UNIVERSITY OF TWENTE.

Automated Failure Diagnosis in Aviation Maintenance Using eXplainable Artificial Intelligence (XAI)

Author:
S.G. ten Zeldam

Supervisors:
Prof. dr. ir. T. Tinga, University of Twente
Dr. ir. R. Loendersloot, University of Twente
Dr. ir. R.G.K.M. Aarts, University of Twente
Dr. A. de Jong, NLR - Netherlands Aerospace Centre

July 9, 2018

University of Twente
Faculty of Engineering Technology
Mechanical Engineering
Maintenance Engineering & Operations

Master thesis
Document number ET.18/TM 5822

Automated Failure Diagnosis in Aviation Maintenance Using eXplainable Artificial Intelligence (XAI)

Sophie ten Zeldam[1,2,3], Arjan de Jong[1], Richard Loendersloot[2] and Tiedo Tinga[2,3]

[1] *Netherlands Aerospace Centre (NLR), Amsterdam, the Netherlands*
Arjan.de.Jong@nlr.nl

[2] *University of Twente, Dynamics based Maintenance group, Enschede, the Netherlands*
r.loendersloot@utwente.nl
t.tinga@utwente.nl

[3] *Netherlands Defence Academy, Military Technical Sciences, Den Helder, the Netherlands*
sg.t.zeldam@mindef.nl

ABSTRACT

A repair card is used in aviation maintenance to report a failure or anomaly and register it in the maintenance management system. An incorrect or incomplete repair card may result in incorrect maintenance and make it very hard to analyse the maintenance data. There are several reasons for this incomplete reporting. Firstly, (part of) the information is often unknown at the moment the maintenance crew fills in the card. Also, the findings on repair cards are generally filled in as free-form text, making it difficult to automatically interpret the findings. An automatically assessed failure description will lead to more complete and consistent repair cards. This will also improve the efficiency of troubleshooting since this failure diagnosis can add information which would otherwise not be at the disposal of the maintenance crew at that time. With this research, a model is developed which is able to automatically diagnose a failure. The model utilises a data driven approach, combining maintenance data and usage data. The model is based on Artificial Intelligence (AI) such that it is no longer necessary to completely understand the physics of a (sub)system or component. A newly proposed XAI (eXplainable AI) methodology, Failure Diagnosis Explainability (FDE), is added to the model to provide transparency and interpretability of the assessed diagnosis. The assessed diagnosis is explained by checking whether a new failure matches the expected values of a certain diagnosis (class). A failure is from class (diagnosis) A, because the features have similar values as class A. Contrary, this failure is not from classes (diagnosis) B and C, because the features have dissimilar values as class B and C. Two techniques are used to check whether a failure matches the expectations: visual representation and the proposed χ-factor. The proposed model and XAI methodology FDE are applied to a case study with a main wheel of the RNLAF (Royal Netherlands Air Force) F-16. This feasibility study already showed the value of this automated failure diagnosis model with an achieved accuracy of 81% of classifying a diagnosis . The proposed XAI methodology FDE was able to explain the diagnosis assessed by the failure diagnosis model both with visual representation and the χ-factor. Thereby the feasibility of this model is proved. This model will also support a repair shop to repair NFF (No Failure Found) components based on their historical usage.

1. INTRODUCTION

A repair card is used in aviation maintenance to report a failure or anomaly and register it in the maintenance management system. An incorrect or incomplete repair card may result in incorrect maintenance and make it very hard to analyse the maintenance data. An example from practise is a helicopter Main Gear Box (MGB) removal due to a leakage found during a 500 hours inspection. The maintenance crew described the complaint as ‘defect, 500 hrs’. The component shop carried out an overhaul when a small repair could also have solved the problem. The overhaul, however, was unforeseen and there was no spare MGB available which resulted in a grounded helicopter. So the consequences of this incomplete repair card were additional maintenance costs and a decrease in availability.

This example is not unique in the aviation sector, it is rather common and there are several reasons behind this. Firstly, (part of) the information is often unknown at the moment the maintenance crew fills in the repair card. Also, the findings

on repair cards are generally filled in as free-form text. As a result, repair cards may contain incorrect information and can be incomplete. A consequence of the incorrect and incomplete repair cards is that they are not practical for data analysis because it is difficult to automatically interpret the findings.

A possible solution for this problem is the use of AI (Artificial Intelligence) combining maintenance data and usage data. To receive the user's trust the algorithms used in AI should be explained with (eXplainable AI). A literature study about AI, XAI and their applications is carried out. AI can be defined by 'the study of how to make computers do things at which, at the moment, people are better'(Rich, 1983). Many data can be fed to AI models and after training they can recognise patterns much better than humans (Koch, 2018). AI is no longer science fiction but it is already applied in everyday technologies. Examples are Netflix suggesting your next movie, Facebook personalising your timeline and recognising friends in pictures, filtering for your email on spam and phone assistants such as Siri and Google Now. The application of AI within the maintenance sector is a potential step forward in the development of the sector. Ultimately, the goal is to predict failures and to automatically plan maintenance activities with AI. But first, the possibilities of diagnosing and explaining the cause of a failure will be researched. In other words, why is a certain diagnosis assessed? This is important because, as an example, AI can be used for diagnosing patients in the healthcare. An example is cancer diagnosis. Some algorithms are able to predict whether the patient has cancer or not, more accurately than doctors. But as long as humans do not understand how this algorithm made the assessment, it will not be used in practise since they are not trusted (Holzinger et al., 2017). Transparent and interpretable explanations are required for trust and acceptance of both doctors and patients. Therefore, the need for XAI is growing. Since a comparable trust problem will be present in the maintenance sector when applying an AI model, XAI will be an option to solve this problem.

Firstly, the current status of AI within the maintenance sector will be discussed. For many industrial and military machines, there is a large amount of historical data. Sometimes data are stored after a system or component failure, but there is also a lot of data for which there was no purpose of collecting it at the time. This data was stored because it might prove useful in the future. The challenge now is to transform this big amount of (historical) data to usable information. Many methods exist for analysing this data and turning it into useful forms for decision making, including statistical correlation/regression methods, fuzzy logic classification and neural-network clustering techniques (Vachtsevanos et al., 2006). Failure search routines can be developed with AI for troubleshooting and diagnosis (Russell & Norvig, 2010). These developments have lead to 'Knowledge Acquisition', the abstraction of knowledge from experts and structuring it for applications in investment projects. This method is also known as Knowledge Based Systems Analysis and Designs Support (KADS) (Smit, 2010).

In part of these publications of AI in maintenance (and also specifically in failure diagnosis), the diagnosis relies on just one or a few features (Yan et al., 2014; Al-Garni et al., 2006) while others requires a massive amount of data and/or knowledge to train the model (Milne et al., 2001; Tarifa et al., 2002; Khoo et al., 2000). For the failure diagnosis model many data is available but not all data are useful due to missing information: when a failure can not be linked to the corresponding usage data of the the failure description is insufficient. Also, this model aims to eliminate the requirement of systems knowledge. Concluding, AI can be a potential solution to the problem were XAI have to be implemented for the trust of the user.

Therefore a model will be generated which can automatically assess a failure diagnosis based on maintenance data and usage data. An automatically assessed failure description will lead to more complete and consistent repair cards. This will improve the efficiency of troubleshooting since this failure diagnosis can add information which would otherwise not be at the disposal of the maintenance crew at that time. Conventional ways to link failures to the usage are physical models, but this research utilises a data driven approach combining maintenance and usage data, and AI into a failure diagnosing model. AI is capable of recognising more patterns and relations than humans can. With this model, it is no longer necessary to completely understand the physics (and failure mechanisms) of a (sub)system or component. This data driven approach makes it difficult to establish causal relations between features. To convince the users of the model, a plausible explanation is needed to understand the cause of the failure. XAI techniques will be implemented in the model to provide transparency and interpretability of the resulting diagnosis.

Figure 1 proposes the steps to achieve an automatic failure diagnosis based on usage data (sensor information from the flights) with XAI. This methodology is newly proposed in this research and will be discussed in the remainder of this thesis. Section 3 describes which data have to be collected and in which form, to develop an automated failure diagnosing model (blocks 1,2). In Section 4, the training of the model is described (blocks 3,4,5). A historical reconstruction will be made and features and algorithms will be selected. The failure diagnosis will be discussed in Section 5 (blocks 6,7,8). The diagnosis will consist of an assessment of the possible causes, the explainability of this assessment and the correctness of the diagnosis. Finally, all these steps will be demonstrated in a case study in Section 6 were the feasibility of the model will be tested. But first, in Section 2, the

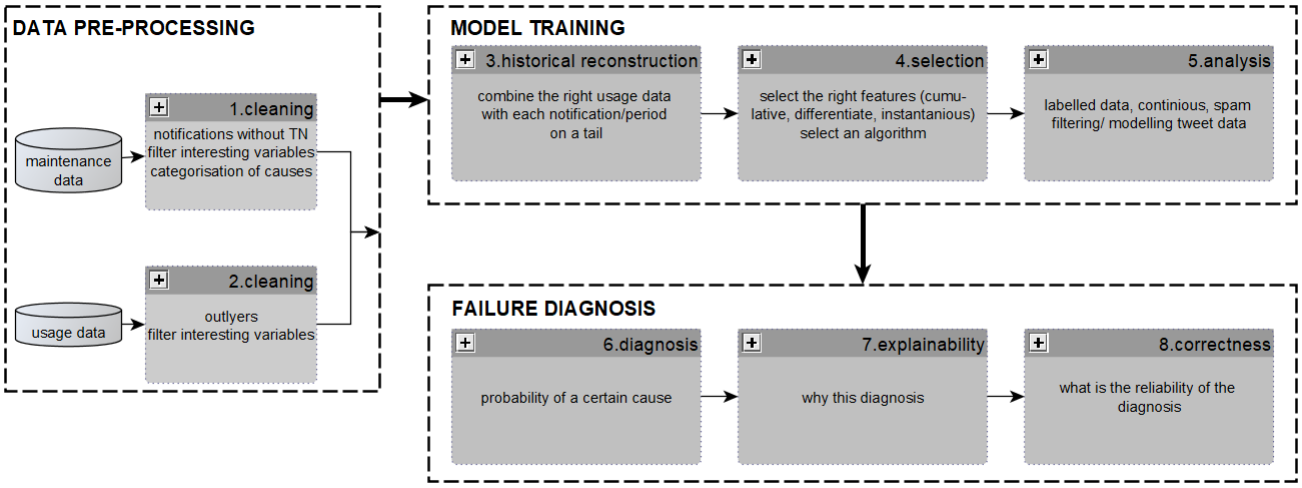

Figure 1. Functional diagram of an automated failure diagnosis model

motivation for this research and the research questions will be discussed.

2. RESEARCH MOTIVATION

Companies, institutes and governments are starting to realise the importance of maintenance more and more. This is especially interesting for companies owning or working with capital assets such as buildings, infrastructures, ships, trains, airplanes and plants. Maintenance technologies are applied for various reasons such as for lifetime extension and for safety reasons.

This research will focus on the increase of availability and reliability of the system. Maintenance has gone through a development over the years which is qualitatively represented in Figure 2. The availability of systems has been increased due to new developed technologies, improved accuracies of methods and a growth of awareness and acceptance of the importance of maintenance.

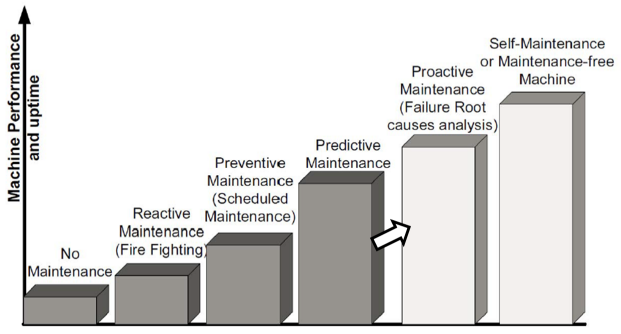

Figure 2. The development of maintenance technologies (Lee & Wang, 2008)

The fourth step, predictive maintenance, is striving for just-in-time maintenance. At the moment, this is usually achieved by either monitoring the status of a system and predicting its status in the future, or by using historic sensor data to predict the system's status in the future based on trends or a physical model. The use of sensors which monitor the parameters to determine the system's status require knowledge about the failure mechanisms of the system and the parameters which indicate the system's status.

Another approach than physics based, is a data driven approach where decisions are made based on data analyses. Since a lot of data are already available but remained unused so far, this is an opportunity. This research will integrate data from multiple sources (maintenance and usage data), also known as data fusion. Data fusion achieves more accuracy, consistency and useful information than that provided by the individual sources. In other words, the whole is greater than the sum of its parts. The usage data used for this research are not measuring the system's status but only the usage of the airplane. Examples are speed and air temperature, so no status indicators such as remaining profile of a tyre or a pump's flow rate.

It will be researched if a diagnosis can be conducted by AI after a failure occurs. Usage data obtained from sensors and maintenance data from repair cards will be used as an input for the AI-model. This attempts to solve the current problems of the incomplete and incorrect repair cards resulting in incorrect maintenance activities and the limitations of data analysis.

2.1. Research questions

The objective of this research is to automatically diagnose failures with a data driven approach. This in order to improve the efficiency of complaint handling and to improve the quality of data (for further analysis). The main question of

this research is:

Which data driven approach gives the best results for automated failure diagnosis using XAI?

The sub-questions that will lead to the answer of the main question are:

- *Which data are required and in which form should it be presented to develop an automated failure diagnosis model based on usage?*
- *How should the model be trained to obtain a failure diagnosis model?*
- *Which methodologies are applicable for eXplainable Artificial Intelligence (XAI)?*
- *How should the data be explained to substantiate the provided failure diagnosis?*

2.2. Research scope and approach

A literature study (Section 1) demonstrated that AI gets bigger and becomes more important in today's technologies. Also in the maintenance industry, the application of AI is growing but not a common practise yet. One of the applications of AI is diagnosing in the medical sector. A lack in these diagnoses is their explainability. As long as doctors cannot explain to their patients why a model assesses a certain diagnosis, the model will not be trusted. A similar issue can be observed in the field of aviation maintenance, where this research focusses on. Explainability will be added so that the user's trust on the diagnosis will increase by gaining more knowledge about the diagnosis assessment.

In this research, different existing methods and technologies which are possibly useful for the application of failure diagnosis will be discussed. Not all the details are discussed about the existing methods and techniques but the focus is mainly on combining and applying them to failure diagnosis. For the explainability, a new methodology is proposed which explains why a new failure is assessed to a certain diagnosis. This methodology is inspired by approaches proposed by DARPA (Defense Advanced Research Projects Agency), the agency of the United States Department of Defense (Gunning & et al., 2016). Finally, in a case study, a specific model is built to demonstrate that failures can be diagnosed with AI and explained with the proposed XAI methodology. Below, the scopes per section will be discussed:

- **Data pre-processing** theoretical guidelines will be given which can be applied to various applications. One example will be elaborated in the case study. All decisions (e.g. which variables will be selected) are made manually and technologies to automate this process are out of the scope of this research.
- **Machine learning model** considerations are given about the selection of features and which algorithms are suitable for this type of problem. Detailed information about the algorithms and optimisation of the features is out of the scope.
- **Failure diagnosis** state of the art approaches of XAI suitable for failure diagnosis are discussed but are not elaborated in detail.

