Article

Pattern Discovery in White Etching Crack Experimental Data Using Machine Learning Techniques

Baher Azzam [1],*, Freia Harzendorf [1], Ralf Schelenz [1], Walter Holweger [2] and Georg Jacobs [1]

[1]   Center for Wind Power Drives, RWTH Aachen University, 52074 Aachen, Germany;
      freia.harzendorf@cwd.rwth-aachen.de (F.H.); ralf.schelenz@cwd.rwth-aachen.de (R.S.);
      georg.jacobs@cwd.rwth-aachen.de (G.J.)
[2]   Schaeffler Technologies AG & Co. KG, 52134 Herzogenaurach, Germany; walter.holweger@t-online.de
*   Correspondence: baher.azzam@cwd.rwth-aachen.de; Tel.: +49-241-80-20543

check for updates

Abstract: White etching crack (WEC) failure is a failure mode that affects bearings in many applications, including wind turbine gearboxes, where it results in high, unplanned maintenance costs. WEC failure is unpredictable as of now, and its root causes are not yet fully understood. While WECs were produced under controlled conditions in several investigations in the past, converging the findings from the different combinations of factors that led to WECs in different experiments remains a challenge. This challenge is tackled in this paper using machine learning (ML) models that are capable of capturing patterns in high-dimensional data belonging to several experiments in order to identify influential variables to the risk of WECs. Three different ML models were designed and applied to a dataset containing roughly 700 high- and low-risk oil compositions to identify the constituting chemical compounds that make a given oil composition high-risk with respect to WECs. This includes the first application of a purpose-built neural network-based feature selection method. Out of 21 compounds, eight were identified as influential by models based on random forest and artificial neural networks. Association rules were also mined from the data to investigate the relationship between compound combinations and WEC risk, leading to results supporting those of previous analyses. In addition, the identified compound with the highest influence was proved in a separate investigation involving physical tests to be of high WEC risk. The presented methods can be applied to other experimental data where a high number of measured variables potentially influence a certain outcome and where there is a need to identify variables with the highest influence.

Keywords: feature selection; premature bearing failure; wind turbine gearbox failure; random forests; artificial neural networks; association rules

1. Introduction

White etching cracks (WECs) are a phenomenon in industrial gearboxes as well as automotive applications leading to a yet unpredictable failure called white structure flaking (WSF) or white etching cracks [1,2]. Such failures are responsible for significant operating and downtime costs, especially for wind turbine gearboxes [3–5]. To put the importance of WEC failures into perspective, WECs can lead to bearing failures within 5%–20% of the bearing's L10 lifespan (DIN ISO 281), leading to unexpected maintenance costs [6–11]. Several recent investigations aimed at understanding the causes behind WEC related bearing failure in wind turbine gearboxes [5,11–20]. However, the causes of WEC bearing failures remain a subject of debate [5,10,21–23].

Several root cause investigations of WEC failures found that lubricants and their components, e.g., additives, can play an important role in leading to WECs [22,24,25]. In one investigation, an experiment

was performed with a lubricant composed of only a base oil (no additives), resulting in no WEC failure even after 1000 h of testing, while another experiment with a lubricant containing over-based calcium sulfonates as rust preventers and short-chain zinc dithiophosphates as antiwear additives resulted in WEC failure after 40 h of testing [22]. Paladugu et al. also performed life tests on cylindrical roller thrust bearings in different oils [26] A so-called 'WEC critical oil' with additives resulted in premature bearing failure within 5% of the lifetime of another bearing that was lubricated with a mineral oil containing no additives [26]. These results not only implicate the so-called 'WEC critical oil', but also indicate that oil additives may have an influence on risk of WECs. Similarly, several other investigations used a specific oil, containing additives, to successfully promote WEC failure [1,10,21,23,27], the most recent of which is the investigation by Gould et al., where lubricant additives were systematically varied to study the effect of different additive combinations on bearing time until failure [24]. The investigation found that the lubricant containing zinc dialkyl-dithiophosphate (ZnDDP) led to WECs sooner than any other tested lubricant under the test conditions [24].

While WECs were produced under controlled conditions in several investigations in the past, converging the findings from the different combinations of factors that led to WECs in different experiments remains a challenge. This challenge could be addressed using machine learning (ML) algorithms that are able to discover patterns in high-dimensional data belonging to several experiments. However, ML algorithms are often criticized for a lack of transparency. Transparency into the drivers of accuracy of ML algorithms are crucial if such algorithms are to be used to identify root causes from experimental data.

This paper addresses these issues by first developing machine learning models that are able to learn patterns from experimental data and demonstrate high skill in identifying risky variable combinations from different experiments. The developed models are then further tested following a technique designed to reveal the inner-workings of the models driving the accuracy of their judgements. More specifically, the models were tested to identify which variables are important for the performance of the models and to what extent, relative to one another.

In order to train and assess the models in identifying risky conditions with respect to WECs from previous experiments, a dataset containing roughly 700 high- and low-reference oil compositions was used. The data was provided by Schaeffler on the condition that the identities of the constituting oil compounds remained anonymized. The dataset was compiled based on physical tests and chemical simulations performed by Schaeffler in collaboration with 4LinesFusion, a supplier of industrial analytics solutions [28,29]. Three data analysis methods were designed and applied to the dataset to identify patterns between high-reference oil compositions leading to knowledge of the constituting chemical compounds, which made a given oil composition high-reference with respect to WECs.

The methods presented in this paper can be applied to other experimental data where a high number of measured variables influence a certain outcome and where there is a need to identify variables with the highest influence. Since this is a common objective of many root-cause investigations in tribology, the authors aim to support the efforts of a large audience in the field of tribology with the outcomes of this paper.

