ISA Transactions, Available online 11 May 2020
DOI: 10.1016/j.isatra.2020.05.001

An Integrated Machine Learning Model for Aircraft Components Rare Failure Prognostics with Log-based Dataset

Maren David Dangut*, Zakwan Skaf,**
Ian K. Jennions***

, **, * Integrated Vehicle Health Management (IVHM) Center*
Cranfield University Bedford, United Kingdom , MK430Al
*(*maren.dangut@cranfield.ac.uk, **z.skaf@cranfield.ac.uk, ***i.jennions@cranfield.ac.uk)*

Abstract

Predictive maintenance is increasingly advancing into the aerospace industry, and it comes with diverse prognostic health management solutions. This type of maintenance can unlock several benefits for aerospace organizations. Such as preventing unexpected equipment downtime and improving service quality. In developing data-driven predictive modeling, one of the challenges that cause model performance degradation is the data-imbalanced distribution. The extreme data imbalanced problem arises when the distribution of the classes present in the datasets is not uniform. Such that the total number of instances in a class far outnumber those of the other classes. Extremely skew data distribution can lead to irregular patterns and trends, which affects the learning of temporal features. This paper proposes a hybrid machine learning approach that blends natural language processing techniques and ensemble learning for predicting extremely rare aircraft component failure. The proposed approach is tested using a real aircraft central maintenance system log-based dataset. The dataset is characterized by extremely rare occurrences of known unscheduled component replacements. The results suggest that the proposed approach outperformed the existing imbalanced and ensemble learning methods in terms of precision, recall, and f1-score. The proposed approach is approximately 10% better than the synthetic minority oversampling technique. It was also found that by searching for patterns in the minority class exclusively, the class imbalance problem could be overcome. Hence, the model classification performance is improved.

Keywords: predictive maintenance, prognostic, imbalance learning, artificial intelligence, aerospace,

Nomenclature

Aircraft Communications Addressing and Reporting System	ACARS
Aircraft Condition Monitoring System	ACMS
Aircraft Functional-Item Number	FIN
A330 –Long-Range Aircraft Family	LR
A320 -Single-Aisle Aircraft Family	SA
Built-in test Equipment	BITE
Central Maintenance System	CMS
Electronic Control Unit/ Electronic Engine Unit	4000KS
Fault Detection, Diagnostics, and Prognostics	FDDP
Flight Warning Computers	FWCs
Flight Deck Effect	FDE
High-Pressure Bleed Valve	4000HA
Imbalanced Ratio	IR
Line Replacement Unit	LRU
Natural Language Processing	NLP
Pressure Regulating Valve	4001HA
Synthetic Minority oversampling Techniques	SMOTE
Term Frequency–Inverse Document Frequency	TF-IDF
The Air Transport Association	ATA
Satellite Data unit	5RV1
Trim Air Valve	438HC

1. Introduction

Airlines are increasingly concerned about the availability and reliability of assets and services. Most of them rely on scheduled maintenance to ensure that equipment is operating correctly in order to avoid unplanned breakdowns. Such types of maintenance are usually carried out on independent targeted components based on its usage without considering the relationship of components working together and influencing each other's lifetime. Moreover, this type of maintenance is labour-intensive and ineffective in identifying and predicting failures, especially in a complex system such as aircraft. In contrast, predictive maintenance help in identifying anomalous behaviour from an extensive historical failure data and turn it into meaningful, actionable insights for proactive maintenance – preventing downtime or accidents. This type of maintenance provides an intelligence forecast when or if equipment will fail, so its maintenance and repair can be scheduled before the failure occurs. Predictive maintenance requires having the working knowledge of the equipment, and this can be achieved by installing sensors to record and monitor target variables. So that alerts are triggered when there is a violation of the defined threshold settings. This approach can sometimes be an effective solution in a simple system. However, it is unfeasible in a complex system since adding sensors to all components is unfeasible, especially in a large fleet, which is cost-intensive and potential regulatory challenges.

Furthermore, Fault detection, diagnostics, and prognostics (FDDP) have a huge potential to improve aircraft operational reliability and stability since the main aim of FDDP is to minimize losses while ensuring the safety of equipment and reduce the risk of unplanned breakdowns [1]. FDDP involves detecting the occurrence of fault as early as possible, classifying the fault type accurately, and predicting the next occurrences of such a fault. FDDP models are designed to detect anomalies of critical components by analyzing historical data to provide actionable alerts to the operators[2]. Since the operational and maintenance datasets generated by modern aircraft have become much larger as both the number of samples and the dimensionality increased. Therefore, implementing the traditional model-based and knowledge-based approaches are becoming too difficult [3].

Moreover, finding abnormal patterns in a large log-based data is extremely challenging due to the complex non-linear relationships among the components process and sub-systems. Component failure resulting in unplanned breakdowns rarely occur during stable operation. The rare occurrence of component failures creates skewness or imbalanced distribution in the generated dataset [2], [3]. The imbalanced data problem has shown to degrade the performance of data-driven models causing unreliable prognostics [4], [5]. The aforementioned challenges have motivated more research in the application of data-driven prognostics for conditioned-based maintenance in the aerospace industry [6].

In recent times, most of the predictive maintenance deployed in the aerospace industry are trends of modelling of some specific data features such as vibration, pressure, and exhaust gas, etc., which are more concentrated on the engine and auxiliary power unit [7], [8]. Whereas, considering predictive modelling at the system level is more efficient because the model will be able to capture the working relationship between components. Therefore, equipment failure logs are fruitful sources of information both for diagnoses and prognostics. However, intensive data pre-processing is required to harness valuable pieces of information. The recent technological advances have made equipment to operate through software applications. For example, modern aircraft are incorporated with advanced technology such as monitoring sensors and various aircraft communication systems (such as ACARS, ACMS) [9], which generates more extensive datasets. This application produces records of their operations, which includes some pre-defined parameters, failure messages, and other valuable target variables representing failures detected during the last operations—exploring such large historical record help in detecting an impending issue in advance. Therefore, relying on these flight failure records to develop predictive modeling for asset health management is a promising technique. That is, the application of advanced analytics to anticipate maintenance needs in order to avoid the risk associated with service disruption [10].

Furthermore, building a predictive model from aircraft central maintenance system CMS data (which is the record of failure messages) in the total absence of digital sensor data measurements poses many challenges that are not yet fully explored. Many problems arise from learning with textual-based datasets. The first problem concerns the multidimensional data treading to be able to identify patterns leading to the component replacement. The second problem is mining patterns from random failure messages generated from different aircraft in a fleet [11]. The third problem is the inherent imbalanced distribution in the dataset. For instance, most of the failure messages in the CMS data are related to component replacement due to planned maintenance or not-fault-found while the minority are related to unplanned component replacement (that is the real unplanned replacements which is our target in this study).

The imbalance classification problem or rare event occurrence is prevalent in many real-life application domains. For instance, detecting fraud in a credit card transaction, where most transactions are legitimate, and few are fraudulent. The fraudulent minority transactions are more important to predict than the legitimate majority because if any fraudulent transaction goes unnoticed, the consequences can be grave [12]. Likewise, in clinical diagnosis where most patients can be healthy while a few diagnose to have a certain rare disease [13]. The costs of misclassifying infected as healthy cannot be tolerated because of the high risks of deterioration and fatality. Similarly, Imbalanced classification can also be applied in aircraft predictive maintenance modeling, where most of the generated failure messages represent false alarm or no fault found, and the minority represents the real faults that resulted in component replacement. The problem of imbalanced data in aircraft predictive maintenance modeling using log-based CMS failure messages is that component failure rarely occurs, which creates imbalanced distribution in the generated dataset. In some cases, the ratio between class in the dataset can be as high as 10000:1, which is known as an extreme imbalanced problem [14]–[16]. Moreover, identifying patterns and learning from extreme imbalanced datasets increases classification challenges in machine learning. Hence, an improved method of accurately recognizing the minority class instances is required process [17], [18].

Several research approaches have been conducted to solve the imbalanced classification problem. The solution for the imbalanced classification problem can be categorized into three main groups; the data level, the algorithm level, and the hybrid approach see Figure 1. The data level approach involves re-sampling the dataset before presenting it as an input to the learning algorithm. The algorithm level approach tackles the imbalanced data problem by modifying the traditional machine learning algorithms to respond favorably to both classes during learning [19]. The hybrid method combines two or more algorithms or data level approaches to achieve better performance.

Although imbalance classification problems have been extensively researched [20], [21], the open literature lacks an exhaustive unified solution to handle the problem for predictive modeling generally. Hence, it is still an open area of research. Therefore, this study aims at developing an actionable prognostics model, which will enable the anticipation of unscheduled maintenance activities relating to aircraft functional items replacements, which can be achieved by identifying predictive signatures in the CMS failure messages. Secondly, to provide a suitable approach to handle the rare occurrence of unplanned component replacements.

