

EXPLAINABLE NATURAL LANGUAGE PROCESSING FOR AIRCRAFT STRUCTURAL INTEGRITY DEFECT CLASSIFICATION

Vlado Kekoc[1], Michael J. Scott[2], Wim J.C. Verhagen[2] & Pier Marzocca[2]

[1]Defence Science and Technology Group, Fishermans Bend, Victoria, Australia
[2]RMIT University, Melbourne, Victoria, Australia

Abstract

Aircraft maintenance reports contain unstructured text that can reveal structural degradation, recurring defects, and emerging airworthiness risks. However, Natural Language Processing (NLP) use in continuing airworthiness is limited by concerns about transparency, interpretability, and trust. This paper presents an explainable NLP framework for classifying structural integrity issues in aircraft defect reports. From 6,453 Civil Aviation Safety Authority (CASA) records, 682 reports were retained across ten condition classes and ten aircraft part classes. Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and DistilBERT models were compared, with TF-IDF Random Forest (RF) providing the best balance of accuracy and interpretability. On 137 held-out test reports, it achieved 94.9% accuracy for part classification and 85.4% accuracy for condition classification. SHapley Additive exPlanations (SHAP) were used to identify the terms contributing to each prediction and were reviewed by two independent Aircraft Structural Integrity (ASI) experts. Given the small expert sample, the review is treated as exploratory. The ASI experts showed strong agreement in blind review and generally found the eXplainable Artificial Intelligence (XAI) outputs useful, but trust varied and did not always follow usefulness. Timing results were also mixed, with one expert showing a clearer review-time reduction than the other. Overall, the findings suggest that explainable lexical NLP models can provide accurate and auditable decision support for aircraft sustainment, while expert oversight remains essential for ambiguous reports, overlapping class boundaries and weak explanations. While the current work focuses on providing the immediate post-hoc interpretability required for practitioner trust in operational processes, it establishes the necessary feature attribution methods to facilitate a future transition toward Neuro-Symbolic Artificial Intelligence (NSAI). The findings show that explainable NLP can support aircraft defect-report classification, but expert validation remains essential where labels are ambiguous, explanations are generic, or model confidence does not align with engineering judgement.

Keywords: Classification, Defects, Explainable AI, Maintenance, Trustworthiness.

1. Introduction

1.1 Motivation & Significance

Across both civil and military aviation, the majority of fleet costs and safety risks are incurred during operations and maintenance rather than during design or manufacture [1], [2], [3], [4]. Maintenance logs, defect reports, pilot incident reports, and safety occurrence reports all contain valuable contextual

information on aircraft health. However, these data are largely unstructured and remain under-utilised at scale. As fleet complexity and reporting volumes increase, manual review processes struggle with timeliness and consistency. This can delay the identification of recurring defects, systemic issues, or early indicators of structural degradation. This challenge is particularly pronounced in Defence and legacy fleets, where aging platforms, service life extensions, and complex supply amplify the operational consequences of missed or delayed action on early warning signals.

Recent advances in natural language processing (NLP), including deep learning and transformer-based architectures, provide mechanisms to automate classification and prioritisation of maintenance and operational reports [5], [6]. However, aviation remains a highly regulated, safety-critical domain in which opaque “black-box” models are unsuitable for operational use. For such methods to be acceptable, they must support certification, regulatory assurance, and engineering judgement through transparent and interpretable behaviour. Adoption of a transparent and interpretable methodologies and tools also ensures alignment with Defence and regulatory authority policies that identify explainability as a core pillar of Responsible AI [7], [8].

This research looks to address these considerations by working towards deployment of trustable automated systems in airworthiness contexts. The authors of this paper have applied NLP and Machine Learning (ML) methods to aircraft maintenance and pilot reports in previous work [9], [10], [11], to identify lead indicators to major defects and towards automated defect risk analysis using fine-tuned Large Language Models (LLMs) on large datasets, however XAI has not been fully realised and critical for aircraft structural integrity where predictions must remain traceable, reviewable and trustworthy to users. In the remainder of this paper, the state of the art in this field is summarised in Section 1.2 and associated gaps are identified, informing several novel contributions as explicated in Section 1.3. Subsequently, the methodology underpinning the research is set out in Section 2 and model results are presented in Section 3. Findings are subsequently discussed around the ASI Expert Review in Section 4, informing the overall conclusions of the work as well as highlighting directions and recommendations for future work.

1.2 State-of-the-art

Despite significant progress in explainable AI (XAI) frameworks [12], a clear gap remains at the domain application level. Within the context of this research, the interest is primarily on the intersection of XAI and aircraft structural integrity. A review of recent XAI literature shows active investigation of explainability and interpretability within NLP, while aircraft-specific XAI research has largely concentrated on sensor-based models. A subset of recent literature in this area shows that few studies explicitly examine how explainable NLP outputs support continuing airworthiness decision-making, regulatory assurance, or confidence during audit and review processes [13]. The EASA AI Roadmap [14], frames aviation AI adoption around a human-centric trustworthiness framework, progressing from human augmentation and human-AI collaboration toward advanced automation, while expanding guidance to techniques such as reinforcement learning and symbolic AI. This is directly relevant to aerospace maintenance applications, where AI tools must support innovation without compromising safety, assurance, or regulatory traceability.

The studies summarised in Table 1 can be grouped into three main streams. First, recent general XAI research has advanced practical methods for explaining complex models and supporting trustworthy organisational decision-making, including emerging work on large language models [15], [16]. Second, aviation and maintenance NLP studies have demonstrated the value of extracting information from unstructured maintenance records, including aircraft maintenance narratives, short-form defect records, and equipment failure descriptions [17], [18], [19]. However, these studies generally focus on predictive or discovery performance rather than explainability, auditability, or human decision support. Third, aerospace XAI and prognostics research has applied methods such as SHAP to improve interpretability

in maintenance-related contexts, but primarily for sensor-based prognostics and health management rather than text-based defect classification [20], [21]. Defence focused AI research further reinforces the importance of trust, explainability, and integration for operational adoption, but remains largely strategic rather than implementation-focused [22]. Taken together, the literature indicates that XAI, aviation NLP, and aerospace maintenance analytics are active research areas, but their intersection remains underdeveloped, particularly for explainable NLP applied to structural integrity defect reports and continuing airworthiness decision-making.

Table 1 – Recent literature relevant to XAI considerations.

Ref.	Domain	Key Contribution	Limitation / Gap
[15]	XAI	A comprehensive guide to XAI for applied LLMs with real-world challenges with python code and addressing the data protection laws that demand 'right to explanation'	Practical Techniques, but a guide for applied research, without specific research to aircraft sustainment domain. Noting, the interpretability of LLMs remains an ongoing area of research filled with open questions.
[16]	Decision Support / XAI	Framed XAI as a prerequisite for trustworthy organisational decision-making across data, method, and application levels.	Not tailored to maintenance or airworthiness contexts.
[18]	Aviation NLP	Applied NLP to KC-135 aircraft maintenance records, demonstrating value of unstructured text.	Lacks explicit explainability and decision-level transparency.
[19]	Maintenance NLP	Used NLP and probabilistic methods to detect No Fault Found (NFF) events in short maintenance records.	Limited explainability and scalability to modern NLP models.
[17]	Maintenance Text Mining	Proposed unsupervised ensemble methods to discover failure mechanisms without labels.	Provides limited instance-level explanations for human decision-makers.
[20]	Aerospace XAI	Applied SHAP to interpret temporal feature importance in aerospace prognostics.	Restricted to sensor data; excludes textual maintenance narratives.
[21]	XAI for PHM	Linked SHAP explanations to classical prognostic metrics, reinforcing interpretability for maintenance decisions.	Does not address NLP or unstructured text.
[22]	Defence AI	Identified trust, explainability, and integration as critical enablers for Defence AI adoption.	Strategic perspective without concrete NLP/XAI implementations.

This research addresses several of the gaps identified in Table 1 by integrating NLP classification with the SHapley Additive exPlanations (SHAP) XAI framework, with a focus on identifying structural integrity indicators within aircraft maintenance reporting, which has broad application in aircraft Condition-Based Maintenance (CBM) [23], [24]. The resulting framework evaluates these explanations using objective, standardised evaluation metrics, including the use of a proposed XAI scoresheets [25], and with reference to aircraft domain language features. By establishing a human-in-the-loop (HITL) verification process, this work aims to bridge the divide between model performance and practitioner trust in environments characterised by high audit, compliance and aerospace industry regulatory requirements.

Feature attribution methods such as SHAP provide means of improving trust in automated maintenance record categorisation by identifying features that align with engineering knowledge and expertise [26]. This supports the adoption of transformer-based approaches [27]. However, a key challenge remains in that this may still be seen as post-hoc rationalisation of a "black-box" which lacks sufficient transparency and auditability required for safety-critical systems [28]. While post-hoc explainability techniques provide insight into the model behaviour, they do not demonstrate that decisions were derived through an engineering reasoning process.

