aerospace

MDPI

Article

Evaluation of Influence Factors on the Visual Inspection Performance of Aircraft Engine Blades

Jonas Aust [1,*], Dirk Pons [1] and Antonija Mitrovic [2]

[1] Department of Mechanical Engineering, University of Canterbury, Christchurch 8041, New Zealand; dirk.pons@canterbury.ac.nz
[2] Department of Computer Science and Software Engineering, University of Canterbury, Christchurch 8041, New Zealand; tanja.mitrovic@canterbury.ac.nz
* Correspondence: jonas.aust@pg.canterbury.ac.nz; Tel.: +64-210-241-3591

Abstract: Background—There are various influence factors that affect visual inspection of aircraft engine blades including type of inspection, defect type, severity level, blade perspective and background colour. The effect of those factors on the inspection performance was assessed. Method—The inspection accuracy of fifty industry practitioners was measured for 137 blade images, leading to N = 6850 observations. The data were statistically analysed to identify the significant factors. Subsequent evaluation of the eye tracking data provided additional insights into the inspection process. Results—Inspection accuracies in borescope inspections were significantly lower compared to piece-part inspection at 63.8% and 82.6%, respectively. Airfoil dents (19.0%), cracks (11.0%), and blockage (8.0%) were the most difficult defects to detect, while nicks (100.0%), tears (95.5%), and tip curls (89.0%) had the highest detection rates. The classification accuracy was lowest for airfoil dents (5.3%), burns (38.4%), and tears (44.9%), while coating loss (98.1%), nicks (90.0%), and blockage (87.5%) were most accurately classified. Defects of severity level S1 (72.0%) were more difficult to detect than increased severity levels S2 (92.8%) and S3 (99.0%). Moreover, visual perspectives perpendicular to the airfoil led to better inspection rates (up to 87.5%) than edge perspectives (51.0% to 66.5%). Background colour was not a significant factor. The eye tracking results of novices showed an unstructured search path, characterised by numerous fixations, leading to longer inspection times. Experts in contrast applied a systematic search strategy with focus on the edges, and showed a better defect discrimination ability. This observation was consistent across all stimuli, thus independent of the influence factors. Conclusions—Eye tracking identified the challenges of the inspection process and errors made. A revised inspection framework was proposed based on insights gained, and support the idea of an underlying mental model.

Keywords: visual inspection; influence factors; impact factors; gas turbine engine blades; defect detection; eye tracking; MRO; aviation; aircraft engine maintenance

Citation: Aust, J.; Pons, D.; Mitrovic, A. Evaluation of Influence Factors on the Visual Inspection Performance of Aircraft Engine Blades. *Aerospace* **2022**, *9*, 18. https://doi.org/10.3390/aerospace9010018

Academic Editor: Daniel Ossmann

Received: 30 November 2021
Accepted: 25 December 2021
Published: 29 December 2021

1. Introduction

Aircraft engine components are subject to various internal and external factors such as vibration, high temperatures, rotational speed, rubbing, corrosion, and foreign objects debris (FOD). All of these factors combined can cause internal stress and material fatigue, which has the potential to cause part failure over time. Hence, to assure safe operation, aircraft engines are inspected frequently either after a set amount of flying hours (e.g., 15,000–20,000) or flying cycles, i.e., the number of flights from start to landing (e.g., 5000–10,000) [1]. Rotating parts such as engine blades are subject to more wear and tear compared to static components. Preventative maintenance and frequent inspection can extend the life of those components, thus saving the airline material costs and reducing downtime. Visual inspection is the most common check-up of aircraft engines, as it accounts for 90% of all non-destructive testing (NDT) [2,3].

In gas turbine maintenance, repair, and overhaul (MRO), engine blades are visually inspected at different levels and under different inspection conditions. The first type of

inspection, the so called borescope inspection, is used for internal examination of the engine. A tubular device with a camera and light source at the tip is inserted into the engine through borescope holes and the blades are inspected while the engine is turned either manually or electronically. It allows for the inspecting of hard-to-reach areas under magnification without requiring disassembly. This inspection type is characterised by a dark environment with the borescope diode being the only light source, limited accessibility and manoeuvrability of the borescope tip due to the engine design, overlapping blades, and distorted images due to the wide angle borescope lens.

If a critical condition is found during borescope inspection, the engine gets subsequently disassembled and the parts are inspected on a work bench. During this piece-part inspection, blades are presented one by one under somewhat ideal inspection conditions, e.g., optimal lighting. It also allows for the inspecting of the blades from different perspectives and on different backgrounds.

This paper contributes to the body of knowledge by addressing several gaps. Firstly, the inspection accuracy in borescope inspection is assessed and compared to piece-part inspection. Secondly, the detectability of defect types that have yet to be quantified was evaluated, including bends, breakage, burns, coating loss, tip curl, and tip rub. Thirdly, the effect of other influence factors such as defect severities, blade perspectives, and background colour was analysed. This was worth doing because the gained insights have the potential to improve the inspection processes and training strategies accordingly. The results show significant differences for the type of inspection, defect type, severity level, and blade perspective. No significance was found for background colour. Furthermore, eye tracking was applied to understand the different search strategies, how the influence factors affect the gaze pattern, and what inspection errors occurred. Lastly, a revised visual inspection framework has been developed, taking into account the recognition and judgement component.

2. Literature Review

The purpose of any inspection is to assure that an organisation's products or services meet a defined standard. Therefore, it is essential to understand how accurate the inspection processes are, i.e., how well deviations from the standard are detected and serviceability decisions are made. The most common measure for evaluating the inspection performance is the inspection accuracy. It takes into account correct and incorrect inspections of defective and non-defective parts. Previous studies analysed the inspection performance as part of both quality assurance processes in manufacturing and serviceability inspections in maintenance operations [4,5]. The achieved inspection accuracies range from 67% to 76% [6–8], and from 52% to 68% [5,9,10], respectively. In the aviation maintenance domain, the inspection of several components including aircraft fuselage, cargo bays, landing gear components and engine parts was addressed, and inspection rates of 42% to 87% were reported [5,11–13]. Not all studies reported a specific detection rate, since it is dependent on the defect type and size. Hence, for cracks and dents—the most common defects investigated in previous studies—the results were commonly reported as probability of detection (PoD) curves, which visualises the likelihood of detecting the defect (y-axis) as a function of the defect size (x-axis) [14–18]. For such PoD curves, a minimum sample size of N = 60 per defect type is required [19].

While it is important to know how accurate the inspection processes are, there is also a need to understand why the performance is not 100%, i.e., what factors influence the inspection performance and to what extent? Several works have identified the various influence factors in visual inspection including individual (subject), physical and environmental, task-related, and organisational factors [20]. An overview is provided in [21], which was later complemented by additional factors by See [20]. Aust and Pons provided an overview of factors affecting borescope inspection specifically, and grouped them by Ishikawa's 6M categories [22,23].

Although the various influence factors were already identified, and their relevance acknowledged, the quantification thereof remains difficult. Previous work on the visual inspection of composite materials focused on assessing the effect of lighting, defect size, surface slant, paint colour and surface finish, part cleanliness, and inspection distance [14–17,24]. Furthermore, the effect of demographic factors on the inspection results was evaluated, taking into consideration the work experience, professional certifications, inspection training, education, visual acuity, gender and age. While the above factors were assessed for dent and crack detection on flat composite panels, there might be differences in blade inspection due to the more complex geometries, different material properties, and surface coatings, leading to different defect types and severities [25].

There are several gaps in the body of knowledge, which will be addressed in the following. A first study by Aust and Pons measured the inspection accuracy in piece-part blade inspection under consideration of the cleanliness factor [13]. However, no previous work was found that quantified the performance of borescope inspection. Moreover, no other impact factors affecting blade inspection were quantitatively analysed.

Dent and crack detection received the most attention, while other defects such as nicks and tears were only recently addressed [13]. There is still a range of defects that have not yet been evaluated, such as bends, breakage, burns, coating loss, tip curl, and tip rub [25]. Most studies created the defects artificially using drop weights [14] and impact testing devices [15]. While this is beneficial to control the defect size, it is limited to dull impact damage such as dents. Other defect types such as tears, burns or coating loss are much more difficult to create artificially as the results would look fairly different to real defects, and thus would not be representative. Hence, there is a gap in assessing operationally introduced defects of different severities.

Another factor that was analysed in previous studies was the surface slant and tilt of composite aircraft panels [14]. The study focused on the effect of lighting and the resulting reflection, shading and shadow cues. Hence, only a single, somewhat ideal perspective was analysed. As proposed by the author, future work could focus on analysing different perspectives from various angles. This might even be more worthwhile to do for parts with more complex geometries such as compressor or turbine blades. Moreover, it could provide additional insights that might be useful for advanced technologies such as continuum robots [26–28].

The effect of colour was analysed in different ways. While one study assessed the influence of the defect colour in woven fabrics [4], another tested whether the paint colour of composite materials affected the inspection performance [14]. However, none of them addresses the effect of background colour.

The different influence factors have yet not been analysed under eye tracking observation, except for cleanliness [13] and part complexity [4]. Eye tracking offers the opportunity to gain a better understanding of the inspection process such as the various search strategies that have been applied and how the different influence factors affected the gaze paths and search patterns. This information cannot be extracted from the performance data (inspection accuracy). Furthermore, eye tracking allows identification of which inspection error occurred. This has the potential to suggest remedies, thereby improving the inspection quality and thus contribute to safety.

An alternative to eye tracking is the think-aloud method, whereby participants are asked to verbally express what they are looking at while inspecting the parts. However, in the light of blade inspection this much more difficult to do than in the typical usability studies where participants navigate through a website with clearly separated elements. Moreover, there is the risk that the verbalisation could hinder their thought process and affect the performance significantly [29]. More drawbacks of this approach are outlined in [29,30]. It should be noted that a substantial number of participants in the present study were non-native English speakers. Thus, verbalising their search approach would have added another layer of complexity. Eye tracking in turn provides a method to capture

relevant data about the eye movement without any additional effort or actions required by the participant.

3. Materials and Methods

3.1. Research Objective and Methodology

The study objective was to analyse the effect of different influence factors on inspection performance. The variables admitted to the study were inspection type, defect type, severity level, blade perspective, background colour, and demographic variables such as expertise. Each one was analysed individually. The study was not designed to seek a correlation between these various factors. This would have been a much larger research work.

The parts under examination were compressor and turbine blades of V2500 turbo fan engines. The research sample comprised images taken of those blades with different settings (background colour and blade perspective), as well as stills from the corresponding borescope videos. The images were presented to industry practitioners under eye tracking observation.

The approach taken in this work was hypothesis testing. The hypotheses are based on the typical industry practice, which is understanding the effect of the factors on the inspection accuracy. The results were analysed statistically and where applicable model building was applied to each factor looking at the demographic variables.

3.2. Research Sample (Stimuli)

3.2.1. Part Selection

For this study, high-pressure compressor (HPC) blades were used, since these parts are exposed to airborne contamination, debris on the runway, and wildlife causing foreign object damage (FOD)—one of the main reasons for unscheduled engine shop visits. Due to the current operational processes at our industry partner, only stages six to 12 were available to us. Of those we collected 524 blades from different engines and airlines for possible inclusion in this study. Only a portion of this catalogue could be used to minimise the impact the study might have on the operational processes and productivity. Purposeful sampling was applied (see Figure 1) which led to N = 80 blades (53 defective and 27 non-defective). Non-defective refers to both undamaged blades and blades with acceptable conditions. In this study we presented the participants with images of the blades to allow for the specific assessment of each factor individually. All variables were held constant and only the one factor to be analysed was varied. Hence, some blades were shown multiple times with different image acquisition settings, i.e., the same part was shown with different background colours. The same principle applies to the other factors. Thus, a total number of 137 images (118 piece-part images and 19 borescope stills) were presented to 50 participants, leading to N = 6850 observations. This large dataset lent itself to statistical analysis.

While the main focus of this study lies on the inspection of HPC blades, we also included a few high-pressure turbine (HPT) blades to analyse other defect types that only occur in the hot section, such as burns, cracks, coating loss, and cooling hole blockage.