The main advantage of our proposed approach over other state-of-the-art log-based rare failure prediction techniques are. Firstly, we propose a novel hybrid model that blends natural language processing techniques and ensemble learning for predicting rare aircraft component failure using imbalanced textual log-based data. The model is based on a log-based pattern identification technique, which involves transforming and integrating well-known natural language processing techniques (the TF-IDF and Word2vec) and ensemble learning for pattern identification and classification. It uses log-based aircraft central maintenance system data, which is not often, use for predictive maintenance modelling. In Addition, our approach helps in tackling the extreme imbalanced classification problem by searching for patterns exclusively in the minority class, which improves model performance. In predictive maintenance, a state-of-the-art ensemble-learning algorithm is adapted as a base classifier; we also show how unscheduled maintenance can be mitigated using reliable and robust prognostic models.

This paper is structured as follows; in section 2, we present related work and description of the datasets. Section 3 explains the methodology and the proposed approach. Section 4 presents the case study and experimental setup and discuss the result. Finally, conclusions and future work are presented in section 5.

2. RELATED WORK

The use of the system's operational logs is well studied in different application domains [22]. Each application domain has its specific requirements that have an impact on the design and development of the corresponding solution. Some researchers have focused on log data for troubleshooting and anomaly detection solutions [23], [24]. Other application domains that have practically shown its use are computer hard-disk failure prediction [25], [26], medical equipment failure [27], [28], and many more [29]. System failure-messages obtained from logs can also be used to understand the behaviours and common failure patterns of equipment. The most closely related work to our approach is failure-messages-based machine learning modelling. Li et al. [30], provides an approach for mining system log files. The authors attempt to provide a way of understanding and categorizing common failure patterns that resulted in system failures. Similarly, Tanguy et al. [31] show the application of NLP in mining text-based aviation incident reports data to identify future threats. However, they did not go further to show how such failure can be predicted in order to prevent its future occurrence. Similarly, the use of aircraft data has been used for modelling. For example, Korvesis et al. [32], use aircraft post-flight report data to developed an event failure prediction system via multi-instance regression. In contrary, we focus on classification and finding a solution to the rare occurrence of failures instead of regression. Another closes work which uses the aircraft CMS dataset to develop the predictive model is Nicchiotti et al. [9], the authors transformed Eigen-face and principal component analysis method which is adapted from image processing, they classify their model using Support Vector Machine (SVM). In our study, we explore a different approach, which is the use of natural language processing techniques to identify patterns related to aircraft component failures. Likewise, Yan et al. [33], proposed a predictive model to predict faults with high priority in advance by exploring the historical data of aircraft maintenance systems. The authors did not take into consideration the problem of the rare failure occurrence, which is part of our focus in this study. Their study also considers single aircraft instead of the fleet. Analysing fleet data can be more challenging; therefore, our methodology considers a fleet-based approach instead of a single aircraft. Verhagen et al. [34] develop an approach to reduce unscheduled maintenance by focusing on identifying operational factors affecting component reliability. The research uses a statistical data-driven approach and, the authors applied a proportional hazard model on aircraft operational and maintenance datasets. Although their work uses an aircraft operational dataset, which is closely related to the one used in our study. However, our approach focuses on the application of machine learning.

Another related category for predicting failure from log-based data is the rule-based expert system [35], [36]. In this type of approach, preconditioned rules are defined; the rules are then marched against the input data. If the predefined condition is met, a failure alert will be triggered. The rules are mainly defined by domain experts, not through data mining. Vilalta et al. [37] described the use of a rule-based approach to detect patterns in the sequence of events. In practice, rule-based approaches are more effective for a small and simple system. Its application in a large and complex system is quite challenging and, in some cases, impractical because domain experts need to continually update the rules in the event of any upgrades or changes, which is cumbersome.

Another related category for processing log-based data is the application of sequential pattern mining [38], [39], which is mainly about extracting interesting, useful, and unexpected patterns across sequential data using a statistical approach. Many studies have shown the applicability of using text sequence mining for failure prediction in a complex system [40]–[42]. We explore the sequence mining techniques and find out that applying sequence pattern mining alone is not suitable for our problem because of the rare occurrence of unplanned aircraft component replacement.

Although there are many existing approaches in the literature, some are suitable for solving failure prediction in specific types of equipment. Hence, the particularities of our data limit us from using out of the shelf approach. Our approach differs from the aforementioned approaches in many aspects. We proposed a new approach of pre-processing the aircraft central maintenance log-based data. In addition, the new approach provides a solution to the imbalanced classification problem, which enhances the model performance. Finally, the proposed hybrid machine learning technique for aircraft component replacement prediction is developed.

Our approach applies a unique combination of TF-IDF and Word2Vec notions from the NLP for extracting patterns and categorization of failure messages into common failure. Considering a text-based aircraft CMS failure messages, sets of patterns in each segment is considered as a document, and each pattern is considered a word. TF-IDF help in pruning out unproductive and redundant patterns, while Word2Vec is used to find the most relevant documents related to the target component it also helps in converting words into a vector of numbers. This approach improves categorization accuracy by considering the temporal characteristics of CMS failure messages to provide overall performance improvement (reduce false positive, increase prediction recall, and precision). As system failures tend to occur very rarely; therefore, the approach also includes a solution to the rare occurrence of target failures by searching for patterns exclusively in the minority class.

3. METHODOLOGICAL APPROACH

This section describes the methodology used in this study.

As seen in Figure 2, the traditional machine learning framework is divided into three phases. The pre-processing phase, the model training, testing and validation phase, and the model deployment phase. In building a machine-learning model from any data source, one must often deal with the fact that data are imperfect. Therefore, we ensure data quality, by cleaning the data, which involves correcting outliers, handling missing values, and aggregating impossible combinations, before carrying out any further analysis. This is necessary because jumping into analyzing data that has not been carefully screened for such problems can produce misleading results [43]. Secondly, we carryout feature engineering to select the best predictors for our problem, at this stage, the datasets are merged, and the right features that best describe our target components are selected using the feature engineering process. The aircraft operational log data contains timestamp and flight circle numbers, which makes it easier for creating windows lags. The third step in the data pre-proceeding phase, which involves identifying component failure patterns and trends. We focus on the component replacement that occurs due to unplanned maintenance. We aim to find the best framework for processing time-series, log-based datasets with rare failures. With a focus on addressing the imbalanced classification problem, to improve the performance of the base learning algorithm.

3.1 Problem description:

We formally describe the log-based rare failure prediction problem as follows. Given a functional item number FIN of a particular aircraft family A, with rare failure occurrence. Using a log-based failure messages $fI_m(A)$ collected from a fleet, can we infer the probability of its replacement$\pi_R(T)$ within a time window T. This problem can be solved using machine learning; hence, we consider it as a binary and multiclass classification problem for predicting aircraft functional item failure with a given period T. The training data contains predictive features extracted from the log of failure messages obtained from a fleet of civil aircraft. The failure labels are provided from the actual aircraft maintenance record. Note that our solution is targeted at specific functional items replacements, not a generic replacement. In addition, the targeted functional items are extremely rare, and our goal is to develop a model that can overcome the challenge of rarity while making predictions. The main aim of the prognostic system is to adequately provide failure alerts early enough to give maintenance engineers enough time to deal with the problem before it actually occurs. Also, the alerts should not come too early to avoid component wastage due to premature replacement. Therefore, the prediction window needs to be defined using domain expert knowledge. In this study, a prediction window is defined as; at least two flights and no more than ten flights in advance. In defining the imbalanced problem, we consider the dataset to be imbalanced if the imbalanced ratio (IR) between classes is approximately 5% to 30%. If IR is less than 5%, we consider it to be an extremely imbalanced problem. Finally, our prediction aim is, for each selected Functional Item Number (FIN), we target to achieve at least 50% prediction of unscheduled maintenance

3.2 Implementation of the proposed approach

This section discusses the implementation of our novel approach. The approach can be applied in multivariate time series, text-based, and imbalanced datasets. Therefore, the raw aircraft

CMS raw is sequential in time series format, and the flight cycles are also in sequence. The failure messages are text-based. Likewise, the record of unplanned components replacements is rare in the dataset. This specification makes it suitable to test our approach. We also focus on solving the extreme imbalanced problem to enhance the reliability and performance of data-driven models. Thus, producing improved predictive models that will mitigate the risk associated with unscheduled maintenance.

We use natural language processing and time series analysis techniques to identify trends and patterns. As shown in Figure 3, a combination of natural language processing -Term Frequency-Inverse Document Frequency (TF-IDF) and word2vec method is applied to detect patterns and trends of target components. To handle the infrequent occurrence of component replacement in the flight dataset, we made some assumptions. Such as aircraft components replacements are characterized by categorical and text-based features and occur in an uneven-intervals. Secondly, we also assume that target components are less represented (highly infrequent). Therefore, we develop our model to search for all patterns preceding each target component exclusively to predict the next replacement. To achieve that, we transform the TF-IDF and word2vec technique to evaluate the importance of each failure or error message [33], [44]. TF-IDF is a machine learning Natural Language Processing (NLP) word embedding technique that weighs words in text mining [45]. This technique provides us with the representation of text in a coordinated system where related error messages, based on the corpus of relationships, are placed closer together. It helps us also to filter out un-related failure messages. The TF-IDF consists of two parts namely,

1. TF- Term Frequency: which calculates the frequency of word appearance in a document. If a given term is t

$$\therefore\ TF(t) = \frac{n_t}{n_d} \qquad (1)$$

Where n_t is the total number of times t appear in a document and n_d is the total number of terms in the document.