2. Data and Methods

2.1. Data Description

Roughly 700 low- and high-reference oil compositions were present in the available dataset. More specifically, 352 oil compositions were present, which were identified by Schaeffler and 4LinesFusion to be low risk with respect to WECs. Additionally, 327 oil compositions identified to be high risk with respect to WECs were present in the dataset. Eight oil compositions were identified as medium risk. These compositions posed a significant class imbalance in the dataset due to their considerably lower number of examples in the available data set compared to the number of examples of high and low risk oil compositions. Such a pronounced class imbalance can negatively impact the

performance and accuracy of the developed ML models later on [30]. Therefore, the 8 oil compositions were neglected in the subsequent analyses. From here on, low- and high-reference oil compositions are referred to as high or low risk oils, respectively.

The oil compositions contained either 1 or 2 additives in addition to the base oil. Additives and base oils, from here on referred to as compounds, were anonymized by compound identification numbers (IDs), e.g., c1, c2, or c21. In total, 21 compound identification numbers were present in the dataset. For clarity, Table 1 shows two oil compositions from the dataset.

Table 1. Examples of oil compositions.

Oil	c6	c11	c16	Risk
195	0%	98%	2%	Low
196	0.5%	99%	0.5%	High

The oil compounds selected for this investigation were used in bearing lubricants in several test benches by project partners to instigate WEC failure. Bearings in wind turbine gearboxes as well as other industrial applications suffer from costly, unplanned maintenance due to WECs [6–11]. In addition, oil additives have been shown to influence risk of WECs [22,24,25]. Therefore, there is high interest in identifying the degree to which the selected oil compounds influence WEC risk.

2.2. Methods Overview

Three methods were used to discover patterns in the available data. First, models based on random forests and artificial neural networks were trained and tested to identify oil compounds that influence the risk level of a given oil composition with respect to WECs. In addition, association rule mining was utilized to investigate the relationship between compound combinations and WEC risk, leading to results supporting those of previous analyses. The methods are explained in more detail in the following subsections.

2.3. Random Forests

In order to discover the pattern in the available data and correctly classify WEC risk level of a given oil composition using the percentages of its constituting compounds as input, a random forest (RF) model was developed. The available data of 679 oils including their respective constituents' percentages and their risk classification (high or low) were used to train and test the RF models.

The random forest [31] model relies on the collective ability of multiple weak classifiers (decision trees) to learn to approximate a function. In this case, the desired function should output the risk level of a given oil composition (high or low) using the percentages of the 21 possible compounds contained in the oil as input variables. Since a random forest is no more than an ensemble of decision trees, Figure 1 illustrates how an example decision tree would classify a given oil based on its constituting compounds. Starting from the root of the tree at the top of the figure, a given oil either follows the left or right path depending on its percentage of c9. It then follows the appropriate path depending on its percentage of c6 or c3 to the so-called leaves of the decision tree, illustrated as pie charts in Figure 1. After a number of oil compositions go from the root of the tree to one of the four leaves depending on their constituting compounds, each leave would have a ratio of high and low risk oil compositions as shown in the figure. This process is referred to as training the decision tree. In this example tree, 90% of the oil compositions that made it to the leftmost leaf are low risk oils. If a new oil composition with unknown risk level reaches the leftmost leaf, then the decision tree estimates with 90% probability that the new oil composition is low risk with respect to WECs. A random forest contains a number of such decision trees with different numbers of branches and different splitting criteria at each branch to collectively reach a more accurate classification.
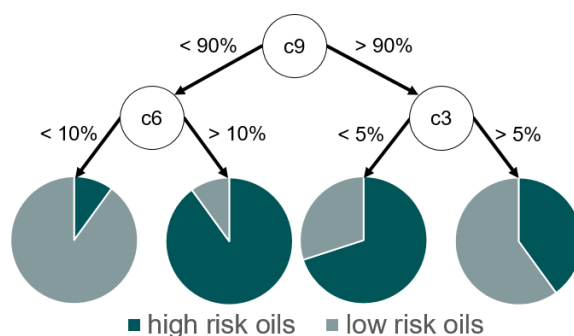
Figure 1. Example decision tree.

To develop a random forest, some design parameters, so-called hyperparameters, need to be decided by the investigator in a process of tuning the RF to reach optimal performance. Some of the most influential hyperparameters on RF performance are [32]:

- Sample size: the size of the sample selected from the total number of oils to be the training data for each tree in the random forest. Decreasing this value will most likely result in less accurate predictions by the individual trees. However, increasing this value can also result in overfitting, where the RF achieves significantly higher performance on the training data, but performs poorly on the test data, i.e., new oil compositions with unknown risk levels.

- Number of tried features at each split (from here on referred to as ftry): the number of randomly selected candidate variables, in this case compound IDs, for each split in a given decision tree in the RF when growing it. A split in a decision tree is every point when a given oil either follows a right or left path. For example, in Figure 1, the first split is performed according to the percentage of c9 in the oil. If two variables are tried with an ftry = 2, then the variable that best splits high and low risk oil compositions is selected. For example, if c1 and c2 are tried and c1 results in a split with the right side of the split containing only high risk oils and the left side containing only low risk oils, and c2 results in a mixture of high and low risk oils on both sides of the split, then c1 is chosen. This is because the split according to the percentage of c1 in the oils, in this example, results in a purer separation of high and low risk oils compared to c2. If ftry is equal to 3, then three compound IDs are instead evaluated at each split. Similar to sample size, decreasing ftry results in worse performance by the individual trees, but increasing it can result in overfitting. Much like the case with sample size, the right balance needs to be found where the highest performance by the RF is reached.

- Node size: the minimum number of oils in a terminal node of any tree in the RF. Without going into more details, the typically used value for classification problems is 1, which was the value chosen for developing the RF in this investigation since it generally provides good results [32]. When attempted, increasing the node size did not lead to higher accuracy.

Probst et al. provide more details on random forest hyperparameters as well as some best practices for tuning RF models [32]. In addition, the pioneering paper by Breiman [31] provides more information about random forests.