3.2.2. Part Preparation

The defective and non-defective parts were provided by our industry partner. The defective parts were all scrap parts and hence could be used for our research without any risk. When scraping a part, it is part of the inspection procedure to mark the defect with a red pen and take a photograph of the blade for evidence. These markings however were not helpful for the purpose of our study, since the part should be presented to the participants in such a way as if it just came off the engine. Hence, these markings had to be cleaned off. To maintain the dirty condition, including deposits and discolouration of the blade, we removed the red markings in a manual process using methylated spirit and cotton buds. A before and after picture is shown in Figure 2.

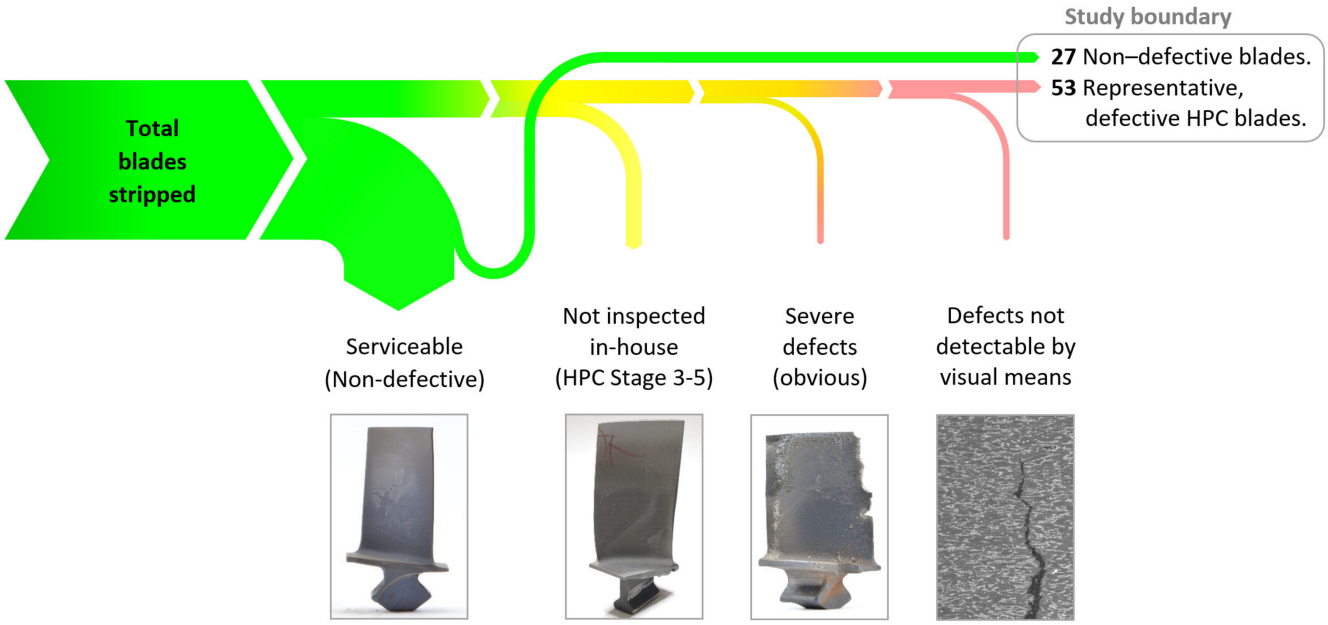

Figure 1. Sample part selection process. The diagram purely serves illustration purposes and is not drawn to scale.

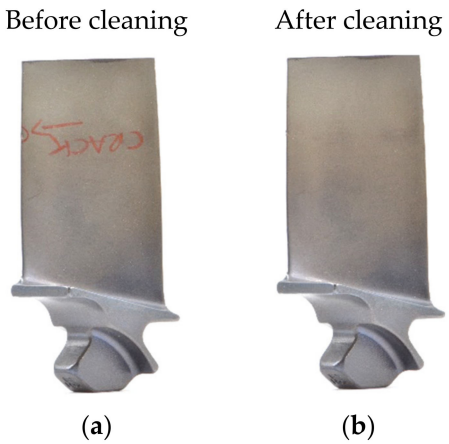

Figure 2. (**a**) Part before and (**b**) after marking removal.

3.2.3. Image Acquisition

There are two levels of inspection we want to analyse, namely borescope inspection and piece-part inspection. For borescope inspection, a set of stills was extracted from the borescope videos that were recorded with a Mentor iQ borescope (manufactured by Waygate Technologies, Shorewood, IL, USA) at our industry partner. The exported stills contained technical and customer-related data such as engine numbers, which was blotted out as it might have influenced the participants' inspection.

To represent the on-bench inspection environment, a second set of images of single blades was acquired in a self-built light tent with somewhat ideal lighting (Figure 3). This provided flexibility to take images from different perspectives and with different coloured backgrounds. A wooden template was used for repeatable positioning of the blade to assure standardised blade orientations.

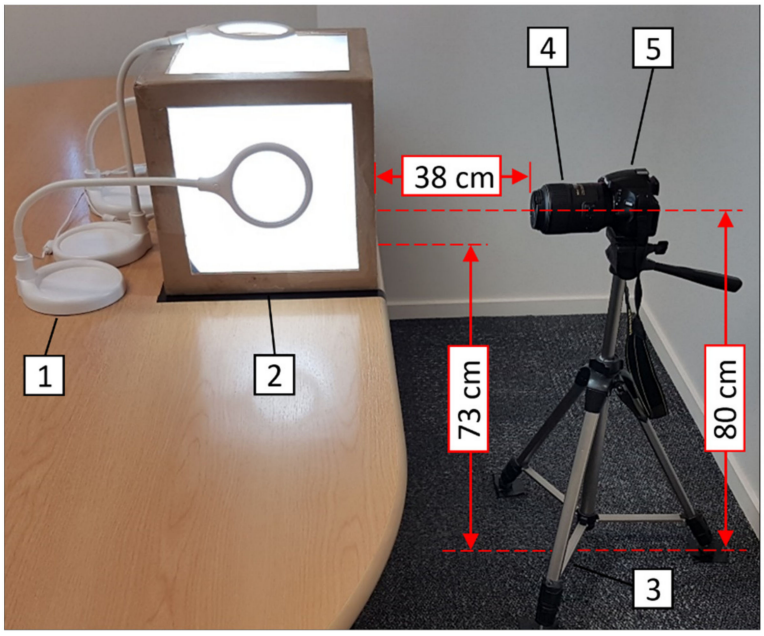

Figure 3. Image acquisition setup comprising (1) three Superlux LSY LED ring lights, (2) self-built light tent, (3) Slik U9000 tripod, (4) Nikon Macro lenses (AF-S Micro Nikkor 105 mm 1:2.8 G), and (5) Nikon D5100 DSLR camera.

The acquired piece-part images and the extracted borescope stills were presented to a bench-inspection and borescope inspection expert respectively to identify all safety-critical defects. This was set as the ground truth for the analysis of the inspection performance. The two experts defining the ground truth did not further participate in the study.

3.3. Research Population

The research population was the same as in [13]. A total of N = 50 participants $\left(M_{Age} = 44.5 \text{ years; } SD_{Age} = 10.33 \text{ years}\right)$ working in a V2500 engine maintenance, repair and overhaul (MRO) shop participated in this study. Participants had 1.5 to 35 of years of work experience in aviation $\left(M_{Work\ Exp.} = 17.7 \text{ years; } SD_{Work\ Exp.} = 9.4 \text{ years}\right)$. A detailed overview of the participants' demographics is provided in [13]. The study included three groups of expertise, namely: inspectors, engineers and assembly operators in descending order. The inspector group includes both borescope inspectors and bench inspectors. While the effect of expertise was already analysed in [13], other research in visual inspection suggests that the performance of individual operators is task-dependent [16]. While the task has not changed much, the inspection environment has.

3.4. Stimuli Presentation under Eye Tracking Observation

This study utilised eye tracking technology to record the participants' eye movement during the inspection task. The eye tracking setup is described in detail in [13]. To present the images, we used PowerPoint 2016 (developed by Microsoft, Redmond, WA, USA), as the pen function was beneficial for the recording of the inspection results, i.e., participants were able to draw a circle around their findings using the mouse, which is somewhat similar to the inspectors' daily job. Moreover, the participants could navigate through the presentation at their own pace. The time was not restricted, but we asked the participants to perform as they would in their daily job.

The research sample (images) was presented in random order with regard to the engine stage, defect type, defect severity, blade perspective, background colour, and inspection type. Any possible learning or fatigue effect is distributed equally across the sample set. However, piece-part images were presented first and borescope stills last, since mixing the two does not represent the operational processes.

3.5. Influence Factors

3.5.1. Inspection Type

The inspection environment varies significantly based on the inspection type (borescope vs. piece-part). While the blades are fully exposed in piece-part inspection and can be inspected one by one, the borescope inspection situation is much different and shows a high level of variability (see Figure 4). For example, the lighting is limited to the integrated borescope LED diode, and is absorbed differently depending on the cleanliness and coating. Moreover, the view is restricted by the engine design and overlapping adjacent blades. While the expectations towards borescope inspectors is to find every single defect in such a difficult environment, this study intended to analyse the feasibility and measure the performance when compared to piece-part inspection in a somewhat ideal environment.

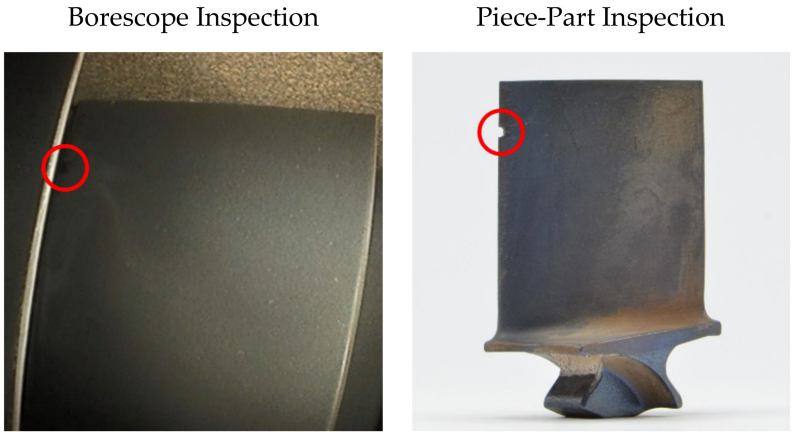

Figure 4. Blade with nick on leading edge at different inspection levels: (**a**) borescope inspection; (**b**) piece-part inspection.

Hypothesis 1 (H1). *The inspection type affects the inspection performance.*

3.5.2. Defect Type

Some defect types are more salient than others. Hence, there was an interest to understand which defect types are most difficult to detect. The research sample included airfoil dents, bends, blockage, breakage, burns, coating loss, cracks, dents, nicks, non-defective blades, tip curl, tip rub, and tears. An overview of the different defect types including a representative image and defect description can be found in [25]. While the focus of this study lies on compressor blades, the inclusion of turbine blades in this sub-study was inevitable in order to analyse additional defect types.

Another metric that is important when it comes to analysing different defect types is the defect classification accuracy, i.e., how many defects have been correctly classified. Correct classification is essential as it determines the service and repair action of that blade. An incorrectly classified defect might be accepted or scrapped without justification depending on the confused defect type. Since the detection of a defect is a prerequisite of the defect classification, the classification accuracy is based on the correctly identified defects, i.e., it provides a measure for the correct classifications of the correctly identified defects.

Hypothesis 2 (H2). *The defect type affects the inspection and classification accuracy.*

3.5.3. Defect Severity

The defect severity is another factor that needs to be assessed. While it might seem obvious that a larger defect size generally leads to a higher detectability, it is unclear yet where the threshold is and whether some defect types are more difficult to detect than others. A simplified defect categorisation was applied with the three defect groups being airfoil defects (surface damage), edge defects with deformation, and edge defects with material loss. Each defect was presented in three severity levels (Figure 5). A tiered severity classification was chosen over a probability of detection (PoD) curve, since the latter would have required a sample size of at least N = 60 per defect, and thus would have exhausted the study. Moreover, the detection rates (probability) for each severity level can be fed back into the defect detection risk framework in [31].