3. DATA PRE-PROCESSING

Before the model can be trained, the data need to be pre-processed so that it can be implemented in a ML (Machine Learning) algorithm. This raises the question which data, in which form and how much is required for proper training. The model will be based on historic maintenance and usage data which is sensor data from historic flights. In this research, the maintenance data will be used to label the failures. The usage data will be added to incorporate the historic usage of a component in the assessment of the diagnosis.

This section will discuss the algorithm prerequisites, the required variables and types of data. The requirements will differ per system and per component. This section provides some guidelines. Later, in the case study, a specific example will be elaborated on.

3.1. Algorithm prerequisites

There are no specific requirements for the dataset that is applied to ML algorithms. Actually Hua et al. (2005) mention that one should be wary of rules-of-thumb generalised from specific cases. The optimum sample size will differ per situation. But there are some general guidelines for the data:

- The optimal amount of features relative to the sample size depends on the algorithm and the feature-output distribution. Hua et al. (2005) show the relation between the number of samples, number of features and the error rate for various algorithms.
- The performance of an algorithm can be greatly influenced by the number of features. Therefore the amount of features used should be close to the optimal amount (Hua et al., 2005).
- A balanced labelled dataset (each class has about the same amount of samples) is preferable to avoid overfitting on the class which contains more examples (Jain & Chandrasekaran, 1982).
- All ML algorithms can only handle single values per case. So when each row is a case, each case must have the same amount of columns.

Since the guidelines are not directional, the data have to be processed by reasoning in which format the data can be implemented in the ML algorithms. Engineering judgement is needed for the amount of features and data to be used.

3.2. Variable selection procedure

This research uses two data sources, namely usage data and maintenance data. The usage data comes from sensors in the airplane and consist of continuous, numeric values. This dataset contains information such as altitude, velocity, acceleration and outside temperature. The maintenance dataset consists mainly of text and categorised variables. This dataset is filled by the maintenance crew and contains information about the failure of a component. Besides the selection of variables from both datasets, the data probably will need cleaning. There are two types of errors in the datasets, namely measurement faults (e.g. an altitude of -100 m) and incompleteness (e.g. when a maintenance notification does not have a tail number (airplane identification number) such that it cannot be linked to the usage data of an airplane).

Maintenance data entails information about the date of installation/removal of a part on a tail, preventive and corrective maintenance and the reason of a notification. The data have to be labelled before training. In order to limit the amount of labels, the failures with similar failure descriptions (and the same failure cause) will be grouped. E.g. 'removal due to being worn out' and 'component x is worn' should be in the same group. This grouping is also done because ML algorithms always require multiple events per group in order to train the model.

Variables are used to describe the usage. More variables can result in a more accurate description, but not all variables contain relevant information on the usage of the system. The usage variables will be different for each system, in each situation. In the case of a new system that has not been built yet, there is more freedom of selecting variables than in the case of an existing system. In general, to select the usage variables, it should be checked whether there is a relation between that variable and the loads on the system or the performance of the system (Tinga, 2010).

Also the features, i.e. the specific parts or details retrieved from a variable or signal, have to be determined. Is data required from the entire flight or only from take-off and landing or only during the flight? From all flights or only from the last one or last 10? This need will depend on the type of failure mechanism i.e. for corrosion, fatigue and wear all flights are required but an overload can probably be seen in (one of) the last flights. For the landing gear failures only the take-off and landings are interesting but for structural failures the entire flight (from take-off to landing) is needed.

3.3. Training and input data

A distinction is made between training data and input data. When this model is trained, data can be entered in the model and the model provides a diagnosis. Subsequently, the maintenance crew validates the diagnosis. The input data are now labelled and can be used to enrich (i.e. train) the model.

Since the usage data will be combined with the maintenance data (repairs, removals etc.), usage data of a certain period can be both training and input data. The usage data of the last year for a specific tail will be used as input data, e.g. when a failed component has been on a tail for one year. If another component, on the same tail failed four months earlier, then the usage data for the months before this failure can be used for training the model for that specific component.

4. MACHINE LEARNING MODEL

After pre-processing the data, the data from the different sources have to be combined and the model has to be trained for failure diagnosis. This section will discuss the data requirements for the ML algorithms and the procedure to select the variables from both maintenance data and usage data. Finally, the destinction between training data and input data will be discussed.

4.1. Historical reconstruction

As can be seen in the functional diagram in Figure 1, the maintenance and usage data will be combined to come to a historical reconstruction where both sources are aligned to one timeline as is shown in Figure 3.

Most air forces and airlines exchange components between airplanes (tails), which is not common in every industry (e.g. process industry). Making a historical timeline of a component may show that a component has been installed on different tails. The usage data, in this case, need to be collected from different tails. Visualisation of the historical timeline is very important. This will give the maintenance crew a quick overview of the history of a component which helps to judge the final diagnosis assessed by the model. Figure 3 shows the air temperature and wing root bending during landing to which the airplane was exposed during the lifetime of a specific component. As can be seen, there was a high wing root bending during one of the landings in the beginning of the wheel's life which probably indicates a hard landing. When this component shows a relatively quick degradation this might can be explained by this hard landing, so it helps the maintenance crew to judge the model's diagnosis assessment.

4.2. Feature selection

Since all ML algorithms can only handle single values per case, the usage of the component has to be expressed in features. Features are individual measurable properties or characteristics of a phenomenon being observed (Bishop, 2006). The features will be selected depending on the failure mechanism and the ML algorithm. E.g. when the velocity during take-off is one of the selected variables it is likely

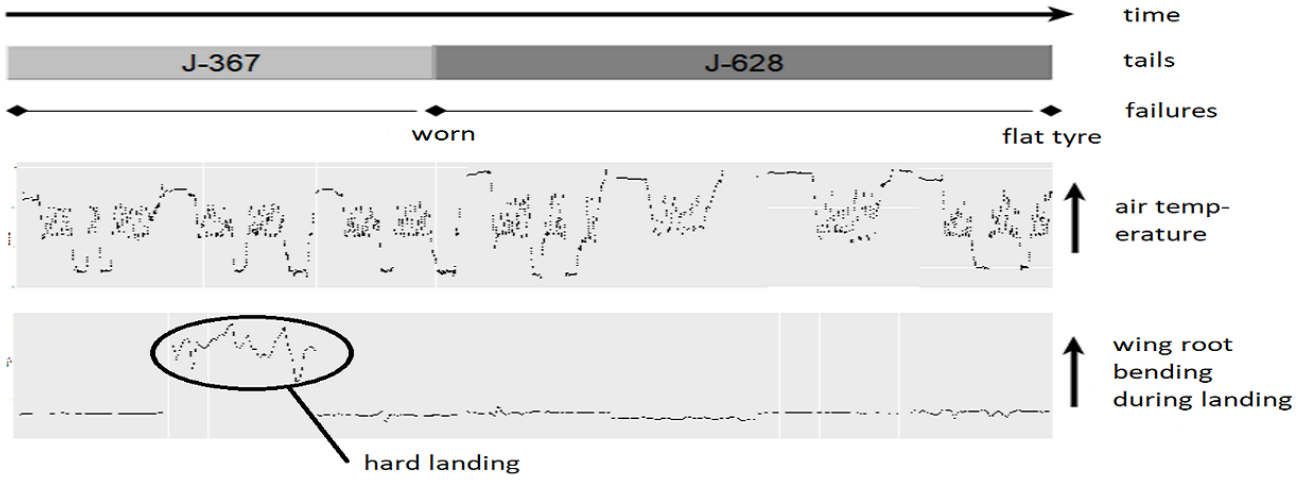

Figure 3. Reconstruction of a historical timeline for one component, installed on two different tails during a certain period of its operational life

that the amount of data points is not equal for every landing and the data can therefore not be implemented directly in a ML algorithm. To get data strings of the same size, several options are possible: choose the lowest amount of data points and delete the superfluous points for the other take-offs; select the largest amount of data points and extrapolate or interpolate between data points for the other take-offs; or determine characteristics such as average, standard deviation and kurtosis for each take-off. Neural networks prefer to work with the whole data string (extrapolation or interpolation) and more simple algorithms require characteristics (features).

4.3. Algorithms

During the training process, the dataset contains supervised (labelled) data, i.e. each notification is classified in a failure diagnosis category. Diagnosing a new failure will be a supervised multi-classification problem, for which there are some suitable algorithms (Wang & Xue, 2014; Kotsiantis, 2007; Wu et al., 2008):

- Support Vector Machines (SVM), finds the best classification function which separates the training examples of both classes. Originally it separates binary classes so a multiclass problem will be decomposed in a series of binary problems (one-versus-one or one-versus-rest).
- k-Nearest Neighbour (kNN), finds the k training examples that are closest to the test object.
- Naive Bayes, is a conditional probability model where the probability of belonging to a class, given certain features, is computed. The disadvantage is the assumption that all features are conditionally independent given the class labels.
- Random forest, contains n decision trees and classifies by sorting an object based on feature values. Each node in the tree represents a feature and each branch a (range of) values that a feature can take.
- Neural networks, used for non-linear separable problems. They consist of input units, a hidden layer and output units.

The above mentioned algorithms are all suitable for the failure diagnosing model and will be compared after applying them individually to the data from the case study. The algorithm with the highest diagnosing accuracy will be selected.

5. FAILURE DIAGNOSIS

The failure diagnosing model (from Figure 1) is trained with the selected algorithm and pre-processed data. When data from a test example (or a new failure) is given to the model as input, it comes up with a probable failure diagnosis. This section will discuss how this diagnosis can be presented to the maintenance crew and the possible ways to explain the diagnosis to help the maintenance crew to verify the reliability of the diagnosis. Finally, the correctness of the model will be discussed.

5.1. Diagnosis

Visualisation is a strong method to transfer data from the model to the user (here the maintenance crew). Therefore the output of the model will be visualised. Commonly, probabilities are shown as simple function plots, with either probability versus data value or value versus cumulative probability. The ubiquity of these representations make them easy to read and interpret, even if the user is unfamiliar with the subject (Potter et al., 2012). This is why the assessed results are presented in a bar graph as shown in Figure 4.

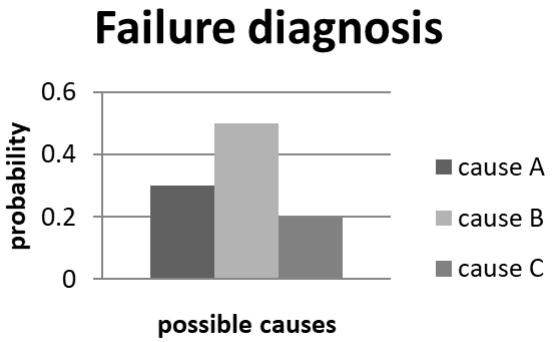

Figure 4. Example of a bar graph showing the outcome of the failure diagnosis model

5.2. Explainability

ML algorithms are often quite reliable, but sometimes not very explainable which results in a lack of trust as discussed in Section 1. Trust can be based on deterrence, on knowledge or on identification according to Lewicki & Bunker (1995). This research will strive to achieve trust based on knowledge. In order to give the maintenance crew this knowledge, it is very important to make the diagnosis assessment explainable to them.

5.2.1. XAI methodologies proposed by DARPA

DARPA, is currently researching the explainability of various AI applications (Gunning & et al., 2016). Depending on the chosen algorithm, some options are suggested which will be discussed below. To refer these options to a failure diagnosis, an example of a tyre failure of an airplane will be used. The three possible diagnoses in this example are wear, impact (overload) and other (for the remaining causes).

Local Interpretable Model-agnostic Explanations (LIME) Ribeiro et al. (2016) state that 'LIME is an algorithm that can explain the assessments of any classifier or regressor in a faithful way'. One of the examples in this paper is the explanation of an image classification assessment as shown in Figure 5. The top three classes predicted were electric guitar, acoustic guitar and labrador. The model also shows which part of the picture is used for each assessment. Although this research is focussing on the explanation of image recognition, it possibly can be applied to airplane component failure diagnosis as well. The methodology should first be transfered from pixels to text/numerical values. The image will be a new failure and in the example of the tyre a possible top three of predicted classes is low profile, high amount of landings/take-offs and moderate or low landing/take-off velocities. This can help the maintenance crew to validate the diagnosis of wear. The model will not perform the diagnosis, only the explainability. The classes are pre-defined but the importance of each class can vary for each failure.

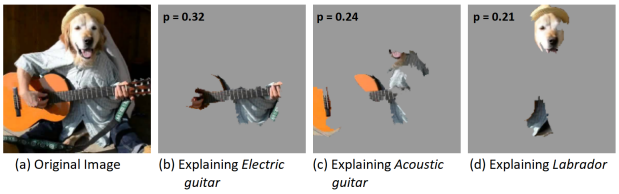

Figure 5. Application of explaining how an image is classified with LIME

Explanatory text In the paper of Hendricks et al. (2016) explanatory text sentences are used to justify an assessment. The research is also focussed on explaining deep visual models. One example is a picture of a bird (Figure 6). The model gives: 'This is a yellow breasted chat because this is a bird with a yellow breast and a grey head and back'. For the tyre example the explanation could be, this tyre is worn because it has lasted for over 200 landings/take-offs and now has a profile below x mm.

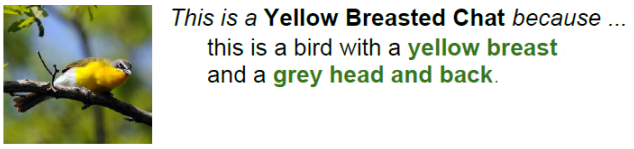

Figure 6. Application of explaining how an image is classified with LIME

Bayesian teaching Bayesian teaching is the optimal selection of examples for machine explanation. So examples of the training data will be selected to explain the assessment of the model. In Figure 7 (Gunning & et al., 2017) a picture of a child is used as input for the model. The output, based on image recognition, is that the face is angry because it is similar to the images of kids with angry faces are given and dissimilar to images of kids with sad and happy faces. For the failed tyre, training notifications will be selected with similar feature values. E.g. the tyre had an impact because the amount of landings is (more or less) equal to the ones from these failures (followed by showing these similar notifications).

