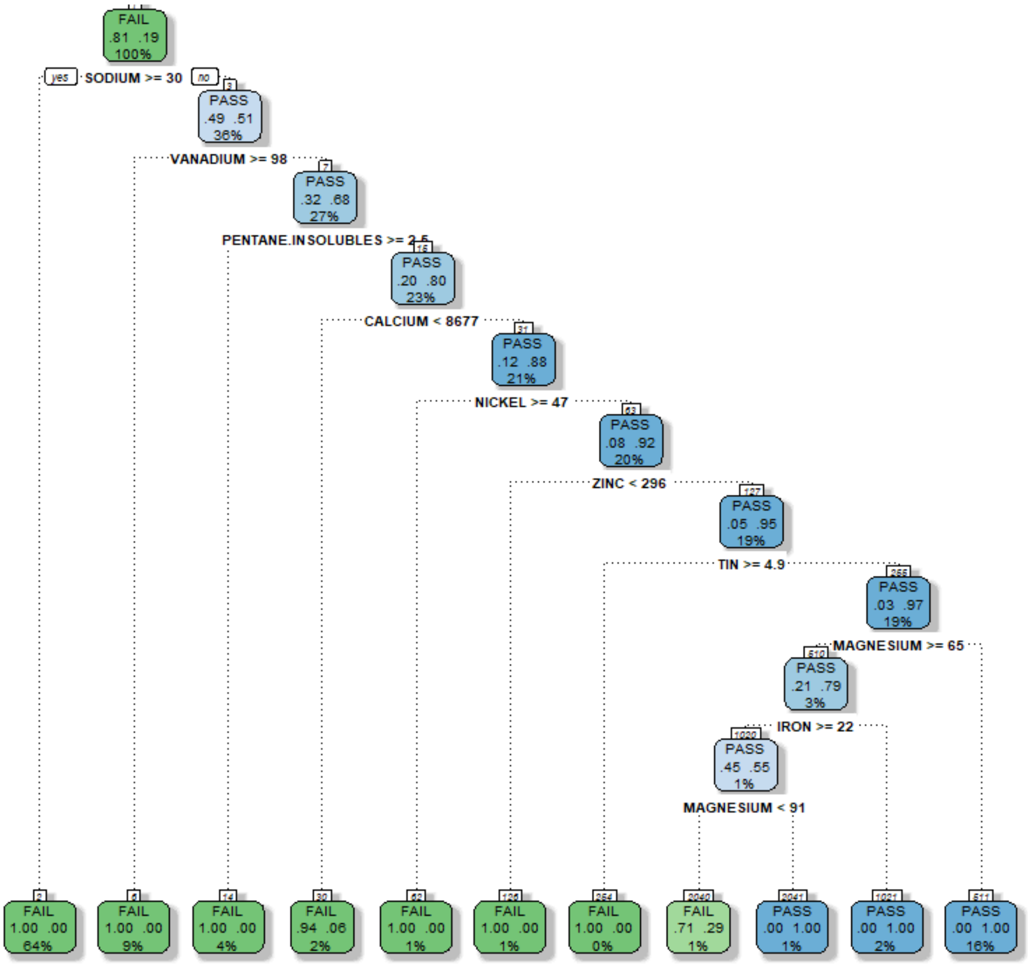

Figure 2: Schematic representation of decision tree generated

The DT modelled above can be used to classify a sample with good interpretation of the variables. Table 2, represent a sample results as an example.While classifying the sample in table 2 using the decision tree from the top node, it exceeds the thresholds in parts per million (ppm) of sodium (< 30ppm), vanadium (< 98ppm), pentane insolubles (< 2.5%), calcium (<8677ppm) and nickel (<47ppm), but fails due to zinc level being less than 296ppm. This indicates the sample represented by the values in Table 2 would fail due tio zinc being higher than allowable limit. This would offer the maintenance team the opportunity to investigate the probable reasons for the depletion of zinc an essential ingredient in the anti-oxidation and anti-wear additive zinc dialkyldithiophosphate (ZDDP), which could lead to increased wear.