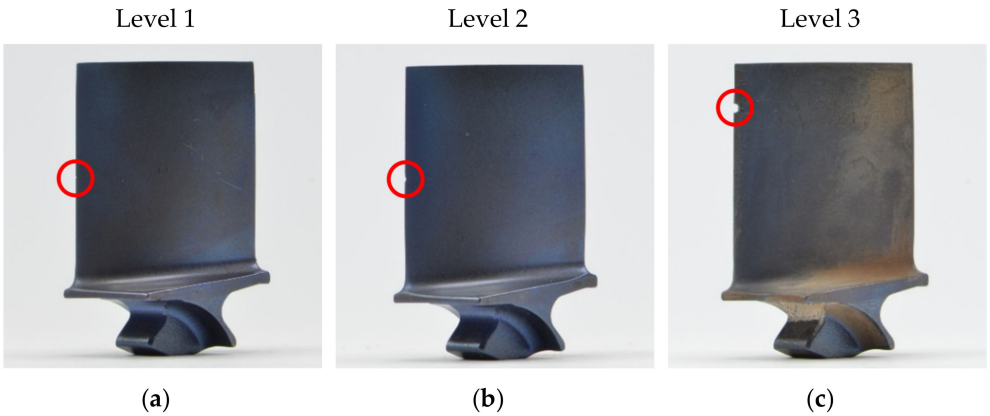

Figure 5. Different levels of defect severity from (**a**) lowest Level 1 to (**c**) highest Level 3. The defects are highlighted by red circles.

Hypothesis 3 (H3). *The defect severity affects the inspection performance.*

3.5.4. Blade Perspective

Different levels of inspection allow for different viewing perspectives of the part. Thus, there is a relationship between the level of disassembly and the defect manifestation [31]. There was an interest to understand which perspective(s) are most beneficial to detect different defect groups (edge defects vs. surface defects). The insights could be useful for future standardisation and automation. Eight different blade perspectives were tested (Figure 6). The perspective nomenclature was adopted from [32].

View	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**
Angle	0°	45°	90°	135°	180°	225°	270°	315°
Sample Image								

Figure 6. Different blade perspectives with 45 degree intervals.

Hypothesis 4 (H4). *The blade perspective affects the inspection performance.*

3.5.5. Background Colour

The motivation for analysing the background colour derived from industry. When reviewing historic data of defective blades, we realised that quite a few of the blade images were taking on a coloured background. When asking staff why they took those images on a coloured background as opposed to a white work bench, they said that the colour made the defects stand out more. This hypothesis that defects are more salient on coloured background and lead to better inspection results was tested. If this hypothesis could be proven true, the coloured background would be an easy and affordable way to improve inspection.

The parts were presented on four different backgrounds including white, red, green, and yellow (Figure 7). In the image acquisition process we used different sheets of coloured paper, rather than post-processing the images and changing the colour artificially. The paper was provided by our industry partner to represent the real situation as close as possible. We also included different defect groups, i.e., surface defects, edge defects and non-defective blades to understand whether certain defects are better visible on one colour than another.

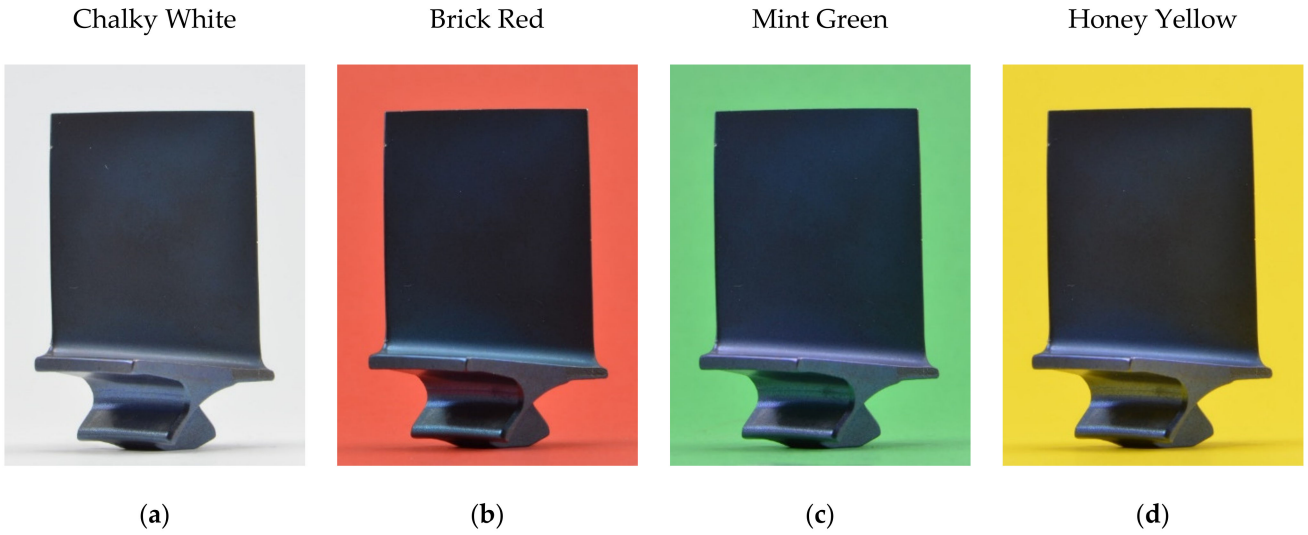

Figure 7. Blade with different coloured backgrounds: (**a**) chalky white (RGB: 229, 230, 232), (**b**) brick red (RGB: 230, 99, 79), (**c**) mint green (RGB: 111, 194, 124), and (**d**) honey yellow (RGB: 241, 212, 59).

Hypothesis 5 (H5). *The background colour affects the inspection performance.*

3.6. Data Analysis

This study applied statistical hypothesis testing to analyse the five hypotheses H1 to H5 as previously outlined. Each hypothesis tests the effect of an influence factor on the inspection performance. The inspection performance can be measured in Inspection Accuracy and Defect Classification Accuracy. The primary measure in this study was Inspection Accuracy, and is defined as the proportion of correct serviceability decisions (true positives and true negatives) and the total amount of blades inspected. For the assessment of the defect type an additional measure was introduced, namely the Defect Classification Accuracy. The purpose was to analyse how accurately participants identified the defect type after successful detection.

The independent variables assigned to this study were the five influence factors and the following demographic variables: Expertise, Education, Previous Inspection Experience, Work Experience, and Visual Acuity (Table 1).

Table 1. Overview of demographic factors and levels thereof.

Demographic Factor	Levels
Expertise	Inspector, Engineer, Assembly Operator
Education	University degree, Diploma, Trade Certificate
Previous Inspection Experience	Yes, No
Work Experience in Aviation	1 to 4, 5 to 9, 10 to 19, 20 and more years
Visual Acuity	Corrected vision, No corrected vision

Statistical methods were selected considering the type of data. Analysis of Variance (ANOVA) was used to analyse each influence factor individually and provide a visual representation of the direction of effects. When two categorical variables were assessed (e.g., inspection type and defect type), a factorial ANOVA was performed. Subsequently, the interaction of the multiple variables was analysed using generalised linear/non-linear model building. This allowed the dependencies between the individual variables to be elucidated. For example, there are complex interactions between blade perspective and defect type. In the case of inspection accuracy, and again for classification accuracy, the dependent variable takes values of 1 and 0, i.e., the decision can be correct (1) or incorrect (0). Hence, logit odds ratio was applied in these cases.

4. Statistical Results

4.1. Inspection Type

The first factor that was analysed is the type of inspection, i.e., borescope and piece-part inspection. Table 2 shows the inspection accuracy for each of them by group of expertise.

Table 2. Inspection accuracies by expertise group and inspection type (in percentages).

Expertise	Borescope Inspection M (SD)	Piece-Part Inspection M (SD)
Inspectors (N = 18)	66.7 (13.0)	83.3 (10.1)
Engineers (N = 16)	63.4 (11.6)	82.1 (14.3)
Assembly Ops. (N = 16)	60.7 (20.5)	82.1 (12.2)
All participants (N = 50)	63.7 (15.3)	82.6 (12.0)

The results show that the average inspection accuracy improved throughout all groups of expertise from borescope inspection to piece-part inspection. This was confirmed by an ANOVA, $F(1, 698) = 33.085$, $p < 0.001$. To see whether this is consistent for all defect groups, a Factorial ANOVA was subsequently performed and showed significance too, $F(3, 692) = 5.093$, $p < 0.002$. Figure 8 highlights that the significant difference was predominantly for non-defective blades and nicks, while the inspection accuracy was less affected by the inspection type for dents and tears.

Dents were equally detectable in both levels of inspection with a tendency towards better performance in borescope inspection. A possible reason could be the advantage of the borescope lighting being quite focused and creating shades on the blade surface, which may allow for better differentiation between dents (indentations) and deposits (offset material), and thus an improved inspection performance.

As later discussed in Section 4.4., the blade perspective in borescope inspection (P3) is most beneficial for detecting airfoil damage such as dents. Hence, this could have further contributed to an equal performance in borescope and piece-part inspection.

It stood out that the inspection accuracy for non-defective blades was quite low in both piece-part and borescope inspection. This could possibly be explained by staff of high reliability organisations (e.g., in aviation) taking a conservative approach in favour of safety, i.e., in the case of uncertainty it would be reasonable to remove the blade from service. Another reason for the low performance might be the skewed research sample, i.e., the amount of defective blades was disproportionately high compared to reality. Thus,

participants may have been biased towards finding defects. Furthermore, participants knew that they were under scrutiny and that their performance was recorded, which might have further contributed towards a risk averse approach resulting in higher false-positive rates.

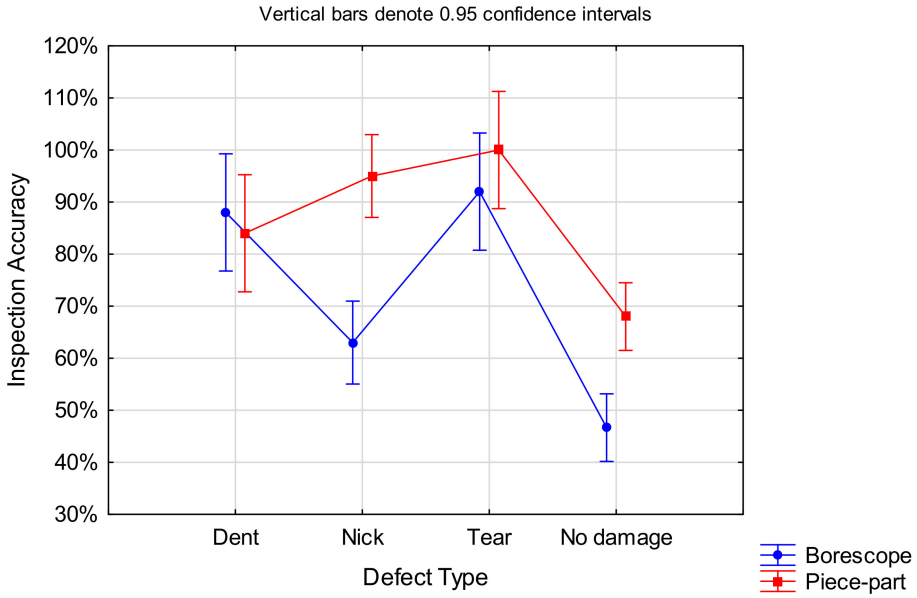

Figure 8. Effect of Inspection Type on the Inspection Accuracy for each defect group.

A Generalised Linear/Non-linear logit model for Inspection Accuracy with categorical factors (Inspection Type, Expertise, Previous Inspection Experience, Education, and Visual Acuity), and continuous predictor of Work Experience, showed the only significant factor was Inspection Type ($X^2(1, 698) = 32.347$, $p < 0.001$), which has already been discussed.

4.2. Defect Type

In this section the factor 'Defect Type' is analysed in detail. The inspection results measured in detection accuracy and classification accuracy are shown in Table 3.

Table 3. Inspection accuracy and classification accuracy by defect type (in percentages).