Neural networks Neural networks are used for classification but if the decisions towards this classification can be made visible it can also be used as an explanation for a diagnosis. A neural network contains an input layer, one or more hidden layers and an output layer. The hidden layers in between are seen as a black box. In Figure 8 (Gunning & et al., 2017) a neural network is shown which predicts the type of animal. The training data contains images of all kinds of animals, the input is an image of a dog. They explain that the first layer neurons respond to simple shapes, the second layer neurons to more complex structures (a tail, paws, head). This will

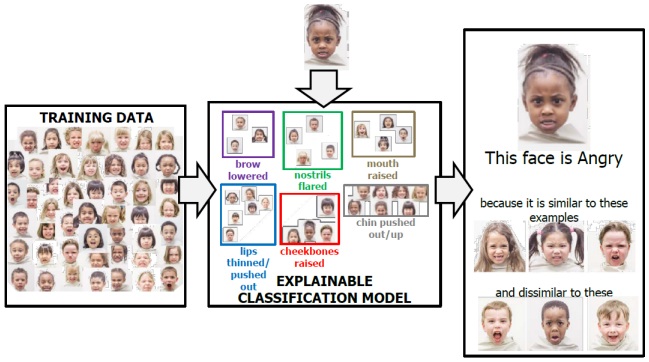

Figure 7. Application of explaining how an image is classified with LIME

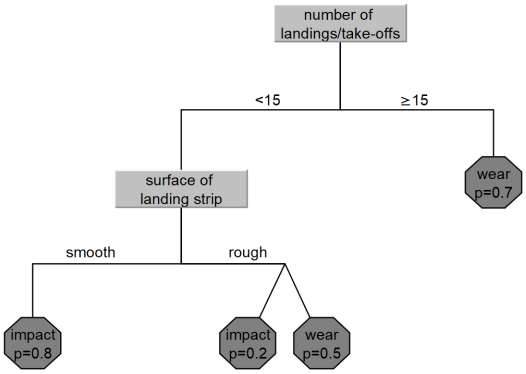

Figure 9. Simplified explanation of a tyre failure diagnosis by a decision tree

further increase until the n^{th} layer were the neurons respond to highly complex, abstract concepts. The output of the model was 10% wolf and 90% dog. For the example of the tyre, the simple shapes will probably be rough and clear separations between data (tail numbers, number of landings/take-offs) more complex layers will include values of velocity, variation in velocity etc.

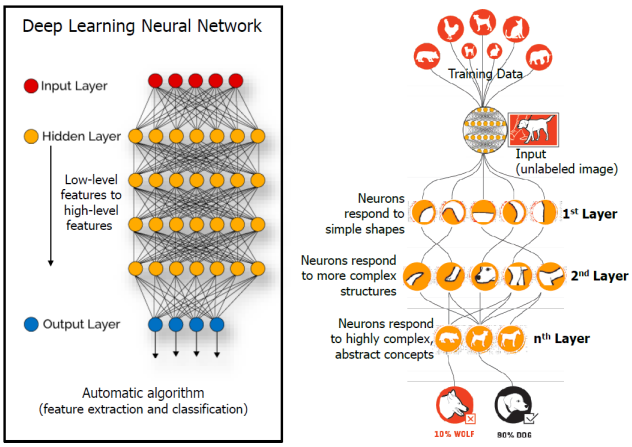

Figure 8. Application of explaining how an image is classified with LIME

Decision trees Decision trees are very powerful, but also simple and efficient for extracting knowledge from data. They are easy to interpret, understand and control by humans (Ertel, 2009). In Figure 9 an example is given for the tyre failure. The decision tree determines the nodes (number of landings/take-offs, surface of landing strip) and the limit values (15 and smooth/rough) by computing the information gain. For each layer in the tree, the feature with the highest information gain will become a node.

5.2.2. Proposed XAI methodology: Failure Diagnosis Explainability (FDE)

As discussed in Section 5.2.1, there are various methodologies to make a ML decision more explainable for the user (in this case the maintenance crew). As these methodologies are still in development and are only conceptual at the moment, detailed information has not been available yet. Apart from that, most of the previously mentioned XAI methodologies are focussing on images. The data used for the failure diagnosis model will consist of text and numerical values, therefore an additional step has to be taken to make it possible to apply the concepts which are focussing on images to a text and numerical problem. Also, since DARPA's methodologies are very new and only applied to a specific example, there is no information available yet about how to compare the performances of these different methodologies. This lack of information makes it, at the moment, not possible to apply these methodologies to failure diagnosis and therefore a new XAI methodology will be proposed here: Failure Diagnosis Explainability (FDE). The newly proposed methodology is based on the methods used for the different options suggested by DARPA. A common method to describe on what grounds the decision was based is with characteristics. In the example of the guitar playing dog, parts of the image were highlighted which were characteristic for a specific decision. For the bird, the characteristics were described in text form and for the example of the child, pictures with similar and dissimilar characteristics were shown. The dog from the neural network was explained by its characteristic shapes and finally in a similar manner, the decision tree from Figure 9 explains tyre failures based on usage characteristics.

The failures are described by different features (as discussed in Section 4.2) which will be used to explain the failure diagnosis assessed by the model. Since many features are implemented in the model, one can wonder which ones are most important: which features characterise a certain diagnosis the most. E.g. a certain value for feature j

is very typical for a worn tyre. The variable importance of a model gives the (relative) importance of each variable (feature) for a trained model. But since this model includes various diagnoses (classes), the importance of each feature per class is not determined: the variable importance shows which feature is most important in deciding which class a failure belongs to, which is not equal to the characterisation of a certain class. To achieve this, new models will be trained in which, in each model, one class is appointed as 1 and the others as 0 (binary). E.g. class A is 1 and classes B and C are 0. Still, this does not precisely determine the variable importance per class, but it does yield the variable importance of a specific class relative to the other classes. A top five of most important features is determined per class (diagnosis) which can be found in Figure 12 in the case study. Before the model is trained, all values will be normalised using feature scaling. Note, the variable importance can only be determined for so called 'white box' algorithms as decision tree, random forest and Naive Bayes.

Finally, this methodology will show the user how much (and for which features) a new failure matches with each diagnosis. The more the features match the expected values of a specific diagnosis, the more likely it is that this failure belongs to that class. Also, if the features are very dissimilar to the values of a certain class, it is very unlikely that this failure belongs to that specific class. So, the reasoning of this methodology is similar to the one from Bayesian teaching from Section 5.2.1. A failure is from class (diagnosis) A, because the features have similar values as class A. Contrary, this failure is not from classes (diagnosis) B and C, because the features have dissimilar values as class B and C (Gunning & et al., 2017).

The expectations for each feature are expressed by boxplots. If the value of a new failure falls within the 50% range (recognised by the box) of the boxplot, there is a good match. If the value falls within the 95% range (recognised by the whiskers of the plot), the values still matches but less compared to the 50% range. In case the value falls outside the 95% range, it is unlikely the failure belongs to that class, based on a 95% confidence interval; 2 sigma limit. The boxplots of the top five features of each class were represented in one graph. The values of a new failure will be added to the graph to give the maintenance crew an indication of which features match with the expectations of a certain diagnosis and which do not. A fit is made through these value points. If this line is equal to the x-axis, the new failure corresponds entirely with the expected values. An example of such a plot can be found in Figures 13 and 14 in the case study.

For a better visual presentation, the medians of all boxplots were aligned with the same x-axis. Also, all values are, per variable, multiplied with their scaled importance $w(f_j)$. This means that a deviation on the most important variable will be enlarged compared to a deviation on the fifth most important variable. Equation 1 shows the computation of all absolute explanation values $S_{i,j}$ (both for the boxplot and for a new failure), where the fraction normalises the value $f_{i,j}$ with f as the sensor data from feature j and failure i. Following the median will be substracted and finally, multiplied with the scaled importance.

$$S_{i,j} = \Big(\frac{f_{i,j} - \min(f_j)}{\max(f_j) - \min(f_j)} - \text{median}(f_j)\Big) * w(f_j) \quad (1)$$

Apart from a visual representation with boxplots, the χ-factor is introduced which is a measure to what extend the value of a new failure (x) from feature j matches the expected values of the same feature from a specific diagnosis. The χ-factor is determined by the value from feature j of a new failure divided by the variance of that feature j per specific diagnosis as shown in Equation 2. The χ-factor is determined per variable but also the maximum and average value from the top five features per class are determined. If the χ-factor is equal to 0, the new failure matches entirely with the expected value. The higher the factor is, the less the new failure matches the expectations of that specific diagnosis.

$$\chi_{x_j} = \frac{x_j}{\text{variance(j)}} \quad (2)$$

This methodology will be shown and tested in Section 6.3.2 in the case study.

5.3. Correctness

The number of correctly classified examples is a performance measure for the diagnosing model. To measure this, part of the data is randomly separated before training and assigned as the test data. The correctness will be expressed in the accuracy of the algorithm, which means the percentage of correctly classified failures. A common practice within ML is to compare the achieved accuracy with a certain baseline. One common baseline for classification problems is the 'most frequent' which always classifies the most frequent label in the data set. When the accuracy of a ML algorithm is below this baseline, the algorithm will not be of any value.

If the training data would be used for performance measurements, overfitting of the model cannot be noticed. After the initial training, when the model is in use, it will continuously enrich itself after every new example. The maintenance crew validates the diagnosis before it is added to the training data, hence the model will learn by updating itself. This process is also called incremental learning (Ertel, 2009). Incremental learning will improve the correctness of the model, unless only failures from the same class are added to the training

data. In which case the dataset can become very unbalanced which has a negative influence for overfitting.

6. CASE STUDY

In this section, the model proposed in this thesis is applied to a case study to test the feasibility. For the case study, data from RNLAF (Royal Netherlands Air Force) F-16s (fighter airplane) main wheels are used. The F-16 is equipped with a nose landing gear and a main landing gear. The nose landing gear consists of one nose wheel, the main landing gear of two main wheels. The RNLAF has a total of 68 F-16s using these main wheels. The tyres are removed for corrective maintenance or in case of other maintenance e.g. the wheel will be removed for when say a shock strut replacement is conducted. After the replacement of the shock strut, the original wheel will be reinstalled on the same tail. Common failures include the tyre being worn out, a flat tyre and a replacement due to contamination (mainly oil). There is no preventive maintenance on the wheels. The RNLAF does not retread their wheels, so a worn tyre will be directly disposed of. Flat tyres will be repaired if possible and contaminated tyres will be cleaned. After repair, the wheels will be mounted on an arbitrary tail. Most likely a different one, but occasionally it could be the same. Since the data from the used case is not public, the steps in data processing will be elaborated to provide more transparency.

6.1. Data pre-processing

The data has to be pre-processed before the model can be trained. The maintenance and usage data are prepared separately prior to being combined. For the pre-processing, the steps described in Section 3 will be followed. The main wheel is chosen for the case study since this component had one of the most removals and has only a few failure modes which are easy to understand.

6.1.1. Maintenance data

The maintenance data (repair cards) is obtained from the Computerised Maintenance Management System (CMMS) SAP (Systems, Applications and Products). A repair card is filled in manually by the maintenance crew. The data is extracted from SAPs database and is filtered by tail and by number of notifications per SN (Serial Number). Failure registrations without a specified tail are deleted since the tail is needed to link a failure to the usage data. If there is only one notification of a specific SN, the component is still in use (only the installation is reported, while no removal is reported). The installations do not mark a failure, therefore only the removals are kept. For this case study it is assumed that all removals are failures since cannibalisation (i.e. removal of sound components to be used on another tail) of wheels is not common for these F-16s. For the case study, only one removal is eliminated due to these filters.

The data is labelled in three categories; ‘flat’, ‘worn’ and ‘other and unclassifiable’ (when the cause is unknown, different from the other two or the failure did not have enough information to be labelled). A distinction between an elimination from the dataset and a label of ‘other and unclassifiable’ is made between the kind of information available. If a failure cannot be linked to the usage, it will be eliminated but a failure with a limited failure description will be labelled as ‘other and unclassifiable’. For labelling, several data fields were checked for words as ‘worn’, ‘due’, ‘flat’, ‘deflating’ etc. This labelling is done automatically, but it is based on words which are found manually by going through the failure descriptions. Text mining may be an option to fully automate this process. There is no contamination class since these failures could not be traced from the repair cards or were not present in the data set. The labelled notifications (242) are combined with a dataset containing information about the tail and installation and removal date of each unique component. While combing the maintenance data with the usage data, only 144 removals are kept due to missing usage data (elaborated in Section 6.2). So the amount of data is big (notifications and usage data together), but there are only a few failure cases which can also be combined with usage data. For data mining, 144 notifications is generally considered as a relatively small number, but in industries such as the aviation sector and in particular in the military sector, these numbers are rather common. A better registration of failures can easily lead to more notifications. Also, the class ‘other and unclassifiable’ is now considerably high (82 of the 144 failures), this is mainly due to many failures which did not have enough information to classify. Preferably, this class is the smallest of the three, but this is a limitation of the data from current practise.

6.1.2. Usage data

Sensor data from almost 7,000 flights are used. The data are extracted via a spool file, as it is stored in a relational database, requiring a considerable amount of computing effort. The variables, which are appointed as related to the loads on the wheels by engineering judgement, are: CAS (Calibrated Air Speed), strain from several strain gauges, weights, longitudinal and lateral accelerations, pitch and air temperature. There are more variables available but these are (for example) related to the engines or electronic components. Also, the phase of the flight (take-off, flight, landing) is assigned to each point. The variable selection is done by logical reasoning. In case of doubt, the variable was selected. This selection is done to avoid that the model will fit on unrealistic relations and to gain more causality in the model. E.g. the model could find a strong relation between maximum altitude and flat tyres but it later turns out that there were nails on the landing strip where flights with high altitude

	n_flights	w_m_CAS_T	w_m_CAS_L	w_m_FS325_T	max_FS325__T	w_m_FS325_L	w_m_TEMP_T	w_m_TEMP_L	days_before_failure	label
1	-0.98139135	-2.43914072	1.1693447432	-0.165336975	-0.22249063	0.574432924	-0.978770180	-1.768879625	0.188979918	other
2	0.07166643	-0.92659475	0.0082763604	-0.206886278	-0.20269187	-0.148377051	-0.518846040	-0.253808124	0.955546346	worn
3	0.07166643	-0.92659475	0.0082763604	-0.206886278	-0.20269187	-0.148377051	-0.518846040	-0.253808124	0.955546346	worn
4	-0.87608557	0.06278203	-0.7465409050	-0.480736948	-0.54068516	-0.634541921	0.128497126	0.604973286	-0.604963882	other
6	-0.34955668	-0.11592674	0.4228300911	-0.196726539	-0.22249063	0.047708314	-0.290571600	-0.962820940	2.954094533	other
8	-0.66547401	0.68826272	0.9062447781	-0.369578308	-0.43179193	-0.250248954	-0.114446469	-0.568176758	0.353244153	worn
9	-0.66547401	0.68826272	0.9062447781	-0.369578308	-0.43179193	-0.250248954	-0.114446469	-0.568176758	0.353244153	worn
10	-0.66547401	0.68826272	0.9062447781	-0.369578308	-0.43179193	-0.250248954	-0.114446469	-0.568176758	0.057416434	worn

Figure 10. Example of data table which is used as input for a ML algorithm

were trained. These type of unrealistic correlations can be prevented by the manual variable selection. The drawback could be, however, that a real but unexpected correlation is not revealed.

6.2. Machine Learning model

The ML algorithms require a dataset in which each row is a notification and the columns are the different features. After the data pre-processing, the maintenance and usage will be combined to put the data in the right format. Every notification should be combined with the usage related to this notification. No usage data was available for 98 of the 242 failures, leaving 144 failures for training and testing. The unavailability of the usage data can have various reasons e.g. the recorder was full or the data are yet not added to the database due to the mission location of the airplane.