Table 2: Example of sample results for classification

Parameter	Value	Parameter	Value
Viscosity @ 40^{o}C	144	Carbon content	0.11
Viscosity @ 100^{o}C	14.7	Iron	11.1
Flash point	180	Chromium	0.16
Magnnesium	22.6	Copper	7.5
Calcium	8460	Tin	0.08
Zinc	238	Aluminium	1.4
Sodium	5.1	Nickel	7.7
Pentane Insolubles	0.01	Vanadium	19.4

Model validation and performance

Classification table

The classification table, table 3, indicates the model prediction utilizing the testing data, was used to calculate the sensitivity, which represent the proportion of events (1 or PASS) predicted as events (1), specificity which indicates the proportion of non-events (0 or FAIL) predicted as non-events (0) and false positive which indicate the number of non-events (0) predicted as events (1). Classification of observations using the prediction where the test data was used, was done based on a cut off value of 0.5 giving Sensitivity of 94.44% ({51/ (51+3)}), Specificity of 98.92% ({274/ (274+3)}) and a false positive of 1.08%, all computed from results in table 3.

Table 3: Classification table -cut-off level of 0.5

	Actual 'FAIL'	Actual 'PASS'
Classified 'FAIL'	274	3
Classified 'PASS'	3	51

AUC (Area under the ROC curve) and classification accuracy

The receiver operator curve (ROC), as shown in figure 3, illustrates the decision tree classification model's predictive power by computing the area under the curve (AUC). A model with a high AUC indicates to be one with a higher predictive power i.e. able to classify a 'FAIL' sample from the 'PASS' samples.The classification accuracy of a model is computed as the proportion of the events and non-events classified as events and non-events accurately to the total number of outcomes. The ROC curve developed is shown on figure 3, while AUC for the decision tree model is 0.9753≃0.98. The classification accuracy which can be computed as {(274+51)/ (274+51+3+3)} giving 0.98.

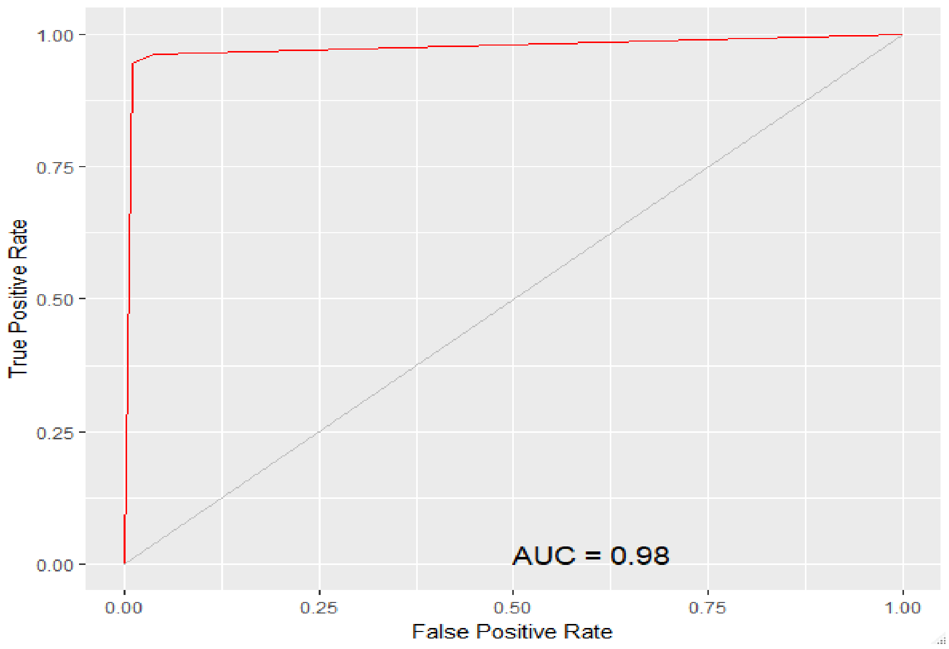

Figure 3: Receiver Operating Curve (ROC) curve

The prediction of the model returns an overall error of 2% and an average class error of 3% which is admissible.

Risk chart

A risk chart or cummulative gain chart, which offers another perspective or aspect of a binary classification model by exposing the gain expected in identifing samples requiring attention using the developed model compared to using random sampling. This aspect offers a additional frame of reference while considering the performance of a binary classification model such as decision tree (DT).