Furthermore, post-hoc models may not be sufficient to enable quick decision making in critical situations, or to provide domain experts with sufficient information required to support or enable decisions required in critical situations [29]. Moreover, XAI model transparency may unintentionally increase overreliance by raising competency through trust in the AI and shifting perceived responsibility toward the system, thereby reducing users' motivation to critically scrutinise AI recommendations [30].

Neuro-symbolic AI, which combines neural learning with symbolic reasoning, has emerged as a promising approach for addressing the limitations of post-hoc techniques by providing greater traceability, explainability, and generalisations [31] and is a key technology to investigate in further expansion of this work, an argument which will be revisited in the conclusions and recommendations section. While [32] argues that Large Reasoning Models (LRMs) remain tenuous for high compliance and regulatory domains, where moderately complex tasks may perform well, but collapse under higher compositional complexity, even when they have sufficient token budget and, in some cases, explicit algorithms. As such, LLM-style reasoning is not a substitute for conventional algorithmic or symbolic methods, and a motivation for neurosymbolic AI approaches that combine neural pattern recognition with explicit planning and reasoning. This supports the need for neurosymbolic and explainable AI approaches that combine learned pattern recognition with explicit rules, traceable reasoning, and verifiable decision pathways.

Keeping these considerations and limitations in mind, this work represents the first stage of a "SHAP now, NSAI later" roadmap. Establishing explainability through feature attribution provides an immediate mechanism for user trust, while creating a pathway toward future neuro-symbolic techniques, which better align with the traceability, and auditability required by aerospace regulators and Defence policy.

1.3 Research Contributions

To operationalise the focus of the current research and realise the novel contributions associated with addressing the formerly identified gaps, this research addresses the following research questions:

1. How effectively can NLP models perform multi-class classification of structural integrity maintenance defect reports?

2. How well do SHAP-based explanations align with domain expert reasoning and continuing airworthiness practice?

3. To what extent do SHAP-based XAI outputs support trust, decision support, and efficiency for continuing airworthiness decision-makers?

The research objective is to develop and validate an explainable NLP multi-classification framework that supports transparent, auditable decision-making in aircraft sustainment and subsequently measure the performance. In Section 2, an overview will be given of the research methodology, which is followed by results in Section 3. After discussion of the implications and limitations of this work in the ASI expert review Section 4, conclusions are drawn and recommendations for future work are addressed.

2. Methodology

To address the objective of this research and answer the associated research questions, a four-stage methodology is proposed and is discussed in detail in Section 2.1. Subsequently, the key elements of the methodology are explored in additional detail in Sections 2.2 – 2.4, covering classification, explainability and associated performance metrics.

2.1 Overview

The proposed methodology comprises four stages, as shown in Figure 1.

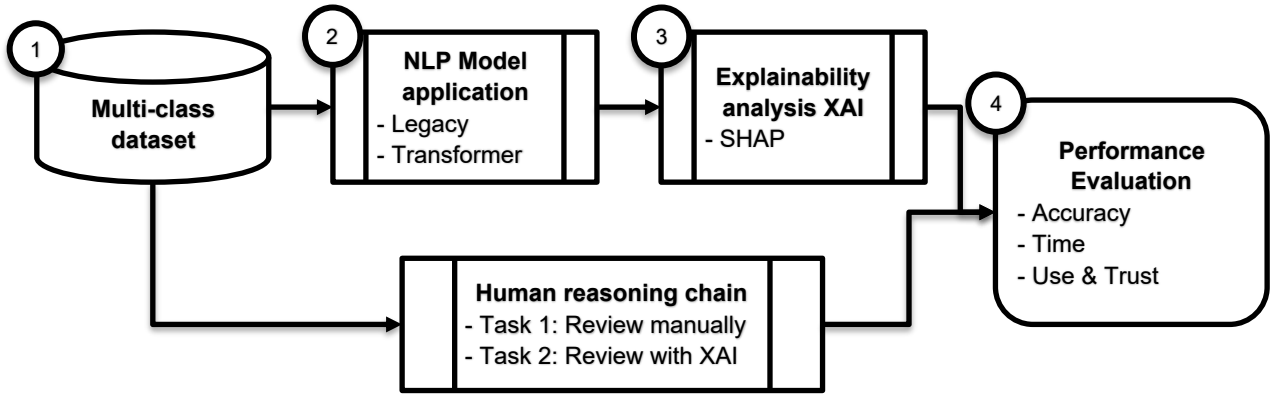


Figure 1 – Explainable NLP multi-classification framework.

The four stages can be briefly described as follows:

1. **Multi-class dataset:** this research addresses datasets comprising (semi-)unstructured technical reports, with multi-class labels to tag phenomena of interest. As an example of such datasets, consider maintenance reports capturing maintenance tasks and interventions on a particular asset such as an aircraft or component. Typically, such reports feature both structured data (e.g., on utilisation metrics such as flight time or cycles, or maintenance labels such as task IDs, aircraft zone(s) of application or ATA / iSpec chapter or JASC codes) and unstructured data, typically in the form of free-text entries where maintainers write down observations. Finally, as noted, these datasets often use multi-class labels to categorise data: for instance, a maintenance occurrence may be related to different failure mechanisms (e.g., wear, fatigue, or corrosion), which in turn may comprise different (sub)labels (e.g., corrosion may comprise galvanic, pitting, fretting, or stress corrosion, and many other subtypes of corrosion besides). These labels are typically applied by maintainers as part of their reporting processes. Finalising this first stage, it is typical for a level of data preprocessing to be required to bring a dataset in the proper format and shape for subsequent processing.

2. **NLP model application:** within the scope of this research, this stage focuses on the ability to perform multi-class text classification by using both legacy and transformer-based NLP models. These models will be set out in more detail in Section 2.2. below.

3. **Explainability analysis:** Explainability in machine learning refers to the ability to provide human-understandable justifications for model predictions and enabling trust for the end-user to verify the output validity. In aviation maintenance and continuing airworthiness, explainability is essential for operational acceptance and regulatory assurance compliance. As discussed, a range of techniques can be applied to provide explainability to the model outputs generated by the previous stage. Within the context of this research, the chosen approach is the SHAP explainable AI framework, which is applied to the multi-class classification outputs generated by the NLP model(s) applied in the previous stage. Section 2.3 details the reason for this selection.

4. **Performance evaluation:** to evaluate how well the models classify the report text (as compared to human-provided labels, which are generated as a ground truth and represented by the 'Human reasoning chain' process in Figure 1) *and* how well the justifications for these model outputs hold up, a range of metrics has been considered. A full set of metrics should cover not only technical performance characteristics such as accuracy but should also address process time in all its aspects (to enable a valid comparison between NLP outputs versus human-reasoned outputs, which includes processing time but also should include aspects such as verification / checking time). These considerations can be further translated into a cost perspective, by calculating the cost impacts of processing time, but also extending to business implications of true and false predictions. Finally, and crucially in the context of this research, explainability can be addressed by measuring decision-maker trust in the generated outputs. A more detailed overview of the considered and selected metrics is provided in Section 2.4.

The overall methodology focuses on establishing a baseline for explainable NLP. The simple and modular flow makes it possible to cater for future integration of additional or alternative models such as transformed-based approaches, explainability frameworks (such as NSAI) and evaluation metrics.

2.2 NLP model application for text classification

This research applies supervised NLP text-classification models to aircraft defect-report text in order to predict two multi-class labels: the reported defect condition and the affected aircraft part. The input text for each model is the codified CASA defect-report description, with the associated source labels used as target classes after filtering and harmonisation. The final evaluated dataset comprised ten condition classes and ten aircraft part classes, and separate classifiers were trained for each target.

Classification on multi-class problems can be performed using a variety of methods. Important considerations are problem complexity, scale and data availability. Given the nature of the dataset and its classification challenge, as well as the possibility to integrate with explainability frameworks, multiple approaches were considered in this work.

Three natural language processing approaches have been employed and tested: 1) a combination of Term Frequency-Inverse Document Frequency (TF-IDF) with a random forest classification; 2) Word2Vec document embeddings with a downstream classifier, where it should be noted that both 1) and 2) are examples of 'legacy' NLP methods; and 3) transformer-based embeddings (DistilBERT), which is an example of a transformer-based method, representing a larger class of Large Language Models (LLMs). These approaches were selected to compare classical sparse text representations, dense word-embedding methods, and contextual transformer-based language models.