2. IDF- Inverse Document Frequency: which measures the importance of each term in the document.

$$\therefore\ IDF(t) = log\ \frac{m_d}{m_t} \qquad (2)$$

Where, m_t is the total number of documents that contain term t and m_d is the total number of documents

Putting it all together

$$\Rightarrow TF - IDF(t) = TF(t).IDF(t) \qquad (3)$$

Therefore, implementing the above approach, we denote document to be each window in the dataset, and term t to be target component and failure messages represent words. For instance, let R_1 be the first component replaced due to unplanned breakdown of equipment, R_2 for second replacement and R_3 for third and so on. Let alphabet (A, B, C, etc.) represent the failure messages. Therefore, all failure message in a window, which precedes each replacement, constitute a pattern and is represented as follows.

R_1 → ABC, YZP, PPB….

R_2 → XYZ, AEP, CDB….

R_3 → CDA, EDM, OPN….

We then identify the pattern for each $\boldsymbol{R_i}$ within that window. For instance, looking at Figure 3, **W1** = $\{R_1$**:** (ABEG)**,** R_2**:** (ECDB), R_3**:** (GBED), R_2**:** (DEAB)$\}$. We then find all patterns that are related to each target component replacement $\boldsymbol{R}$ across all the datasets. Finally, the extracted patterns are then used to train the model.

Therefore, during model training, taken for example, all failure messages related to R_1 are identified and all the possible combinations of failure messages related to R_1 are created, which produces more new different patterns. This is done to increase more patterns that are related to each replacement, which will contribute to addressing the imbalanced problem. The combination and creation of new patterns are achieved using bootstrapping techniques. To avoid the overfitting problem, we use select with replacement approach.

Furthermore, the model is developed to flag-up component replacement prognostic alert when a pattern is detected. The features, such as date-time and flight cycle numbers, play a vital role in defining when in advance, the model should flag up prognostic alerts.

Furthermore, our pattern recognition strategy is similar to the one developed by Vilalta et al. [37]. However, our approach differs in the learning strategy instead of using the rule-based model; we make use of supervised learning (classification technique) to build a data-driven model. The imbalanced classification problem is overcome by searching for patterns on the minority class exclusively. The strategy is shown in Algorithm 1. In addition, having known the patterns. The next step is we represent the features into a vector space using the word2vec method. Prior to that, categorical features are handled using the one-hot-encoding technique [46]. As shown in Figure 3, all the terms in the pattern are selected and then converted into a vector space dimension. To illustrate, considering the windows w_i and patterns ABC ... leading to components replacements F_i.

W_1= ABEGR_1 −CBDER_3 −DEABR_1 −EDCBR_2

W_2= AEDBR_1 −BEAGR_1 −CDCBR_2-DEBCR_3

W_3= EDCBR_2 −ABEGR_1 −EDBCR_2 −CBDER_3

Using TF-IDF and word2vec approach to identify all failure messages related to each target component replacement.

$$Boolean\, frequencies = m(t) \begin{cases} 1, & if\ t\ occurs\ in\ window \\ 0, & otherwise \end{cases} \tag{4}$$

Where $m(t)$ represents the term frequency of patterns of failure messages leading to each component replacements. All corresponding replacements in each window are then sum up. tf (t,w) is the total number of patterns present in each window.

The inverse document frequency measures how much each failure message provides in relation to components replaced in each window. That is if it is common or it is rare across all windows.

$$Idf\ (t, W) = log\ \frac{N}{|\{d \in W : t \in d\}|} \tag{5}$$

Where N is the total number of failure messages in a window N = $|W|$, and $|\{d\ \in W: t\ \in d\}|$, number of windows where term *t* appears.

Therefore $tf - idf(t, d, W)\ =\ td(t, d).idf(t, W)$ (6)

We then create our feature victor using the following equation.

$$\overrightarrow{\boldsymbol{vw_n}} = tf(t_1, w_n), tf(t_2, w_n), tf(t_3, w_n), \ldots, tf(t_n, w_n) \tag{7}$$

Using continues bag of words strategy in word2vec, each dimension of the feature vector is represented by the pattern, for example, $tf(t_1, w_n)$ represents the frequency of term 1. For example, using equation 4 and Table I, patterns for window 1 to 3 are represented as victors as follows:

$$\overrightarrow{vw_1} = tf(t_1, w_1), tf(t_2, w_1), tf(t_3, w_1), tf(t_4, w_1), \ldots, tf(t_n, w_1)$$

$$\overrightarrow{vw_1} = (2{,}1{,}1 \ldots \mathrm{n})$$

$$\overrightarrow{vw_2} = tf(t_1, w_2), tf(t_2, w_2), tf(t_3, w_2), tf(t_4, w_2), \ldots, tf(t_n, w_2)$$

$$\overrightarrow{vw_2} = (2{,}1{,}1 \ldots \mathrm{n})$$

$$\overrightarrow{vw_3} = tf(t_1, w_2), tf(t_2, w_3), tf(t_3, w_3), tf(t_4, w_3), \ldots, tf(t_n, w_3)$$

$$\overrightarrow{vw_3} = (1,1,1 \ldots \text{n})$$

The resulting vectors show that window one $\overrightarrow{vw_1} = (2,1,1 \ldots)$ has two patterns that prompt replacement of the component R_1 , one pattern for R_2, and one for R_3 We then represent it in a general matric with the shape $|w| * l$ where $|w|$ is the cardinality of the feature vector space in each window, and l is the total number of pattern vectors.

$$M|W| * L = \begin{vmatrix} 2 & 1 & 1 \ldots \text{n} \\ 2 & 1 & 1 \ldots \text{n} \\ 1 & 1 & 1 \ldots \text{n} \\ 1 & 2 & 0 \ldots \text{n} \end{vmatrix}$$

3.2 Algorithm 1: Detecting patterns and trends of target components

- Find the pattern of failure message preceding the target component within a given fixed window.
- Carry out validation of characters that uniquely identify the target component.
- Combine the characteristic to build a data-driven predictive model.

The pseudocode:

```
INPUT:
 Imbalanced time series dataset
{
F = Sequence of failure messages
fm =failure message
W = window size
r = Target replacements
T= Time
}
OUTPUT:  P = Pattern for Target Replacement
TARGET_PATTERN ( F, W, r, T)
        Step 1: Get the Data D
                Initialize variables G= 0, H=0
                Define window size W
        Step 3: loop through the series of event in each W to identify a component
                 replacement.
                FOREACH  F,  $f_{m(i)}$ = ( $r_i, t_i$ ) ∈ $F$
                 (where $t_i = current\ time$)
        Step 4:  Identify a pattern preceding the component replacement.
                        FOREACH F,  $f_{m(j)}$ = ( $r_j, t_j$ ) ∈ $H$
                                If ($current\ time - t_j$) > W ; Remove $f_{m(i)}$ from H
                        END
        Step 5: Generate a pattern for each event that occurs together, leading to the
                 replacement of the component.
                        IF $f_{m(i)}$ is a target replacement
                                G = G ∪ {$r_j$| $r_{j,}$ ...}
                                H ∪ $f_{m(i)}$
                        Invoke Aprior Algorithm[51] on G
                END
        Step 6: Next window: Go-to step 3
                Use TF-IDF and Word2Vec on G to find all related pattern P
        Step 7: Output P
```

Algorithm 1 transverse through the sequence of failure messages, which is in time-series format. The algorithm store patters of failure messages related to each target functional item failure in memory. The identified patterns are then used for fault prediction.

4. Case Study

This section describes the case study and the experimental setup.

Data Description: This study uses more than seven years' worth of data. The datasets are collected from two databases. The first database is the aircraft Central Maintenance System (CMS) data, which comprises of error messages from BIT (built-in test) equipment (that is aircraft fault report(s) record) and the flight deck effect (FDE). These messages are generated at different stages of flight phases (that is take-off, cruise, and lading). The second database is the logs of aircraft maintenance activities -that is, the comprehensive description of all aircraft maintenances recorded over time. These databases are associated with a fleet of civil aircraft. In aircraft, the main purpose of CMS is to facilitate maintenance activities by directly alerting fault message, that can be used by pilots and maintenance engineers; to at the main base-perform troubleshooting or at the line stop level -perfume component removal[47]. The primary function of aircraft CMS is to acquire, and store messages transmitted by the connected system Built-In Test Equipment (BITE) or by Flight Warning Computers (FWCs) as seen in Figure 4.