A notification can only include single values and not a (varying) range of data points. Therefore the variables of the usage data were translated to characteristics (as discussed in Section 4.2) with single values such as number of flights, average speed etc. Figure 10 shows an example of the input data after pre-processing and the feature selection. The selected features are:

- number of flights
- average CAS during 1) take-off and 2) landing
- average and maximum strain gauge (called FS325) measuring the wing root bending during 1) take-off and 2) landing
- average and maximum strain gauge (FS374) measuring the fuselage bending during 1) take-off and 2) landing
- average and maximum strain gauge (BL120) measuring the wing tip bending during 1) take-off and 2) landing
- average lateral acceleration during 1) take-off and 2) landing
- average longitudinal acceleration during 1) take-off and 2) landing
- average total weight during 1) take-off and 2) landing
- average air temperature during 1) take-off and 2) landing
- the amount of calendar days between failures (i.e. the component age)

Table 1. Accuracies obtained from various ML algorithms

Algorithm	Accuracy
Naive Bayes	62%
Neural network	43%
Support Vector Machine	69%
Random forest	81%

All these characteristics are calculated over the period of time since the first installation of the considered component. Now that the datasets are combined, a historical reconstruction can be made. This example has already been shown in Figure 3.

Since the algorithms can only handle numerical and categorical values, all text fields were deleted. These data were already taken into account, since these fields were already used for the labelling of the notifications. Several algorithms (from Section 4.3) were applied to the training data. Table 1 shows the accuracies of the different algorithms. The first algorithm, Naive Bayes, probably performs moderately since this algorithm is simply too weak or too simple to detect the patterns in this dataset. The neural network and SVM are probably overfitting on the training set. Overfitting can be seen when the accuracy on the test set is remarkably lower than on the training set. Since random forest gave the highest accuracy, this algorithm is used for the remaining steps.

6.3. Failure diagnosis

The model is now trained with the training set. After that, one failure from the test set will be fed into the model to assess a diagnosis for this failure. This single failure from the test set simulates a new failure. The results (diagnosis) of two new failures and how they should be interpreted (explained) are discussed in this subsection.

6.3.1. Diagnosis

As discussed in Section 5.1 the assessed diagnosis is represented in a bar graph. Figure 11 shows the results for two specific failures which are most likely a 'worn' tyre (case A) and a 'flat' tyre (case B). This figure only shows the probability of each diagnosis yet does not explain the user why each diagnosis has a certain probability.

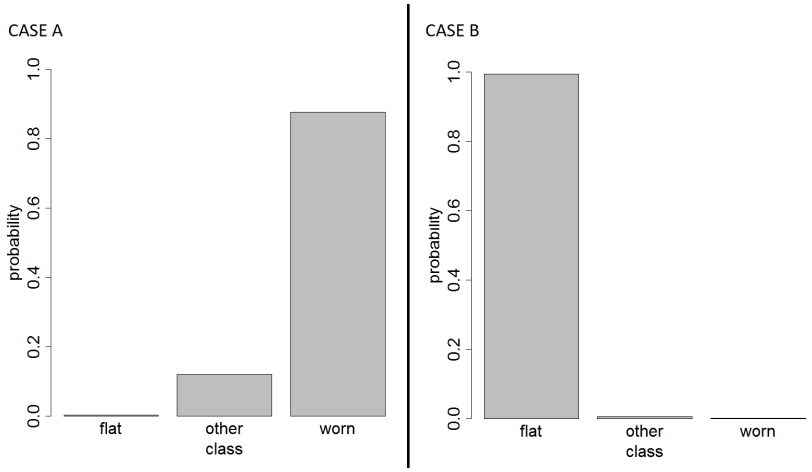

Figure 11. Probability of each class plotted for two specific cases, A and B

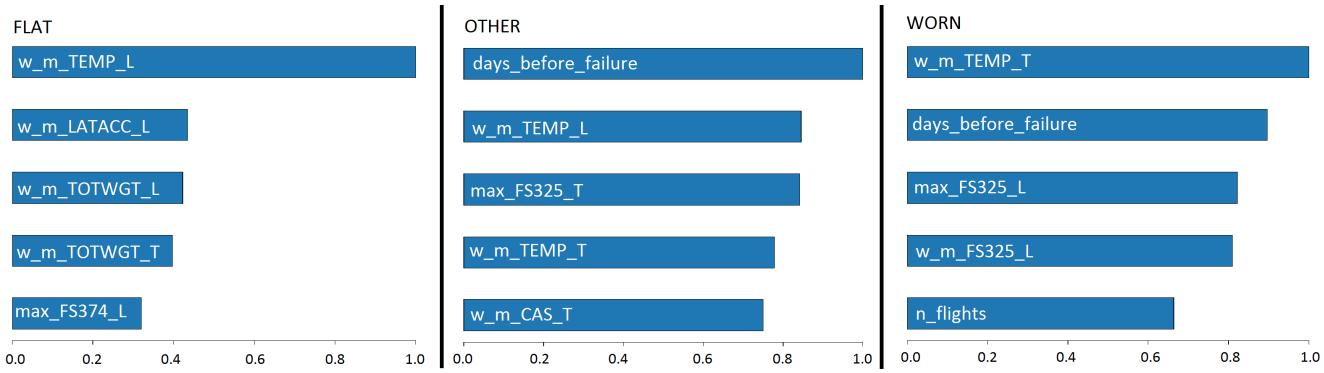

Figure 12. Variable importance plot for the different features

6.3.2. Failure Diagnosis Explainability (FDE)

As discussed in Section 5.2.2, the variable importances per class relative to the others had to be determined. The top five of scaled importances per class are shown in Figure 12. As can be seen, the weighted mean air temperature during landing is the most important feature to determine whether the tyre is 'flat' or not and the maximum fuselage bending during landing is the fifth. These top five features are most distinctive for the classification of the diagnosis (class) 'flat' versus the other two classes 'other and unclassifiable' and 'worn'. A complete overview of all the abbreviations used for the features can be found in Table 2.

Figure 13 and Figure 14 present the results of a 'worn' (case A) and a 'flat' tyre (case B) (as shown in Figure 11). As can be seen in Figure 13, the values of this new failure (case A) fall within the 50% range for four of the five features for the 'flat' diagnosis. The first variable is on the edge of the 95% interval. For the diagnosis 'other and unclassifiable', only three values fall within the 50% range. Therefore it is unlikely that the new failure belongs to the class 'flat' or 'other and unclassifiable'. For the diagnosis of a 'worn' tyre,

Table 2. Features from top fives of important variables

Features	Meaning
w_m_TEMP_L	weighted average air temperature during landing
w_m_TEMP_T	weighted average air temperature during take-off
w_m_LATACC_L	weighted average lateral acceleration during landing
w_m_TOTWGT_L	weighted average total weight during landing
w_m_TOTWGT_T	weighted average total weight during take-off
max_FS374_L	maximum fuselage bending during landing
days_before_failure	number of days before failure
max_FS325_T	maximum wing root bending weight during take-off
max_FS325_L	weighted average wing root bending during landing
w_m_FS325_L	weighted average wing root bending during landing
w_m_CAS_T	weighted average calibrated air speed during take-off
n_flights	number of flights before failure

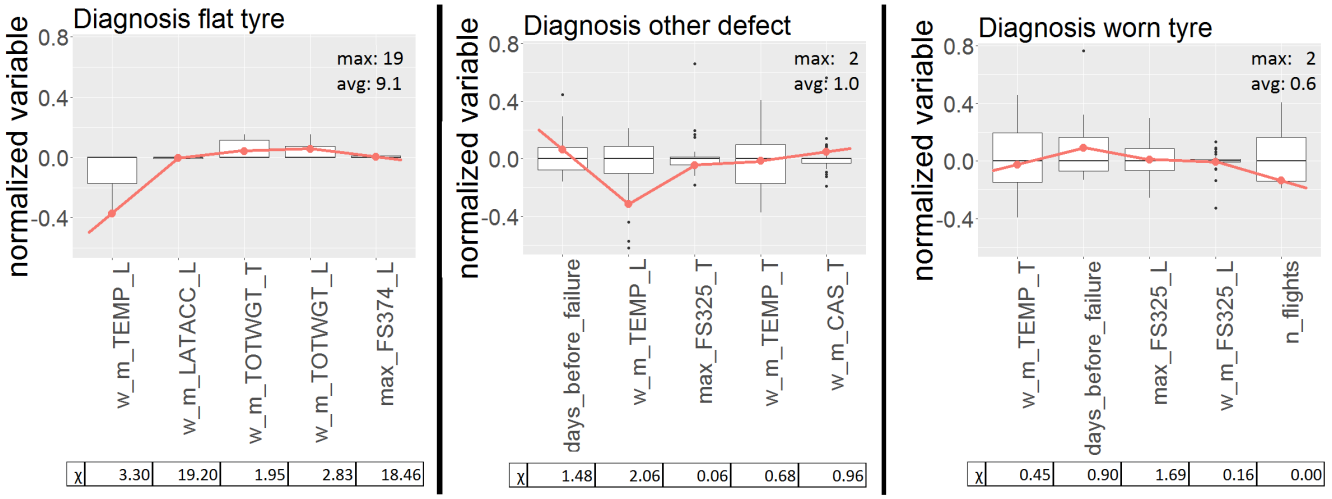

Figure 13. Boxplots explaining a specific case of a 'worn' tyre (case A)

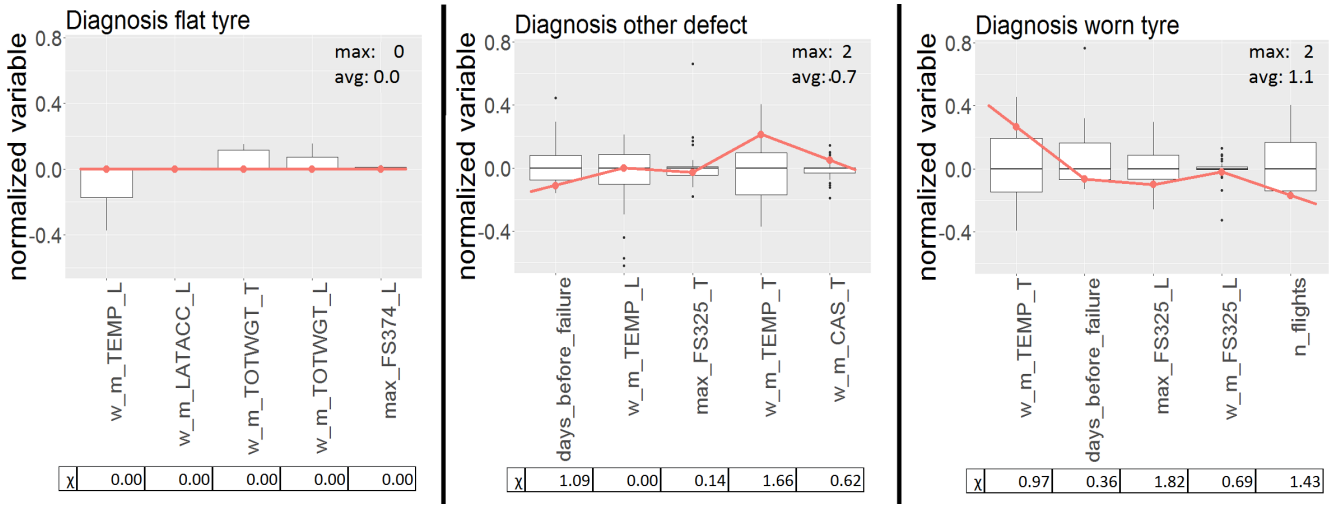

Figure 14. Boxplots explaining a specific case of a 'flat' tyre (case B)

the new failure falls within all 50% ranges and two variables even match the median (the third and the fourth variable). So it is likely that this new failure is 'worn' because all values correspond with the expectations. Even the fourth variable (the weighted average wing root bending during landing), which has a very tight range, matches. The χ-factors per variable are presented in the bottom of the graph. Both the maximum and the average χ-factor indicate case A is from class 'worn' since the maximum and average are lowest for 'worn'. Concluding, case A is 'worn' because the features have similar values as class 'worn'. Contrary, this failure is not 'flat' or 'other and unclassifiable' because the features have dissimilar values as classes 'flat' and 'other and unclassifiable'.

In Figure 14, the values of case B entirely correspond with the expectations of a 'flat' tyre failure. This makes it very likely, the assessed class will be 'flat'. On the contrary, the new failure does not correspond with most of the variables from the classes 'other and unclassifiable' and 'worn'. The χ-factor comes to the same conclusion since both the maximum and average are lowest for 'flat'.

In conclusion both, from case A and B, assessed diagnoses (shown in Figure 11) can be explained with the proposed XAI methodology FDE both with visual representation and with the χ-factor.

6.3.3. Correctness

After training the model, it achieved an accuracy of 81% on the test set which means that the notifications from the test set were classified to the correct diagnosis ('flat','other and unclassifiable' or 'worn') in 81% of the cases. As discussed in Section 5.3 this result will be compared to

the 'most frequent' baseline, which is the percentage of occurrence of the biggest class. In this case the class 'other and unclassifiable' is the biggest class. Of the in total 42 cases from the test set, 28 are labelled as 'other and unclassifiable'. If the model would (without any knowledge or training) assess every case as 'other and unclassifiable', it has an accuracy of 67%. The model showed its value already with the first test within this feasibility study, with a relative improvement of 21% compared to the baseline. Figure 15 shows the confusion matrix of the test data. This matrix gives an overview of the correctly and wrongly classified failures and shows, for example, 24 of the 28 failures which are from the class 'other and unclassifiable' were diagnosed by the correct class. The others are diagnosed with 'worn'. It is remarkable that most misclassifications are caused by a 'flat' or 'worn' tyre which are predicted as 'other and unclassifiable'. This can be explained by the fact that the class 'other and unclassifiable' most likely contain also many failures which are actually from the classes 'flat' or 'worn' but were simply unclassifiable due to the lack of information on the repair card.

Figure 15. Confusion matrix of the test set applied to the model

7. DISCUSSION

Different methods and technologies have been discussed and tested to come to the model proposed in this research. This model is able to diagnose a failure and also to explain the assessement of this diagnosis with XAI. A selection is made of algorithms which were suitable for the type of data and the type of problem (a classification). The algorithms are not discussed in detail since they are commonly used algorithms and the focus of this research was not on developing a new algorithm. The algorithm with the highest accuracy was selected. In case another algorithm was chosen, the model's diagnosing accuracy would have been considerably lower. The accuracy of the model will never become 100% since the model will not be able to predict when a tyre is punted by a nail. The focus for this research was more on the addition of XAI to the model. The conceptual methodologies for XAI proposed by DARPA are currently not reproducible since detailed information is not available yet. Even if they are reproducible, a translation has to be made from images to text and numerical values. Therefore a new XAI methodology is proposed, called FDE. The selected ML algorithm was from influence on the proposed XAI methodology, since the variable importence can only be determined for, so called, 'white box' algorithms as decision tree, random forest and Naive Bayes. So the selection of another algorithm could have required another XAI methodology than FDE.