1) TF-IDF was applied to represent the text data as numerical vectors. Each term in a document is assigned a weight based on its frequency in the document and its inverse frequency across the dataset, capturing both the importance of a word within a specific record and across all records. TF-IDF is a well-established technique for text classification tasks in its own right, but is typically used with other methods to increase its applicability. This is necessary as TF-IDF does not account for word order or context, which is crucial for understanding technical jargon and phrases in maintenance logs. In this research, TF-IDF was coupled with a subsequent random forest model to classify records into various defect categories (see Section 3.1). However, TF-IDF does not explicitly model word order or deeper semantic context. To perform classification, the TF-IDF vectors were passed to a Random Forest classifier, trained separately for condition and part prediction. Random Forest was selected because it is robust for tabular sparse features, handles nonlinear interactions between terms, provides class-probability outputs, and integrates effectively with SHAP explainability. In the final evaluation and ASI expert review tasks, this TF-IDF Random Forest model was the model used to generate the XAI-supported predictions, confidence values, and top contributing SHAP terms.

2) Word2Vec embeddings were considered as a classical dense-representation alternative. Word2Vec learns vector representations of words from their co-occurrence patterns in the corpus, meaning that words used in similar contexts are positioned closer together in the embedding space. For report classification, word vectors can be aggregated into document-level vectors and used with a downstream classifier. This approach has the advantage of capturing some semantic similarity between related maintenance terms, but it also loses detailed word order and can be sensitive to limited domain-specific training data, especially where rare technical phrases are important. In this study, Word2Vec was treated as a comparative "classical NLP" baseline alongside TF-IDF.

3) DistilBERT is a lightweight transformer-based model and was used to generate contextual embeddings for the textual entries in the dataset studied in this research. DistilBERT captures the semantic meaning of the text, providing a more nuanced representation than simple frequency-based approaches. DistilBERT is therefore in theory more suitable to apply towards complex, technical language. DistilBERT was fine-tuned on the preprocessed maintenance logs to classify the text directly. However, transformer models require greater computational resources, are more data-hungry, and are less straightforward to explain at the term level. Although SHAP can be applied to transformer models, this is computationally more expensive and less direct than SHAP explanations for Random Forest models.

As covered in more detail in Section 3, evaluation of these methods has shown that the combination of TF-IDF and Random Forest classification consistently yielded the best results. Results in Section 3 are discussed on the basis of using this approach. Importantly, this model also supported the HITL XAI study: for each held-out report, it produced predicted condition and part labels, probability-based confidence scores, and SHAP-derived terms that ASI experts could inspect when judging whether the model output was useful, trustworthy, or required correction.

2.3 Explainability analysis

A range of explainability techniques has been considered to explain model behaviour. These include rules-based decision trees, DeepLIFT, Integrated Gradients, attention-based methods, SHAP, LIME and NSAI. SHAP and LIME are typically seen as standard approaches for post-hoc interpretability and are both more broadly applied in aerospace contexts, as shown below.

1) SHapley Additive exPlanations (SHAP) provide a theoretically grounded, model-agnostic approach to feature attribution by quantifying each input feature's contribution to an individual prediction, using principles from cooperative game theory [33]. Within the aerospace context, SHAP has been applied to, amongst others, remaining useful life prediction for simulated aero-engine C-MAPSS data [34] and fault diagnosis for incident cognition [35] or specific components such as landing gears [33].

2) Local Interpretable Model-agnostic Explanations (LIME) approximate complex models locally using simpler surrogate models to explain individual predictions. Applications of LIME within aerospace include fault diagnosis [33], health monitoring [36] and air traffic management [37].

2.4 Performance evaluation

To evaluate the performance of applicable NLP models, a number of metric categories and associated metrics are considered for incorporation into this research.

- **Accuracy metrics for classification problems:** as this research focuses on classification performance, the 'traditional' metrics for classification apply; at its core, the confusion matrix delineates true and false positives and negatives, and a variety of ratios give rise to commonly used metrics such as accuracy, precision, recall and F1-score [38], [39]. In this study, the filtered dataset remained imbalanced, macro-averaged and weighted metrics were included alongside overall accuracy. Precision-recall curves were used as they better show model behaviour in imbalanced data, particularly for minority classes than overall accuracy alone [40], [41].

- **Time metrics:** while computational requirements for the NLP models are an important technical constraint to consider, the primary focus of time metrics in the context of this research is on overall processing time in relation to the manual and automatic forms of the problem applications. In concrete terms, this means that for manual tasks, the full process has to be mapped in terms of its constituent elements, such as preparation of data inputs, manual processing, labelling, and verification and validation as required.

  For the automated process, many of these steps would revert to the required computational time, but verification and validation may require additional attention if a model has been found to be susceptible to higher rates of false positives or negatives when compared to the manual alternative, or when a model is applied to a dataset with changes in characteristics when compared to training sets. The primary focus of time metrics in this research is on overall processing time in relation to manual and automated forms of the task. In concrete terms, manual processing was assessed through spreadsheet change-log intervals for ASI Expert A and B, including preparation, review, labelling, confirmation, and usefulness/trust responses where available.

- **Explainability and trust metrics:** In safety-critical aviation contexts, trustworthiness extends beyond model accuracy to encompass transparency, consistency, and the ability to support engineering judgement. For NLP-based decision support tools to be operationally acceptable, their outputs must be explainable in terms that align with established airworthiness practices and domain knowledge. This includes clear attribution of model decisions to relevant language features, predictable behaviour when exposed to new or evolving data, and traceability suitable for audit and regulatory review.

  Trustworthiness, in this sense, is achieved when engineers can understand, interrogate, and justify model-assisted assessments with the same rigour applied to traditional analytical methods, enabling informed use rather than blind reliance. Trust in NLP-based decision support systems can be assessed using a combination of qualitative and quantitative metrics that reflect both model behaviour and user confidence. Quantitative measures may include explanation stability across similar inputs, consistency of feature attributions under data perturbation, and alignment between model-highlighted terms and domain-relevant structural integrity indicators. Qualitative assessment can be supported through expert review, examining whether explanations are coherent, traceable, and usable within airworthiness decision workflows.

  Together, these metrics provide an evidence-based means of evaluating trustworthiness beyond classification performance alone. In this study, the XAI output provided the predicted condition and part labels, model confidence values, and top contributing SHAP terms for expert inspection. A confidence threshold of 0.8 was used operationally: predictions with predicted-label probability ≥ 0.8 were labelled “Certain”, while predictions below this threshold were labelled “Unsure”. Expert trust was then assessed quantitatively through agreement/disagreement with model labels, usefulness responses, and trust responses, and qualitatively through free-text expert rationale.

In the remainder of this work, classification accuracy metrics are applied in tandem with explainability and trust metrics to investigate whether XAI-supported NLP classification can improve the effectiveness of ASI report review while preserving expert oversight. This is particularly important where the model performs well statistically but may still produce explanations that experts judge as incomplete, generic, or inconsistent with aviation maintenance reasoning.

3. Model Results

The model results are based on a Civil Aviation Safety Authority (CASA) aircraft Defect Report Service (DRS) Portal[1] dataset containing a total of 6,453 DSRs, comprising of Airbus A320 (1,135 DSRs) and Boeing B737 (5,318 DSRs) maintenance defect records for the period of occurrence dates from 1 November 2004 to 15 November 2024, with each report including a DR number, occurrence date/time, aircraft type, part name, part condition, title, JASC system information, and a free-text defect description. The raw data contained 276 unique condition labels and 769 unique part labels. For modelling, labels were harmonised by merging smoke/fumes and odour into smoke/fumes/odour, and check valve into valve, reducing the harmonised label space to 275 conditions and 768 parts.

For this study, the analysis then filtered to structurally relevant/high-frequency categories: candidate conditions included cracked, leaking, corroded, contaminated, smoke/fumes/odour, seized, loose, worn, delaminated, missing, with bird strike excluded from the final paired modelling set because it did not remain after intersecting with the selected parts. It should be noted, that filtering to a subset of the data may limit the generalisability of the model to new broader maintenance reports, however the study is primarily aimed at supporting the analysis of high value and high frequency structural integrity defects. The final modelling dataset contained 682 reports across ten condition classes and ten-part classes: air conditioning, bearing, bleed air, bolt, oven, pump, seal, skin, valve, and window. The model results therefore provide a baseline for how well text-based machine learning can classify maintenance reports before comparing the outputs with ASI expert manual review and XAI SHAP supported HITL assessment.

The text input was the codified report description, which was lowercased, numeric and non-alphabetic characters were removed, whitespace was normalised, and the cleaned text was tokenised where required. For the TF-IDF Random Forest model, the cleaned report descriptions were represented using unigram and bigram TF-IDF features. TF-IDF was fitted only on the training split to avoid information leakage. The final training vocabulary contained 3,265 TF-IDF features after applying filters for only words or terms that appear in at least two reports, so a word that appears in only one document is removed and remove words or terms that appear in more than 95% of documents. So common words across the corpus are removed because they are not very informative.