Generally, in aircraft, sensors, and monitoring systems are install and configured to monitor components. Based on the configured rules, failure messages are generated when there is a violation of any configured rules. As explained in Airbus training manuals [48], each time a fault is detected and isolated, a failure message is generated by system BITE. The message is memorized in the BITE memory and transmitted to the CMS. Each failure message is made up of 48 characters long, which is composed of a faulty line replaceable unit (which is made up of one or more parts depending on the type) and ATA 6-digit reference number. A message might contain several Line Replacement Unit (LRU), but only one suspected element is faulty. Each message syntax is of the form B-FIN-BUSNAME; **B** (Most probable suspected component) – **FIN** (Functional Item Number) – **BUS NAME** (complementary information) as seen in Figure 5.

All the CMS failure messages are recorded in a logbook. Based on the failure messages, unplanned maintenance can be scheduled for the malfunctioned items. After the maintenance, the engineers update maintenance records with the repair details. The maintenance record provides detailed information about each component or item replaced (such as date of repair, part identification number, and time spent on troubleshooting, etc.). In this study, the CMS log data is collected from a fleet of civil aircraft over seven years. The data is unique in many aspects; It is temporal, and it can be seen as numeric time-series or symbolic sequence with features extracted from failure message or with event occurrences over some segments or window period. It contains categorical values, both text and numerical, as in the case of failure source, failure type, and ATA number. The old traditional approach to predictive maintenance using this type of data is to use experience domain experts to examine the historical failure message to identify abnormalities. Then based on such observation, predictive patterns will be manually formulated for a targeted functional item using some pre-conditioned rules. Such an approach is highly expert experience-based and time-consuming. However, such an approach provides an important concept that system failure can be predicted by analysing its historical failure history. The concept motivates this study and serves as bases for our problem formulation. So far, the CMS data have only been used for short time troubleshooting, anomaly detection, Line Replacement Unit (LRU) removal, and system failure analysis or test, no much work have been found to use this type of data for building predictive maintenance models.

Validation: to validate the performance of predicting aircraft components failure from imbalanced log-based data with the proposed approach. We modelled it in two categories, binary classification, and multi-class classification. In the first scenario, we modelled it as a multi-class classification problem that is predicting all the targeted component failure at the same time. Secondly, we modelled it as a binary classification problem that is predicting individual functional items. In both instances, we use ensemble-learning algorithms as base-classifier. We choose to evaluate the approach using ensemble learning because of its capability of combining more than once classifiers to achieve better results, which has an advantage over a single classifier, especially in a skewed data distribution context. To evaluate the model in terms of imbalance classification, we compare our proposed approach with the existing synthetic minority oversampling technique (SMOTE).

As shown in Table I, the data is group into two categories representing different types of aircraft in the fleet. The A330 –long-range (LR) and the A320 -Single-aisle (SA) aircraft. The dataset ranging from the year 2011 to 2015 is used for training the model, while from 2016 to 2018 is used for testing. After the pattern identification-using algorithm 1, the resulting dataset is then divided into two (for training and testing). Data ranging from the year 2011 to 2015 is used for model training, while from 2016 to 2018 is used for evaluation and testing.

The effectiveness of the proposed approach was demonstrated on the log-based CMS dataset. For each aircraft family, we choose a target functional item Number (FIN) of high practical value with an adequate number of known failure cases. We select out of many, the following aircraft functional items to be used in the experiment. The target components selected for this study are based on some group of common failures in an aircraft subsystem that happens with a frequency of 0.1 - 1% over some time.

LRU for A330 –long-range (LR) aircraft family: 4000KS - Electronic Control Unit/ Electronic Engine Unit**,** 4001HA – Pressure Regulating Valve**,** 5RV1 – Satellite Data unit, and 438HC – Trim Air Valve.

LRU for A320 -Single-aisle (SA) aircraft family: 11HB – Flow control valve, 10HQ - Avionics equipment ventilation computer, 1TX1 - Air traffic service unit, and 8HB - Flow control valve 2.

Imbalanced Ratio (IR): In the A330 aircraft family, the size of the training dataset is 360575, and the A320 family size is 389829. The frequency of functional items replacement emanating from unscheduled maintenance is as follows. In the A330 aircraft family, 4001HA is replaced 17 times giving us the imbalance ratio (IR) of 360558: 17, 4000KS is replaced 15 times given us IR of 360560: 15, 5RV1 is replaced 16 times given us IR of 360559: 16, and 438HC is replaced 25 times given us IR of 360550: 25. Similarly, in the A320 aircraft family, 11HB is replaced 11 times giving us the imbalance ratio (IR) of 389818: 11, 10HQ is replaced 12 times given us IR of 389817: 12, 1TX1 is replaced 25 times given us IR of 389804: 25, and 8HB is replaced 14 times given us IR of 389815: 14.

$$\mathbf{IR} = \frac{Minority\ class}{Majority\ class} * 100 \tag{8}$$

4.1 Evaluation

The performance of the model is measured using precision, recall, f1-score, and ROC curves. The experimental results are displayed in Table II and Figures 10-13.

Prognostic Criteria: Prognostics alerts for component replacement should flag up in a reasonable time, not too early to avoid wasting the useful component life due to premature removal. Also, not too close to failure to give enough time to prepare for maintenance. As seen in Figure 6, in this study, we denote a maximum wasted life as (MWL), which are alerts that flag off not more than ten flight cycles before replacement. Similarly, the minimum notice period (MNP), as alerts that are flag off not less than two flight cycles before replacement. We define the metrics as follows.

True Positive Rate- TPR: alerts that flag off in-between MWL and MNP and truly replacement occur.

$$\boldsymbol{TPR} = \frac{TP}{TP+FN} = recall \tag{9}$$

False Negative Rate – FNR: alerts that flag off in-between MWL and MNP, and no replacement occur (no fault found).

$$\boldsymbol{FNR} = (TP + TN) \,/\, (TP + FN + TN) = 1 - FPR \tag{10}$$

False Negative –FN: – alerts triggered much earlier before failure occur, which are ten flight cycles away from replacement (less than MWL) and alerts too close to replacement (more than MNP), and truly replacement is needed.

True Negative – TN: alerts triggered much earlier before failure occur, which are ten flight cycles away from replacement (less than MWL) and alerts too close to replacement (more than MNP), and no replacement is needed.

Sensitivity and specificity: The goal of our method is to make sure actual positives are not overlooked, which is to minimize false negatives to the best acceptance tolerance level. Also, the effort is to make sure actual negatives are classified as negatives that are achieving low false negatives. Perfect precision will mean no false positive (FP =0), and perfect recall means no false-negative (FN=0).

Precision: Measure of classifier exactness, the percentage of true positive predictions made by the classifier that is truly correct. So, low precision indicates a large number of False Positives.

$$\boldsymbol{Precision} = TP/\,(TP + FP) \tag{11}$$

Recall: Measure of classifier Completeness recall is defined as the percentage of true positives that are correctly detected by the classifier. So, low recall indicates many False Negatives.

$$\boldsymbol{Recall} = TP/\,(TP + FN) \tag{12}$$

F1-Score: is the mean average between the precision and the recall

$$\boldsymbol{F1\_Score} = (2 * ((precision) * (recall)))/\,((precision) + (recall)) \tag{13}$$

Receiver Operating Characteristic Curve (ROC) Curves: it is a graphical representation that illustrates the diagnostic ability of the classifier as a discriminant threshold is varied.

Area under the curve AUC = 1/2(TP1/ (TP +F) + TN / (TN + FP)) (14)

To test the performance of our approach for extreme imbalance problem and in log-based failure prediction. We set the experiments based on the aircraft CMS dataset. We consider data from two categories of aircraft family, -A330 and -A320. We compare the performance of our approach with the existing imbalance learning method (we choose synthetic minority oversampling techniques because of its wide industrial application). In each case, five different ensemble machine learning algorithms are considered as base classifiers.

Scenario 1: multiclass approach

We make a prediction for all FIN and compare it against the baseline imbalanced learning algorithm -SMOTE.

1. SMOTE + *Random Forest (RF), XGboost (XGB), Decision Tree (DT), Naïve Bayes (NB), Light Gradient Boosting Machine (LGBM), Gradient Booting Decision Tree (GBDT), and Support Vector Machine (SVM):* After cleaning the data. We divided the data into training and testing. The training data was resampled using SMOTE. Then the different machine learning algorithms are used to train the classifier.

2. Our approach + *Random Forest (RF), XGboost (XGB), Decision Tree (DT), Naïve Bayes (NB), Light Gradient Boosting Machine (LGBM), Gradient Booting Decision Tree (GBDT), and Support Vector Machine (SVM):* After cleaning the data. We carry out behavioral pattern analysis. We then divided the data into training and testing. We train the model without applying any existing imbalanced learning method. Then the different machine learning algorithms are used to train the classifier.

In the first instance, we consider all failure related to the aforementioned targeted FIN. During evaluation, accuracy, recall, and precision is used as performance metrics. The comparison result of the two cases is shown in Figures 7 and 8. Random forest outperformed other ensemble classifiers. Therefore, in the second scenario, which is predicting individual functional items (binary classification approach), we use only random forest.

Scenario 2: Binary classification approach- Individual component failure prediction model:

We make a prediction for each FIN and compare it against the baseline imbalanced learning algorithm -SMOTE.