Defect Type	Mean Inspection Accuracy % Mean (SD)	Mean Classification Accuracy % Mean (SD)
Airfoil dent	19.0 (31.8)	5.3 (25.8)
Bend	79.0 (30.5)	58.8 (41.7)
Blockage	8.0 (25.5)	87.5 (44.7)
Burn	79.5 (18.7)	38.4 (38.2)
Coating Loss	85.3 (24.4)	98.1 (8.7)
Crack	11.0 (29.1)	73.3 (50.5)
Dent	36.0 (32.0)	54.3 (49.9)
Nick	100.0 (0.0)	90.0 (30.3)
No damage	59.0 (23.0)	100.0 (0.0)
Tear	95.5 (6.1)	44.9 (22.2)
Tip Curl	89.0 (20.9)	84.1 (29.0)
Tip Rub	83.3 (16.8)	76.0 (30.6)

The statistical analysis using a Factorial ANOVA (Figure 9) revealed statistical significance of defect type on the inspection accuracy, $F(11, 1738) = 112.11$, $p < 0.001$. The results show that blockage, cracks, and airfoil dents were the most difficult ones to detect with 8%, 11% and 19%, respectively. The highest detection rates were achieved for nicks (100.0%), tears (95.5%), and tip curls (89.0%). The defect classification accuracy was also dependent

on the defect type, F(11, 1130) = 31.273, $p < 0.001$. Airfoil dents (5.3%), burns (38.4%), and tears (44.9%) were the most difficult defects to classify, while the highest classification accuracy was noted for coating loss (98.1%), followed by nicks (90.0%), blockage (87.5%) and tip curl (84.1%). An interesting finding is that blockage had the lowest detection rate (8.0%) while having one of the highest classification accuracies (87.5%). This implies that participants who found this defect type knew exactly what to look for. The same effect can be seen for cracks. Contrarily, tears had the second highest detection rate (95.5%), but a low classification accuracy (44.9%). This is further discussed below.

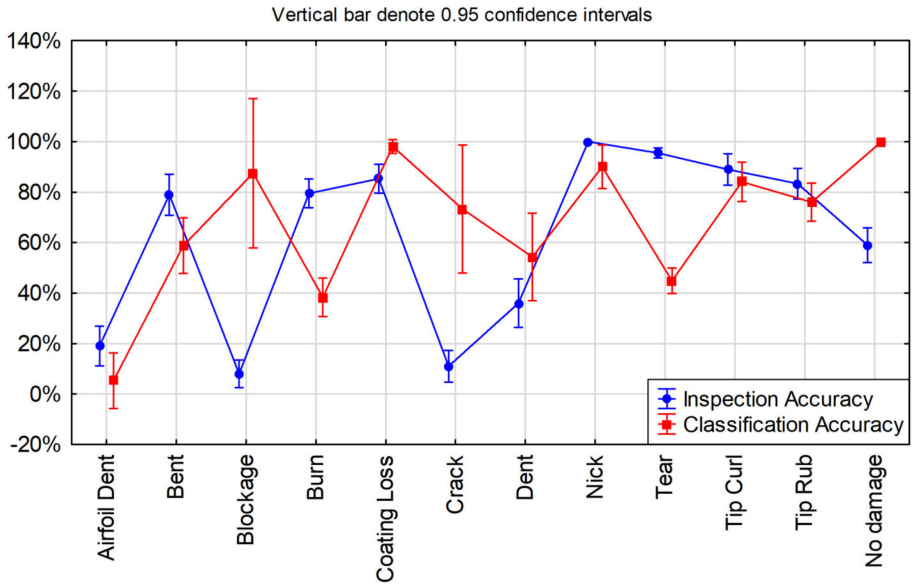

Figure 9. Mean inspection accuracy and defect classification accuracy for each defect type.

Generalised linear/non-linear logit modelling around inspection accuracy and classification accuracy with Defect Type, Expertise, Visual Acuity, Education, and Previous Inspection Experience as categorical factors and Work Experience as a continuous variable showed that there was no correlation between the demographic variables and the inspection or classification accuracy, respectively.

There was a need to understand which defects were misclassified and with what other defect types they were confused. Figure 10 provides an overview of the actual defect type (ground truth) and the classification thereof made by the participants. The correct classifications of each defect type are highlighted in grey. The highest miss-classification occurred for airfoil dents, burns and tears.

Airfoil dents (surface) are most commonly confused with dents (edges) potentially due to their similarity in terminology and characteristics, while the difference might not be known by inspectors. Even if adding the two types of dents, the accumulative classification rate is still below 50% and offers great potential for improvement.

Burns are often confused with coating loss (31.6%) and breakage (28.2%), depending on the amount of missing material. Tears in contrast have the widest range of confused defect types with breakage (21.9%) and cracks (21.3%) being the most common ones. This appears to indicate that defects were classified based on their visual appearance rather than on contextual knowledge or by taking into account the potential root causes.

Breakage was added to Figure 10 under predicted class as a significant amount of participants predicted this defect class. Since it was not pre-defined and there was no breakage in the dataset, this class does not appear as a column under the ground truth. The 'no damage' column was added to the defect list as it provides additional insights of what participants supposedly detected on non-defective blades, e.g., in 7.5% of the time a non-defective blade was incorrectly removed from service because participants supposedly detected an airfoil

dent. Most non-defective blades were incorrectly classified because of suspect tip rub (22.5%) and nicks (18.0%), followed by coating loss (12.5%) and tip curls (12.0%).

		True Class (Ground Truth) [%]											
		Airfoil Dent	Bend	Blockage	Burn	Coating Loss	Crack	Dent	Nick	Tear	Tip Curl	Tip Rub	No damage
Predicted Class [%]	Airfoil Dent	5.3	12.8	-	-	-	-	-	3.3	1.0	1.1	-	7.5
	Bend	21.1	58.8	-	-	-	-	-	-	1.8	11.5	-	4.5
	Blockage	-	-	87.5	-	-	-	-	-	-	-	0.9	-
	Burn	-	-	12.5	38.4	1.9	-	-	-	0.5	-	1.8	8.0
	Coating Loss	15.8	-	-	31.6	98.1	20.0	-	-	-	-	0.9	12.5
	Crack	-	-	-	0.6	-	73.3	-	-	21.3	-	-	4.5
	Dent	42.0	21.1	-	0.6	-	-	54.3	3.3	4.3	3.3	0.9	10.5
	Nick	-	2.6	-	-	-	-	45.7	90.1	4.6	-	3.6	18.0
	Tear	-	-	-	0.6	-	6.7	-	-	44.9	-	-	-
	Tip Curl	15.8	-	-	-	-	-	-	-	-	84.1	5.9	12.0
	Tip Rub	-	-	-	-	-	-	-	-	-	-	76.0	22.5
	Breakage	-	1.3	-	28.2	-	-	-	3.3	21.9	-	10.1	-

Figure 10. Classification matrix showing the distribution of defect types that are often confused.

4.3. Severity

Three levels of severity (S1 to S3) were analysed and the inspection results of each level grouped by the different defect categories are presented in Table 4.

Table 4. Inspection accuracies by defect group and severity level (in percentages).

Defect Group	Severity Level 1 M (SD)	Severity Level 2 M (SD)	Severity Level 3 M (SD)
Airfoil defects (surface damage)	30.0 (46.3)	76.0 (43.1)	98.0 (14.1)
Edge defects (deformation)	83.3 (37.4)	98.0 (14.1)	98.0 (14.1)
Edge defects (material loss)	74.7 (43.6)	96.7 (18.0)	100 (0.0)
All defects	72.0 (45.0)	92.8 (25.9)	99.0 (10.0)

It comes as no surprise that the inspection accuracy improved with increasing severity, $X^2(2, 797) = 102.5$, $p < 0.001$. The difference was measured between severity level one and level two (Odds Ratio = 5.01, $p < 0.001$), and between level one and level three (Odds Ratio = 38.5, $p < 0.001$). There was no significant difference between level two and three (Odds Ratio = 7.68, $p = 0.624$).

Generalised linear/non-linear logit modelling revealed that both Defect Type ($X^2(2, 797) = 37.889$, $p < 0.001$) and Expertise ($X^2(2, 797) = 9.585$, $p < 0.01$) were correlated with inspection accuracy. Surface defects were more difficult to detect than both types of edge defects, i.e., with and without material loss. On average and across all three severity levels, engineers showed the lowest inspection accuracy of 80.5% followed by assembly operators (86.7%) and inspectors (88.2%). Figure 11 highlights that the difference specifically occurred in severity level S1, i.e., engineers had an inspection accuracy of 61.6%, while both assembly operators (73.2%) and inspectors (80.2%) performed better on smaller defects. No other demographic factors were significant (all $p > 0.129$).

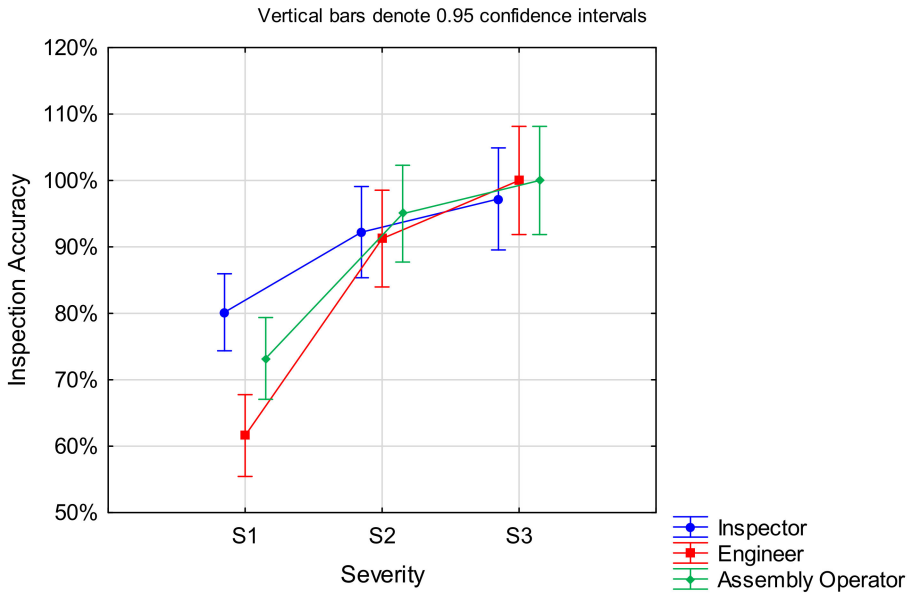

Figure 11. Effect of defect severity on inspection accuracy for each group of expertise.

4.4. Blade Perspective

Eight different blade perspectives covering a 360-degree view of the blade were tested. The inspection rates of each perspective are presented in Table 5.

Table 5. Inspection accuracies by blade perspective and defect group (in percentages).

Blade Perspective	Airfoil Defects (Surface Damage) M (SD)	Edge Defects (Deformation) M (SD)	Edge Defects (Material Loss) M (SD)	No Damage (Non-Defective) M (SD)	All Blades M (SD)
P1	98.0 (14.1)	100.0 (0.0)	10.0 (3.0)	58.0 (49.9)	66.5 (47.3)
P2	84.0 (37.0)	98.0 (14.1)	100.0 (0.0)	44.0 (50.1)	81.5 (38.9)
P3	100.0 (0.0)	96.0 (19.8)	100.0 (0.0)	54.0 (50.4)	87.5 (33.2)
P4	80.0 (40.1)	98.0 (14.1)	100.0 (0.0)	44.0 (50.1)	80.5 (39.7)
P5	38.0 (6.9)	20.0 (5.7)	100.0 (0.0)	46.0 (50.4)	51.0 (50.1)
P6	34.0 (47.9)	100.0 (0.0)	100.0 (0.0)	64.0 (48.5)	74.5 (43.7)
P7	64.0 (48.5)	46.0 (7.1)	100.0 (0.0)	66.0 (47.9)	69.0 (46.4)
P8	86.0 (35.1)	96.0 (2.8)	98.0 (2.0)	44.0 (50.1)	81.0 (39.3)

The results were statistically analysed, and the One-way ANOVA confirmed that the blade perspective is significant for the inspection performance, $F(7, 1592) = 14.772$, $p < 0.001$. Figure 12 shows that the inspection accuracy in P5 (51.0%) was significantly lower than in any other perspective. Perspective P3 shows the highest accuracy with 87.5% and is notably higher than P1 (66.5%), P6 (74.5%), and P7 (69.0%). Good results were also achieved in perspectives P2, P8, and P4 with 81.5%, 81.0%, and 80.5%, respectively.