The data was split using an 80-20 train-test design, randomly using a state variable set to 42. This produced 545 training reports and 137 unseen test reports. Stratification was performed by condition label because the combined condition-part label distribution was too sparse for reliable joint stratification. Model validation was performed using 5-fold stratified cross-validation on the training set only. Class imbalance was addressed using random oversampling within the training pipeline. The final Random Forest models used 200 trees, maximum depth of 20, and separate classifiers for condition and part prediction. The same train-test split was also used for the Word2Vec and DistilBERT comparisons to ensure a consistent evaluation across models.

Comparative classification results shown in Table 2, that the TF-IDF Random Forest model provided the strongest overall performance across both prediction tasks. For condition classification, TF-IDF Random Forest achieved the highest accuracy and weighted F1-score, indicating that sparse lexical features were highly effective for distinguishing defect condition labels in the maintenance report text. Word2Vec Random Forest also performed comparatively well, particularly for part classification, but showed a modest reduction in performance compared with TF-IDF, suggesting that averaged word embeddings captured useful semantic patterns but lost some discriminative term-level detail.

[1] Civil Aviation Safety Authority, *Defect Report Service Portal* (Web Page) https://drs.casa.gov.au/

DistilBERT produced substantially lower results for both condition and part classification, with weaker macro-average scores indicating difficulty in predicting minority classes. This is likely due to the relatively small and imbalanced labelled dataset, where transformer-based models require more data, tuning, or class-balancing strategies to generalise effectively. Overall, these results supported the selection of TF-IDF Random Forest as the most reliable baseline model for this dataset, while Word2Vec provided a reasonable alternative and DistilBERT would require further optimisation before deployment.

Table 2 – Models multi-class classification performance for the condition and part classification tasks on the held-out evaluation set of 137 maintenance reports.

Model	Feature Type	Target	Accuracy	Macro F1	Weighted Precision	Weighted Recall	Weighted F1
Random Forest	TF-IDF	Condition	0.85	0.79	0.86	0.85	0.85
Random Forest	TF-IDF	Part	0.95	0.96	0.95	0.95	0.95
Random Forest	Word2Vec	Condition	0.83	0.71	0.85	0.83	0.83
Random Forest	Word2Vec	Part	0.91	0.86	0.92	0.91	0.90
DistilBERT	Transformer	Condition	0.72	0.38	0.60	0.72	0.65
DistilBERT	Transformer	Part	0.74	0.51	0.69	0.74	0.67

The TF-IDF model achieved stronger performance for part classification than condition classification, as shown in Table 3. On the 137-report held-out test set, the part classifier achieved 94.9% accuracy, with weighted precision, recall, and F1-score of 0.952, 0.949, and 0.948. Macro F1 for part classification was 0.96. The condition classifier achieved 85.4% accuracy, with weighted precision, recall, and F1-score of 0.860, 0.854, and 0.854. Macro F1 for condition classification was 0.79. The combined exact-match accuracy, requiring both condition and part to be correct for the same report, was 83.2%.

Table 3 – TF-IDF Random Forest multi-class classification performance for condition and part.

Condition class	P	R	F1	N	Part class	P	R	F1	N
Contaminated	0.76	0.92	0.83	24	Airconditioning	0.94	0.79	0.86	19
Corroded	0.82	0.90	0.86	10	Bearing	1.00	1.00	1.00	3
Cracked	0.91	0.84	0.88	25	Bleed Air	0.85	0.96	0.90	23
Delaminated	0.67	0.67	0.67	3	Bolt	1.00	0.89	0.94	9
Leaking	1.00	1.00	1.00	14	Oven	0.92	1.00	0.96	12
Loose	0.60	0.60	0.60	5	Pump	1.00	1.00	1.00	5
Missing	0.80	0.80	0.80	5	Seal	1.00	1.00	1.00	9
Seized	0.95	0.95	0.95	19	Skin	1.00	1.00	1.00	24
Smoke/Fumes/Odour	0.88	0.77	0.82	30	Valve	0.95	1.00	0.98	21
Worn	0.50	0.50	0.50	2	Window	1.00	0.92	0.96	12
Overall Accuracy			**0.85**	**137**	**Overall Accuracy**			**0.95**	**137**
Macro average	**0.79**	**0.79**	**0.79**	**137**	**Macro average**	**0.97**	**0.96**	**0.96**	**137**
Weighted average	**0.86**	**0.85**	**0.85**	**137**	**Weighted average**	**0.95**	**0.95**	**0.95**	**137**

Note: P = precision, R = recall, F1 = F1-score, and N = class support. Metric definitions follow the scikit-learn `classification_report` convention, where support denotes the number of true instances for each class, macro average denotes the unweighted mean across labels, and weighted average denotes the support-weighted mean across labels [42].

The comparative results indicate that the TF-IDF Random Forest model provided the strongest balance between predictive performance and interpretability. Its superior performance suggests that the classification of aviation maintenance reports is driven largely by specific technical terms and short domain-specific phrases, which TF-IDF represents directly and effectively. In contrast, Word2Vec reduced these lexical signals into averaged dense embeddings, which may have diluted important class-discriminating terms, while DistilBERT was likely constrained by the limited and imbalanced training data available for fine-tuning across multiple condition and part classes.

The precision-recall curves shown in Figure 2 indicate that the TF-IDF Random Forest XAI models performed well above the one-vs-rest prevalence baseline of 0.10, with stronger and more consistent performance for part classification than condition classification. Precision-recall analysis was used because it directly evaluates the trade-off between precision, i.e., the proportion of predicted positives that are correct, and recall, i.e., the proportion of true positives retrieved, which is especially informative for imbalanced classification tasks [41]. Average Precision (AP) was used to summarise each precision-recall curve as a single score by weighting precision at each decision threshold by the corresponding increase in recall. For the multiclass one-vs-rest setting, micro-AP was calculated globally by pooling all class-label decisions, so it reflects overall retrieval performance and is more influenced by frequent classes. Macro-AP was calculated as the unweighted mean of the class-specific AP values, so each class contributes equally and the metric is more sensitive to rare class performance (these definitions follow standard implementations of AP and multi-class averaging[2]).

The part model achieved very high AP values, with micro-AP = 0.973 and macro-AP = 0.993, indicating that most component classes were retrieved with high precision across recall levels. The condition model was also strong overall, with micro-AP = 0.914 and macro-AP = 0.856, but showed greater variation across defect types: classes such as seized, leaking, cracked, corroded, and smoke/fumes/odour performed strongly, whereas rarer or more ambiguous classes such as delaminated, worn, and loose were less stable. For the XAI outputs, a confidence threshold of 0.8 was used operationally: predictions with predicted-label probability ≥ 0.8 were labelled “Certain”, and predictions below 0.8 were labelled “Unsure”. This threshold therefore provided a conservative HITL cue, helping ASI experts distinguish high-confidence model predictions from cases requiring closer manual review.

These findings have important implications for XAI in maintenance-report classification. The best-performing model was also the most transparent: TF-IDF Random Forest allows explanations to be traced back to individual words and bi-grams that contribute to each prediction. This makes SHAP-based explanations more accessible to domain experts, as influential features can be inspected as meaningful maintenance terms rather than abstract embedding dimensions or transformer token interactions. By contrast, although Word2Vec and DistilBERT offer richer semantic representations, their lower performance and reduced explanation clarity make them less suitable for this dataset without further optimisation.

Overall, the results support the use of TF-IDF Random Forest as the preferred modelling approach for this study. It provides strong classification accuracy, robust performance on the held-out test set, and interpretable feature level explanations that align with the needs of aviation maintenance decision support. The TF-IDF Random Forest model outputs were then used in the subsequent ASI expert review for comparison and SHAP assisted terms.