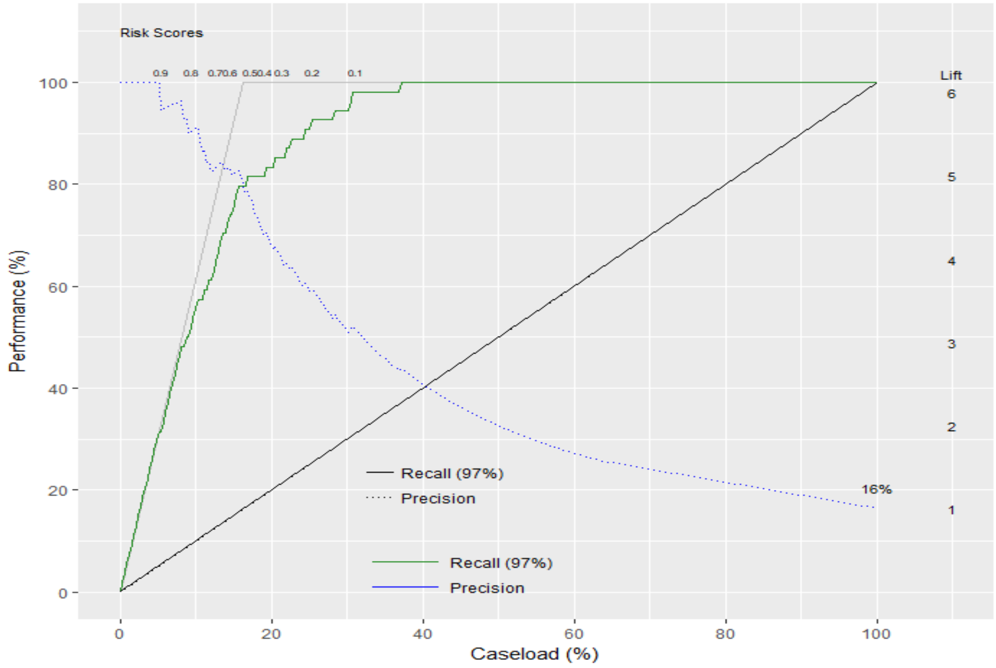

Figure 4: A risk chart for a decision tree on the plant data

Suppose each year 100 samples are tested, the strike rate of 16% would give 16 out of 100 samples as being of interest. If after evaluation, only 50 samples could be tested, 8 samples of the 16 would be identified requiring attention. A random 50% case load would deliver 50% performance hence only half of the classified samples would be found. The performance achieved in classifying the samples while adopting the developed model is illustrated by the dashed green line of the plot. Adopting the model, the practitioner expects to identify 95% of the samples needing attention or intervention and depicted by the approximately 95% performance on a 50% case load. So 15 of the 16 samples classified, are anticipated to be correctly classified from the 50 samples, a consequential change beyond the 8 randomly selected. The blue line in the plot depicts the lift in performance, in this case a value over 3. This means that the maintenance practitioner is able to identify almost thrice as many samples that require intervention than he could expect if sample classification was performed randomly. This offers an enhanced response with respect to the population as a whole, exposure eventually improving diagnosis and prediction in terms of the lubricant performance. At 95% case load the model achieves 100% performance, hence all samples requiring attention will have been identified by the time 95 samples have been classified. Hence, using the model ensures almost all samples required i.e 16 would be classified, further with a 5% savings in the effort expended to classify all of the required samples previously. This aspect provides a higher accuracy in classifyng the samples in the LCM program.

4. CONCLUSION

The variable selection using RF uncovered the important and admissible variables for the classification based on the dataset. The important variables could be treated as critical variables that keen interest could be turned to for instance investigation on the causes of deviations. Decision tree was modelled and tuned which improved its predictive power from 96.61% to 97.53% hence improving classification output score towards high accuracy thus high reliability. Classification or scoring of data requires thresholding, which defines probability intervals for each class or score hence making it completely adaptable for the UOA sample classification. The schematic representation of the DT provides meaningful insights for maintenance decision making, due to its interpretability. The use of different data set for model building (training) and testing, measures the performance of the model cases not visible previously, moreover alludes to the importance of historical data to the future classification. This makes DT modelling a tool that is easily adoptable by the maintenance engineers to evaluate samples as the model is updated using new generated data, moreover, this will reduce the time to make maintenance decisions and significantly reduce errors that may arise due to manual classification of in-service oil samples.

Varying various model parameters in the DT can effectively increase the scope of information for example use of cp = 0, would produce a full decision tree which management could review the various limits susceptible to causing the sample to fail which could be used for UOA parameter threshold limit setting in specific LCM programs.

Subsequent work would require root cause analysis on the identified significant variables and comparison of the DT model with other "black box" models.