The combination of maintenance and usage data with (X)AI showed its value. An accuracy improvement of 21% (to a total of 81%) for classifying a diagnosis compared to the baseline is achieved already with this feasibility study. Even though the model was trained with only a small amount of failures, which is very common in this industry. Also, with FDE, the model was able to explain the assessed diagnosis to the user. This transparency was given both visually and numerically with the χ-factor. The usefulness of automatically diagnosing a failed tyre based on usage data can be questioned. After all, given that the tyre is failed, the distinction between a flat and worn tyre is easy to make. However, this case study served as a feasibility study. The tyres were one of the components with most data available and the failures were easy to understand which made validation of the model possible. Applying this model to components which are hard to diagnose (e.g. when the component is physically hard to reach) can really improve the efficiency of troubleshooting. With this research the requirements for the data and for such an analysis have become clear.

The current technology is already able to (based on trend analysis and physical models) predict a component or system failure quite accurate. However, after a component is uninstalled from the system and tested on a test bench it will receive a NFF (No Failure Found) so that no specific maintenance can be applied. The proposed model from this research is able to assess a diagnosis based on the component's historical usage. This can prevent overmaintenance on the component which was about to fail.

The greater goal is to predict failures. To implement this to the current model, the model also has to predict a time to failure besides assessing (classifying) a diagnosis. So besides a classification problem there is also a regression problem (determining the time to failure). Regression problems require different ML algorithms than classification problems. Still this can be based on usage data. For predicting the time to failure, features such as time since operating, time to failure have to be added to the model. Cumulative features

need to be implemented besides the average and maximum values of the variables which were already used. Given a certain history, the time to failure can be predicted per diagnosis for a specific component. E.g. given the history of a tyre (amount of landings above a certain speed and/or temperature, total force (cumulative) etc.), the time to a flat tyre and to a worn tyre will be predicted. The diagnosis with the lowest time to failure will be most likely to happen first.

In the case study, the model was applied to a F-16 main wheel but likely this model can be applied to other components as well. As long as the systems are equipped with usage monitoring sensors which are able to store the data and the component failures are reported and stored in a database. Since the model links the failures to the usage, there has to be some relation between the failures and the usage, even though this reason does not have to be known already. Therefore the model might not be applicable to all components yet. New sensors should be added to the system so that the usage of those components can be recorded as well. Within the RNLAF, it is likely that the model can, apart from the application to the components of the F-16, be applied to other weapon systems (other airplanes or helicopters) as well. An important requirement for application is measuring and storing of usage data. The Dutch air force is the only one of which all F-16 are equipped with this usage sensing and storing system. Therefore the model will, in its current status, not be applicable to the F-16s of other air forces. Since civil airplanes are more and more equipped with sensors monitoring the systems usage, the application will not be limited to the defence sector. Also it can be applied more widely than in the aviation sector alone, e.g. maritime, rail, energy generation and distribution and process industries.

8. CONCLUSION

A literature study showed that AI is getting bigger and more and more involved in today's technologies. The use of AI can be a potential step in the development of the maintenance sector to participate in this change. As a start, this research focusses on diagnosing failures with AI based on usage data. Besides, to provide knowledge and trust to the user, the assessed diagnoses have to be transparent and explainable. Therefore, XAI has been implemented in the model. Following the answers on the sub-questions of this research will be given. Finally the main question is answered, which data driven approach gives the best results for automated failure diagnosis?

Which data are required and in which form should it be presented to develop an automated failure diagnosis model based on usage?
This question is answered in Section 3. First of all, there are no strict requirements for the amount of notifications (number of failures) and amount of variables c.q. features. But a high amount of notifications is preferred. Before the maintenance and usage data must be combined, they have to be cleaned and prepared so that the ML algorithm can handle the data. ML algorithms require a dataset were each row is a failure and each column a feature. The algorithms which were proposed can only handle numerical and categorical data so all important text fields should be in some way converted to numerical or categorical values. The usage has to be expressed in features since the ML algorithms can only handle single values.

How should the model be trained to obtain a failure diagnosis model?
As can be seen in Section 4 there are three important steps in training the model. First, the maintenance and usage data need to be combined. Therefore the usage data from a specific tail from a specific period need to be linked to a certain failure. Secondly, the feature selection is depended on the failure mechanism and the ML algorithm (neural networks can handle whole data strings while more simple networks require only features). Finally, a ML algorithm needs to be chosen based on the best performances (correctness). There are various options available, also for this type of data and type of problem. Guidance can be found in the mentioned literature from Section 4 and a selection can be based on comparing the performance (accuracy) of the various algorithms.

Which methodologies are applicable for eXplainable Artificial Intelligence (XAI)?
DARPA is currently researching the explainability of various AI applications. The options suggested to explain a failure diagnosis are LIME, explanatory text, Bayesian teaching, neural networks and decision trees. Since detailed information about all these methodologies is not available (yet), a new XAI methodology is proposed, Failure Diagnosis Explainability. Apart from this sub-question's objective, there is no information available yet about how to compare the performances of the different methodologies since untill now they are only conceptual.

How should the data be explained to substantiate the provided failure diagnosis?
By using the variable importance per class, a top five was made of the variables which characterise a certain class the most. The features of a new failure need to be compared with the expectations per feature. Apart from a visual representation of this comparison, the χ-factor is introduced which is a measure to what extend the value of a new failure matches the expected value from a specific diagnosis. The lower the χ-factor, the more the feature matches the expected value. The transparency in the diagnosis assessment gives the user additional information about the history of the component and helps the user to trust the assessed diagnosis.

Which data driven approach gives the best results for

automated failure diagnosis using XAI?
For the selection of an AI algorithm, the pre-selected algorithms have been applied to the case study. This resulted in the highest assessment accuracy for random forest. For the addition of XAI to the model, an overview is presented with the applicable XAI methodologies. FDE is proposed as a new XAI methodology next to the existing conceptual methodologies. The proposed model showed its value with an improves diagnosing accuracy of 21% (from 67% to 81%). Aside from this achievement, the model was also able to make the diagnosis assessment transparent and explainable for the user, the maintenance crew. Since there is insufficient information available abouth the other methodologies and how to compare them, it can not be determined whether this option is the best or not. But this feasibility study thereby demonstrated that the application of AI to failure diagnosis is possible and that XAI can support by gaining trust of the users.

9. RECOMMENDATIONS

With the implementation of the case study, many difficulties arose. Most of them concerned the preparation of the data in such a way that the data could be implemented in the ML algorithm. Since many of these difficulties will be common for data-driven approaches in this industry, some general recommendations can be drawn. First of all, when the data are stored in a database, it does not mean they are also available for the preferred analysis. In case it is not available it should be made clear which steps need to be taken. The sensor data from the case study were stored in a relational database. Therefore it took significant effort to extract the data from the database. Secondly, the maintenance and usage data have to be combined. To do this, there must be a possibility to link these datasets with each other (by tail, by date, by pilot, by airport or anything else). Thirdly, the data should be prepared for the algorithm (cleaning, labelling, inventing features etc.). This process can take a lot of effort and logical reasoning by selecting the features can add causality to the model. The way the data is recorded and stored will partly have influence on this effort. E.g. for cleaning and labelling a certain way of recording and storing can help, but it does not have influence on the effort to select the features.

For further research, the features can be changed, the data can be more balanced (optimal is equal amount per class) and more data can be used (also with incremental learning). These changes may lead to improvement in the accuracy (correctness) of the model. Other features which could be taken into consideration are the number of exceedances of certain values and the variance or distribution of the data. For this research, a first set of features is used and optimisation of the features was out of the scope of this research. By using other labelling techniques, consulting more data sources or ensuring the repair cards are filled in more completely and more consistently the number of unclassifiable diagnoses will decrease which makes the dataset more balanced. The model is now tested on one component, a logical next step is testing the model on other components as well. For example more complex components with more failure mechanisms, and components which are repaired multiple times during their lifetime. For a prove of concept of the proposed XAI methodology FDE hypothesis testing can be used were a value from a certain variable from a new failure will be tested versus the expected values of that variable for a specific diagnosis. The expected values can be expressed in different ways such as average, variance, standard deviation etc.

REFERENCES

Al-Garni, A., Jamal, A., Ahmad, A., Al-Garni, A., & Tozan, M. (2006). Neural network-based failure rate prediction for de havilland dash-8 tires. *Engineering Applications of Artificial Intelligence*(19), 681-691.

Bishop, C. (2006). *Pattern recognition and machine learning*. Springer.

Ertel, W. (2009). *Introduction to artificial intelligence*. Springer. (Section 8.4, 9.4)

Gunning, D., & et al. (2016). *Explainable artificial intelligence XAI* (Tech. Rep. No. DARPA-BAA-16-53). Defense Advanced Research Projects Agency (DARPA).

Gunning, D., & et al. (2017). *Explainable artificial intelligence XAI '- program update* (Tech. Rep. No. DARPA-BAA-16-53). Defense Advanced Research Projects Agency (DARPA). (slide 4, 15)

Hendricks, L., Akata, Z., Rohrbach, M., & et al. (2016). Generating visual explanations. In *European conference on computer vision.*

Holzinger, A., Biemann, C., Pattichis, C., & Kell, D. (2017). *What do we need to build explainable AI systems for the medical domain?* (Tech. Rep.).

Hua, J., Xiong, Z., Lowey, J., Suh, E., & Dougherty, E. (2005). Optimal number of features as a function of sample size for various classification rules. *Bioinformatics*, *21*(8), 1509-1515.

Jain, A., & Chandrasekaran, B. (1982). Dimensionality and sample size considerations in pattern recognition practice. *Handbook of Statistics*, *2*, 835-855.

Khoo, L., Ang, C., & Zhang, J. (2000). A fuzzy-based genetic approach to the diagnosis of manufacturing systems. *Engineering Applications of Artificial Intelligence*(13), 303-310.

Koch, M. (2018). Artificial intelligence is becoming natural. *Cel, Elsevier Inc.l*(173), 531-533.

Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques. *Frontiers in Artificial Intelligence and Applications*, *160*, 3-24.

Lee, J., & Wang, H. (2008). *Complex system maintenance handbook*. Springer. (Figure 3.1 in Section New technologies for maintenance on page 51)

Lewicki, R., & Bunker, B. (1995). *Trust in relationships* (Tech. Rep. No. RS 95-7). Business Research College of Business.

Milne, R., Nicol, C., & Trave-Massuyes, L. (2001). Tiger with model based diagnosis: initial deployment. *Knowledge-Based Systems*(14), 213-222.

Potter, K., Kirby, R., Xiu, D., & Johnson, C. (2012). Interactive visualization of probability and cumulative density functions. *International Journal for Uncertainty Quantification*, *4*(2), 397-412.

Ribeiro, M., Singh, S., & Guestrin, C. (2016). why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.*

Rich, E. (1983). *Artificial intelligence*. McGraw-Hill.

Russell, S., & Norvig, P. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.

Smit, K. (2010). *Onderhoudskunde [maintenance science*. VSSD. (Section 9, p.538)

Tarifa, E., Humana, D., Franco, M. S., S., & et al. (2002). Fault diagnosis for a msf using neural networks. *Desalination*(152), 215-222.

Tinga, T. (2010). Application of physical failure models to enable usage and load based maintenance. *Reliability Engineeringand System Safety*(95), 1061-1075.

Vachtsevanos, G., Lewis, F., Roemer, M., Hess, A., & Wu, B. (2006). *Intelligent fault diagnosis and prognosis for engineering systems*. John Wiley & Sons,Inc. (Section 5, p. 181-191)

Wang, Z., & Xue, X. (2014). *Support vector machines applications*. Springer. (Chapter 2)

Wu, X., Kumar, V., Quinlan, J., Ghosh, J., & et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1-37.

Yan, C., Wang, H., Zhou, L., & Li, Z. (2014). Fault diagnosis expert system of turbine generator sets based on rule reasoning and case reasoning. *Applied Mechanics and Materials*, *513-517*, 4443-4448.

Appendix

This paper is published and presented during the fourth European conference of the Prognostics and Health Management Society 2018 in Utrecht, the Netherlands from 3-6 July

Automated Failure Diagnosis in Aviation Maintenance Using eXplainable Artificial Intelligence (XAI)

Sophie ten Zeldam[1,2,3], Arjan de Jong[1], Richard Loendersloot[2] and Tiedo Tinga[2,3]

[1] *Netherlands Aerospace Centre (NLR), Amsterdam, the Netherlands*
Arjan.de.Jong@nlr.nl

[2] *University of Twente, Dynamics based Maintenance group, Enschede, the Netherlands*
r.loendersloot@utwente.nl
t.tinga@utwente.nl

[3] *Netherlands Defence Academy, Military Technical Sciences, Den Helder, the Netherlands*
sg.t.zeldam@mindef.nl

ABSTRACT

An incorrect or incomplete repair card, typically used in aviation maintenance for reporting failures, may result in incorrect maintenance and make it very hard to analyse the maintenance data. There are several reasons for this incomplete reporting. Firstly, (part of) the information is often unknown at the moment the maintenance crew fills in the card. Also, the findings on repair cards are generally filled in as free-form text, making it difficult to automatically interpret the findings. An automatically assessed failure description will lead to more complete and consistent repair cards. This will also improve the efficiency of troubleshooting since this failure diagnosis can add information which would otherwise not be at the disposal of the maintenance crew at that time. This research utilises a data driven approach combining maintenance and usage data. The model will be based on Artificial Intelligence (AI) such that it is no longer necessary to completely understand the physics of a (sub)system or component. A newly proposed XAI (eXplainable AI) methodology, Failure Diagnosis Explainability (FDE), will be added to the model to provide transparency and interpretability of the assessed diagnosis. The assessed diagnosis is explained by checking whether a new failure matches the expected values of a certain diagnosis (class). On the other hand, when a failure is dissimilar to the expected values of a certain diagnosis (class), it is unlikely to be the actual diagnosis. The different steps towards this failure diagnosing model are applied to a case study with a main wheel of the RNLAF (Royal Netherlands Air Force) F-16. This feasibility study already showed the value of this automated failure diagnosis model with an achieved accuracy of 81% of classifying a diagnosis . The proposed XAI methodology was able to explain the diagnosis assessed by the failure diagnosis model.

1. INTRODUCTION

A repair card is used in aviation maintenance to report a failure or anomaly and register it in the maintenance management system. An incorrect or incomplete repair card may result in incorrect maintenance and make it very hard to analyse the maintenance data. An example from practise is a helicopter Main Gear Box (MGB) removal due to a leakage found during a 500 hours inspection. The maintenance crew described the complaint as 'defect, 500 hrs'. The component shop carried out an overhaul when a small repair could also have solved the problem. The overhaul, however, was unforeseen and there was no spare MGB available which resulted in a grounded helicopter. So the consequences of this incomplete repair card were additional maintenance costs and a decrease in availability.