[2] ‘average_precision_score - scikit-learn 1.9.0 documentation’ https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html



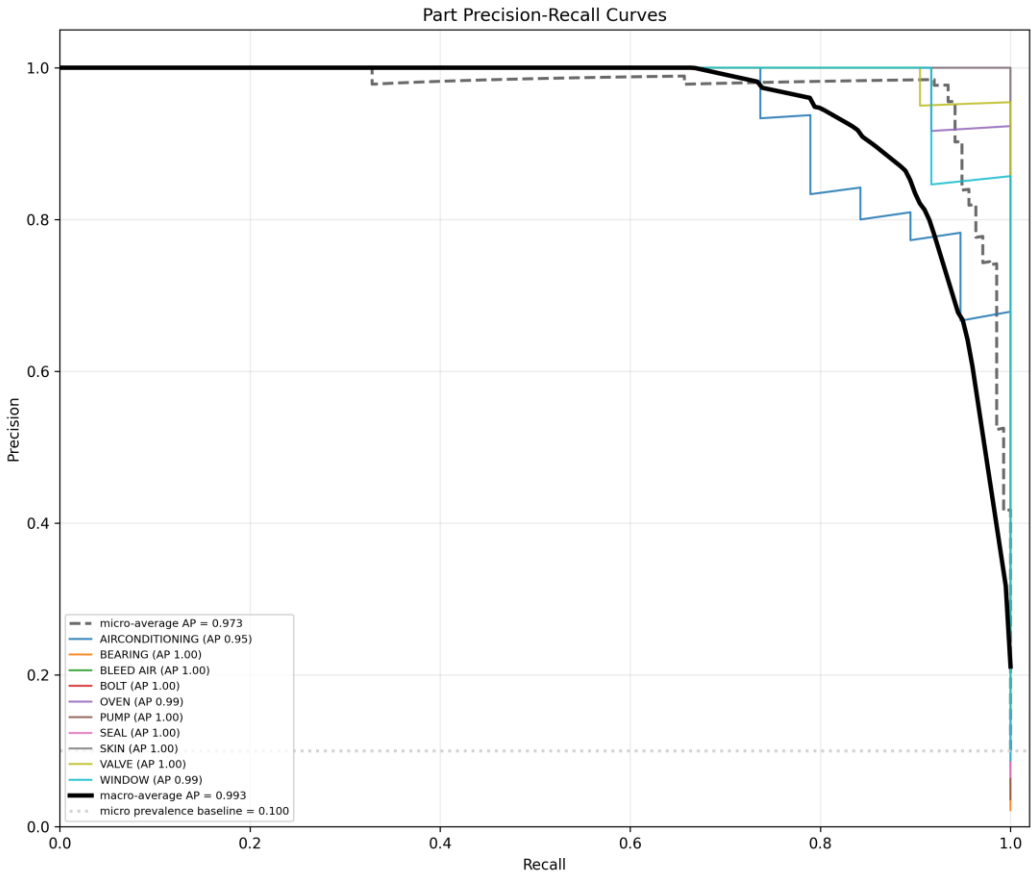


Figure 2 – XAI model outputs Precision-Recall charts for Condition and Part classifications.

4. ASI Expert Review

Explainable NLP approaches can bridge a critical gap between AI capability and operational trust in aviation maintenance contexts. This section examines the independent ASI expert reviews conducted to provide a domain informed benchmark for assessing the XAI-assisted defect report classification outputs. The two ASI experts were aircraft structural integrity domain specialists who were independent of the study team and not involved in model development, providing an external expert assessment of the defect reports and model explanations. Two datasets were prepared. In Task 1, the experts completed a blind review of defect reports by manually assigning the most appropriate structural integrity condition and part labels, rating their confidence, identifying the top terms that informed their judgement, and documenting their reasoning. In Task 2, the experts reviewed the XAI model outputs, including predicted labels, confidence ratings, and SHAP-derived explanatory terms, then indicated whether they agreed or disagreed with the model classifications, provided corrected labels where required, and assessed whether the XAI evidence was useful and trustworthy.

This process allowed the study to compare automated model predictions and explanations against independent expert reasoning. The expert review therefore served two purposes: i) to establish a human domain benchmark for condition and part classification; and ii) to evaluate whether the XAI outputs aligned with the evidence that ASI experts considered meaningful in the report text. By separating blind expert labelling from the subsequent XAI review task, the process helped distinguish expert judgement based solely on maintenance report content from expert judgement informed by model predictions and explanations. Moreover, by providing interpretable outputs that align with airworthiness reasoning, the approach supports earlier identification of structural risks and improves confidence in the use of automated decision-support tools.

Taken together, the model evaluation and expert study indicate that TF-IDF Random Forest classifiers can classify aircraft maintenance reports effectively, particularly for affected part labels. SHAP explanations made the model outputs inspectable by exposing the key textual cues behind each prediction. However, condition classification remains more difficult, especially for rare or semantically overlapping defect categories. The expert validation therefore supports the use of XAI as decision support rather than full automation: the model can prioritise likely labels and provide evidence, while ASI experts remain essential for resolving ambiguous condition interpretations, while part classification is clearer.

The models are described as XAI-supported because each prediction was accompanied by SHAP-based feature attribution. For every held-out report, SHAP TreeExplainer[3] was applied to the Random Forest classifier. The explanation was class-specific: SHAP values were calculated for the predicted condition class and predicted part class. The system then ranked the TF-IDF terms present in the report by their absolute SHAP contribution and selected the top three terms for each target. This reduced thousands of model features to a small, human reviewable explanation consisting of the three strongest uni- or bi-gram contributors.

These SHAP terms were the basis of the expert-facing XAI task. Experts did not only see the predicted label; they also saw the model confidence and the top three explanatory terms. Model confidence was derived from the predicted class probability and translated into “Certain” when probability was at least 0.80, otherwise “Unsure”. This confidence signal was meaningful in the expert review: “Certain” predictions had higher expert agreement than “Unsure” predictions for both condition and part classification.

[3] SHAP Developers, “shap.TreeExplainer,” SHAP Latest Documentation, https://shap.readthedocs.io/en/latest/generated/shap.TreeExplainer.html

The comparison, in Table 4, suggests that the XAI model aligned most closely with the harmonised source labels, achieving 85.4% condition agreement, 94.9% part agreement, and 83.2% exact agreement across both labels. This is expected to some extent, because the model was trained to reproduce the source-data labelling structure, whereas the ASI experts were applying independent domain judgement. The strongest and most consistent result across all sources is part classification: the XAI model, Expert A, and Expert B, all performed better on part labels than condition labels. This indicates that affected components are generally more explicitly signalled in the report text, while defect conditions require more interpretation.

Table 4 – Comparison of Condition and Part actual, XAI model, Expert A, Expert B labels for match rates against source actual labels in dataset.

Source	Condition labels available	Condition matches actual	Condition match	Part labels available	Part matches actual	Part match	Both labels available	Both labels match actual	Both-label match
XAI Model	137	117	85.4%	137	130	94.9%	137	114	83.2%
ASI Expert A - Task 1	80	57	71.2%	76	64	84.2%	76	53	69.7%
ASI Expert B - Task 1	60	45	75.0%	60	50	83.3%	60	41	68.3%
ASI Expert A - Task 2	69	52	75.4%	52	52	100%	51	42	82.4%
ASI Expert B - Task 2	60	45	75.0%	60	50	83.3%	60	41	68.3%

The expert results show that manual classification is not simply a matter of matching keywords to labels. Expert A and Expert B matched the source condition labels in 71.2% and 75.0% of their Task 1 responses, respectively, while part label agreement was higher at 84.2% and 83.3%. The lower condition alignment likely reflects ambiguity in defect descriptions, especially around overlapping categories such as contamination, fumes/odour, corrosion, seizure, looseness, and wear, as shown in Figure 3. These are cases where an expert may choose a different but defensible label from the original source classification.

This performance pattern was reflected in Task 2, both experts agreed more often with the model's part predictions than its condition predictions. Expert A agreed with 87.7% of confirmed part predictions compared with 71.8% of confirmed condition predictions. Expert B agreed with 80.0% of confirmed part predictions compared with 66.7% of confirmed condition predictions. Thus, the human review confirmed the same trend observed in the held-out test metrics: part classification was more reliable than condition classification. The lower condition performance appears to be associated with semantic overlap between condition classes and class imbalance. High-performing condition classes included leaking, seized, cracked, corroded, and contaminated, while lower performance was observed for sparse classes such as worn, loose, and delaminated. Expert disagreements in Task 2 were also concentrated around ambiguous condition boundaries, particularly cases where the model predicted contaminated, but experts preferred smoke/fumes/odour.

Table 5 shows that Expert A analysed a larger share of the 137 reports than Expert B in both tasks: 58.4% versus 43.8% in Task 1, and 52.6% versus 43.8% in Task 2. However, Expert A's Task 2 data were less complete, with 56 condition-and-part confirmations and 44 usefulness/trust responses. Expert B completed all 60 reports in both tasks. Normalised by task-specific report counts, Task 2 was faster for both experts, but the magnitude differed substantially: Expert A improved only slightly from 1.71 to 1.68 min/report, while Expert B reduced from 1.92 to 1.40 min/report, a 27.1% reduction. This suggests the XAI-supported workflow produced the clearest efficiency gain for Expert B, while Expert A's Task 2 timing should be interpreted cautiously because the response set was incomplete.

Table 5 – ASI Expert review and analysis timings breakdown.