The number of trees in the random forest is also a design decision when developing a random forest. The degree of influence of this hyperparameter is controversial with the research consensus favoring setting it to a computationally feasible large number [32,33]. In this investigation, increasing the number of trees above 500 trees did not lead to higher accuracy, so the number of trees was set to 500.

In order to identify the optimal ftry and sample size values, hyperparameter tuning was performed by trying different combinations of the two hyperparameters and assessing the performance of the resulting random forests. Ultimately, the combination resulting in the random forest with the least classification error was selected. In case of ties, the parameters requiring less computational effort was

selected. Since there were only 21 compound IDs present in the data, ftry values could only range from 1 to 21. For the sample size, it was decided to try the range from 1 to 469 with steps of 26, since the training set contained 475 oils. All possible ftry values were tried. The combination resulting in the best performance was ftry = 5 and sample size = 53. This combination led to a 10-fold cross validation accuracy of 98.51%. The R package by Meyer et al. was utilized for tuning the hyperparameters of the random forest models in this paper [34].

It is worth noting that the dataset was initially split into a 70%, 30% split before tuning the RF. The division was performed randomly. The tuning was performed using only the 70% set (containing 476 oils). Ten-fold cross validation (CV) was used to estimate the error of each RF model. The benefit of this method is that it allows for testing the machine learning algorithm with the chosen hyperparameters on every available oil in the dataset. Therefore, this was the method of choice for validating the generalizability of all machine learning algorithms in this investigation. For a more detailed explanation of how 10-fold cross validation was applied in this investigation, the reader is referred to publication [28].

The following steps provide an overview of the analysis performed on the oil compositions using random forests:

1. Splitting the 679 available oil compositions randomly into two smaller datasets: 70% of oils are selected as the training set and 30% are selected as the test set.
2. Hyperparameter tuning: different combinations of sample size and ftry are used to train a random forest model using the training set. Ten-fold cross validation is used to estimate the classification performance of each resulting random forest model. The combination resulting in the top performance is identified as the optimal combination.
3. Developing a tuned RF classifier: the optimal hyperparameter combination is used to develop an RF classifier, trained using the training set.
4. Testing the tuned RF classifier: the developed tuned RF classifier is tested on the test set to verify its accuracy on unseen data.
5. Reaching a more representative estimate of model accuracy: use the optimal hyperparameter pair to perform 10-fold cross validation on all 679 oils. This is done to reach an estimate of accuracy that involves testing every available oil rather than only the 30% of the available oils in the testing set.

After developing a random forest classifier to accurately classify the WEC risk level of oil compositions, the focus shifted to reveal the inner workings of the developed ML model and gain an understanding of what drives the accuracy of its classifications. In other words, the task was identifying which compound IDs had an influence on WEC risk of a given oil composition and to what extent. This was achieved by following the Boruta algorithm [35]; 21 randomly shuffled versions (so called shadows) of the compounds were added to the data, and a statistical test was used to iteratively remove the compounds proven to be less important in WEC risk classification than the random shadows. A compound was considered unimportant if, on average over several iterations, it was found to be less important than the most important shadow compound. Each shadow was a randomly shuffled copy of one of the 21 compound identification numbers present in the dataset. Kursa and Rudnicki also provide more details on the Boruta algorithm and the calculation of the importance values [35].

2.4. Artificial Neural Networks

Artificial neural network (ANN) models were trained to classify the WEC risk (high or low) of an oil, taking the identities of its constituting chemical compounds and their respective percentages as input. The available dataset of oil compositions and their risk classification were used to train and test the ANN models. Similar to the random forest model, developing an ANN model involved selecting and tuning hyperparameter values to improve model accuracy. Eleven neural networks were developed, gradually increasing the 10-fold cross validation classification accuracy on unseen test oils

to 99.8% by tuning the hyperparameters of the networks [28]. Changing the following hyperparameters proved most influential on model performance: the number of hidden layers, the number of nodes per layer, the types of activation functions, the type and parameters of regularization, the type of loss function, and the parameters of the optimizer function.

The network delivering the highest accuracy of 99.8% contained 3 hidden layers with L2 regularization applied only after the first hidden layer to help prevent overfitting. Ng provides more details on L2 regularization [36]. The 3 hidden layers contained 19, 15, and 9 nodes, respectively. The activation function used after every hidden layer was the leaky rectified linear unit (leaky ReLU) [37] as a countermeasure against the vanishing gradient problem. The adaptive moment estimation (Adamax) optimization function [38] was used to optimize the neural network during training with the exponential decay rates for the first and second moment estimates set to 0.93 and 0.98, respectively, and the learning rate set to 0.0018. Categorical cross entropy was used as the loss function. Finally, the output layer consisted of two nodes corresponding to high or low risk with respect to WECs. Softmax [39] was used as the activation function following the output layer in order to facilitate the determination of the target classification, high or low risk, of a given oil composition.

ANN models, through the process of training, approximate a desired function taking in the available input and producing the desired output. In this case, the input was the composition of each oil under investigation with respect to the 21 possible compounds in the dataset; i.e., there were 21 input variables. The output was the WEC risk level of the lubricant, high or low risk. More complex problems require more complex neural networks, and the aforementioned hyperparameters allowed for modularity in constructing the ANN model to meet the required complexity of the problem at hand. The process of tuning the hyperparameters involved iterations of trial and testing guided by previous experience and domain knowledge. Schmidhuber provides a more detailed overview of artificial neural networks and deep learning [40].

The neural networks were tested further to identify the most influential oil compounds in terms of risk of WECs. Twenty-one compound identities were present in the available data, so 10-fold cross validation was performed 21 times on each network architecture. Each time, a network with the previously selected hyperparameters was trained on correct data but tested on data with one of the compound's information shuffled. If the average 10-fold cross validation classification accuracy did not significantly decrease (>97%) as a result of distorting the data of a compound, then it was concluded that this compound was not influential in classifying WEC risk of an oil. As far as we know, this method of feature selection was developed during this investigation, inspired by the fundamental idea of the Boruta analysis. Additional work will be done to test its capabilities with different datasets before the release of a more detailed publication concerning the method. For ease of reference, this method is named Neural Network-based Feature Selection (hereafter referred to as "NN-based FS").