This example is not unique in the aviation sector, it is rather common and there are several reasons behind this. Firstly, (part of) the information is often unknown at the moment the maintenance crew fills in the repair card. Also, the findings on repair cards are generally filled in as free-form text. As a result, repair cards may contain incorrect information and can be incomplete. A consequence of the incorrect and incomplete repair cards is that they are not practical for data analysis because it is difficult to automatically interpret the findings.

Therefore a model will be generated which can automati-

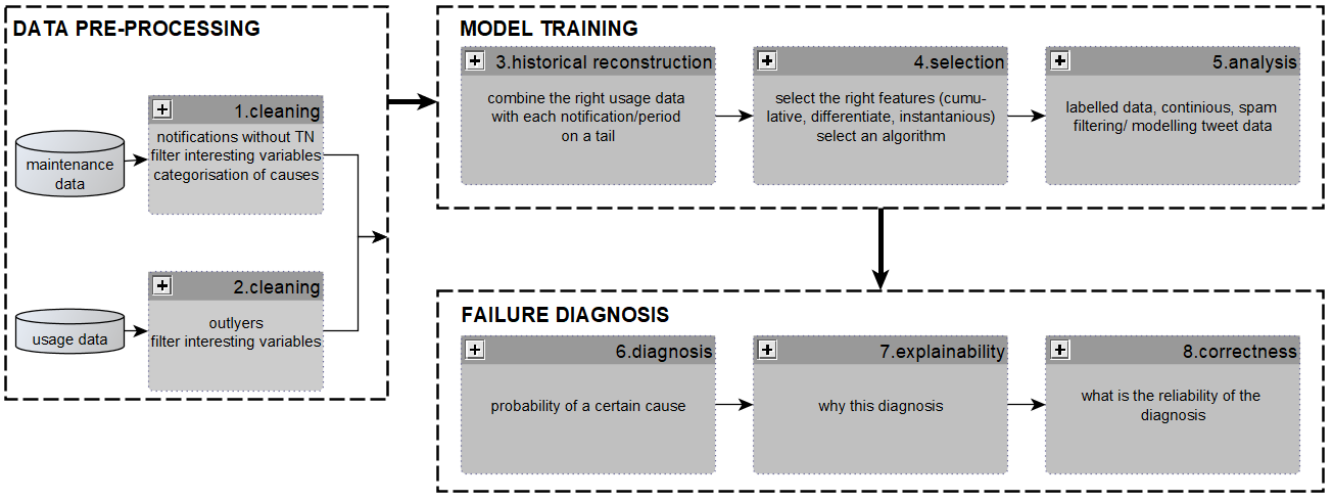

Figure 1. Functional diagram of an automated failure diagnosis model

cally assess a failure diagnosis based on usage data. An automatically assessed failure description will lead to more complete and consistent repair cards. This will improve the efficiency of troubleshooting since this failure diagnosis can add information which would otherwise not be at the disposal of the maintenance crew at that time. Conventional ways to link failures to the usage are physical models, but this research utilises a data driven approach combining maintenance and usage data, and Artificial Intelligence (AI) into a failure diagnosing model. AI is capable of recognising more patterns and relations than humans can. With this model, it is no longer necessary to completely understand the physics of a (sub)system or component. This data driven approach makes it difficult to establish causal relations between features. To convince the users of the model, a plausible explanation is needed to understand the cause of the failure. XAI (eXplainable AI) techniques will be implemented in the model to provide transparency and interpretability of the resulting diagnosis.

Figure 1 proposes the steps to achieve an automatic failure diagnosis based on usage data (sensor information from the flights) with XAI. This methodology is newly proposed in this research and will be discussed in the remainder of this paper. Section 2 describes which data have to be collected and in which form, to develop an automated failure diagnosing model (blocks 1,2). In Section 3, the training of the model is described (blocks 3,4,5). A historical reconstruction will be made and features and algorithms will be selected. The failure diagnosis will be discussed in Section 4 (blocks 6,7,8). The diagnosis will consist of an assessment of the possible causes, the explainability of this assessment and the correctness of the diagnosis. Finally, all these steps will be demonstrated in a case study in Section 5 were the feasibility of the model will be tested.

2. DATA PRE-PROCESSING

Before the model can be trained, the data need to be pre-processed so that it can be implemented in a ML (Machine Learning) algorithm. This raises the question which data, in which form and how much is required for proper training. The model will be based on historic maintenance and usage data which is sensor data from historic flights. In this research, the maintenance data will be used to label the failures. The usage data will be added to incorporate the historic usage of a component in the assessment of the diagnosis.

This section will discuss the algorithm prerequisites, the required variables and types of data. The requirements will differ per system and per component. This section provides some guidelines. Later, in the case study, a specific example will be elaborated on.

2.1. Algorithm prerequisites

There are no specific requirements for the dataset that is applied to ML algorithms. Actually Hua et al. (2005) mention that one should be wary of rules-of-thumb generalised from specific cases. The optimum sample size will differ per situation. But there are some general guidelines for the data:

- The optimal amount of features relative to the sample size depends on the algorithm and the feature-output distribution. Hua et al. (2005) show the relation between the number of samples, number of features and the error rate for various algorithms.
- The performance of an algorithm can be greatly influenced by the number of features. Therefore the amount of features used should be close to the optimal amount (Hua et al., 2005).
- A balanced labelled dataset (each class has about the same amount of samples) is preferable to avoid overfit-

ting on the class which contains more examples (Jain & Chandrasekaran, 1982).

Since the guidelines are not directional, the data have to be processed by reasoning in which format the data can be implemented in the ML algorithms. Engineering judgement is needed for the amount of features and data to be used.

2.2. Variable selection procedure

This research uses two data sources, namely usage data and maintenance data. The usage data comes from sensors in the airplane and consist of continuous, numeric values. This dataset contains information such as altitude, velocity, acceleration and outside temperature. The maintenance dataset consists mainly of text and categorised variables. This dataset is filled by the maintenance crew and contains information about the failure of a component. Besides the selection of variables from both datasets, the data probably will need cleaning. There are two types of errors in the datasets, namely measurement faults (e.g. an altitude of -100 m) and incompleteness (e.g. when a maintenance notification does not have a tail number (identification number) such that it cannot be linked to the usage data of an airplane).

Maintenance data entails information about the date of installation/removal of a part on a tail, preventive and corrective maintenance and the reason of a notification. The data have to be labelled before training. In order to limit the amount of labels, the failures with similar failure descriptions (and the same failure cause) will be grouped. E.g. 'removal due to being worn out' and 'component x is worn' should be in the same group. This grouping is also done because ML algorithms always require multiple events per group in order to train the model.

Variables are used to describe the usage. More variables can result in a more accurate description, but not all variables contain relevant information on the usage of the system. The usage variables will be different for each system, in each situation. In the case of a new system that has not been built yet, there is more freedom of selecting variables than in the case of an existing system. In general, to select the usage variables, it should be checked whether there is a relation between that variable and the loads on the system or the performance of the system (Tinga, 2010).

Also the features, i.e. the specific parts or details retrieved from a variable or signal, have to be determined. Is data required from the entire flight or only from take-off and landing or only during the flight? From all flights or only from the last one or last 10? This need will depend on the type of failure mechanism i.e. for corrosion, fatigue and wear all flights are required but an overload can probably be seen in (one of) the last flights. For the landing gear failures only the take-off and landings are interesting but for structural failures the entire flight (from take-off to landing) is needed.

2.3. Training and input data

A distinction is made between training data and input data. When this model is trained, data can be entered in the model and the model provides a diagnosis. Subsequently, the maintenance crew validates the diagnosis. The input data are now labelled and can be used to enrich (i.e. train) the model.

Since the usage data will be combined with the maintenance data (repairs, removals etc.), usage data of a certain period can be both training and input data. The usage data of the last year for a specific tail will be used as input data, e.g. when a failed component has been on a tail for one year. If another component, on the same tail failed four months earlier, then the usage data for the months before this failure can be used for training the model for that specific component.

3. MACHINE LEARNING MODEL

After pre-processing the data, the data from the different sources have to be combined and the model has to be trained for failure diagnosis. This section will discuss the data requirements for the ML algorithms and the procedure to select the variables from both maintenance data and usage data. Finally, the destinction between training data and input data will be discussed.

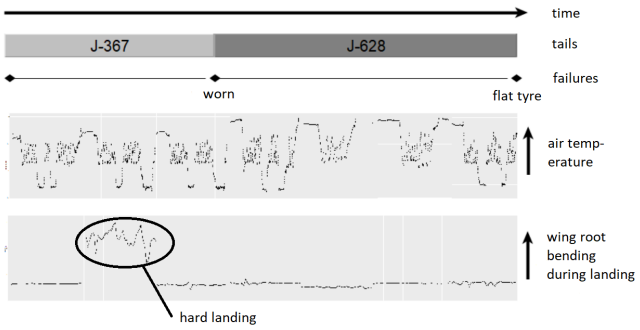

Figure 2. Reconstruction of a historical timeline for one component, installed on two different tails during a certain period of its operational life

3.1. Historical reconstruction

As can be seen in the functional diagram in Figure 1, the maintenance and usage data will be combined to come to a historical reconstruction where both sources are aligned to one timeline as is shown in Figure 2.

Most air forces and airlines exchange components between airplanes (tails), which is not common in every industry (e.g. process industry). Making a historical timeline of a component may show that a component has been installed on different tails. The usage data, in this case, need to be collected from different tails. Visualisation of the historical timeline is very important. This will give the maintenance

crew a quick overview of the history of a component which helps to judge the final diagnosis assessed by the model. Figure 2 shows the air temperature and wing root bending during landing to which the airplane was exposed during the lifetime of a specific component. As can be seen, there was a high wing root bending during one of the landings in the beginning of the wheel's life which probably indicates a hard landing.

3.2. Feature selection

All ML algorithms can only handle single values per case, therefore the usage of the component has to be expressed in features. Features are individual measurable properties or characteristics of a phenomenon being observed (Bishop, 2006). The features will be selected depending on the failure mechanism and the ML algorithm. E.g. when the velocity during take-off is one of the selected variables it is likely that the amount of data points is not equal for every landing and the data can therefore not be implemented directly in a ML algorithm. To get data strings of the same size, several options are possible: choose the lowest amount of data points and delete the superfluous points for the other take-offs; select the largest amount of data points and extra- or interpolate between data points for the other take-offs; or determine characteristics such as average, standard deviation and kurtosis for each take-off. Neural networks prefer to work with the whole data string (extra- or interpolation), but more simple algorithms can utilise characteristics.

3.3. Algorithms

During the training process, the dataset contains supervised (labelled) data, i.e. each notification is classified in a failure diagnosis category. Diagnosing a new failure will be a supervised multi-classification problem, for which there are some suitable algorithms (Wang & Xue, 2014; Kotsiantis, 2007; Wu et al., 2008):

- Support Vector Machines (SVM)
- k-Nearest Neighbour (kNN)
- Naive Bayes
- Random forest
- Neural networks

The above mentioned algorithms are all suitable for the failure diagnosing model and will be compared after applying them individually to the data from the case study.

4. FAILURE DIAGNOSIS

The failure diagnosing model (from Figure 1) is trained with the chosen algorithm and pre-processed data. When a test example (or a new failure) is given to the model as input, it comes up with a probable failure diagnosis. This section will discuss how this diagnosis can be presented to the maintenance crew and the possible ways to explain the diagnosis to help the maintenance crew to verify the reliability of the diagnosis. Finally, the correctness of the model will be discussed.

4.1. Diagnosis

Visualisation is a strong method to transfer data from the model to the user (here the maintenance crew). Therefore the output of the model will be visualised. Commonly, probabilities are shown as simple function plots, with either probability versus data value or value versus cumulative probability. The ubiquity of these representations make them easy to read and interpret, even if the user is unfamiliar with the subject (Potter et al., 2012). This is why the assessment results will be presented in a bar graph as shown in Figure 3.

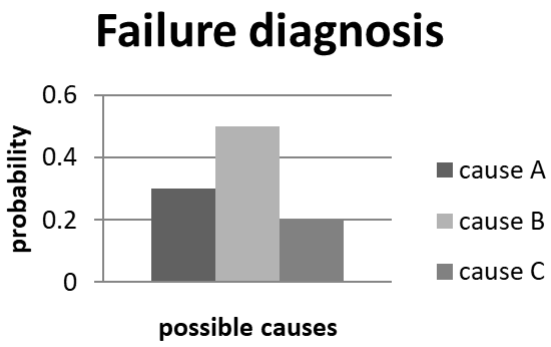

Figure 3. Example of a bar graph showing the outcome of the failure diagnosis model

4.2. Explainability

ML algorithms are often quite reliable, but sometimes not very explainable. An example is cancer diagnosis. Some algorithms are able to predict whether the patient has cancer or not, more accurately than doctors. But as long as humans do not understand how this algorithms made the assessment, it will not be used in practise since they do not trust them (Holzinger et al., 2017). Trust can be based on deterrence, on knowledge or on identification according to Lewicki & Bunker (1995). This research will strive to achieve trust based on knowledge. In order to give the maintenance crew this knowledge, it is very important to make the diagnosis assessment explainable.

4.2.1. XAI methodologies proposed by DARPA

DARPA (Defense Advanced Research Projects Agency), the agency of the United States Department of Defense (Gunning & et al., 2016), is currently researching the explainability of various AI applications (Gunning & et al., 2016). Depending

on the chosen algorithm, some options are suggested which will be discussed below. To refer these options to a failure diagnosis, an example of a tyre failure of an airplane will be used. The three possible diagnoses in this example are wear, impact (overload) and other (for the remaining causes).

Local Interpretable Model-agnostic Explanations (LIME) Ribeiro et al. (2016) state that 'LIME is an algorithm that can explain the assessments of any classifier or regressor in a faithful way'. One of the examples in this paper is the explanation of an image classification assessment. The image is a person, with the head of a dog, playing an acoustic guitar. The top three classes predicted were electric guitar, acoustic guitar and labrador. The model also shows which part of the picture is used for each assessment. Although this research is focussing on the explanation of image recognition, it can be applied to airplane component failure diagnosis as well. The original image will be a new failure and in the example of the tyre a possible top three of predicted classes is low profile, high amount of landings/take-offs and moderate or low landing/take-off velocities. This can help the maintenance crew to validate the diagnosis of wear. The model will not perform the diagnosis, only the explainability. The classes are pre-defined but the importance of each class can vary for each failure.

Explanatory text In the paper of Hendricks et al. (2016) explanatory text sentences are used to justify an assessment. The research is also focussed on explaining deep visual models. One example is a picture of a bird. The model gives: 'This is a yellow breasted chat because this is a bird with a yellow breast and a grey head and back'. For the tyre example the explanation could be, this tyre is worn because it has lasted for over 200 landings/take-offs and now has a profile below x mm.

Bayesian teaching Bayesian teaching is the optimal selection of examples for machine explanation. So examples of the training data will be selected to explain the assessment of the model. In Gunning & et al. (2017) a picture of a child is used as input for the model. The output is that the face is angry because it is similar to certain examples (examples of kids with angry faces are given) and dissimilar to other examples (examples of kids with sad and happy faces are given). For the failed tyre, training notifications will be selected with similar feature values. E.g. the tyre had an impact because the amount of landings is (more or less) equal to the ones from these failures (followed by showing these similar notifications).