Metric	ASI Expert A	ASI Expert B
Task 1 reports analysed	80, 58.4% of 137	60, 43.8% of 137
Task 1 time	137 min	115 min
Task 1 min/report	1.71	1.92
Task 2 reports analysed	72, 52.6% of 137	60, 43.8% of 137
Task 2 time	121 min	84 min
Task 2 min/report	1.68	1.4
Task 2 complete condition+part	56, 77.8% of Task 2	60, 100.0% of Task 2
Task 2 usefulness/trust responses	44, 61.1% of Task 2	60, 100.0% of Task 2
Final total active time	258 min, 4h 18m	182 min, 3h 02m

Shown in Figure 3, the expert confusion matrices show similar patterns to the XAI model, which is important. ASI experts also tended to confuse or reinterpret contaminated vs smoke/fumes/odour, seized vs corroded, and bleed air vs air conditioning. This suggests that some disagreements are caused by label ambiguity rather than model failure alone. Overall, the matrices support the interpretation that the XAI model is reliable for part identification, more challenged by condition classification, and best used as a decision-support tool rather than as an unchecked final classifier.

The XAI-supported task appears to have helped most clearly for Expert A. Expert A's derived Task 2 final labels improved from 69.7% both-label agreement in Task 1 to 82.4% in Task 2, approaching the model's 83.2% exact-match result. Part classification was especially strong for Expert A in Task 2, with all available final part labels matching the source actual values. However, this should be interpreted cautiously because Task 2 final labels depend on completed confirmation fields, and Expert A did not complete all rows. Expert B's Task 2 final agreement was effectively similar to their Task 1 performance, suggesting that the XAI output did not substantially shift their final classifications.

Shown in Table 6, for Task 2, both experts generally found the XAI-supported outputs useful, but usefulness did not always translate into trust. Expert A rated 79.5% of responses as useful but trusted only 45.5%, producing a large usefulness-trust gap of 34.1 percentage points. This suggests the XAI evidence often supported review but was not sufficient for confidence in the model decision. Expert B showed stronger trust alignment, rating 81.7% useful and 71.7% trusted, with only a 10.0 percentage-point gap. This indicates Expert B was more willing to accept useful XAI-supported predictions as trustworthy, while Expert A treated them more cautiously as decision support.

Table 6 – ASI Expert review and analysis usefulness and trustworthiness ratings.

Evaluation	ASI Expert A		ASI Expert B	
	count	*percentage*	*count*	*percentage*
Task 2 reports analysed	72/137	52.6%	60/137	43.8%
Usefulness/trust response pairs analysed	44	61.1%	60	100%
Rated useful	35/44	79.5%	49/60	81.7%
Rated not useful	9/44	20.5%	11/60	18.3%
Rated trusted	20/44	45.5%	43/60	71.7%
Rated not trusted	24/44	54.5%	17/60	28.3%
Useful and trusted	20/44	45.5%	43/60	71.7%
Useful but not trusted	15/44	34.1%	6/60	10.0%

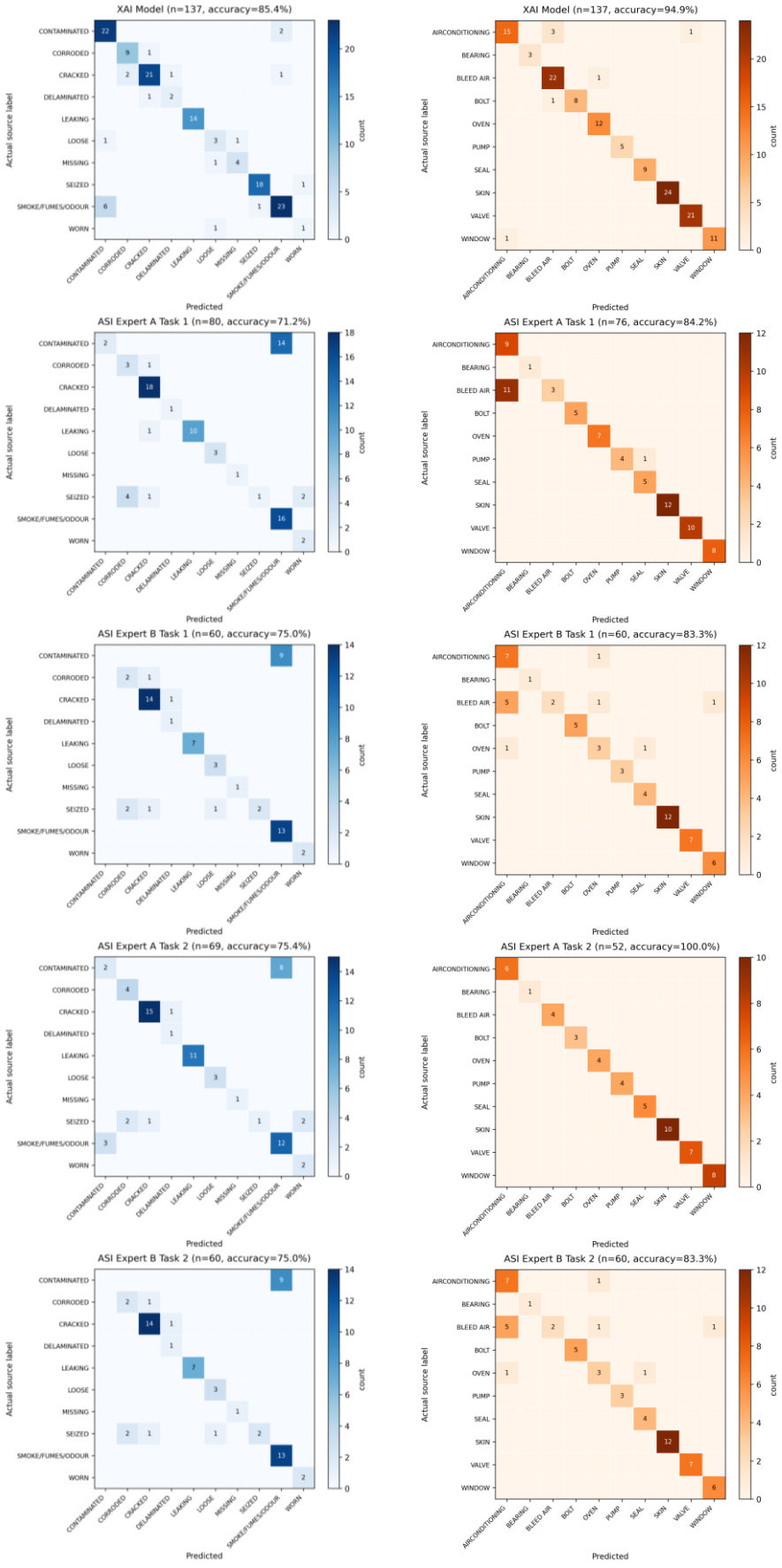


Figure 3 – Confusion Matrices for Condition and Part comparing the actual classifications with predicted for XAI Model and ASI Experts.

The examples in Table 7 show why the HITL review was essential. In straightforward cases, the experts accepted the XAI-supported labels as useful and trustworthy when SHAP terms aligned with the report evidence. However, several cases show that source-label correctness alone did not guarantee trust: experts sometimes used the XAI output as decision support but overrode the model when the SHAP terms appeared generic, when the model seemed to default to contaminated/seized/corroded labels, or when a more specific maintenance interpretation was needed. The HITL process therefore acted as a safeguard by identifying model errors, challenging weak explanations, and exposing mismatches between source labels, model outputs, and expert engineering judgement.

Table 7 – Examples of reports with ambiguity.