In choosing the base-classifier for binary classification, any machine-learning algorithm for classification can be used. Our choice of an ensemble-learning algorithm as a base-classifier is because it is effective in improving predictive performance, especially in classifying skew dataset. In addition, because RF is an ensemble bagging technique that combines multiple decision trees to achieve a better result. The trees in RF create high variance and low bias, making it a suitable choice. Also, since data is distributed over different trees in the forest and each tree sees a different set of data, therefore in general, RF does not over-fit, and also because they are made of low bias trees, it does not suffer from the under-fitting problem. Thus, among the ensemble algorithm, we choose a random forest because it gives better precision and recall compared to others.

We use algorithm 1 to generate patterns related to each targeted FIN. We then adept the RF algorithm to crate the individual failure prognostic model. RF is an ensemble learning method where the training data is divided into several subsamples, and each subsample is trained using a decision tree classifier know as a weaker learner. The result is then aggregated by majority voting providing a stronger base learning algorithm. Apart from sampling on the dataset, trees are randomized by using boosting and bagging techniques to generate splits [49], [50]. This approach enhances the performance of the model.

In predicting targeted individual functional items, their failures are extremely rare. Normally, accuracy is mostly considered as an important metric to evaluate the performance of a classifier. However, the use of accuracy to evaluate performance under extreme imbalanced problems can be misleading because classifies will be biased towards the majority class to achieve high overall accuracy. Therefore, to evaluate the performance of the classifiers more precisely, some alternative metrics are adapted, which include precision, recall, f1-score, and area under the curve.

4.2 Result and Discussion

As shown in Figures 7 and 8, it is observed that comparing our approach with SMOTE using different ensemble learning algorithms as base-classifier. The performance of all the base-classifiers is better with the proposed approach compared to SMOTE. Furthermore, RF outperformed other ensemble algorithms; it shows comparative performance in recall and precision, which means RF is able to identify more faults compared to other base-classifiers. Although, the multiclass approach produced a significant improvement, however, the majority of predictions fall close to the defined maximum wasted life.

As shown in Table II. For individual FIN prediction. It can be observed that for all the functional items, our model has a precision of more than 70%. It means whenever the model predicts aircraft failure that leads to component replacement. It is correct 70% of the time. In other words, this indicates that out of the total prediction, the model prognoses more than seventy percent of failures that lead to LRU replacement. The precision score also shows the model produces less than thirty percent of false-positive alerts. Similarly, an average recall of more than 60% is achieved in all the considered FIN's. Indicating that the model correctly predicts more than sixty percent of actual failure that leads to LRU replacement. It is important to note that for individual prediction (binary classification), the majority of prediction fall close to the defined the minimum notice period, which means component will be adequately utilized. This means binary classification has an advantage over multiclass prediction. Since high cost associated with false-negative is the main concern in this study- that is a misclassifying real failure as not failure, especially for safety-critical equipment were the consequence is grave. Therefore, the recall score shows that the model trigger 60% of the actual failure alert that leads to LRU replacement.

The goal is to obtain both a high percentage of precision and recall in all cases. However, more than 20% of false-positives rate and 30% false-negative rate is still recorded. Nevertheless, our approach achieved our target, which is to predict more than 50% of aircraft component replacement within the desired define range (in-between MNP and MWL), this can be seen by the overall percentage F1-score, which is approximately 65% in all cases. Similarly, to obtain the trade-off between the model sensitivity (TPR) and specificity (1-FPR), ROC Curves of each target component replaced is acquired. The graphical representation of the average result obtained is presented in Figure 10 to 13, as seen in most of the cases the area under the curve (AUC for the testing dataset is above 70%. Indicating good overall sensitivity of classifier to predicting component replacement due to unscheduled maintenance). Note that the ROC curve does not depend on data distribution. This makes it useful in evaluating classifiers predicting imbalanced datasets.

Furthermore, although the proposed approach achieved approximately 20% of the overall percentage of the false-positive rate, in contrast, SMOTE achieved an approximately overall false-positive rate of 30%. This shows a difference of 10%, indicating that our approach achieved a significant improvement compared to synthetic minority oversampling techniques. Furthermore, it can be observed that the imbalanced ratio has an impact on performance. For instance, in cases with extreme IR, we obtain a lower precision and recall compared to the ones with higher IR. Despite the extreme imbalance ratio in all the cases considered, our approach still achieved better performance compared to SMOTE, which indicates its robustness in handling extreme imbalanced datasets.

5. Conclusion

This paper proposes an integrated data-driven learning technique for predicting aircraft component failure using imbalanced, textual, and log-based data. A hybrid model involves blending natural language processing techniques and ensemble prediction is developed to tackle extreme imbalanced classification problem and forecast aircraft component failures. We utilize real-life aircraft Central Maintenance System (CMS) data to develop a predictive maintenance model for predicting aircraft component replacement in advance to avoid unscheduled maintenance. A well-known natural language processing technique, the TF-IDF and Word2vec, are transformed for pattern identification and text vectorization. Then an ensemble random forest algorithm was successfully adapted for individual functional item prediction. In predictive maintenance, we show how unscheduled maintenance can be mitigated using the proposed robust prognostic model. The model can flag off component replacement alerts within the desired define range. In evaluation, we suggest an evaluation criterion that combines the prognostics alerts with the precision and recall within a reasonable timeframe. We compare the performance of our proposed approach against state-of-the-art imbalanced learning techniques (SMOTE). The proposed approach is approximately 10% better than SMOTE. It was also found that by searching for patterns in the minority class exclusively, the class imbalance problem can be overcome. Hence, the model classification performance is improved. Finally, even though the proposed method can predict more than 50% of unscheduled aircraft component failure, it did not go further to determine the root cause of the failure. Therefore, this work can be extended to enhancing aircraft failure diagnosis using proactive logging data. Future work will also aim to increase the performance of the model by exploiting information from a variety of sources, such as sensors and other related variables.

ACKNOWLEDGMENT

This work has been performed under the AIRMES (Airline Maintenance Operations implementation of an E2E Maintenance Service Architecture and its enablers) project. This project has received funding from the Clean Sky 2 Joint Undertaking under the European Union's Horizon 2020 research and innovation programme under grant agreement No 681858. This project has also received funding from the Petroleum Technology Development Fund (PTDF) of Nigeria.

REFERENCE

[1] X. Dai and Z. Gao, "From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis," *IEEE Trans. Ind. Informatics*, vol. 9, no. 4, pp. 2226–2238, 2013, doi: 10.1109/TII.2013.2243743.

[2] S. R. Saufi, Z. A. Bin Ahmad, M. S. Leong, and M. H. Lim, "Challenges and Opportunities of Deep Learning Models for Machinery Fault Detection and Diagnosis: A Review," *IEEE Access*, vol. 7, pp. 122644–122662, 2019, doi: 10.1109/access.2019.2938227.

[3] P. Park, P. Di Marco, H. Shin, and J. Bang, “Fault detection and diagnosis using combined autoencoder and long short-term memory network,” *Sensors (Switzerland)*, vol. 19, no. 21, pp. 1–17, 2019, doi: 10.3390/s19214612.

[4] B. S. Raghuwanshi and S. Shukla, “UnderBagging based reduced Kernelized weighted extreme learning machine for class imbalance learning,” *Eng. Appl. Artif. Intell.*, vol. 74, no. July, pp. 252–270, 2018, doi: 10.1016/j.engappai.2018.07.002.

[5] Z. Wu, W. Lin, and Y. Ji, “An Integrated Ensemble Learning Model for Imbalanced Fault Diagnostics and Prognostics,” *IEEE Access*, vol. 6, pp. 8394–8402, 2018, doi: 10.1109/ACCESS.2018.2807121.

[6] B. Jinsong, G. Yuan, Z. Xiaohu, Z. Jianguo, and J. Xia, “A Data Driven Model for Predicting Tool Health Condition in High Speed Milling of Titanium Plates Using Real-Time SCADA,” *Procedia CIRP*, vol. 61, pp. 317–322, 2017, doi: 10.1016/j.procir.2016.11.191.

[7] G. Nicchiotti and J. Rüegg, “Data-Driven Prediction of Unscheduled Maintenance Replacements in a Fleet of Commercial Aircrafts,” pp. 1–10, 2014.

[8] J. Austin, T. Jackson, M. Fletcher, M. Jessop, P. Cowley, and P. Lobner, “Predictive Maintenance : Distri buted Ai rcraft Engi ne Diagnostics,” 2114.

[9] G. Nicchiotti and J. Rüegg, “Data-Driven Prediction of Unscheduled Maintenance Replacements in a Fleet of Commercial Aircrafts,” pp. 1–10, 2018.

[10] C. V. Oster, J. S. Strong, and C. K. Zorn, “Analyzing aviation safety: Problems, challenges, opportunities,” *Res. Transp. Econ.*, vol. 43, no. 1, pp. 148–164, 2013, doi: 10.1016/j.retrec.2012.12.001.