A subsequently performed generalised linear/non-linear logit model with Inspection Accuracy as the dependent variable, Blade Perspective, Defect Type, Expertise, Previous Inspection Experience, Education, and Visual Acuity as categorical factors, and Work Experience as the continuous predictor revealed that besides the Blade Perspective the Defect Type was a significant factor affecting the inspection accuracy, $X^2(7, 1592) = 94.771$, $p < 0.001$. No other demographic factor was significant. Thus, there was an interest to understand what perspectives are preferable for which defect type. The mean accuracies of all participants for the different groups of defects and perspectives are presented in Figure 13. Edge defects with material loss (e.g., nicks) were the most detected defects in any perspective except for P1 (10.0% detection rate). Airfoil defects (e.g., dents) and edge deformation, (e.g., bends) in contrast, showed one of the highest detection rates in

the P1 perspective with 98.0% and 100.0%, respectively. This highlights the difficulty of standardising the inspection process, especially when no previous work has analysed the perspective factor.

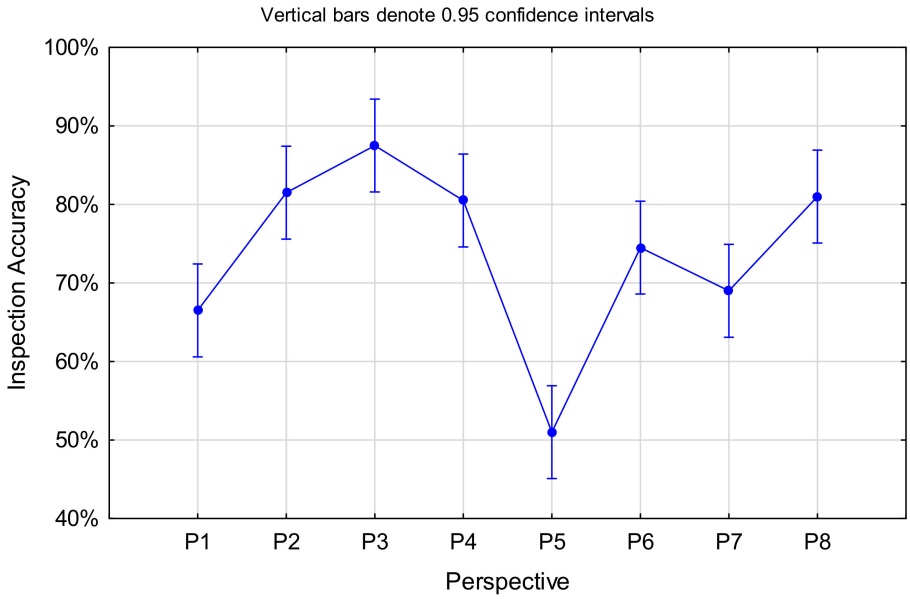

Figure 12. Effect of blade perspective on inspection accuracy.

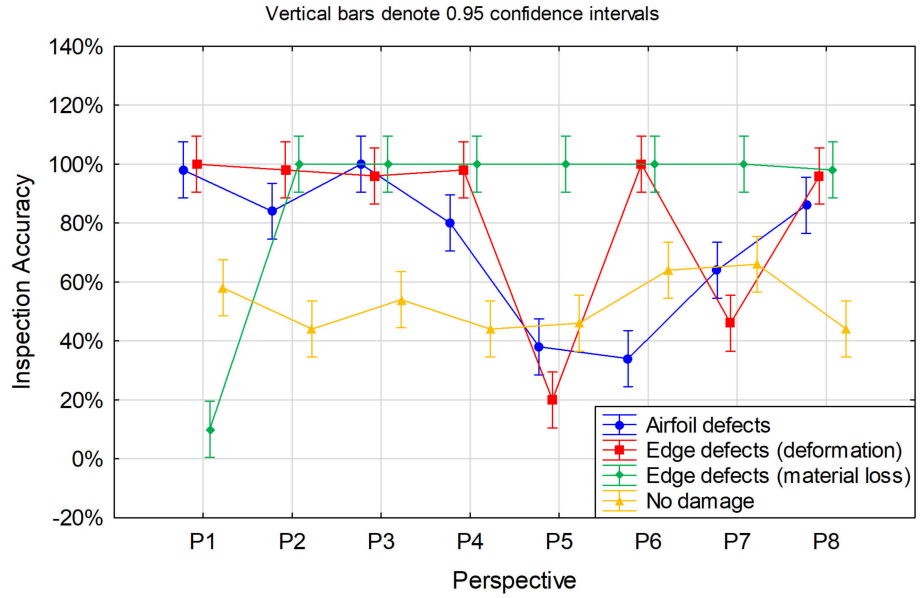

Figure 13. Effect of blade perspective on inspection accuracy for each defect group.

4.5. Background Colour

Four different background colours were assessed to evaluate whether colour had any effect on the inspection performance. The results for each colour is presented in Table 6.

There was no significant difference in inspection accuracy between the different colours for the sample as a whole, $F(3, 796) = 0.759$, $p = 0.517$. A generalised linear/non-linear logit model around inspection accuracy with Background Colour, Expertise, Previous Inspection Experience, Education, and Visual Acuity as categorical factors and Work Experience as the continuous predictor confirmed also that none of the demographical factors was significant.

Table 6. Inspection accuracies by defect type and severity (in percentages).

Defect Type	White M (SD)	Red M (SD)	Green M (SD)	Yellow M (SD)
Surface defects	80.0 (40.4)	86.0 (35.1)	86.0 (35.1)	82.0 (38.8)
Edge defects	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100 (0.0)
No damage	30.0 (46.3)	38.0 (49.0)	22.0 (41.9)	46.0 (50.4)
All defects	77.5 (41.9)	81.0 (39.3)	77.0 (42.2)	82.0 (38.5)

Similar to previous influence factors, there was interest in understanding whether this effect is consistent across all defect types, or whether one benefits more from a specific colour than another. The factorial ANOVA in Figure 14 with inspection accuracy as the dependent variable, and background colour and defect type as categorical factors shows that for both airfoil defects and edge defects, the background colour had no significant effect. For non-defective blades, however, the background colour affects the inspection performance, $F(6, 788) = 2.5287$, $p < 0.02$. Green had the lowest mean inspection accuracy of 22.0%, followed by white (30.0%), red (38.0%), and yellow (48.0%). The difference was measured between green and yellow backgrounds (Odds Ratio = 3.02, $p < 0.05$), while the performance improvement for white and red backgrounds was not significant. Yellow was the brightest colour and while some participants stated it highlighted defects the best (which was actually not the case for the research population as a whole), others complained about the bright colour and that it was fatiguing for the eyes. This might explain why there was a tendency towards less findings on yellow backgrounds, as well as for defective blades, although not significant.

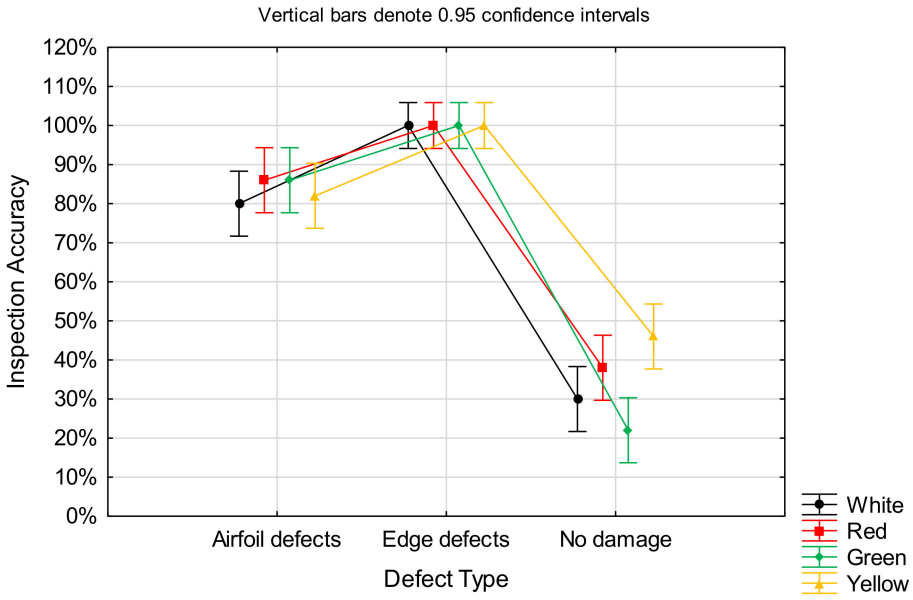

Figure 14. Effect of defect type on inspection accuracy for each background colour.

Another unanticipated finding was that participants tended to perform better on one colour over the others, although this was not consistent across the research population. Thus, it was hypothesised that there was a personal preference (intentionally or unintentionally) that one or more colours are preferred, while others are not. This raises the possibility that colour perception, including colour blindness, may be a factor. Individual inspection results for each participant and colour are plotted in Figure 15. The results show that 23 participants (46%) performed best on a specific colour, while the other three colours lead to equally poor performances. Contrarily, 15 participants (30%) tended to have a least preferred colour on which they performed worst. This finding was underpinned by the

feedback provided by the participants, who had a clear opinion towards specific colours (positively as well as negatively).

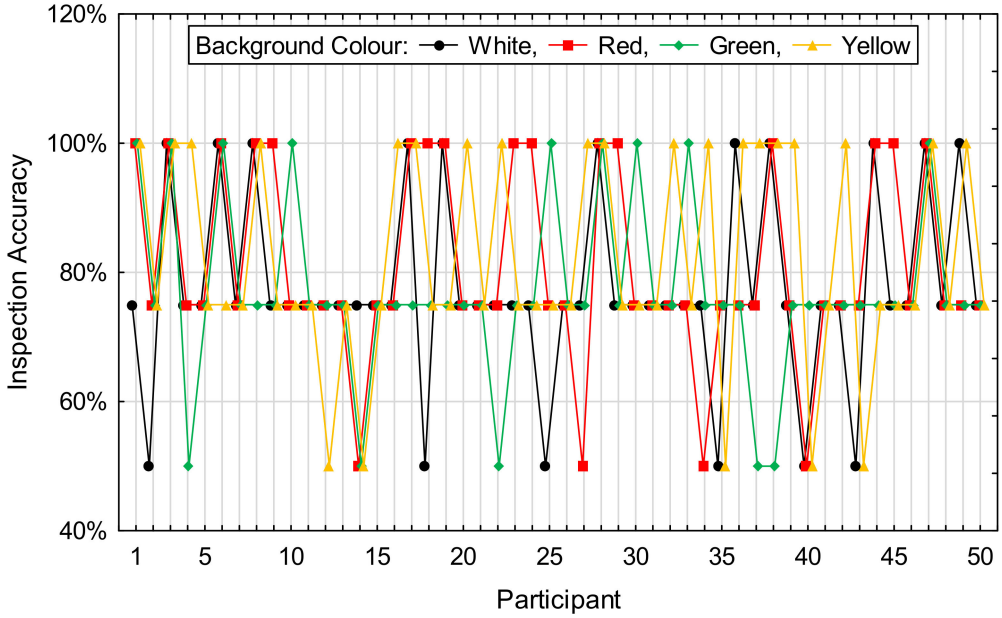

Figure 15. Effect of background colour on the individual inspection accuracy (without standard deviation for legibility purposes). The diagram shows high personal preference for specific colours.

5. Evaluation of the Eye Tracking Data

The eye tracking data were analysed to gain insights into the inspection and influence factors. More specifically, we wanted to understand why defects were missed, what search strategies were applied, and what inspection errors occurred. Background colour was the only factor that was not statistically significant and screening of the eye tracking recordings showed no striking conspicuousness. Hence, the gaze plots and heat maps are not presented here.

The way the two eye tracking outputs can be interpreted is as follows. Heat maps highlighted the areas that attracted most attention and we used a traffic light colour scheme whereby green indicates areas receiving scant attention (short dwell times) and red highlighting indicates areas with the most attention (long dwell times). Areas that were not looked at are not coloured. Gaze plots in contrast show the scan path, order of fixations, and dwell times. A larger gaze plot diameter indicates a longer dwell time, and vice versa for smaller gaze points.

One of the limitations of eye tracking is the restricted ability to quantitatively compare heat maps and gaze plots between different sample groups and subject groups [13]. For this reason and to present the findings in the most concise way, some representative samples were selected and semi-quantitatively analysed in the following sections.

5.1. Inspection Type

The statistical analysis in Section 4 revealed that, overall, piece-part inspection led to better inspection performance than borescope inspection of the same blade. A detailed analysis of the inspection results revealed that this was true for 26.9% of the cases, while 65.1% of the time the same results were achieved in both inspection types. In 8.0% of inspections, borescope led to better results than piece-part. This was predominantly for non-defective blades, i.e., participants incorrectly marked more non-defective blades as defective during piece-part inspection than in borescope inspection.