2.5. Association Rule Mining

In addition to identifying individual oil compounds that are influential to WEC risk classification, an analysis was performed to investigate the relationship between frequently occurring combinations of compounds in the available data and WEC risk. The motivation behind this analysis came from previous investigations, such as [5], which concluded that certain additive combinations resulted in WECs, while others did not. The algorithm used to perform this task is called the Apriori algorithm [41].

The algorithm searches for frequently occurring sets, or combinations of compounds in the oil dataset in an unsupervised manner based on user-defined minimum criteria. The minimum criteria ensure a standard for the quality of rules with quality referring to the strength of the identified associations and their frequency of occurrence in the dataset. The algorithm then generates association rules based on the identified frequent sets that shed light on which compounds are likely to join which other compound or which other groups of compounds. For example, the two association rules shown in Table 2 describe the likelihood of finding c12 in an oil that already has c16 (rule number 1) and the likelihood of finding c12 in an oil that already has the compound combination of c8 and c9 (rule

number 2). The four main metrics used to describe the likelihood of a given association rule can also be used, for example, to define the minimum criteria by the user to filter out rarely occurring associations or association rules with low confidence.

Table 2. Association rule examples.

Rule No.	Association Rules
1	{c16} ≥ {c12}
2	{c8,c9} ≥ {c12}

Several metrics are used to describe the likelihood or the quality of a given association rule. The metrics of the rules in Table 2 are listed in Table 3. These metrics are explained below [42]:

1. Support: the proportion of oils in the dataset that contain all the compounds in a given association rule. For example, the support of rule number 1 from Table 3 is calculated by dividing the number of oils containing both c16 and c12 in the dataset (36 oils) by the total number of oils in the dataset (679 oils); 36/679 = 0.0530.

2. Confidence: the proportion of oils that contain the compound(s) on the left hand side (LHS) of an association rule divided by the support of the rule. Using rule number 2 from Table 3 as an example, confidence is calculated by dividing the proportion of oils in the dataset that contain both c8 and c9 (0.0133) by the support of the rule (0.0133), which would equal 1. This essentially means that all oils that contain both c8 and c9 also contain c12.

3. Lift: the confidence of a rule divided by the proportion of oils in the dataset that contain the compound(s) on the right hand side (RHS) of an association rule. This metric indicates how surprising an association rule is given the expected probability of finding the RHS compound(s) in an oil in the dataset. For instance, rule number 3 from Table 3 has a lift of almost 1, which indicates that the probability of finding c16 in any oil in the dataset is almost identical to the probability of finding c16 in an oil that already contains c3. This means that the association suggested by rule number 3 is weak. In contrast, rule number 2 from Table 3 has a lift of 5.853, which indicates that the association indicated by the rule is strong. For rule number 1, the lift value is below 1, which means that it is more likely not to find c12 in an oil that contains c16 than it is to find c12 in an oil that contains c16.

4. Count: the number of oils in the dataset that contain all the compounds in a given association rule. Using rule number 2 from Table 3 as an example, the count is 9, which means that the number of oils in the dataset that contain the combination of c8, c9, and c12 is 9.

Table 3. Association rule examples.

Rule No.	Association Rules	Support	Confidence	Lift	Count
1	{c16} ≥ {c12}	0.0530	0.136	0.798	36
2	{c8,c9} ≥ {c12}	0.0133	1	5.853	9
3	{c3} ≥ {c16}	0.0177	0.4	1.029	12

Hahsler et al. provide more details on the Apriori algorithm used in this investigation and the metrics of association rules [42].

The defined minimum criteria for this investigation were chosen to be a confidence of 50% and a support of 0.1%, since confidence and support are the best-known constraints for this algorithm [42]. After identifying association rules from the oils dataset using these criteria, the focus shifted to the goal of identifying and comparing association rules from low risk oils and high risk oils. This was performed by splitting the dataset into two datasets consisting of low risk oils and high risk oils, respectively, and mining each of these two datasets for association rules using the same minimum

criteria. Finally, the generated rules and their respective metrics were compared to investigate the relationship between compound combinations and WEC risk.

3. Results

3.1. Random Forests

An RF model was used to classify risk level of oil compositions after the dataset was randomly split into training (70%) and testing (30%) sets. The accuracy of the test set was 99.03% with two misclassifications. The chosen combination, which led to the previously mentioned 99.03% accuracy value, was of ftry = 5 and sample size = 53. Using these hyperparameters, 10-fold cross validation was applied to the entire dataset leading to a slightly more pessimistic but more representative estimated accuracy of 98.51%.