Neural networks Neural networks are used for classification but if the decisions towards this classification can be made visible it can also be used as an explanation for a diagnosis. A neural network contains an input layer, one or more hidden layers and an output layer. The input layer is fed with training data and the unlabelled input and the output gives the assessment. But the hidden layers in between are seen as a black box. In Gunning & et al. (2017) a neural network is shown which predicts the type of animal. The training data contains images of all kinds of animals, the input is an image of a dog. They explain that the first layer neurons respond to simple shapes, the second layer neurons to more complex structures (a tail, paws, head). This will further increase until the n^{th} layer were the neurons respond to highly complex, abstract concepts. The output of the model was 10% wolf and 90% dog. For the example of the tyre, the simple shapes will probably be rough and clear separations between data (tail numbers, number of landings/take-offs) more complex layers will include values of velocity, variation in velocity etc.

Decision trees Decision trees are very powerful, but also simple and efficient for extracting knowledge from data. They are easy to interpret, understand and control by humans (Ertel, 2009). In Figure 4 an example is given for the tyre failure. The decision tree determines the nodes (number of landings/take-offs, surface of landing strip) and the limit values (15 and smooth/rough) by computing the information gain. For each layer in the tree, the feature with the highest information gain will become a node.

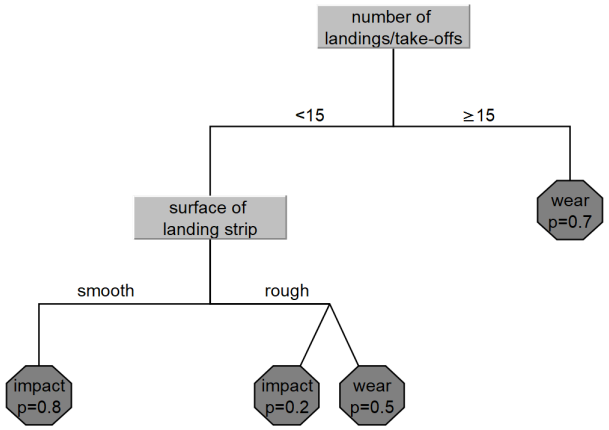

Figure 4. Simplified explanation of a tyre failure diagnosis by a decision tree

4.2.2. Proposed XAI methodology: Failure Diagnosis Explainability (FDE)

As discussed in Section 4.2.1, there are various methodologies to make a ML decision more explainable for the user (in this case the maintenance crew). As these methodologies are still in development and no detailed information has been available yet, a new XAI methodology will be proposed here: Failure Diagnosis Explainability (FDE). This methodology is based on the methods used for the different options suggested by DARPA. A common method to describe on what grounds the decision was based is with characteristics. In the example of the guitar playing dog, parts of the image were highlighted which were characteristic for a specific decision. For the

bird, the characteristics were described in text form and for the example of the child, pictures with similar and dissimilar characteristics were shown. The dog from the neural network was explained by its characteristic shapes and finally in a similar manner, the decision tree from Figure 4 explains tyre failures based on usage characteristics.

The failures are described by different features (as discussed in Section 3.2) so these will be used to explain the failure diagnosis assessed by the model. Since many features are implemented in the model, one can wonder which ones are most important: which features characterise a certain diagnosis the most. E.g. a certain value for feature f is very typical for a worn tyre. The variable importance of a model gives the (relative) importance of each variable (feature) for a trained model. But since this model includes various diagnoses (classes), the importance of each feature per class is not determined: the variable importance shows which feature is most important in deciding which class a failure belongs to, which is not equal to the characterisation of a certain class. To achieve this, new models will be trained in which, in each model, one class is appointed as 1 and the others as 0 (binary). E.g. class A is 1 and classes B and C are 0. Still, this does not precisely determine the variable importance per class, but it does yield the variable importance of a specific class relative to the other classes. A top five of most important features is determined per class (diagnosis) which can be found in Figure 6 in the case study. Before the model is trained, all values will be normalised using feature scaling. Note, the variable importance can only be determined for so called white box models as decision tree, random forest and Naive Bayes.

Finally, this methodology will show the user how much (and on which points) a new failure match with each diagnosis. The more the features matches the expected values of a specific diagnosis, the more likely it is that this failure belongs to that class. Also, if the features are very dissimilar to the values of a certain class, it is very unlikely that this failure belongs to that specific class. So, the reasoning of this methodology is similar to the Bayesian teaching methodology from Section 4.2.1. A failure is from class (diagnosis) A, because the features have similar values as class A. Contrary, this failure is not from classes (diagnosis) B and C, because the features have dissimilar values as class B and C (Gunning & et al., 2017).

The expectations for each feature are expressed by boxplots. If the value of a new failure falls within the 50% range (recognised by the box) of the boxplot, there is a good match. If the value falls within the 95% range (recognised by the whiskers of the plot), the values still matches but less compared to the 50% range. In case the value falls outside the 95% range, it is unlikely the failure belongs to that class, based on a 95% confidence interval; 2 sigma limit. The boxplots of the top five features of each class were represented in one graph. The values of a new failure will be added to the graph to give the maintenance crew an indication of which variables match with the expectations of a certain diagnosis and which do not. A fit is made through these value points. If this line is equal to the x-axis, the new failure corresponds completely with the expected values. An example of such a plot can be found in Figures 7 and 8 in the case study.

For a better visual presentation, the medians of all boxplots were aligned with the same x-axis. Also, all values are, per variable, multiplied with their scaled importance $w(f_j)$. This means that a deviation on the most important variable will be enlarged compared to a deviation of the fifth most important variable. Equation 1 shows the computation of all absolute explanation values $S_{i,j}$ (both for the boxplot and for a new failure), where the fraction normalises the value $f_{i,j}$ where f is the sensor data from feature j and failure i. Following the median will be substracted and finally, multiplied with the scaled importance. This methodology will be shown and tested in the case study in Section 5.3.2.

$$S_{i,j} = \Big(\frac{f_{i,j} - \min(f_j)}{\max(f_j) - \min(f_j)} - \text{median}(f_j)\Big) * w(f_j) \quad (1)$$

4.3. Correctness

The number of correctly classified examples is a performance measure for the diagnosing model. To measure this, part of the data is randomly separated before training and assigned as the test data. If the training data would be used for performance measurements, overfitting of the model cannot be noticed. After the initial training, when the model is in use, it will continuously enrich itself after every new example. The maintenance crew validates the diagnosis before it is added to the training data, hence the model will learn by updating itself. This process is also called incremental learning (Ertel, 2009). Incremental learning will improve the correctness of the model, unless only failures from the same class are added to the training data. In which case the dataset can become very unbalanced.

The correctness will be expressed in the accuracy of the algorithm, which means the percentage of correctly classified failures. A common practice within ML is to compare the achieved accuracy with a certain baseline. One common baseline for classification problems is the 'most frequent' which always classifies the most frequent label in the data set. When the accuracy of a ML algorithm is below this baseline, the algorithm will not be of any value.

5. CASE STUDY

In this section, the techniques proposed in this paper are applied to a case study to test the feasibility of these techniques. For the case study, the data from RNLAF (Royal Netherlands

Air Force) F-16s (fighter airplane) main wheels are used. The F-16 is equipped with a nose landing gear and a main landing gear. The nose landing gear consists of one nose wheel, the main landing gear of two main wheels. The RNLAF has a total of 68 F-16s using these main wheels. The tyres are removed for corrective maintenance or in case of other maintenance e.g. the wheel will be removed for when say a shock strut replacement is conducted. After the replacement of the shock strut, the original wheel will be installed again on the same tail. Common failures include the tyre being worn out, a flat tyre and a replacement due to contamination (mainly oil). There is no preventive maintenance on the wheels. The RNLAF does not retread their wheels, so a worn tyre will be directly disposed of. Flat tyres will be repaired if possible and contaminated tyres will be cleaned. After repair, the wheels will be mounted on an arbitrary tail. Most likely a different one, but occasionally it could be the same. Since the data from the used case is not public, the steps in data processing will be elaborated to provide more transparency.

5.1. Data pre-processing

The data has to be pre-processed before the model can be trained. The maintenance and usage data are prepared separately prior to being combined. For the pre-processing, the steps described in Section 2 will be followed. The main wheel is chosen for the case study since this component had one of the most removals and has few failure modes which are easy to understand.

5.1.1. Maintenance data

The maintenance data is obtained from the Computerised Maintenance Management System (CMMS) SAP (Systems, Applications and Products). A repair card is filled in manually by the maintenance crew. The data is extracted from SAPs database and compiled to a CSV (Comma Separated Values) file. Thereafter the data is filtered by tail and by number of notifications per SN (Serial Number). Failure registrations without a specified tail are deleted since the tail is needed to link a failure to the usage data. If there is only one notification of a specific SN, the component is still in use (only the installation is reported, while no removal is reported). The installations do not mark a failure, therefore only the removals are kept. For this case study it is assumed that all removals are failures since cannibalisation (i.e. removal of sound components to be used on another tail) of wheels is not common for these F-16s. Also, for the case study, only one removal is eliminated due to these filters.

The data is labelled in three categories; ‘flat’, ‘worn’ and ‘other and unclassifiable’ (when the cause is unknown, different from the other two or the failure did not have enough information to be labelled). For labelling, several data fields were checked for words as ‘worn’, ‘due’, ‘flat’, ‘deflating’ etc. This labelling is done automatically, but it is based on words which are found manually by going through the failure descriptions. Text mining may be an option to fully automate this process. There is no contamination class since these failures could not be traced from the repair cards or were not present in the data set. The labelled notifications (242) are combined with a dataset containing information about the tail and installation and removal date of each unique component. While combing the maintenance data with the usage data, only 144 removals are kept (elaborated in Section 5.2). So the amount of data is big (notifications and usage data together), but there are only a few failure cases which can also be combined with usage data. For data mining, 144 notifications is generally considered as a relatively small number, but in industries such as the aviation sector and in particular in the military sector, these numbers are rather common. A better registration of failures can easily lead to more notifications. Also, the class ‘other and unclassifiable’ is now considerably high (82 of the 144 failures), this is mainly due to many failures which did not have enough information to classify the notification. Preferably, this class is the smallest of the three, but this is a limitation of the data from current practise.

5.1.2. Usage data

Sensor data from almost 7,000 flights are used. The data is extracted via a spool file, as it is a relational database, requiring a considerable amount of computing effort. The variables, which are appointed as related to the loads on the wheels, are: CAS (Calibrated Air Speed), strain from several strain gauges, weights, longitudinal and lateral accelerations, pitch and air temperature. There are more variables available but these are (for example) related to the engines or electronic components. Also, the phase of the flight (take-off, flight, landing) is assigned to each point. The variable selection is done by logical reasoning. In case of doubt, the variable was selected. This selection is done to avoid that the model will fit on unrealistic relations and to gain more causality in the model. E.g. the model could find a strong relation between maximum altitude and flat tyres but it later turns out that there were nails on the landing strip where flights with high altitude were trained. These type of unrealistic correlations can be prevented by the manual variable selection. The drawback could be, however, that a real but unexpected correlation is not revealed.

5.2. Machine Learning model

The ML algorithms require a dataset in which each row is a notification and the columns are the different features. After the data pre-processing, the maintenance and usage will be combined to put the data in the right format. Every notification should be combined with the usage related to this notification. No usage data was available for 98 of the 242 failures, leaving 144 failures for training and testing. The

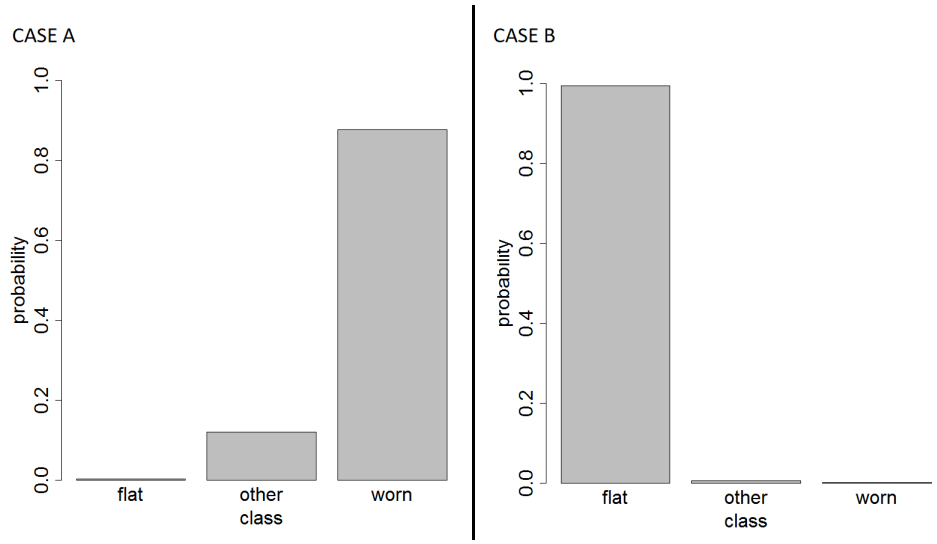

Figure 5. Probability of each class plotted for two specific cases, A and B

unavailability of the usage data can have various reasons e.g. the recorder was already full or the data are not added to the database due to the mission location of the airplane.

A notification can only include single values and not a (varying) range of data points. Therefore the variables of the usage data were translated to characteristics (as discussed in Section 3.2) with single values such as number of flights, average speed etc. The selected features are:

- number of flights
- average CAS during 1) take-off and 2) landing
- average and maximum strain gauge (called FS325) measuring the wing root bending during 1) take-off and 2) landing
- average and maximum strain gauge (FS374) measuring the fuselage bending during 1) take-off and 2) landing
- average and maximum strain gauge (BL120) measuring the wing tip bending during 1) take-off and 2) landing
- average lateral acceleration during 1) take-off and 2) landing
- average longitudinal acceleration during 1) take-off and 2) landing
- average total weight during 1) take-off and 2) landing
- average air temperature during 1) take-off and 2) landing
- the amount of calendar days between failures (i.e. the component age)

All these characteristics are calculated over the period of time since the first installation of the considered component. Now that the datasets are combined, a historical reconstruction can be made. This example has already been shown in Figure 2.

Since the algorithms can only handle numerical and categorical values, all text fields were deleted. Since these fields were already used for the labelling of the notifications, these data were already taken into account. Several algorithms (from Section 3.3) were applied to the training data. Table 1 shows the accuracies of the different algorithms. The first algorithm, Naive Bayes, probably performs moderately since this algorithm is simply too weak or too simple to detect the patterns of this data set. The neural network and SVM are probably overfitting on the training set. Overfitting can be seen when the accuracy on the test set is remarkably lower than on the training set. Since random forest gave the highest accuracy, this algorithm is used for the remaining steps.