The eye tracking data were further analysed to better understand this effect. The resulting heat maps for piece-part inspection in Figure 16 indicate that all participants detected the defect during piece-part inspection. This was confirmed by their recorded

defect markings. As expected, the bench inspector—having the most experience in blade inspection—performed the fastest with 3.799 s. The borescope inspector took significantly longer (15.402 s) and their inspection time was fairly similar to the engineer's (14.912 s). Thus, the latter two took around four times longer than the bench inspector. The longest inspection time was measured for the assembly operator (39.224 s), which was 10 times as long as the one of the bench inspector and 2.5 longer than the borescope inspector's and engineer's.

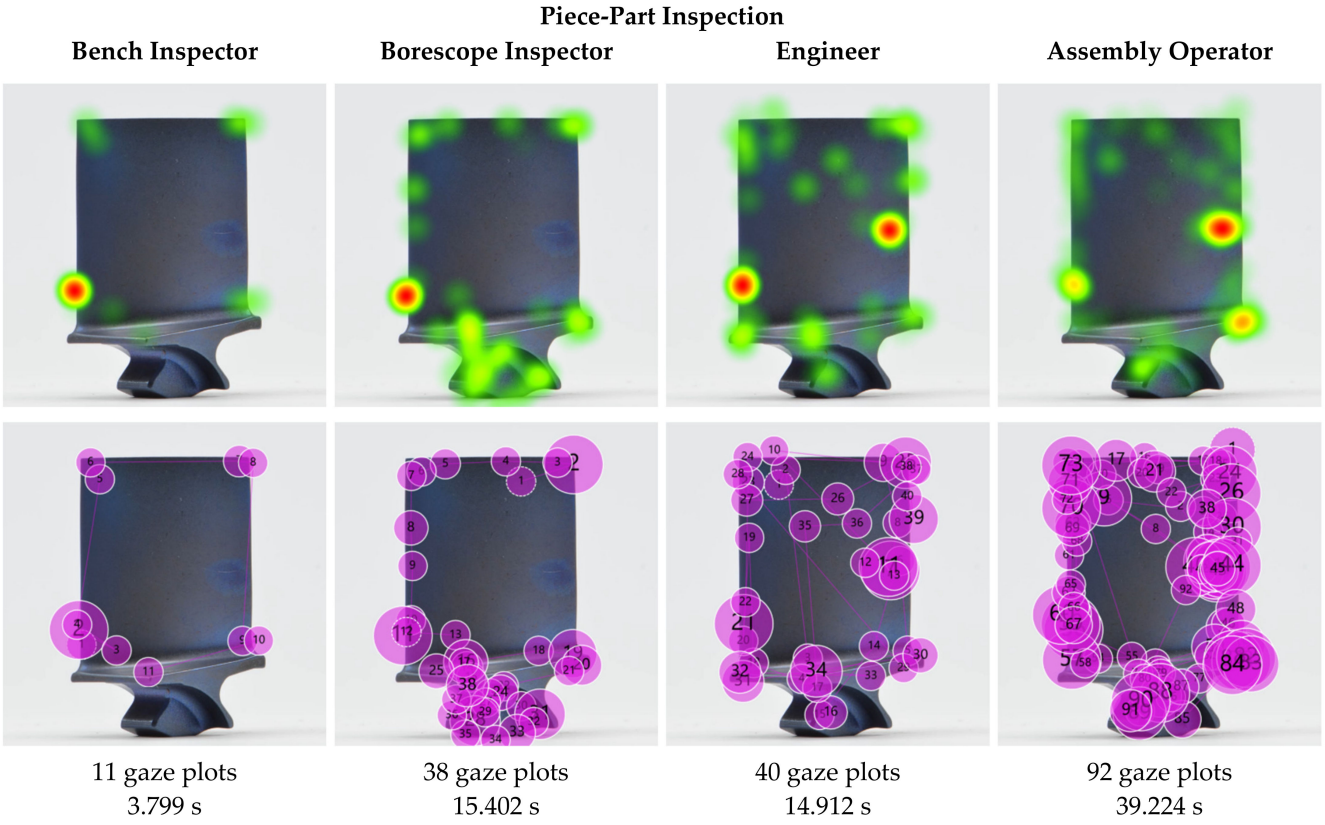

Figure 16. Eye tracking results for piece-part inspection by expertise.

The gaze plots show the search path of each participant. The bench inspectors' eye tracking results show 11 fixations predominantly on the edges of the airfoil. The number of fixations on the airfoil edges in the borescope inspector's recording are similar to the bench inspector. However, while the bench inspector did not even look at the root, the borescope inspector spent half of the time inspecting the platform and root.

This is interesting for two reasons: firstly, because the root is exposed during piece-part inspection and thus one would expect the bench inspector to inspect it. This might reveal another issue with blade inspection, i.e., the term 'blade' is confused with the term 'airfoil', and thus only the latter was inspected, while the root and platform was left out. Secondly, this finding is interesting because the root is not visible during borescope inspection. It might indicate the unfamiliarity of the borescope inspector with inspecting this part of the blade, which might have led to a more detailed and thus longer inspection.

Furthermore, the gaze plots show that both inspectors' focus lay on the edges, whereas the engineer and assembly operator also inspected the airfoil surface. Some discolouration and a negligible surface scratch close to the middle right edge particularly attracted their attention, as can be seen in the heat maps. Additionally, the assembly operator focused on the right corner of the platform with some acceptable chafing (within limits). Thus, we conclude that with increasing experience, smaller, acceptable conditions are being ignored and the focus is rather set on critical locations and defect types. Moreover, less experienced

staff in visual inspection (engineers and assembly operators) tended to inspect the blade in more detail and returned multiple times to re-inspect areas of interest.

Next, the eye tracking recordings of the same blade shown as for the borescope image were evaluated. The resulting heat maps and gaze plots are shown in Figure 17. The defect markings show that only the borescope inspector found and marked the defect. None of the others classified the blade as unserviceable. However, the heat maps show that all participants looked at the defective area for a while. Hence, it is likely that the irregularity (defect) was noticed, but an incorrect decision was made in regards to the serviceability of the blade.

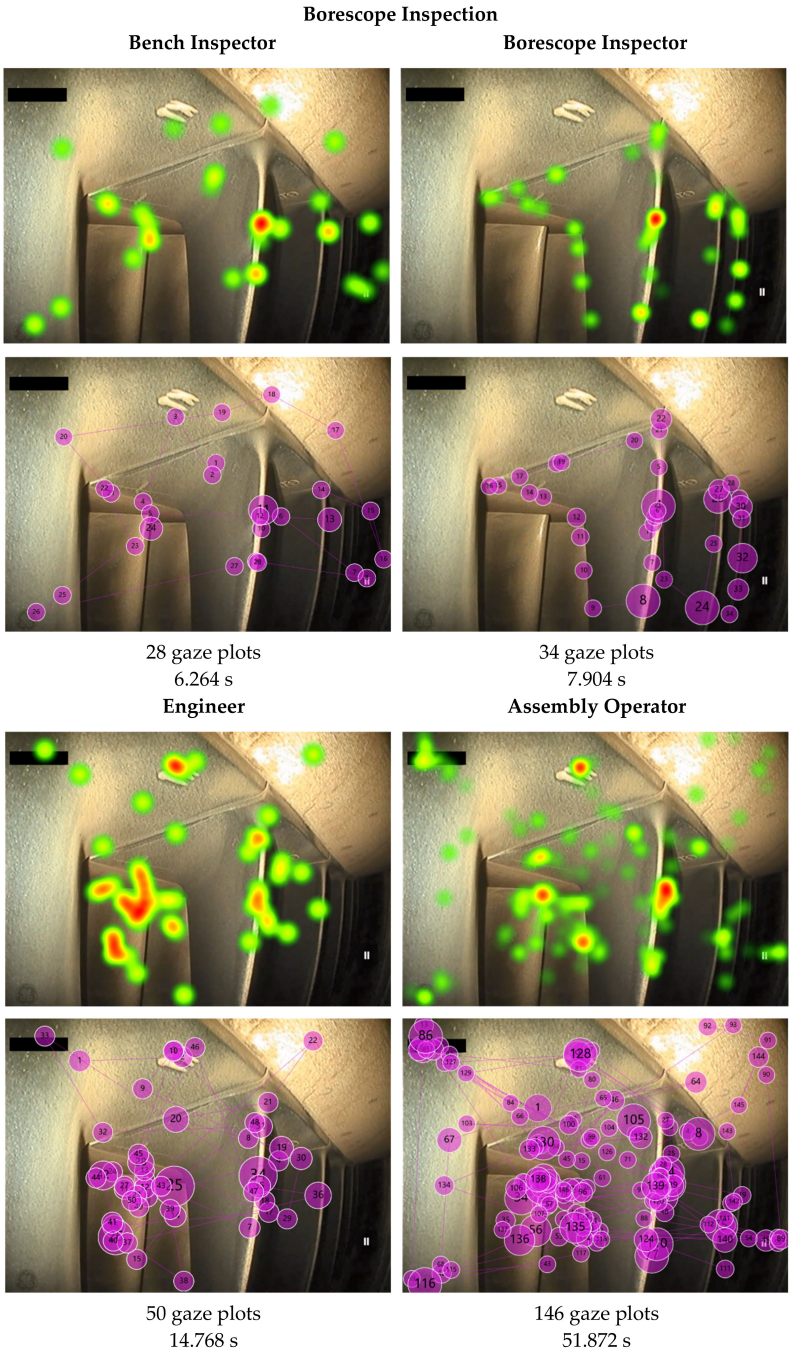

Figure 17. Eye tracking results by expertise for the same blade shown in Figure 16 but now as borescope image.

The inspection times show that the borescope inspector was 1.640 s (21%) slower than the bench inspector, followed by the engineer who was 2.4 times slower than the bench inspector and 1.9 times slower than the borescope inspector. Similarly to piece-part inspection, the assembly operator was significantly slower than the engineer, borescope inspector and bench inspector, and required 3.5, 6.6 and 8.3 times longer, respectively.

The gaze plots of both the bench and borescope inspectors showed a similar number of fixations, with 28 and 34, respectively. However, their scan paths differed significantly from another, i.e., the bench inspector scanned the image in a zig-zag pattern, while the borescope inspector followed a clear search strategy (further described below). The engineer made 50 fixations, and their gaze plots form two clouds: one around the blades in the centre of the image and one around the vanes in the subsequent row. The assembly operator's gaze plot showed 146 fixations with no patterns or clusters, but an unstructured back and forth eye movement.

The eye gaze visualisations show that shiny areas (reflections of the borescope light) attracted the attention of the bench inspector, engineer and assembly operator, and that those areas were inspected multiple times. This was different to the borescope inspector and might be explained by pre-existing knowledge of the latter, i.e., there is no defect information retrievable from reflections. It is further apparent that all participants except the borescope inspector looked at the ceramic liners (top right corner). Equally, their eyes dwelled at the manufacturing stamps, which suggests that the participants were not familiar with those markings and unsure whether it was a defect. Finally, it is visible in the gaze visualisation that the borescope inspector focused on the edges only, while the other participants inspected surfaces such as the airfoil or the aforementioned liners.

The gaze plot of the borescope inspector was further analysed, as it indicated a quite specific search approach that is worthwhile to highlight. As shown in Figure 18a, the borescope inspector's eyes fixated on the defective area straight away (gaze plot 2 to 4). The search continued along the leading edge with some focus around gaze plot 8 (Figure 18b). After the leading edge, the participant continued their search on the trailing edge shown by gaze plots 9 to 12 (Figure 18c). This was followed by inspecting the platform (gaze plot 13 to 22 in Figure 18c,d). Finally, the leading edges of the subsequent blades were analysed (Figure 18e,f). The borescope inspector specifically focused on the trajectory of the foreign object after finding a defect on the foremost blade, i.e., from experience the participant would expect damage on subsequent blades as well. This behaviour was observed for all borescope inspectors and thus seems typical.

Overall, it stood out that for both inspection types (borescope and piece-part) the inspectors had a more systematic and structured search compared to engineers and assembly operators. The latter two showed a larger number of fixations, distributed across the stimuli. The long saccades (distances) in combination with several revisits of the same area led to long inspection times for both the engineer and particularly the inspector.