The Boruta algorithm [35] was used to identify important compounds for classifying WEC risk levels of oils. Broadly speaking, the importance of a given variable, in this case compound, relates to the potential loss in accuracy if that variable was excluded from the input. Kursa and Rudnicki provide more details on the definition of importance in the context of the Boruta algorithm [35]. That led to the identification of eight significantly important compounds: $c16$, $c9$, $c6$, $c21$, $c14$, $c7$, $c8$, and $c11$. Shown in Figure 2, 13 of 21 compounds were found to be important. However, the last eight compounds on the right hand side of the figure (shown in a circle) were clearly significantly important in comparison with the other compounds. Shown in the figure are also the mean, minimum, and maximum importance values of the shadows, indicated as s.Mean, s.Min, and s.Max, respectively. The unimportant compounds were those with less importance than s.Max. Therefore, they are arranged to the left of the s.Max in Figure 2.
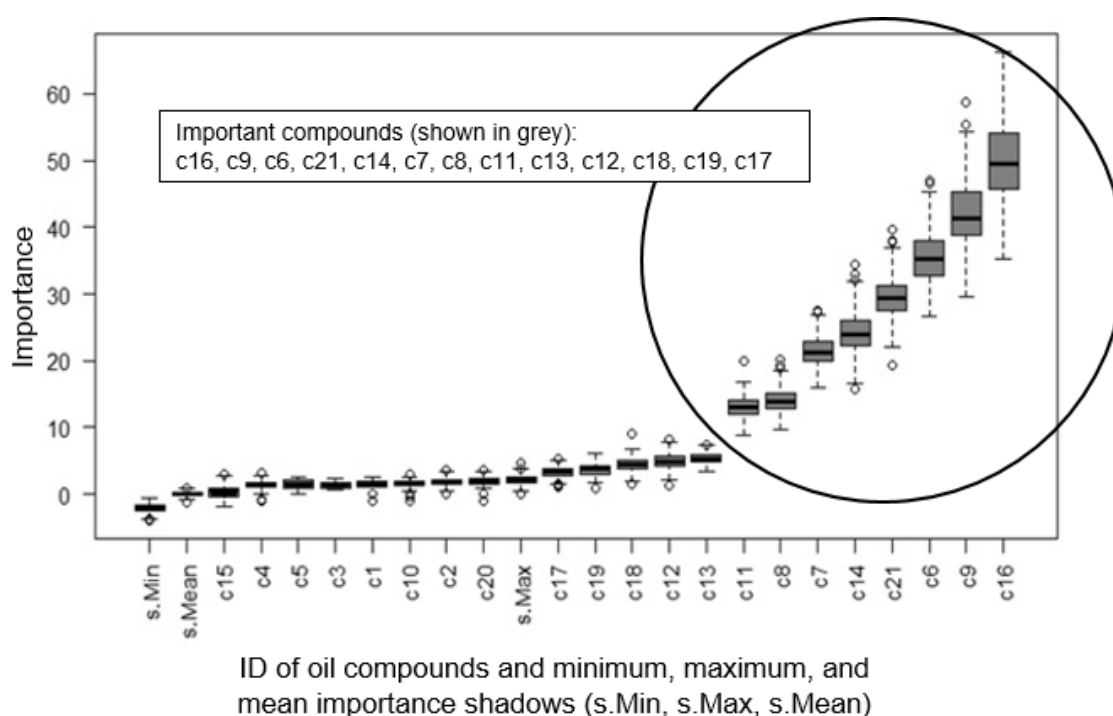
Figure 2. Importance of oil compounds including the mean, minimum, and maximum importance values of the shadows using random forests and the Boruta algorithm [35].
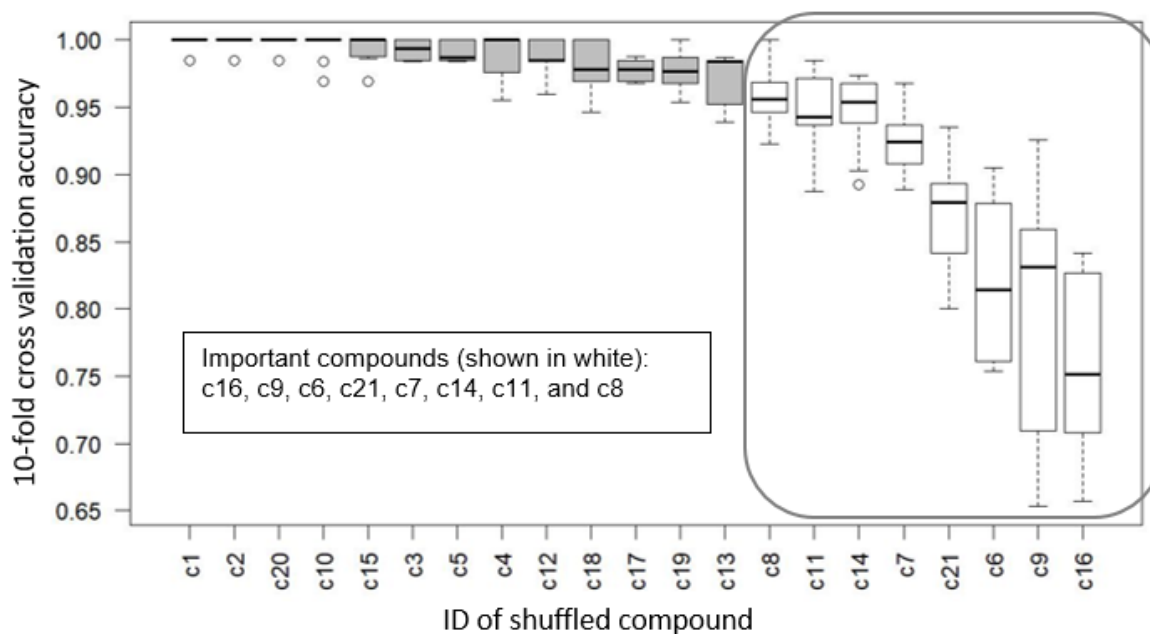
3.2. Artificial Neural Networks

As mentioned earlier, ANN models were trained to classify the WEC risk (high or low) of an oil. Eleven neural networks were developed, gradually increasing the 10-fold cross validation classification

accuracy to 99.8% by altering the networks architecture [28]. In addition to increasing the number of hidden layers and adjusting the number of nodes per layer, using the leaky rectified linear unit (ReLU) activation function and an adaptive moment estimation (Adamax) optimizer proved useful in increasing model accuracy. The top performing ANN consisted of three hidden layers. The R package by Allaire and Chollet was utilized to implement the neural network algorithms [43].

The ANN models were tested further to identify the most influential oil compounds of WEC risk following the NN-based FS method described in the methods section. Shown in Figure 3, eight compounds were found to be influential. Figure 3 shows box plots where the average of each box represents the 10-fold cross validation accuracy for the respective compound ID.