[11] S. Alestra *et al.*, “Rare event anticipation and degradation trending for aircraft predictive maintenance,” in *11th World Congress on Computational Mechanics,*

WCCM 2014, 5th European Conference on Computational Mechanics, ECCM 2014 and 6th European Conference on Computational Fluid Dynamics, ECFD 2014, 2014, pp. 6571–6582.

[12] L. T. Nghiem, “MASI : Moving to Adaptive Samples in Imbalanced Credit Card Dataset for Classification,” *2018 IEEE Int. Conf. Innov. Res. Dev.*, no. May, pp. 1–5, 2018.

[13] T. Gao *et al.*, “Predicting pathological response to neoadjuvant chemotherapy in breast cancer patients based on imbalanced clinical data,” *Pers. Ubiquitous Comput.*, pp. 1–9, 2018, doi: 10.1007/s00779-018-1144-3.

[14] Z. H. Janjua, M. Vecchio, M. Antonini, and F. Antonelli, “IRESE: An intelligent rare-event detection system using unsupervised learning on the IoT edge,” *Eng. Appl. Artif. Intell.*, vol. 84, no. September 2018, pp. 41–50, 2019, doi: 10.1016/j.engappai.2019.05.011.

[15] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[16] S. M. A. Elrahman and A. Abraham, “A Review of Class Imbalance Problem,” *Netw. Innov. Comput.*, vol. 1, pp. 332–340, 2013.

[17] J. L. Olmo, A. Cano, J. R. Romero, and S. Ventura, “Binary and multiclass imbalanced classification using multi-objective ant programming,” *Int. Conf. Intell. Syst. Des. Appl. ISDA*, pp. 70–76, 2012, doi: 10.1109/ISDA.2012.6416515.

[18] M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” *Neural Networks*, vol. 106, pp. 249–259, 2018, doi: 10.1016/j.neunet.2018.07.011.

[19] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning

from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, 2017, doi: 10.1016/j.eswa.2016.12.035.

[20] P. Branco, L. Torgo, and R. P. Ribeiro, "A Survey of Predictive Modelling under Imbalanced Distributions Adapting Resampling Strategies for Dependency-Oriented Data in Imbalanced Domains View project International Workshop on Cost-Sensitive Learning View project A Survey of Predictive Modelling ," 2015. Accessed: Sep. 13, 2018. [Online]. Available: https://www.researchgate.net/publication/275968092.

[21] J. Bi and C. Zhang, "An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme," *Knowledge-Based Syst.*, vol. 158, pp. 81–93, 2018, doi: 10.1016/j.knosys.2018.05.037.

[22] F. Salfner, M. Lenk, and M. Malek, "A survey of online failure prediction methods," *ACM Comput. Surv.*, vol. 42, no. 3, pp. 1–68, 2010, doi: 10.1145/1670679.1670680.

[23] D. Gorinevsky, B. Matthews, and R. Martin, "Aircraft anomaly detection using performance models trained on fleet data," *Proc. - 2012 Conf. Intell. Data Understanding, CIDU 2012*, pp. 17–23, 2012, doi: 10.1109/CIDU.2012.6382196.

[24] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data Mining and Analytics in the Process Industry: The Role of Machine Learning," *IEEE Access*, vol. 5, pp. 20590–20616, 2017, doi: 10.1109/ACCESS.2017.2756872.

[25] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado, "Machine learning methods for predicting failures in hard drives: A multiple-instance application," *J. Mach. Learn. Res.*, vol. 6, pp. 783–816, 2005.

[26] J. Son, Q. Zhou, S. Zhou, X. Mao, and M. Salman, "Evaluation and comparison of mixed effects model based prognosis for hard failure," *IEEE Trans. Reliab.*, vol. 62, no. 2, pp. 379–394, 2013, doi: 10.1109/TR.2013.2259205.

[27] R. Sipos, D. Fradkin, F. Moerchen, and Z. Wang, “Log-based predictive maintenance,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 1867–1876, doi: 10.1145/2623330.2623340.

[28] Y. Yuan, S. Zhou, C. Sievenpiper, K. Mannar, and Y. Zheng, “Event log modeling and analysis for system failure prediction,” *IIE Trans. (Institute Ind. Eng.*, vol. 43, no. 9, pp. 647–660, 2011, doi: 10.1080/0740817X.2010.546385.

[29] K. Zhang, J. Xu, M. R. Min, G. Jiang, K. Pelechrinis, and H. Zhang, “Automated IT system failure prediction: A deep learning approach,” *Proc. - 2016 IEEE Int. Conf. Big Data, Big Data 2016*, pp. 1291–1300, 2016, doi: 10.1109/BigData.2016.7840733.

[30] T. Li, S. Ma, F. Liang, and W. Peng, “An integrated framework on mining logs files for computing system management,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 776–781, 2005, doi: 10.1145/1081870.1081972.

[31] L. Tanguy, N. Tulechki, A. Urieli, E. Hermann, and C. Raynal, “Natural language processing for aviation safety reports: From classification to interactive analysis,” *Comput. Ind.*, vol. 78, pp. 80–95, 2016, doi: 10.1016/j.compind.2015.09.005.

[32] P. Korvesis, S. Besseau, and M. Vazirgiannis, “Predictive maintenance in aviation: Failure prediction from post-flight reports,” *Proc. - IEEE 34th Int. Conf. Data Eng. ICDE 2018*, pp. 1423–1434, 2018, doi: 10.1109/ICDE.2018.00160.

[33] W. Yan and J.-H. Zhou, “Predictive modeling of aircraft systems failure using term frequency-inverse document frequency and random forest,” in *IEEE International Conference on Industrial Engineering and Engineering Management*, 2018, vol. 2017-Decem, pp. 828–831, doi: 10.1109/IEEM.2017.8290007.

[34] W. J. C. Verhagen and L. W. M. De Boer, “Predictive maintenance for aircraft components using proportional hazard models,” *J. Ind. Inf. Integr.*, vol. 12, no.

October 2017, pp. 23–30, 2018, doi: 10.1016/j.jii.2018.04.004.

[35] K. L. Butler, “Expert system based framework for an incipient failure detection and predictive maintenance system,” *Proc. Int. Conf. Intell. Syst. Appl. to Power Syst. ISAP*, pp. 321–326, 1996, doi: 10.1109/isap.1996.501092.

[36] P. Kumar and R. K. Srivastava, “An expert system for predictive maintenance of mining excavators and its various forms in open cast mining,” *2012 1st Int. Conf. Recent Adv. Inf. Technol. RAIT-2012*, pp. 658–661, 2012, doi: 10.1109/RAIT.2012.6194607.

[37] R. Vilalta and Sheng Ma, “Predicting rare events in temporal domains,” *2002 IEEE Int. Conf. Data Mining, 2002. Proceedings.*, pp. 474–481, 2003, doi: 10.1109/icdm.2002.1183991.

[38] S. Abbasghorbani and R. Tavoli, “Survey on sequential pattern mining algorithms,” *Conf. Proc. 2015 2nd Int. Conf. Knowledge-Based Eng. Innov. KBEI 2015*, pp. 1153–1164, 2016, doi: 10.1109/KBEI.2015.7436211.

[39] T. Truong-Chi and P. Fournier-Viger, “A Survey of High Utility Sequential Pattern Mining,” vol. 1, no. 1, pp. 97–129, 2017, doi: 10.1007/978-3-030-04921-8_4.

[40] X. Fu, R. Ren, S. A. Mckee, J. Zhan, and N. Sun, “Digging deeper into cluster system logs for failure prediction and root cause diagnosis,” *2014 IEEE Int. Conf. Clust. Comput. Clust. 2014*, no. 2, pp. 103–112, 2014, doi: 10.1109/CLUSTER.2014.6968768.

[41] W. Chang, Z. Xu, M. You, S. Zhou, Y. Xiao, and Y. Cheng, “A Bayesian failure prediction network based on text sequence mining and clustering,” *Entropy*, vol. 20, no. 12, 2018, doi: 10.3390/e20120923.

[42] H. K. Lim, Y. Kim, and M. K. Kim, “Failure Prediction Using Sequential Pattern

Mining in the Wire Bonding Process," *IEEE Trans. Semicond. Manuf.*, vol. 30, no. 3, pp. 285–292, 2017, doi: 10.1109/TSM.2017.2721820.

[43] Z. Shichao, Z. Chengqi, and Y. Qiang, *Data Preparation for Data Mining*, no. March 2012. 2007.

[44] D. Dubin, "The Most Influential Paper Gerard Salton Never Wrote," *Libr. Trends*, vol. 52, no. 4, pp. 748–764, 2004.

[45] U. Kamath, J. Liu, and J. Whitaker, *Deep Learning for NLP and Speech Recognition*. 2019.

[46] P. Cerda, G. Varoquaux, and B. Kégl, "Similarity encoding for learning with dirty categorical variables," *Mach. Learn.*, vol. 107, no. 8–10, pp. 1477–1494, 2018, doi: 10.1007/s10994-018-5724-2.

[47] Airbus, "CMS.pdf." 2000.

[48] A. training Manuals, "Trouble shooting philosophy with A320 CFDS / A340 CMS," no. May, 2005.