The bench inspector took 2.465 s (65%) longer in borescope inspection compared to piece-part inspection. Similarly, the assembly operator required 12.648 s (32%) more time in borescope inspection. Less surprising was the observation that the borescope inspector was faster in borescope inspection. However, it was indeed surprising that piece-part inspection took them twice as long, although only one blade was presented. Another interesting finding was that the inspection times of the engineer for borescope and piece-part inspection were almost identical, with only 0.144 s of a difference (<1%).

After reviewing the eye tracking data it can be concluded that the reason for the lower inspection performance in borescope inspection did not stem from overlooking the defect, but from incorrect decisions made, i.e., the defect was not recognised as a defect, but as an acceptable condition. The gaze plots further indicate that non-inspecting staff looked at more irrelevant features that might be visually more salient but not critical from a safety perspective.

The direct comparison of the piece-part and borescope image of the same blade revealed that the defect appears much bigger in borescope inspection than in piece-part

inspection. Herein lies the possibility that either the defect is more likely to be detected because it appears bigger, or, contrarily, the damage might not be marked as a defect because the operator does understand the magnification effect and that everything appears bigger in borescope inspection. Our results suggest that the magnification helped with recognising the defect by all levels of expertise. However, the reason for the incorrect decision remains unclear. Future work could address this question by using the think-aloud method to gain additional insights.

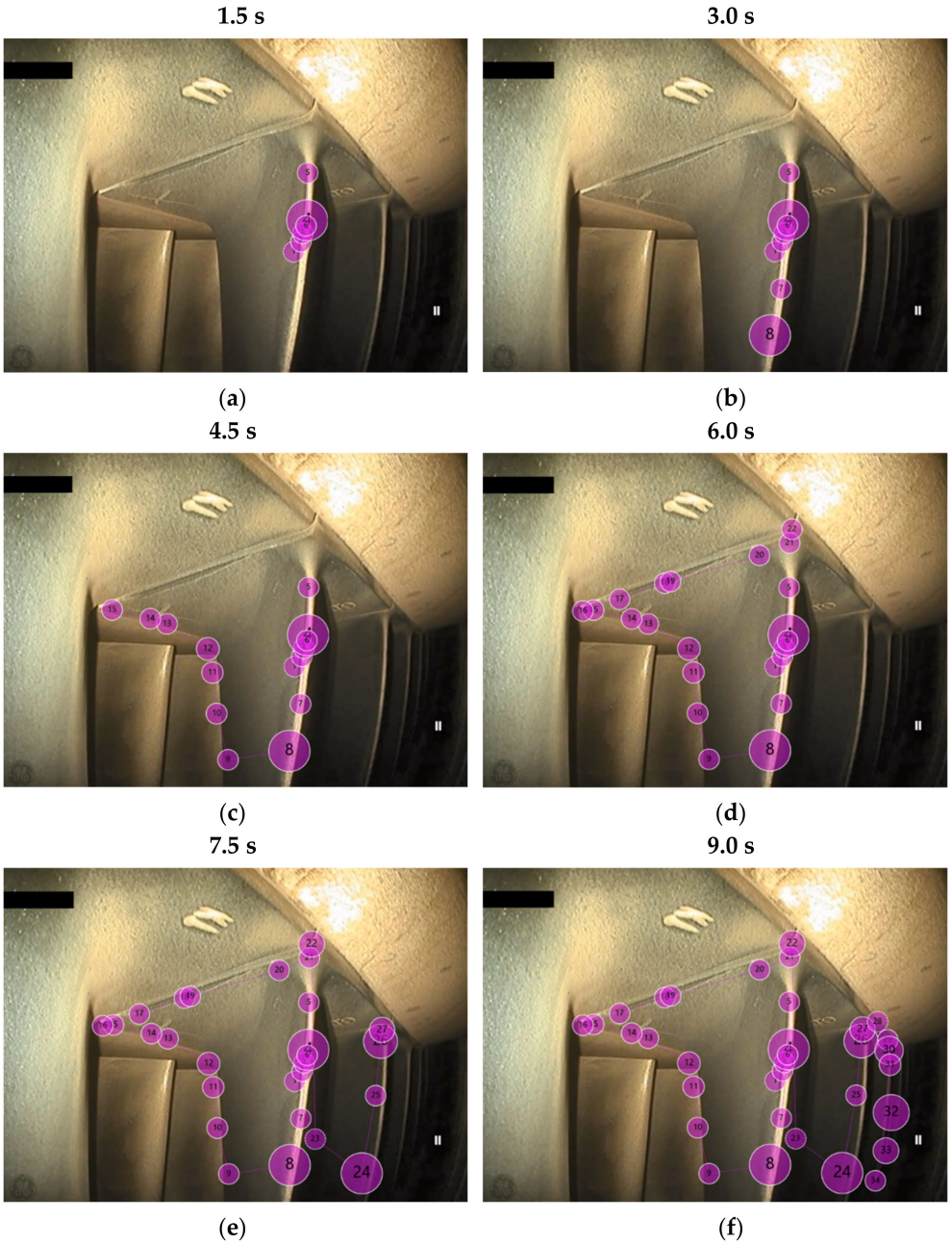

Figure 18. Gaze plot of an experienced borescope inspector in 1.5-s intervals.

5.2. Defect Type

As the results in Section 4.2 show, the lowest detection rates were achieved for dents, airfoil dents, cracks and blockage in descending order. While blockage and airfoil dents are less critical from a safety perspective, the successful detection of cracks and dents on the edges, however, is crucial. Hence, eye tracking was used to understand what caused the low detection rates of cracks and dents, respectively.

The cracked blade (Figure 19) showed a second defect, namely a burn. It is likely that participants focused on this more salient feature and hence overlooked the crack. The heat map and gaze plot in Figure 19 (left) confirm that the participant who missed the crack spent quite some time focusing on the top left corner where the burn was located. Furthermore, it becomes apparent from the gaze plot that the search was aborted once the first defect was found. While this is generally an effective approach and favourable from an operational perspective, it entails the risk that other defects that are more critical or require a different repair action are being missed. While the serviceability decision might be the same, the blade could have been removed from service for the wrong reason.

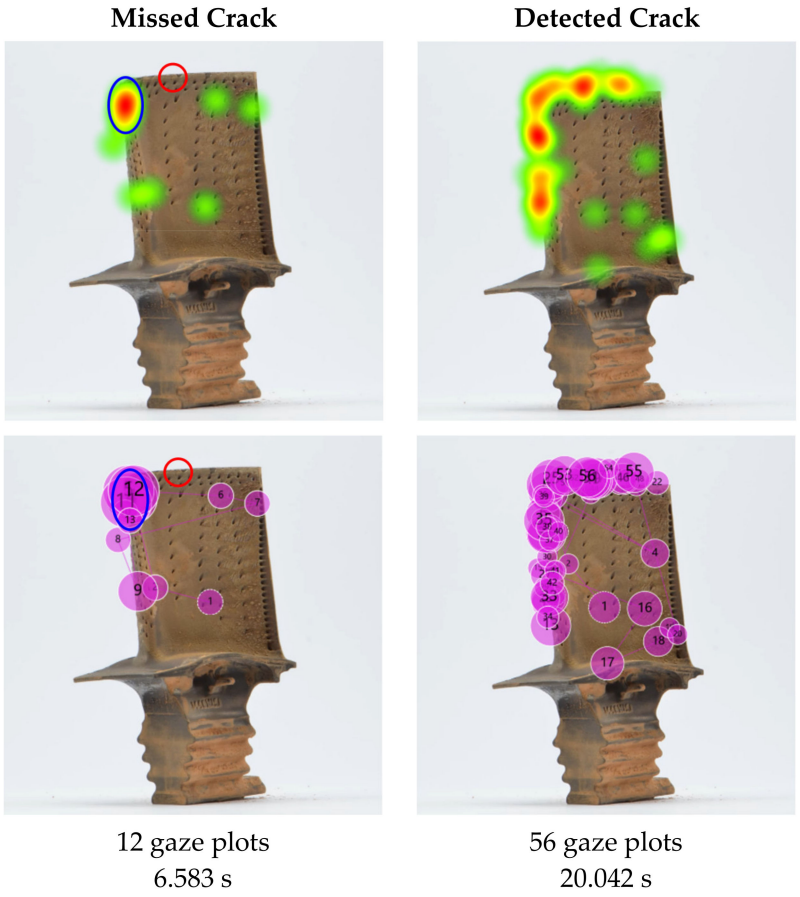

Figure 19. A blade with a crack (red circle) and burn (blue circle) was presented to participants. Heat maps and gaze plots were created for a missed crack (**left**) and a detected crack (**right**).

The eye tracking data of a participant who detected both defects (including the crack) shows that a more systematic and detailed search was performed with continuously inspection of the leading edge and tip of the blade from bottom to top, from left to right, and back repeatedly (Figure 19, right). The detailed search required more time as the eyes had to fixate on almost five times more areas as evident in the gaze plot.

In order to improve the reliability of visual inspection, it is necessary to understand why the defect was missed and what inspection error occurred. Therefore the framework introduced by Aust et al. [13] was utilised to determine the inspection error that might have occurred. There are three types of inspection errors: (1) search error, (2) recognition error, and (3) decision error. A search error occurs when the defective areas was not looked at, i.e., the eye tracking data shows no gaze plots in that area. When the defective area shows fixation points but they were below a set threshold, e.g., 600ms [33], then it can be concluded that a recognition error occurred. If the gaze plot on the defective area is above the threshold but the participant did not mark their finding and classify it as defect, then a decision error arose.

The dented blade in Figure 20 had a detection rate of 62% (31 participants), i.e., 19 participants (38%) missed the defect. The evaluation of the eye tracking data shows that all three types of inspection errors occurred. Of the 19 participants who missed the defect, four (21.1%) did not look at the defective area (search error). The eyes of another eight participants (42.1%) wandered over the defect location, but without recognising the defect (recognition error). The eyes of the remaining seven participants (36.8%) dwelled for a significant amount of time at the defective area with some returning to re-inspect the indication multiple times. However, they decided that the finding is acceptable (did not mark the defect) and thus a decision error occurred.

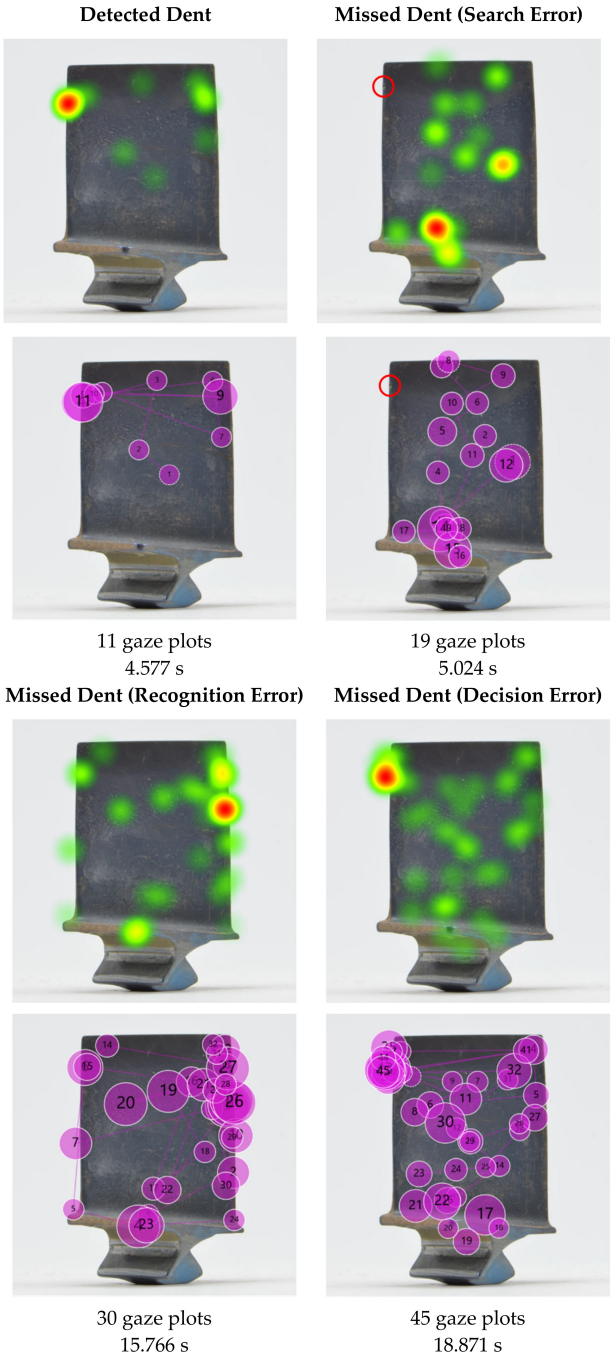

Figure 20. Blade with dent on the trailing edge (indicated by red circle). Heat maps and gaze plots were created, highlighting the different inspection errors.