Figure 3. Artificial neural network (ANN) 10-fold cross validation accuracy of white etching crack (WEC) risk classification after shuffling every oil compound [28].

In order to verify the importance of the eight identified important compounds shown in Figure 3, a new ANN classifier was developed. The classifier was trained to classify WEC risk level of oils based only on the data of the eight identified important compounds. In other words, the input to the new classifier did not include the composition data of the remaining 13 compounds available in the dataset. The developed classifier was able to achieve 10-fold CV accuracy of 98.5% [28].

3.3. Association Rule Mining

The Apriori algorithm was used to investigate the relationship between compound combinations and WEC risk. As discussed in the methods section, the minimum criteria implemented to mine association rules were a confidence of 50% and a support of 0.1%. Twenty-two rules were mined using these criteria, and afterwards the available dataset was split into two datasets: a dataset containing only high risk oils and one containing only low risk oils. Rules were mined from each of the two datasets separately using the same minimum criteria resulting in 62 rules from the high risk set and only seven rules from the low risk set. The R package by Hahsler et al. was utilized to implement the Apriori algorithm [44].

Two main findings were obtained from this analysis. The first finding was that high risk oils were more heterogeneous than low risk oils. In other words, high risk oils were much more likely to contain more than two compounds as compared to low risk oils, which almost always contained a maximum

of two compounds, as shown in Figure 4. This was indicated by the significantly lower number of association rules obtained from low risk oils (seven rules) compared to high risk oils (62 rules).
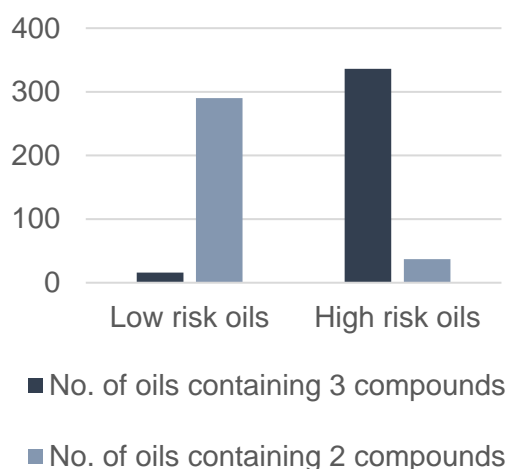

Figure 4. Bar chart of quantities of high and low risk oils with 2 and 3 compounds.

The second finding from the resulting association rules was that occurrence of certain compound groups in low risk oils was different compared to that of high risk oils. This finding is clearly visible in Table 4, which lists association rules mined from the high risk oils dataset that were not in common with the association rules mined from the low risk oils datasets as well as their respective metrics. Despite relaxing the minimum criteria of low risk oils to attempt to extract more, albeit weaker, association rules in common with the ones mined from high risk oils, many association rules from high risk oils were still unique to high risk oils. This further supported the finding that the occurrence of compound combinations was significantly different in high and low risk oils. Table 4 lists a selection of these rules with confidence above 80%.

Table 4. Association rules from high risk oils (minimum confidence of 80%).

Rules	Confidence
$\{c15\} \geq \{c12\}$	1
$\{c17,c21\} \geq \{c16\}$	1
$\{c19,c21\} \geq \{c16\}$	1
$\{c3,c16\} \geq \{c6\}$	1
$\{c3,c6\} \geq \{c16\}$	1
$\{c4,c16\} \geq \{c6\}$	1
$\{c4,c6\} \geq \{c16\}$	1
$\{c5,c16\} \geq \{c6\}$	1
$\{c5,c6\} \geq \{c16\}$	1
$\{c6,c10\} \geq \{c16\}$	1
$\{c8,c9\} \geq \{c12\}$	1
$\{c9,c17\} \geq \{c6\}$	1
$\{c9,c19\} \geq \{c6\}$	1
$\{c9,c18\} \geq \{c6\}$	0.82

4. Discussion

The similar results of different methods help verify the validity of the results reached using the aforementioned data analyses. The important compounds found using ANNs through the newly developed neural network-based feature selection algorithm (NN-based FS) and those found using random forests through the Boruta algorithm are listed in Table 5 in the order of importance. Comparing the identified important compounds from the two methods, it becomes clear that the two results

are in agreement with a slight difference in the order of importance towards the relatively less important compounds. This helps validate the two results since two different methods led to an almost identical conclusion.

Table 5. Identified important compounds.

Method	Important Compounds
Neural Network-based Feature Selection (NN-based FS)	c16, c9, c6, c21, c7, c14, c11, c8
Boruta [35]	c16, c9, c6, c21, c14, c7, c8, c11

A significant observation was made after reaching the order of important compounds listed in Table 5 using the NN-based FS method. Based on chemical domain knowledge, if those compounds were to be ordered based on their respective ability to release hydrogen, that order would match the order of importance identified using the NN-based FS method, as listed in Table 5. This indicates that the results of this investigation are in agreement with previous investigations [2,45], which found the release of hydrogen and its diffusion into the bearing steel to be a driver of WEC formation.

As for the investigation of the relationship between combinations of compounds and WEC risk, two important observations are visible from the resulting association rules. Firstly, the compound associations, listed in Table 4, which were found only in high risk oils, had one thing in common: they almost always, with one exception, contained one or more of the top three important compounds identified by the other analyses to be influential to WEC risk classification. This result further supports the results of the other analyses that pointed at these compounds as risky. In addition to the first observation, the fact that low risk oils generally contain less compounds than high risk oils, as shown in Figure 4, indicates a possibility that oils with more compounds may be more likely to result in WEC failure compared to oils with fewer compounds. It may also be the case that having more compounds in an oil increases the likelihood that a high risk compound is present in the oil. Future investigations might also use this observation as a starting point to examine these possibilities.