Combination	Expert	Complete Report	Actual Source Labels	XAI Model Output	ASI Expert Input
Useful + trusted; labels accepted	A	LH and RH wing upper and lower trailing edge skins extensively corroded. Found during inspection iaw SB A320-57-1154 and SB A320-57-1155.	Condition: CORRODED; Part: SKIN	Condition: CORRODED; SHAP: wing, corroded, lower. Part: SKIN; SHAP: lower, skins, wing.	Condition: Agree; Part: Agree; Useful: Yes; Trust: Yes
Useful + trusted; expert refines condition	B	5th stage check valve failed inspection as per AD 2020-16-51. Valve flappers had corrosion around the bushes, thus, reducing bushing clearance below 0.004". Defective check valve will be replaced, and part will be sent out for repair/overhaul.	Condition: SEIZED; Part: VALVE	Condition: SEIZED; SHAP: valve, check valve, th stage. Part: VALVE; SHAP: valve, th, stage.	Condition: Disagree, corrected to CORRODED; Part: Agree; Useful: Yes; Trust: Yes
Useful + not trusted; condition disagreement	B	Flight Crew reported dual bleed light illuminated on taxi to gate followed by electrical burning smell. Smell dissipated after Engines Shutdown. Troubleshooting carried out could not reproduce smell. Engines Ground Run carried out using Engines and APU as bleed sources in accordance with AMM. Different bleed sources selected with no abnormal smells noted. There is no history and there have been no further occurrences. Considered transitory. Event monitoring will continue.	Condition: SMOKE/FUMES/ODOUR; Part: BLEED AIR	Condition: CONTAMINATED; SHAP: smell, crew, crew reported. Part: BLEED AIR; SHAP: bleed, smell, crew.	Condition: Disagree, corrected to SMOKE/FUMES/ODOUR; Part: Agree; Useful: Yes; Trust: No
Useful + not trusted; source label correct but expert wanted specificity	B	Electrical burning smell from aft galley. No1 and No2 ovens quarantined for further investigation. Both ovens were sent for repair and both ovens were found to have damaged door seals. The oven units were repaired with new seals.	Condition: SMOKE/FUMES/ODOUR; Part: OVEN	Condition: SMOKE/FUMES/ODOUR; SHAP: galley, smell, oven. Part: OVEN; SHAP: oven, galley, the oven.	Condition: Agree; Part: Disagree, corrected to SEAL; Useful: Yes; Trust: No
Not useful + not trusted; condition disagreement	B	Cabin Crew reported burning smell, possibly originating from the AFT galley. Smell disappeared when Galley Oven #1 selected off. Flight continued with no further issues. Troubleshooting traced smell to a small piece of plastic label found behind the oven insert on the Fan Grill. Plastic removed and area cleaned. Oven operated satisfactory with nil further issues. There is no history of defect and there have been no repetitive occurrences. Considered isolated.	Condition: SMOKE/FUMES/ODOUR; Part: OVEN	Condition: CONTAMINATED; SHAP: smell, crew, crew reported. Part: OVEN; SHAP: oven, galley, galley oven.	Condition: Disagree, corrected to SMOKE/FUMES/ODOUR; Part: Agree; Useful: No; Trust: No
Not useful + not trusted; condition and part disagreement	B	Cabin crew harness attachment bolt came out. Bolt found and refitted.	Condition: LOOSE; Part: BOLT	Condition: CONTAMINATED; SHAP: crew, cabin, bolt. Part: BLEED AIR; SHAP: bolt, crew, cabin.	Condition: Disagree, corrected to LOOSE; Part: Disagree, corrected to BOLT; Useful: No; Trust: No

Overall, the results support the value of XAI as a decision-support tool rather than a replacement for expert judgement. The model is effective at reproducing the source labelling scheme, particularly for parts, and the SHAP terms appear to help experts validate or challenge predictions. However, expert disagreement with source labels should not automatically be treated as error; it may reveal ambiguity or limitations in the original labels. The strongest insight is that XAI can improve transparency and accelerate review, but condition classification remains the key challenge because it depends on nuanced maintenance interpretation rather than surface text alone.

5. Conclusion

This paper developed and evaluated an explainable NLP framework for classifying aviation structural integrity defect reports. The results show that TF-IDF Random Forest provided the strongest overall performance, achieving approximately 0.85 accuracy for condition classification and 0.95 accuracy for part classification on the held-out test set. This suggests that the task is driven largely by specific maintenance terms and short technical phrases, which TF-IDF represents effectively and SHAP can explain directly.

The findings answer the three research questions: Firstly, by showing that NLP models can classify structural integrity reports effectively, but that model complexity does not guarantee better performance. Word2Vec performed reasonably but lost some discriminative detail by averaging word embeddings. DistilBERT performed worse, particularly for minority classes, indicating that the available dataset was too small and imbalanced for reliable transformer fine-tuning. Secondly, SHAP explanations were useful because they linked predictions to interpretable words and bi-grams. This made the TF-IDF Random Forest model more transparent than embedding or transformer-based alternatives. However, SHAP explanations were not always sufficient: ambiguous reports, overlapping labels, and weak explanatory terms still required ASI expert judgement. Thirdly, the results indicate that XAI can support trust and decision-making, but only as a decision aid rather than an autonomous authority. SHAP outputs can improve auditability and help experts assess whether predictions are plausible, but trust remains conditional on explanation quality, model confidence, and domain context.

Overall, the study finds that the most accurate model was also the most explainable. TF-IDF Random Forest with SHAP therefore offers the most practical approach for this dataset, balancing performance, transparency, and usability. The key limitation is that the expert review and labelled dataset remain modest in size, limiting generalisability and the performance of more complex models.

Future work should expand the number of independent ASI experts and reviewed reports, measure inter-rater agreement, and test whether XAI explanations genuinely improve decision quality and review efficiency. Larger and more balanced datasets should also be used to reassess transformer and LLM-based approaches, which may become more competitive with sufficient domain-specific training data. This XAI framework is positioned as a foundational baseline for a long-term research roadmap: "SHAP now, NSAI later", with future aims to move beyond statistical pattern recognition with hybrid architectures that integrate symbolic reasoning and deterministic constraints, eventually providing rigorous traceability required to meet regulatory compliance as credible data.

Acknowledgements

Thank you to the Capability Acquisition and Sustainment Group (CASG) for their ongoing funding support in this collaborative research. The authors would also like to thank the Australian Defence Science and Technology Group (DSTG) acknowledging the contributions of Andrew Walliker & Douglas Williams for providing their independent ASI domain expertise.

6. Contact Author Email Address

mailto: vlado.kekoc@defence.gov.au

7. Copyright Statement

The authors confirm that they, and/or their company or organization, hold copyright on all of the original material included in this paper. The authors also confirm that they have obtained permission, from the copyright holder of any third party material included in this paper, to publish it as part of their paper. The authors confirm that they give permission, or have obtained permission from the copyright holder of this paper, for the publication and distribution of this paper as part of the ICAS proceedings or as individual off-prints from the proceedings.

References

[1] U.S. Government Accountability Office, 'Weapon System Sustainment: Aircraft Mission Capable Rates Generally Did Not Meet Goals and Cost of Sustaining Selected Weapon Systems Varied Widely', Washington, DC, Nov. 2020. [Online]. Available: https://www.gao.gov/assets/gao-21-101sp.pdf

[2] F. Camm *et al.*, 'Inclusion of Operating and Support Costs in Competitive Air Force Source Selections: Challenges and Tools for Implementation', 2025, Accessed: Jul. 01, 2026. [Online]. Available: www.rand.org/about/research-integrity.

[3] T. Mofokeng, P. T. Mativenga, and A. Marnewick, 'Analysis of aircraft maintenance processes and cost', *Procedia CIRP*, vol. 90, pp. 467–472, Jan. 2020, doi: 10.1016/j.procir.2020.01.115.

[4] J. Baker, A. FRAeS MIEAust CPEng, J. Cockram BEng CEng FRAeS, Gmh. BEngCEng FRAeS, and R. Hornby, 'Solving the Mystery of the Disappearing Symptom and Why it Benefits the Product Life-Cycle', in *17th Australian International Aerospace Congress : AIAC 2017*, Barton: Engineers Australia, Royal Aeronautical Society, 2017, pp. 738–744. doi: 10.3316/INFORMIT.741329838238681.

[5] J. Ricketts, D. Barry, W. Guo, and J. Pelham, 'A Scoping Literature Review of Natural Language Processing Application to Safety Occurrence Reports', *Safety*, vol. 9, no. 2, p. 22, Jun. 2023, doi: 10.3390/safety9020022.

[6] K. R. Chowdhary, 'Natural Language Processing', in *Fundamentals of Artificial Intelligence*, K. R. Chowdhary, Ed., New Delhi: Springer India, 2020, pp. 603–649. doi: 10.1007/978-81-322-3972-7_19.

[7] Australian Government Department of Defence, 'Policy Settings for Responsible Use of Artificial Intelligence in Defence', 2025. Accessed: Jul. 01, 2026. [Online]. Available: https://www.defence.gov.au/sites/default/files/2026-03/Policy-Settings-for-Responsible-Use-of-Artificial-Intelligence-in-Defence-%5BOFFICIAL%5D.pdf

[8] Federal Aviation Administration, 'Roadmap for Artificial Intelligence Safety Assurance', Jul. 2024. Accessed: Jul. 01, 2026. [Online]. Available: https://www.faa.gov/aircraft/air_cert/step/roadmap_for_AI_safety_assurance

[9] M. J. Scott, V. Kekoc, W. J. C. Verhagen, B. Teunisse, and P. Marzocca, 'Natural Language Processing of Pilot and Maintenance Reports for Proactive Aircraft Sustainment', in *ADSTAR 2024 Conference Proceedings*, Canberra, Australia: Defence Science and Technology Group, Sep. 2024, p. 3.