[49] C. Gutschi, N. Furian, J. Suschnigg, D. Neubacher, and S. Voessner, "Log-based predictive maintenance in discrete parts manufacturing," *Procedia CIRP*, vol. 79, pp. 528–533, 2019, doi: 10.1016/j.procir.2019.02.098.

[50] H. Inoue and R. Inoue, "A very large platform for floating offshore facilities," *Coast. Ocean Sp. Util. III. Proc. Symp. Genoa, 1993*, pp. 533–551, 1995.

[51] R. Agrawal, "Fast Algorithms For Mining Association Rules In Datamining," *Int. J. Sci. Technol. Res.*, vol. 2, no. 12, pp. 13–24, 2013.

List of Figure Captions

TABLES

Table I. Sample of the pre-processed aircraft CMS dataset

Date	Time	Flight circle	A/C No	Window lag	FM pattern	FIN Rplmt
10-03-15	09.03	-91	1	W_1	ABEG	R_1
10-03-15	10.03	-88	2	W_1	DEAB	R_1
11-03-15	10.00	-81	8	W_1	EDCB	R_2
11-03-15	11.05	-80	21	W_1	CBED	R_3
13-04-15	09.08	-79	12	W_2	AEDB	R_1
13-04-15	10.03	-76	9	W_2	BEAG	R_1
14-04-15	22.00	-73	23	W_2	EDCB	R_2
15-04-15	09.05	-71	2	W_2	CBED	R_3
16-04-15	09.02	-70	3	W_3	BEAH	R_1
16-04-15	21.08	-65	18	W_3	ABCG	R_3
17-04-15	13.00	-64	28	W_3	EDBC	R_2

Table II. Showing experiment results using binary classification approach with, RF as base classifier

A330 Aircraft											
		RF+ SMOTE				**RF + Our approach**					
IR	**FIN**	**Precision**	**Recall**	**F1**	**AUC**	**Precision**	**Recall**	**F1**	**AUC**	**TPR**	**FPR**
0.0047	4001HA	0.83	0.62	0.70	0.72	0.94	0.79	0.86	0.87	0.79	0.21
0.0043	4000KS	0.80	0.60	0.68	0.69	0.90	0.76	0.82	0.83	0.76	0.24
0.0044	5RV1	0.80	0.60	0.68	0.69	0.91	0.77	0.83	0.84	0.77	0.23
0.0069	438HC	0.90	0.85	0.87	0.88	0.96	0.85	0.84	0.86	0.85	0.15
A320 Aircraft											
0.0028	11HB	0.70	0.59	0.64	0.65	0.81	0.70	0.75	0.76	0.70	0.30
0.0031	10HQ	0.75	0.62	0.68	0.67	0.86	0.72	0.78	0.79	0.72	0.28
0.0064	1TX1	0.88	0.80	0.83	0.84	0.91	0.82	0.86	0.87	0.82	0.18
0.0036	8HB	0.80	0.66	0.72	0.73	0.88	0.74	0.80	0.81	0.74	0.26

FIGURES

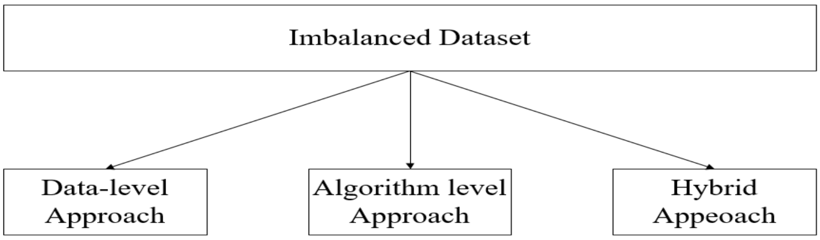


Figure 2. Shows the basic three approaches to solving the imbalanced dataset problem

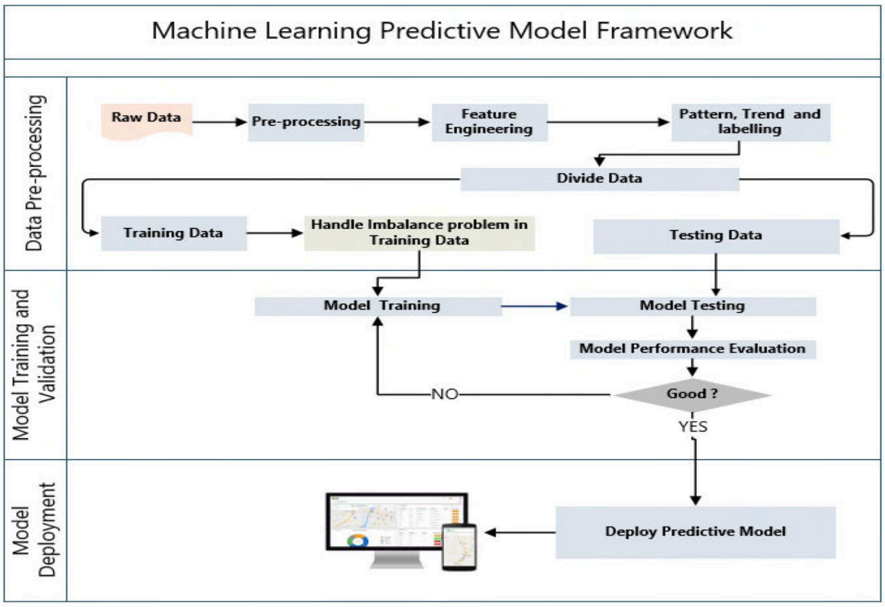


Figure 2. The pipeline for developing predictive model using imbalanced dataset

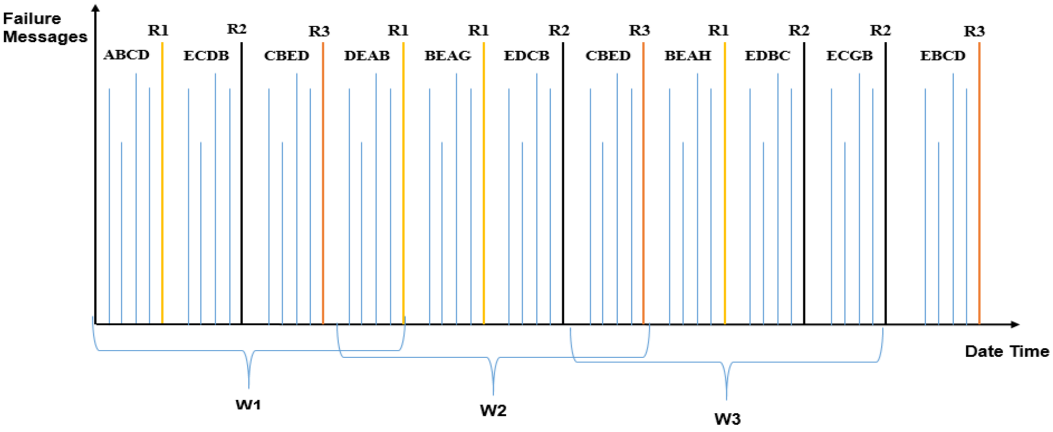


Figure 3. Failure message patterns- A, B, C... represents CMS failures messages and R1, R2... represents LRU replacements

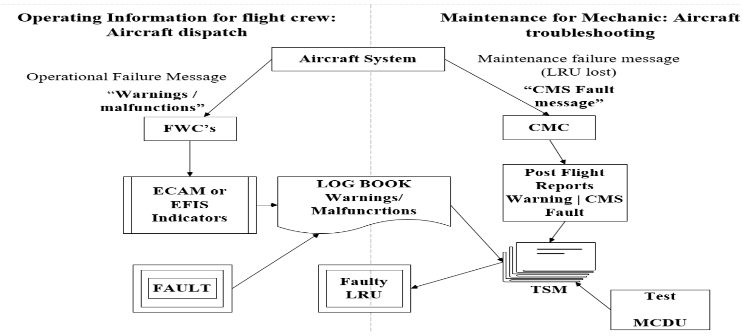
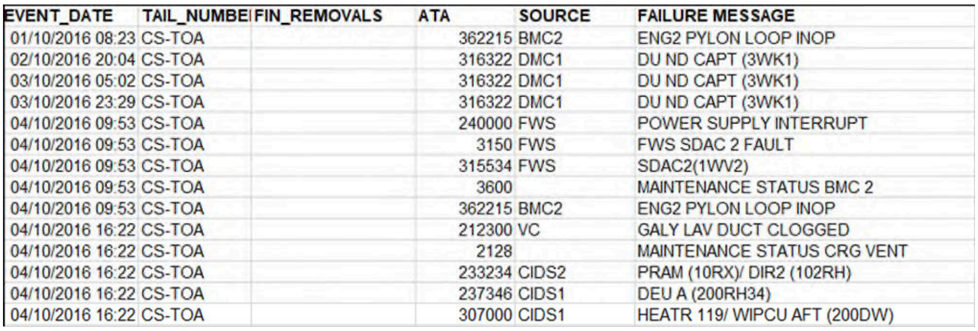