A possibility still remains that certain combinations of compounds that are not risky on their own may become risky when combined. Since the compounds in the association rules in Table 4 are not even weakly associated in low risk oils yet strongly associated in high risk oils, they may be, pending further investigations, risky combinations with respect to WEC failure.

This investigation shows the applicability of data analytics approaches on phenomena where several factors seem suspicious for having an influence on a certain outcome. With the help of these methods, it is possible to identify the influential factors out of a number of suspicious factors. An investigation [24] involving a number of tests with different oils led to a conclusion consistent with the results of the data analyses presented in this paper. The completed data analyses on the available dataset pointed to c16 as the most important oil compound for classifying WEC risk of an oil. The project partner Schaeffler also reported that several tests pointed to c16 as a high risk oil compound with respect to WECs. This agreement between the results of the analysis performed and the results from the project partner corroborates the findings and applications presented in this paper.

5. Conclusions

This paper presented applications of three machine learning techniques to tackle the challenge of pattern discovery in high-dimensional data belonging to multiple experiments on WEC bearing failure. This includes the first application of the purpose-built Neural Network-based Feature Selection (NN-based FS) method. The main conclusions are as follows:

1. It is possible to converge findings from multiple experiments using the presented ML models to discover patterns and conduct root-cause analyses on WECs using only historic data from previous experiments.

2. It is possible to reach said patterns via ML models while maintaining transparency into the drivers of accuracy of the ML models using the techniques presented in this paper.
3. The presented techniques are able to identify patterns to classify a given oil composition as high- or low-risk with respect to WECs with high accuracy using data from previous experiments.
4. The presented techniques are able to identify influential oil compounds on WEC risk using data from previous experiments.
5. NN-based FS was developed and applied during this investigation as a method of feature selection based on neural networks. Since this is the first application of the method, the authors aim to test its capabilities with different datasets before releasing a more detailed publication of the method.

References

1. Gould, B.; Greco, A. Investigating the Process of White Etching Crack Initiation in Bearing Steel. *Tribol. Lett.* **2016**, *62*, 26. [CrossRef]
2. Evans, M.-H. White structure flaking (WSF) in wind turbine gearbox bearings: Effects of 'butterflies' and white etching cracks (WECs). *Mater. Sci. Technol. Ser.* **2013**, *28*, 3–22. [CrossRef]
3. Spinato, F.; Tavner, P.; van Bussel, G.; Koutoulakos, E. Reliability of wind turbine subassemblies. *IET Renew. Power Gen.* **2009**, *3*, 387–401. [CrossRef]
4. Sheng, S. Wind Turbine Gearbox Reliability Database, Condition Monitoring, and Operation and Maintenance Research Update. In Proceedings of the National Renewable Energy Laboratory (NREL), Drivetrain Reliability Collaborative Workshop, Golden, CO, USA, 16–17 February 2016.
5. Haque, T.; Korres, S.; Carey, J.T.; Jacobs, P.W.; Loos, J.; Franke, J. Lubricant Effects on White Etching Cracking Failures in Thrust Bearing Rig Tests. *Tribol. Trans.* **2018**, *61*, 979–990. [CrossRef]
6. Grabulov, A.; Petrov, R.; Zandbergen, H. EBSD investigation of the crack initiation and TEM/FIB analyses of the microstructural changes around the cracks formed under Rolling Contact Fatigue (RCF). *Int. J. Fatigue* **2010**, *32*, 576–583. [CrossRef]
7. Harada, H.; Mikami, T.; Shibata, M.; Sokai, D.; Yamamoto, A.; Tsubakino, H. Microstructural Changes and Crack Initiation with White Etching Area Formation under Rolling/Sliding Contact in Bearing Steel. *ISIJ Int.* **2005**, *45*, 1897–1902. [CrossRef]
8. Greco, A.; Sheng, S.; Keller, J.; Erdemir, A. Material wear and fatigue in wind turbine Systems. *Wear* **2013**, *302*, 1583–1591. [CrossRef]
9. Evans, M.-H. An updated review: White etching cracks (WECs) and axial cracks in wind turbine gearbox bearings. *Mater. Sci. Technol. Ser.* **2016**, *32*, 1133–1169. [CrossRef]
10. Franke, J.; Carey, J.T.; Korres, S.; Haque, T.; Jacobs, P.W.; Loos, J.; Kruhoeffer, W. White Etching Cracking—Simulation in Bearing Rig and Bench Tests. *Tribol. Trans.* **2018**, *61*, 403–413. [CrossRef]
11. Iso, K.; Yokouchi, A.; Takemura, H. Research Work for Clarifying the Mechanism of White Structure Flaking and Extending the Life of Bearings, SAE Technical Paper 2005-01-1868. In Proceedings of the SAE 2005 World Congress & Exhibition, Detroit, MI, USA, 11–14 April 2005.
12. Loos, J.; Bergmann, I.; Goss, M. Influence of Currents from Electrostatic Charges on WEC Formation in Rolling Bearings. *Tribol. Trans.* **2016**, *59*, 865–875. [CrossRef]
13. Garabedian, N.T.; Gould, B.J.; Doll, G.L.; Burris, D.L. The Cause of Premature Wind Turbine Bearing Failures: Overloading or Underloading? *Tribol. Trans.* **2018**, *61*, 850–860. [CrossRef]

14. Errichello, R.; Budny, R.; Eckert, R. Investigations of Bearing Failures Associated with White Etching Areas (WEAs) in Wind Turbine Gearboxes. *Tribol. Trans.* **2013**, *56*, 1069–1076. [CrossRef]

15. Evans, M.-H.; Richardson, A.; Wang, L.; Wood, R. Serial sectioning investigation of butterfly and white etching crack (WEC) formation in wind turbine gearbox bearings. *Wear* **2013**, *302*, 1573–1582. [CrossRef]

16. Arnaud, R.D.C. Tribological Analysis of White Etching Crack (WEC) Failures in Rolling Element Bearings. Ph.D. Thesis, INSA de Lyon, Villeurbanne, France, 2014.