Table 1. Accuracies obtained from various ML algorithms

Algorithm	Accuracy
Naive Bayes	62%
Neural network	43%
Support Vector Machine	69%
Random forest	81%

5.3. Failure diagnosis

The model is now trained with the training set. After that, one failure from the test set will be fed into the model to assess a diagnosis for this failure. This single failure from the test set simulates a new failure. The results (diagnosis) of two new failures and how they should be interpreted (explained) are discussed in this subsection.

5.3.1. Diagnosis

As discussed in Section 4.1 the assessed diagnosis is represented in a bar graph. Figure 5 shows the results for two specific failures which are most likely a 'worn' tyre (case A) and a 'flat' tyre (case B). This figure only shows the

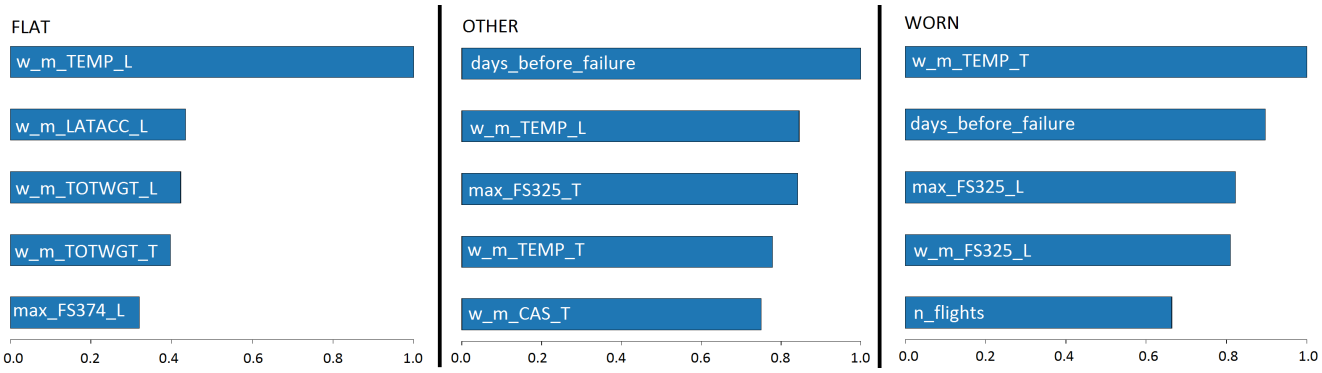

Figure 6. Variable importance plot for the different features

probability of each diagnosis yet does not explain the user why each diagnosis has a certain probability.

5.3.2. Failure Diagnosis Explainability (FDE)

As discussed in Section 4.2.2, the variable importances per class relative to the others had to be determined. The top five of scaled importances per class are shown in Figure 6. As can be seen, the weighted mean air temperature during landing is the most important feature to determine whether the tyre is 'flat' or not and the maximum fuselage bending during landing is the fifth.

Figure 7 and Figure 8 present the results of a 'worn' (case A) and a 'flat' tyre (case B) (as shown in Figure 5). As can be seen in Figure 7, the values of this new failure (case A) fall within the 50% range for four of the five variables for the 'flat' diagnosis. The first variable is on the edge of the 95% interval. For the diagnosis 'other defect', three values fall within the 50% range. Therefore it is unlikely that the new failure belongs to the class 'flat' or 'other and unclassifiable'. For the diagnosis of a 'worn' tyre, the new failure falls within all 50% ranges and two variables even match the median (the third and the fourth variable). So it is likely that this new failure is 'worn' because all values correspond with the expectations. Even the fourth variable (the weighted average wing root bending during landing), which has a very tight range, matches.

In Figure 8, the values of case B completely correspond with the expectations of a 'flat' tyre failure. This makes it very likely, the assessed class will be 'flat'. On the contrary, the new failure does not correspond with most of the variables from the classes 'other and unclassifiable' and 'worn'.

in conclusion both, from case A and B, assessed diagnoses (shown in Figure 5) can be explained with the proposed XAI methodology FDE.

5.3.3. Correctness

After training the model, it achieved an accuracy of 81% on the test set which means that the notifications from the test set were classified to the correct diagnosis ('flat','other and unclassifiable' or 'worn') in 81% of the cases. As discussed in Section 4.3 this result will be compared to the 'most frequent' baseline, which is the percentage of occurrence of the biggest class. In this case the class 'other and unclassifiable' is the biggest class. 28 of the in total 42 cases from the test set are labelled as 'other and unclassifiable'. If the model would (without any knowledge or training) assess every case as 'other and unclassifiable', it has an accuracy of 67%. The model showed its value already with the first preliminary test within this feasibility study, with a relative improvement of 21% compared to the baseline. For further research, the features can be changed, the data can be more balanced (optimal is equal amount per class) and more data can be used (also with incremental learning). These changes may lead to improvement of the model. Other features which could be taken into consideration are the number of exceedances of certain values and the variance or distribution of the data. By using other labelling techniques, consulting more data sources or filling in the repair cards more complete will decrease the number of unclassifiable diagnoses.

6. CONCLUSION

The combination of maintenance and usage data with (X)AI showed its value. An accuracy improvement of 21% (to a total of 81%) for classifying a diagnosis compared to the baseline is achieved already with this feasibility study. Even though the model was trained with a small amount of failures, which is very common in this industry. The usefulness of automatically diagnosing a failed tyre based on usage data can be questioned. After all, given that the tyre is failed, the distinction between a flat and worn tyre is easy to make. However, this case study served as a feasibility study. The tyres were one of the components with most data available and the failures were easy to understand which made validation

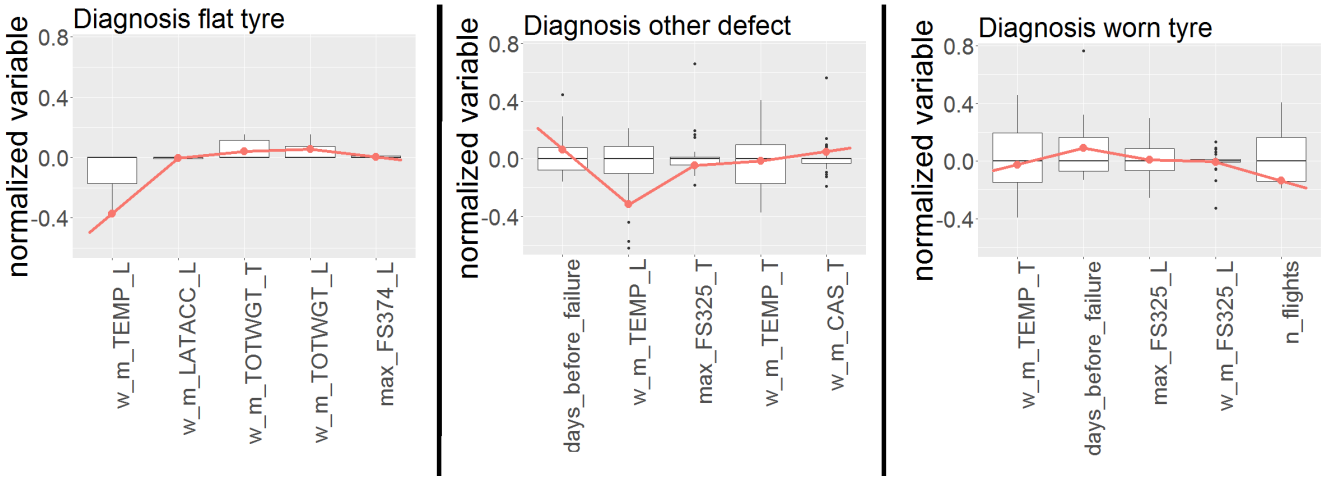

Figure 7. Boxplots explaining a specific case of a 'worn' tyre (case A)

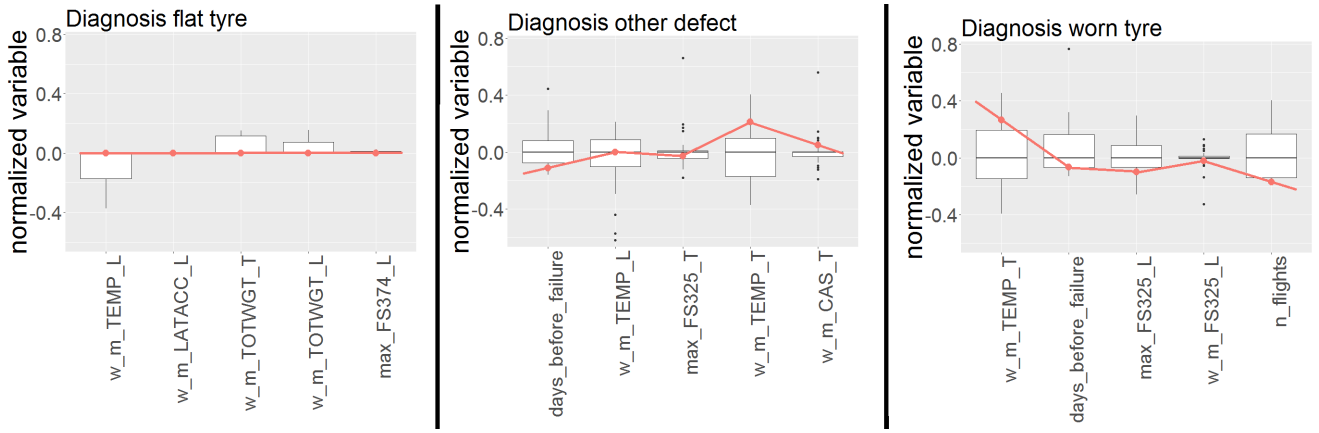

Figure 8. Boxplots explaining a specific case of a 'flat' tyre (case B)

of the model possible. Applying this model to components which are hard to diagnose (e.g. when the component is physically hard to reach) can really improve the efficiency of troubleshooting. With this research the requirements for the data and for such an analysis have become clear.

7. RECOMMENDATIONS

With the implementation of the case study, many difficulties arose. Most of them concerned the preparation of the data in such a way that the data could be implemented in the ML algorithm. Since many of these difficulties will be common for data-driven approaches in this industry, some general conclusions can be drawn. First of all, when the data are stored in a database, it does not mean they are also available for the preferred analysis. In case it is not available it should be made clear which steps need to be taken. The sensor data from the case study were stored in a relative database. Therefore it took significant effort to extract the data from the database. Secondly, the maintenance and usage data have to be combined. To do this, there must be a possibility to link these datasets with each other (by tail, by date, by pilot, by airport or anything else). Thirdly, the data should be prepared for the algorithm (cleaning, labelling, inventing features etc.). This process can take a lot of effort and logical reasoning by selecting the features can add causality to the model. The way the data is recorded and stored will partly have influence on this effort. E.g. for cleaning and labelling a certain way of recording and storing can help, but it does not have influence on the effort to select the features.

REFERENCES

Bishop, C. (2006). *Pattern recognition and machine learning*. Springer.

Ertel, W. (2009). *Introduction to artificial intelligence*. Springer. (Section 8.4, 9.4)

Gunning, D., & et al. (2016). *Explainable artificial intelligence XAI* (Tech. Rep. No. DARPA-BAA-16-53). Defense

Advanced Research Projects Agency (DARPA).

Gunning, D., & et al. (2017). *Explainable artificial intelligence XAI '- program update* (Tech. Rep. No. DARPA-BAA-16-53). Defense Advanced Research Projects Agency (DARPA). (slide 4, 15)

Hendricks, L., Akata, Z., Rohrbach, M., & et al. (2016). Generating visual explanations. In *European conference on computer vision.*

Holzinger, A., Biemann, C., Pattichis, C., & Kell, D. (2017). *What do we need to build explainable AI systems for the medical domain?* (Tech. Rep.).

Hua, J., Xiong, Z., Lowey, J., Suh, E., & Dougherty, E. (2005). Optimal number of features as a function of sample size for various classification rules. *Bioinformatics*, *21*(8), 1509-1515.

Jain, A., & Chandrasekaran, B. (1982). Dimensionality and sample size considerations in pattern recognition practice. *Handbook of Statistics*, *2*, 835-855.

Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques. *Frontiers in Artificial Intelligence and Applications*, *160*, 3-24.

Lewicki, R., & Bunker, B. (1995). *Trust in relationships* (Tech. Rep. No. RS 95-7). Business Research College of Business.

Potter, K., Kirby, R., Xiu, D., & Johnson, C. (2012). Interactive visualization of probability and cumulative density functions. *International Journal for Uncertainty Quantification*, *4*(2), 397-412.

Ribeiro, M., Singh, S., & Guestrin, C. (2016). why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.*

Tinga, T. (2010). Application of physical failure models to enable usage and load based maintenance. *Reliability Engineeringand System Safety*(95), 1061-1075.

Wang, Z., & Xue, X. (2014). *Support vector machines applications*. Springer. (Chapter 2)

Wu, X., Kumar, V., Quinlan, J., Ghosh, J., & et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1-37.

BIOGRAPHIES

Sophie ten Zeldam obtained her Bachelor degree Military Systems & Technology, research group Maintenance Technology, at the Netherlands Defence Academy in 2016. Currently she is working on a Master degree Mechanical Engineering, specialization in Maintenance Engineering & Operations, at the University of Twente. This research is part of her Master degree's final assignment which she will finish in July 2018.

Arjan de Jong received his BEng in Aeronautical Engineering from Haarlem Polytechnic and his Master Degree in Aviation Management from the University of Southampton. He worked for the Schreiner Aviation Group in the Netherlands and the CHC Helicopter Corporation in Vancouver, BC, Canada. He held various operational and strategic positions in maintenance, engineering and asset management. Arjan received his PhD in Aerospace Engineering & Management from the Technical University Delft. He now works at the Netherlands Aerospace Centre - NLR as Maintenance Engineering, Management & Technology team leader.

Richard Loendersloot obtained his Master degree in Mechanical Engineering, research group Applied Mechanics, at the University of Twente in 2001. He continued as a PhD student for the Production Technology group of the University of Twente, researching the flow processes of resin through textile reinforcement during the thermoset composite production process Resin Transfer Moulding. He obtained his PhD degree in 2006, after which he worked in an engineering office on high end FE simulations of a variety mechanical problems. In 2008 he returned to the University of Twente as part time assistant professor for Applied Mechanics. From September 2009 on he holds a fulltime position. Since then, his research started to focus on vibration based structural health and condition monitoring, being addressed in both research and education. He became part of the research chair Dynamics Based Maintenance upon its initiation in 2012. His research covers a broad range of applications: from rail infra structure monitoring, to water mains condition inspection and aerospace health monitoring applications, using both structural dynamics and ultrasound methods. He is involved in a number of European and National funded research projects.

Tiedo Tinga received a Master degree in Applied Physics (1995) and a PDEng degree in Materials Technology (1998) at the University of Groningen. He did his PhD research during his work with the National Aerospace Laboratory NLR on the development of a multi-scale gas turbine material model. He received his PhD degree in 2009 from Eindhoven University of Technology. In 2007 he was appointed associate professor at the Netherlands Defence Academy and in 2016 became a full professor Life Cycle Management. In 2012 he also became a part-time full professor in dynamics based maintenance at the University of Twente. He now leads a research program on predictive maintenance and life cycle management in both institutes.