REFERENCES

Aldrich, C., & Auret, L. (2010). *Fault detection and diagnosis with random forest feature extraction and variable importance methods*. *IFAC Proceedings Volumes (IFAC-PapersOnline)* (Vol. 13). IFAC. https://doi.org/10.3182/20100802-3-ZA-2014.00020

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees, *1*(February), 368. https://doi.org/10.1371/journal.pone.0015807

Cerrada, M., Zurita, G., Cabrera, D., Sánchez, R.-V., Artés, M., & Li, C. (2016). Fault diagnosis in spur gears based on genetic algorithm and random forest. *Mechanical Systems and Signal Processing*, *70–71*, 87–103. https://doi.org/10.1016/j.ymssp.2015.08.030

Chitra Devi, J. (2014). Binary Decision Tree Classification based on C4.5 and KNN Algorithm for Banking Application. *International Journal of Computational Intelligence and Informatics*, *4*(2). Retrieved from http://www.periyaruniversity.ac.in/ijcii/issue/Vol4No2September2014/IJCII-4-2-145.pdf

De Rivas, B. L., Vivancos, J.-L., Ordieres-Meré, J., & Capuz-Rizo, S. F. (2017). Determination of the total acid number (TAN) of used mineral oils in aviation engines by FTIR using regression models. *Chemometrics and Intelligent Laboratory Systems*, *160*(October 2016), 32–39. https://doi.org/10.1016/j.chemolab.2016.10.015

Freyhult, E., Gustafsson, M. G., & Strömbergsson, H. (2015). A machine learning approach to explain drug selectivity to soluble and membrane protein targets. *Molecular Informatics*, *34*(1), 44–52. https://doi.org/10.1002/minf.201400121

Guan, H., Yu, Y., Ji, Z., Li, J., & Zhang, Q. (2015). Deep learning-based tree classification using mobile LiDAR data. *Remote Sensing Letters*, *6*(11), 864–873. https://doi.org/10.1080/2150704X.2015.1088668

Hapfelmeier, A., Hothorn, T., Ulm, K., & Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Statistics and Computing*, *24*(1), 21–34. https://doi.org/10.1007/s11222-012-9349-1

Hutengs, C., & Vohland, M. (2016). Downscaling land surface temperatures at regional scales with random forest regression. *Remote Sensing of Environment*, *178*, 127–141. https://doi.org/10.1016/j.rse.2016.03.006

Ide, D., Ruike, A., & Kimura, M. (2015). Extraction of causalities and rules involved in wear of machinery from lubricating oil analysis data. In *The Second International Conference on Digital*

Information Processing, Data Mining, and Wireless Communications (DIPDMWC2015) (pp. 16–22). Wilmington, New Castle, DE 19801, USA: The Society of Digital Information and Wireless Communications (SDIWC).

Janitza, S., Tutz, G., & Boulesteix, A. L. (2016). Random forest for ordinal responses: Prediction and variable selection. *Computational Statistics and Data Analysis*, *96*, 57–73. https://doi.org/10.1016/j.csda.2015.10.005

Jotheeswaran, J., & Koteeswaran, S. (2016). Feature Selection using Random Forest Method for Sentiment Analysis. *Indian Journal of Science and Technology*, *9*(3). https://doi.org/10.17485/ijst/2016/v9i3/86387

Nasridinov, A., Ihm, S.-Y., & Park, Y.-H. (2013). A Decision Tree-Based Classification Model for Crime Prediction (pp. 531–538). Springer, Dordrecht. https://doi.org/10.1007/978-94-007-6996-0_56

Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering*, *6*(5), 551–560. https://doi.org/10.4236/jbise.2013.65070

Omiotek, Z., Burda, A., & Wójcik, W. (2013). The use of decision tree induction and artificial neural networks for automatic diagnosis of Hashimoto's disease. *Expert Systems with Applications*, *40*(16), 6684–6689. https://doi.org/10.1016/j.eswa.2013.03.022

Ray, P. K., Mohanty, S. R., Kishor, N., & Catalao, J. P. S. (2014). Optimal Feature and Decision Tree-Based Classification of Power Quality Disturbances in Distributed Generation Systems. *IEEE Transactions on Sustainable Energy*, *5*(1), 200–208. https://doi.org/10.1109/TSTE.2013.2278865

Rokach, L., & Maimom, O. (2007). *Data mining with decision trees: theory and applications*. *Data Mining and Knowledge Discovery*. https://doi.org/10.1007/978-0-387-09823-4

Rutkowski, L., Jaworski, M., Pietruczuk, L., & Duda, P. (2014). The CART decision tree for mining data streams. *Information Sciences*, *266*, 1–15. https://doi.org/10.1016/j.ins.2013.12.060

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., & Khovanova, N. (2015). Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomedical Signal Processing and Control*, 1–7. https://doi.org/10.1016/j.bspc.2017.01.012

Wakiru, J., Pintelon, L., Muchiri, P. N., & Chemweno, P. 1. (2017). A statistical approach for analyzing used oil data and enhancing maintenance decision making : Case study of a thermal power. In *2nd International Conference on Maintenance Engineering (IncoME-II 2017)* (pp. 117–128).