17. Evans, M.-H.; Walker, J.; Ma, C.; Wang, L.; Wood, R. A FIB/TEM study of butterfly crack formation and white etching area (WEA) microstructural changes under rolling contact fatigue in 100Cr6 bearing steel. *Mater. Sci. Eng. A* **2013**, *570*, 127–134. [CrossRef]

18. Gegner, J.; Nierlich, W. The Bearing Axial Cracks Root Cause Hypothesis of Frictional Surface Crack Initiation and Corrosion Fatigue Driven Crack Growth. In Proceedings of the NREL 2011 Wind Turbine Tribology Seminar, Broomfield, CO, USA, 15–17 November 2011.

19. Grabulov, A.; Ziese, U.; Zandbergen, H. TEM/SEM investigation of microstructural changes within the white etching area under rolling contact fatigue and 3-D crack reconstruction by focused ion beam. *Scripta Mater.* **2007**, *57*, 635–638. [CrossRef]

20. Uyama, H.; Yamada, H.; Hidaka, H.; Mitamura, N. The Effects of Hydrogen on Microstructural Change and Surface Originated Flaking in Rolling Contact Fatigue. *Tribol. Online* **2011**, *6*, 123–132. [CrossRef]

21. Richardson, A.D.; Evans, M.-H.; Wang, L.; Wood, R.J.K.; Ingram, M.; Meuth, B. The Evolution of White Etching Cracks (WECs) in Rolling Contact Fatigue-Tested 100Cr6 Steel. *Tribol. Lett.* **2017**, *66*, 6. [CrossRef]

22. Holweger, W.; Wolf, M.; Merk, D.; Blass, T.; Goss, M.; Loos, J.; Barteldes, S.; Jakovics, A. White Etching Crack Root Cause Investigations. *Tribol. Trans.* **2015**, *58*, 59–69. [CrossRef]

23. Gould, B.; Greco, A. The Influence of Sliding and Contact Severity on the Generation of White Etching Cracks. *Tribol. Lett.* **2015**, *60*, 29. [CrossRef]

24. Gould, B.; Demas, N.; Pollard, G.; Rydel, J.J.; Ingram, M.; Greco, A. The Effect of Lubricant Composition on White Etching Crack Failures. *Tribol. Lett.* **2019**, *67*, 7. [CrossRef]

25. Wang, L.; Holweger, W. White Etching Cracks (WECs) in Wind Turbine Bearings. In Proceedings of the World Tribology Congress 2017, Beijing, China, 17–22 September 2017.

26. Paladugu, M.; Lucas, D.R.; Hyde, R.S. Effect of lubricants on bearing damage in rolling-sliding conditions: Evolution of white etching cracks. *Wear* **2018**, *398–399*, 165–177. [CrossRef]

27. Kruhöffer, W.; Loos, J. WEC Formation in Rolling Bearings under Mixed Friction: Influences and "Friction Energy Accumulation as Indicator. *Tribol. Trans.* **2017**, *60*, 516–529. [CrossRef]

28. Azzam, B.; Harzendorf, F.; Schelenz, R.; Holweger, W.; Jacobs, G. White Etching Crack Bearing Failure: Identification of High Risk Lubricant Compounds using Artificial Neural Networks. In Proceedings of the International Conference on Gears 2019, Garching bei München, Germany, 18–19 September 2019.

29. Brandt, J.M.; Benedek, M.; Guerin, J.S.; Fliege, J. Industrial Analytics-based Bearing Risk Classification Systems investigated with Advanced Mathematical Models. *IoT - Eng. Cyber Phys. Hum. Syst.*. (under review).

30. Japkowicz, N. The class imbalance problem: Significance and strategies. In Proceedings of the Int'l Conf. on Artificial Intelligence, Las Vegas, NV, USA, 15 May 2000.

31. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

32. Probst, P.; Wright, M.; Boulesteix, A.-L. Hyperparameters and Tuning Strategies for Random Forest. *arXiv* **2018**, arXiv:1804.03515. [CrossRef]

33. Probst, P.; Boulesteix, A.-L. To tune or not to tune the number of trees in random forest? *arXiv* **2017**, arXiv:1705.05654.

34. Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wienl; R Package Version 1.7-0. Available online: https://CRAN.R-project.org/package=e1071 (accessed on 22 November 2018).

35. Kursa, M.B.; Rudnicki, W.R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *1*, 1–13.

36. Ng, A. Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance. In Proceedings of the 21st International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004.

37. Glorot, X.; Bengio, A.B.U.Y. Deep Sparse Rectifier Neural Networks. In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011.

38. Kingma, D.; Ba, J.L. ADAM: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.

39. Bridle, J.S. Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. In *Neurocomputing*; Springer: Berlin/Heidelberg, Germany, 1990; pp. 227–236.

40. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]

41. Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; Verkamo, A. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*; AAAI Press: Menlo Park, CA, USA, 1996; pp. 307–328.

42. Hahsler, M.; Grün, B.; Hornik, K. arules—A Computational Environment for Mining Association Rules and Frequent Item Sets. *J. Stat. Softw.* **2005**, *1*, 1–25. [CrossRef]

43. Allaire, J.; Chollet, F. keras: R Interface to 'Keras'; R Package Version 2.2.0. 2018. Available online: https://CRAN.R-project.org/package=keras (accessed on 22 November 2018).

44. Hahsler, M.; Buchta, C.; Grün, B.; Hornik, K. arules: Mining Association Rules and Frequent Itemsets; R Package Version 1.6-2. 2018. Available online: https://CRAN.R-project.org/package=arules (accessed on 9 January 2019).

45. Tamada, K.; Tanaka, H. Occurence of brittle flaking on bearings used for automotive electrical instruments and auxiliary devices. *Wear* **1996**, *199*, 245–252. [CrossRef]