[10] M. J. Scott *et al.*, 'Application of Natural Language Processing for Aircraft Defect Tracking in Maintenance Operations', in *Proceedings of the 34th Congress of the International Council of the Aeronautical Sciences (ICAS)*, Florence, Italy, Sep. 2024, p. 12. [Online]. Available: https://www.icas.org/ICAS_ARCHIVE/ICAS2024/data/papers/ICAS2024_0083_paper.pdf

[11] M. J. Scott *et al.*, 'Automated Aircraft Defect Risk Analysis Utilising Maintenance and Pilot Reports', in *AIAC 2025: 21st Australian International Aerospace Congress*, Melbourne, Australia, 2025.

[12] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, 'Explaining Explanations:

An Overview of Interpretability of Machine Learning', Feb. 2019, *arXiv*. doi: 10.48550/arXiv.1806.00069.

[13] K. Feng, W. Ma, D. Lu, H. Li, and D. Li, 'Analysis of Emergency Decision-Making Patterns in Civil Aviation Risk Events Based on the Observe-Decide-Act Model', *Reliab. Eng. Syst. Saf.*, p. 112177, Dec. 2025, doi: 10.1016/j.ress.2025.112177.

[14] European Union Aviation Safety Agency, 'EASA's First Regulatory Proposal on Artificial Intelligence for Aviation Is Now Open for Consultation', Nov. 2025. [Online]. Available: https://www.easa.europa.eu/en/newsroom-and-events/news/easas-first-regulatory-proposal-artificial-intelligence-aviation-now-open

[15] W. Hsieh *et al.*, 'A Comprehensive Guide to Explainable AI: From Classical Models to LLMs', Dec. 2024, *arXiv*. doi: 10.48550/arXiv.2412.00800.

[16] K. Coussement, M. Z. Abedin, M. Kraus, S. Maldonado, and K. Topuz, 'Explainable AI for enhanced decision-making', *Decis. Support Syst.*, p. 114276, Jun. 2024, doi: 10.1016/j.dss.2024.114276.

[17] A. S. Bhardwaj and D. Veeramani, 'An unsupervised multi-class ensemble classifier for identifying equipment failure mechanisms from maintenance records', *Reliab. Eng. Syst. Saf.*, p. 111410, Jul. 2025, doi: 10.1016/j.ress.2025.111410.

[18] E. Joback *et al.*, 'Application of Natural Language Processing Techniques to KC-135 Aircraft Maintenance Data', *Military Operations Research*, vol. 28, no. 3, pp. 67–86, 2023, [Online]. Available: https://www.jstor.org/stable/27254916

[19] J. G. Pelham and C. Hockley, 'Analysis of Short form Maintenance Records for NFF Using NLP, Phrase Matching, and Bayesian Learning', *Procedia CIRP*, vol. 59, pp. 257–262, Jan. 2017, doi: 10.1016/j.procir.2016.10.123.

[20] Y. Alomari and M. Andó, 'SHAP-based insights for aerospace PHM: Temporal feature importance, dependencies, robustness, and interaction analysis', *Results in Engineering*, vol. 21, p. 101834, Mar. 2024, doi: 10.1016/j.rineng.2024.101834.

[21] M. L. Baptista, K. Goebel, and E. M. P. Henriques, 'Relation between prognostics predictor evaluation metrics and local interpretability SHAP values', *Artif. Intell.*, vol. 306, p. 103667, May 2022, doi: 10.1016/j.artint.2022.103667.

[22] G. Moy, S. Shekh, M. Oxenham, and S. Ellis-Steinborner, 'Recent Advances in Artificial Intelligence and their Impact on Defence DST-Group-TR-3716', Canberra, Apr. 2020. Accessed: Jul. 11, 2026. [Online]. Available: https://www.dst.defence.gov.au/sites/default/files/publications/documents/DST-Group-TR-3716_0.pdf

[23] H. Bokinsky *et al.*, 'Application of natural language processing techniques to marine V-22 maintenance data for populating a CBM-oriented database', pp. 463–472, Jan. 2013.

[24] D. A. Nguyen, S. Jose, T. P. K. Nguyen, and K. Medjaher, 'Explainable multimodal learning for predictive maintenance of steam generators', *PHM Society Asia-Pacific Conference*, vol. 4, no. 1, Sep. 2023, doi: 10.36001/phmap.2023.v4i1.3783.

[25] M. Winikoff, J. Thangarajah, and S. Rodriguez, 'A Scoresheet for Explainable AI', Feb. 2025, *arXiv*. doi: 10.48550/arXiv.2502.09861.

[26] S. M. Lundberg and S. I. Lee, 'A Unified Approach to Interpreting Model Predictions', *Adv. Neural Inf. Process. Syst.*, vol. 2017-December, pp. 4766–4775, May 2017, Accessed: Jul. 02, 2026. [Online]. Available: https://arxiv.org/pdf/1705.07874

[27] D. Valcamonico *et al.*, 'A systematic procedure for the analysis of maintenance reports based on a taxonomy and BERT attention mechanism', *Reliab. Eng. Syst. Saf.*, vol. 257, p. 110834, May 2025, doi: 10.1016/j.ress.2025.110834.

[28] A. Nanyonga, H. Wasswa, K. Joiner, U. Turhan, and G. Wild, 'Explainable Supervised Learning Models for Aviation Predictions in Australia', *Aerospace 2025, Vol. 12, Page 223*, vol. 12, no. 3, p. 223, Mar. 2025, doi: 10.3390/AEROSPACE12030223.

[29] S. Sutthithatip, S. Perinpanayagam, and S. Aslam, '(Explainable) Artificial Intelligence in Aerospace Safety-Critical Systems', *IEEE Aerospace Conference Proceedings*, vol. 2022-March, 2022, doi: 10.1109/AERO53065.2022.9843612.

[30] T. Matschak, I. Nastjuk, S. Trang, and J. Benitez, 'Unveiling the hidden layer: Ambivalent

cognitive effects of XAI in augmented decision-making under uncertainty', *Decis. Support Syst.*, p. 114721, Jun. 2026, doi: 10.1016/j.dss.2026.114721.

[31] P. Hitzler, A. Eberhart, M. Ebrahimi, M. K. Sarker, and L. Zhou, 'Neuro-symbolic approaches in artificial intelligence', *Natl. Sci. Rev.*, vol. 9, no. 6, Jul. 2022, doi: 10.1093/NSR/NWAC035.

[32] P. Shojaee, I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, and M. Farajtabar, 'The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity', Nov. 2025, *arXiv*. doi: 10.48550/arXiv.2506.06941.

[33] K. N. Kadripathi, A. Perrusquia, and D. Ignatyev, 'Integrating Explainable AI into Two-Tier ML Models for Trustworthy Aircraft Landing Gear Fault Diagnosis', *AIAA Science and Technology Forum and Exposition, AIAA SciTech Forum 2025*, 2025, doi: 10.2514/6.2025-1928.

[34] G. Youness and A. Aalah, 'An Explainable Artificial Intelligence Approach for Remaining Useful Life Prediction', *Aerospace*, vol. 10, no. 5, p. 474, May 2023, doi: 10.3390/aerospace10050474.

[35] M. Xiong, H. Wang, C. Che, and R. Lin, 'Toward safer aviation: Application of GA-XGBoost-SHAP for incident cognition and model explainability', *Proc. Inst. Mech. Eng. O J. Risk Reliab.*, vol. 238, no. 6, pp. 1195–1208, Dec. 2024, doi: 10.1177/1748006X231205498.

[36] P. Narayan and J. R. Martin, 'Evaluating the Reliability of Explainable Machine Learning for Spacecraft Actuator Fault Detection', *2026 Regional Student Conferences*, Jan. 2026, doi: 10.2514/6.2026-110084.

[37] Y. Xie, N. Pongsakornsathien, A. Gardi, and R. Sabatini, 'Explanation of machine-learning solutions in air-traffic management', *Aerospace*, vol. 8, no. 8, p. 224, Aug. 2021, doi: 10.3390/AEROSPACE8080224/S1.

[38] M. Sokolova and G. Lapalme, 'A systematic analysis of performance measures for classification tasks', *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/J.IPM.2009.03.002.

[39] D. Powers, 'Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation', *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011, Accessed: Jul. 09, 2026. [Online]. Available: https://researchnow.flinders.edu.au/en/publications/evaluation-from-precision-recall-and-f-measure-to-roc-informednes/

[40] J. Davis and M. Goadrich, 'The relationship between precision-recall and ROC curves', *ACM International Conference Proceeding Series*, vol. 148, pp. 233–240, 2006, doi: 10.1145/1143844.1143874.

[41] T. Saito and M. Rehmsmeier, 'The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets', *PLoS One*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/JOURNAL.PONE.0118432.

[42] F. Pedregosa FABIANPEDREGOSA *et al.*, 'Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot', *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, Accessed: Jul. 09, 2026. [Online]. Available: http://scikit-learn.sourceforge.net.