Operating Information for flight crew: Aircraft dispatch
Maintenance for Mechanic: Aircraft troubleshooting
Aircraft System
Operational Failure Message
"Warnings / malfunctions"
Maintenance failure message (LRU lost)
"CMS Fault message"
FWC's
CMC
ECAM or EFIS Indicators
LOG BOOK Warnings/ Malfuncrtions
Post Flight Reports Warning | CMS Fault
FAULT
Faulty LRU
TSM
Test
MCDU

Figure 4. Traditional Troubleshooting Philosophy in A330 CMS



EVENT_DATE	TAIL_NUMBE	FIN_REMOVALS	ATA	SOURCE	FAILURE MESSAGE
01/10/2016 08:23	CS-TOA		362215	BMC2	ENG2 PYLON LOOP INOP
02/10/2016 20:04	CS-TOA		316322	DMC1	DU ND CAPT (3WK1)
03/10/2016 05:02	CS-TOA		316322	DMC1	DU ND CAPT (3WK1)
03/10/2016 23:29	CS-TOA		316322	DMC1	DU ND CAPT (3WK1)
04/10/2016 09:53	CS-TOA		240000	FWS	POWER SUPPLY INTERRUPT
04/10/2016 09:53	CS-TOA		3150	FWS	FWS SDAC 2 FAULT
04/10/2016 09:53	CS-TOA		315534	FWS	SDAC2(1WV2)
04/10/2016 09:53	CS-TOA		3600		MAINTENANCE STATUS BMC 2
04/10/2016 09:53	CS-TOA		362215	BMC2	ENG2 PYLON LOOP INOP
04/10/2016 16:22	CS-TOA		212300	VC	GALY LAV DUCT CLOGGED
04/10/2016 16:22	CS-TOA		2128		MAINTENANCE STATUS CRG VENT
04/10/2016 16:22	CS-TOA		233234	CIDS2	PRAM (10RX)/ DIR2 (102RH)
04/10/2016 16:22	CS-TOA		237346	CIDS1	DEU A (200RH34)
04/10/2016 16:22	CS-TOA		307000	CIDS1	HEATR 119/ WIPCU AFT (200DW)

Figure 5. An example of a real CMS messages with event date, aircraft tail number, LRU, ATA reference number, and maintenance message.

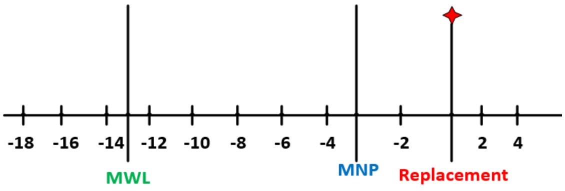


Figure 6. Representation of flight cycles from replacement

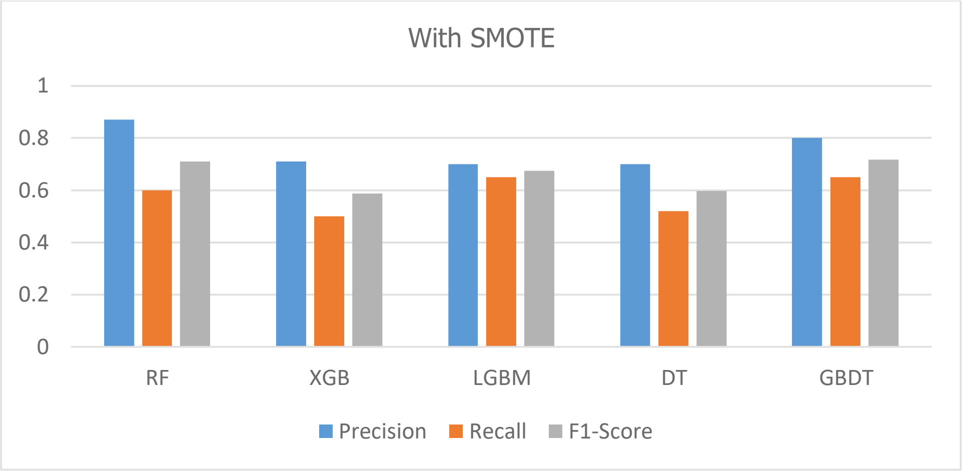


Figure 7. Showing the performance of ensemble-classifiers with SMOTE Using a multiclass approach

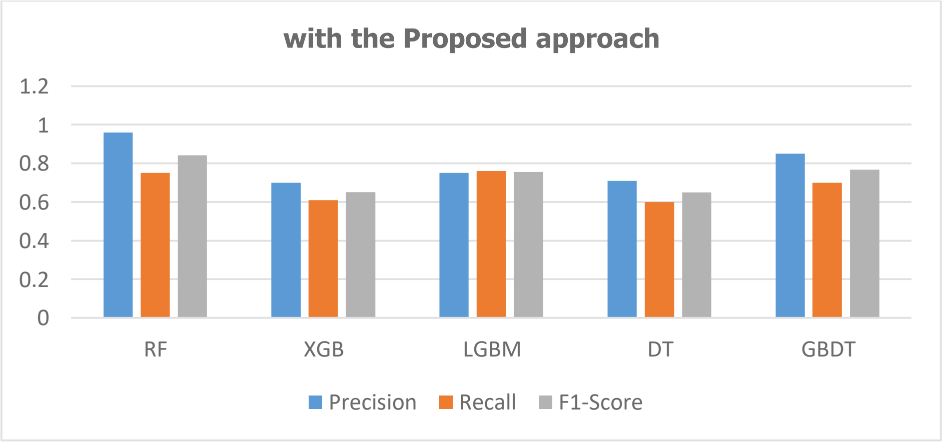


Figure 8. Showing the performance of ensemble-classifiers with the proposed approach Using a multiclass approach

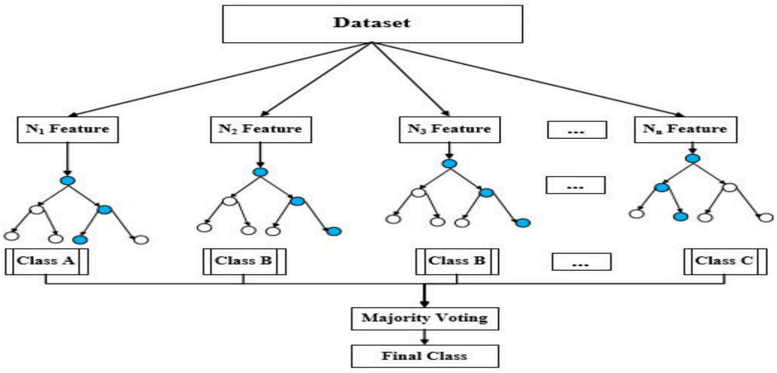
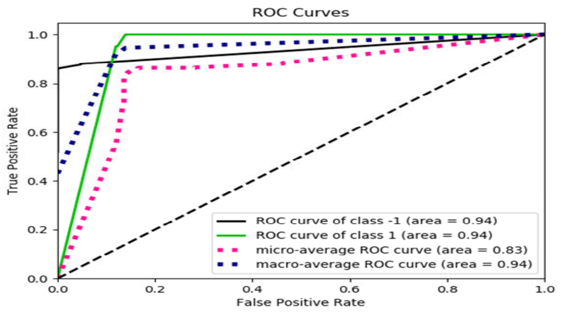


Figure 9. Random Forest Ensemble



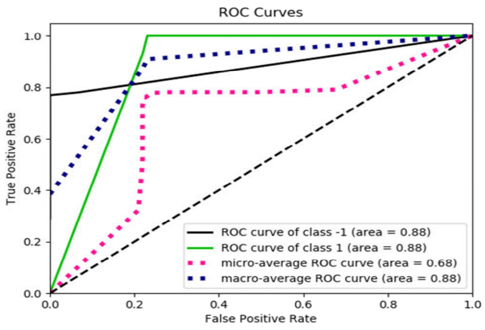
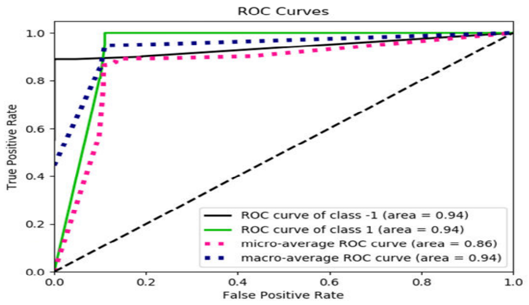


Figure 10 . ROC for 4001HA

Figure 11. ROC for 4000KS



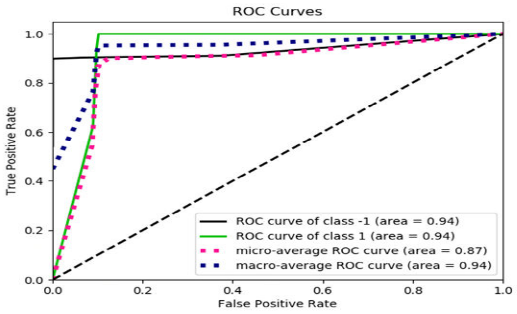


Figure 12. ROC for 5RV1

Figure 13. ROC for 438HC