The gaze plot analyses in Figure 20 indicates that the number of fixations and the inspection time T increases for each inspection error, i.e., $T(search\ error) < T(recognition\ error) < T(decision\ error)$. To test whether this was true for the entire research population, an ANOVA was performed around inspection time and inspection error. It was found that there is a strong correlation between the two, F(2, 19) = 44.182, $p < 0.001$. As shown in Figure 21, participants who made a search error needed on average 3.710 s, while inspections leading to recognition errors took 8.659 s, and participants who could not decide whether the finding was a defect or acceptable required 18.449 s. A generalised linear/non-linear normal log model was constructed around inspection time. The results are presented in Table 7 and show that the inspection times associated with all three types of inspection errors differ significantly from each other. Hence, the inspection time might indicate what inspection error occurred.

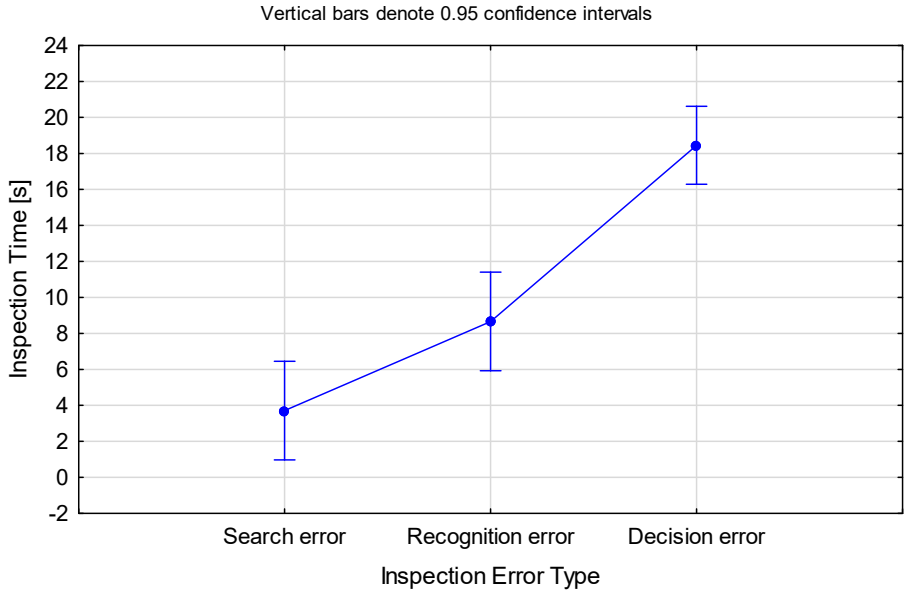

Figure 21. Correlation between inspection time and inspection error.

Table 7. Statistical model around the inspection time (Wald test and parameter estimates).

Effect	Reference Level	Level of Effect	Wald. Stat	Estimate	Lower CL 95%	Upper CL 95%	*p*
Inspection Error	Recognition Error	Search Error	14.2811	−0.817130	−1.24093	−0.393333	<0.001
Inspection Error	Recognition Error	Decision Error	43.3729	0.786754	0.55261	1.020895	<0.001

5.3. Severity

As expected, the inspection accuracy decreased with decreasing defect severity. An interesting finding, however, was that the inspection performance in the lowest severity level S1 differed significantly between the expertise groups. As the analysis in Section 4.3 showed, inspectors were more accurate than assembly operators, who in turn performed better than engineers.

The evaluation of the eye tracking recordings for the different expertise groups revealed that there is a predominant error occurring in each group, and interestingly it is a different type of error for each of them. Figure 22 shows that most inspectors made a search error (57.1%), while engineers struggled to recognise the defect 80.0% of the time, and two out of three assembly operators incorrectly classified the defect as acceptable (decision error).

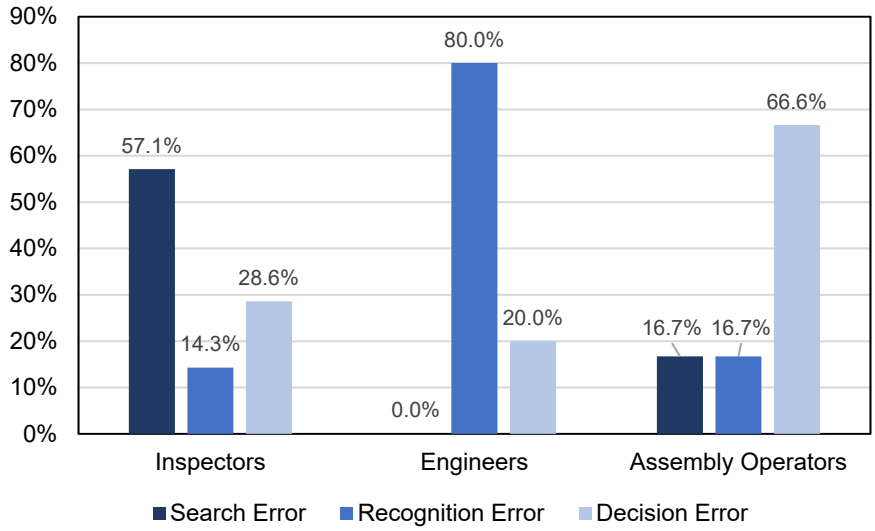

Figure 22. Distribution of inspection errors made by expertise group.

The eye tracking data of the participants who missed the defect were further analysed, and heat maps and gaze plots were created for a representative sample of each group and for the predominant inspection error. The results are shown in Figure 23.

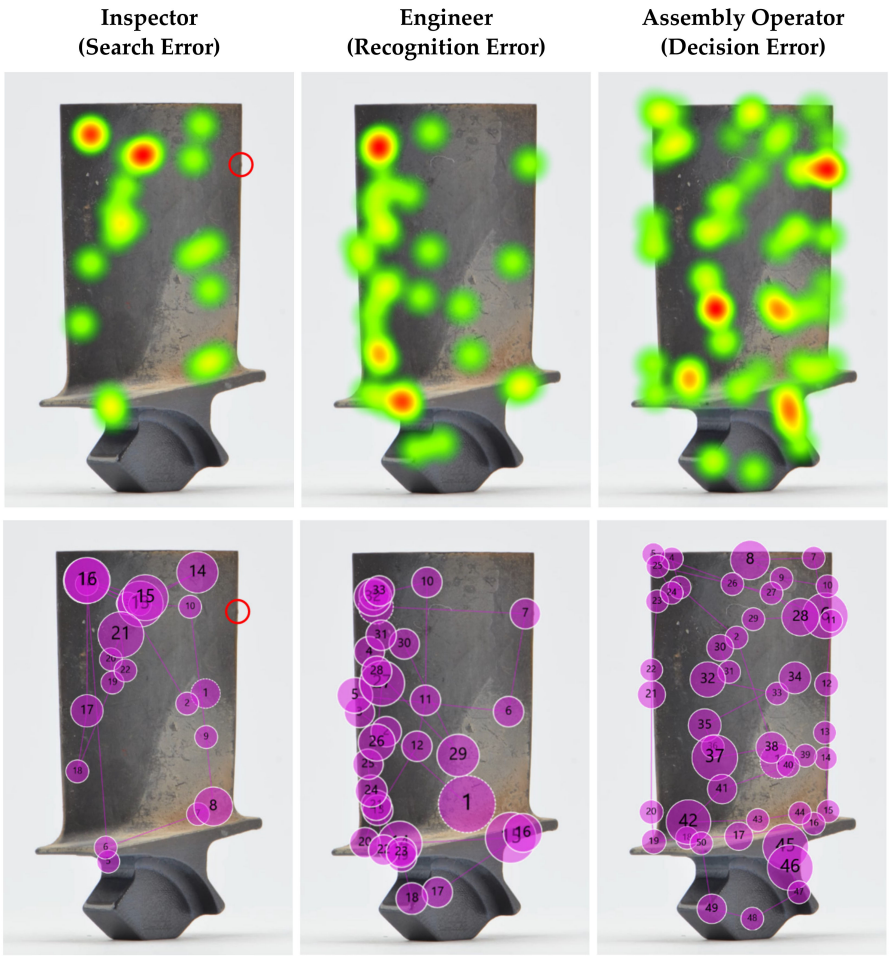

Figure 23. Eye tracking results for each level of expertise, highlighting the predominant inspection error. Defect indicated by red circle.

The eye tracking recordings of inspectors who missed the defect showed that a search error occurred and that other features attracted their attention. This can be deduced from the gaze plot in Figure 23, which clearly shows that the defective area was not looked at, i.e., not showing any fixations.

One inspector stood out from the others, as the recording showed an extreme short inspection time of 2.224 s. The eye tracking results revealed that this participant only skimmed the blade, which is evident by the few fixations (N = 9), each under 283 ms and thus below the recognition threshold. This phenomenon of an initial holistic scan prior to a detailed search is already described in [13,33]. However, in this case the inspector did not continue with a detailed search but rather continues inspecting the next blades and thus missing the defect. This search behaviour was only observed for one of the inspectors and thus is not representative for the expertise group.

The engineer's gaze plot indicated that the participant focused predominantly on the leading edge and thus missed the defect on the trailing edge. Their eyes scanned the defective area early on (gaze plot 7) for 197 ms, but never returned to inspect it in more detail. This infers that the participant did not recognise the defect and a recognition error occurred.

The heat map and gaze plot of the assembly operator highlighted that the participant looked at and recognised the defective area (red colour in heat map and large gaze plot #6). However, they did not mark it as defective, which leads to the conclusion that a decision error occurred. The presented blade is relatively dirty with deposits on in the airfoil. A possible explanation for the decision error might be that the participants identified the irregularity as a deposit on the edge as opposed to an edge defect.

Another interesting finding from the eye tracking data is a notable difference in inspection time between the three groups. The inspector was 3.232 s (31.5%) faster than the engineer, who was 6.952 s (40.4%) faster than the assembly operator. Consequently, the inspector was 10.184 s (59.2%) faster than the assembly operator. This aligns with previous studies [13]. It also matches the findings in Section 5.2, whereby search errors are associated with short inspection times, while decision errors are linked to long inspection times.

When comparing the eye tracking results of the three expertise groups it becomes apparent that the gaze plots varied significantly. While the scan path of the inspector showed fewer gaze plots and a more systematic search in a counter clockwise circle, the engineer's gaze plot exhibited more fixations, predominantly on the leading edge. The assembly operator in turn inspected the entire airfoil and root of the blade in an unorganised, almost chaotic way. Nonetheless, the more detailed search led to recognition of the defect, although the serviceability decision was incorrect, i.e., the defect was classified as acceptable (decision error).

5.4. Blade Perspective

The statistical analysis revealed that perspective P3 is overall the best viewing perspective for defective and non-defective blades. However, the inspection accuracy is still not 100%, and the eye tracking data were further analysed to understand why even the optimum perspective led to missing the damage. Figure 24 shows the heat maps and gaze plots of two representative participants who missed a bend on the leading edge. The inspection times and number of gaze plots of the two are fairly similar and do not stand out from other participants who detected the defect. Thus, no conclusions can be drawn from these quantitative numbers.

The first participant looked several times and for quite a while at the defective area, as evident in the heat map and gaze plot (Figure 24, left). This means that a decision error occurred, which could have been caused by insufficient training or inexperience. In the second case, the heat map (Figure 24, right) shows that the first area the participant looked at was the defect. The participant did however continue the search immediately without spending much time on it. However, based on a dwell time of 517 ms it can be assumed that the feature was recognised.

Both cases confirm that perspective P3 is beneficial for detecting deviations and that missing the defect was caused by incorrect decision making rather than due to visual search capabilities or inappropriate search strategies.

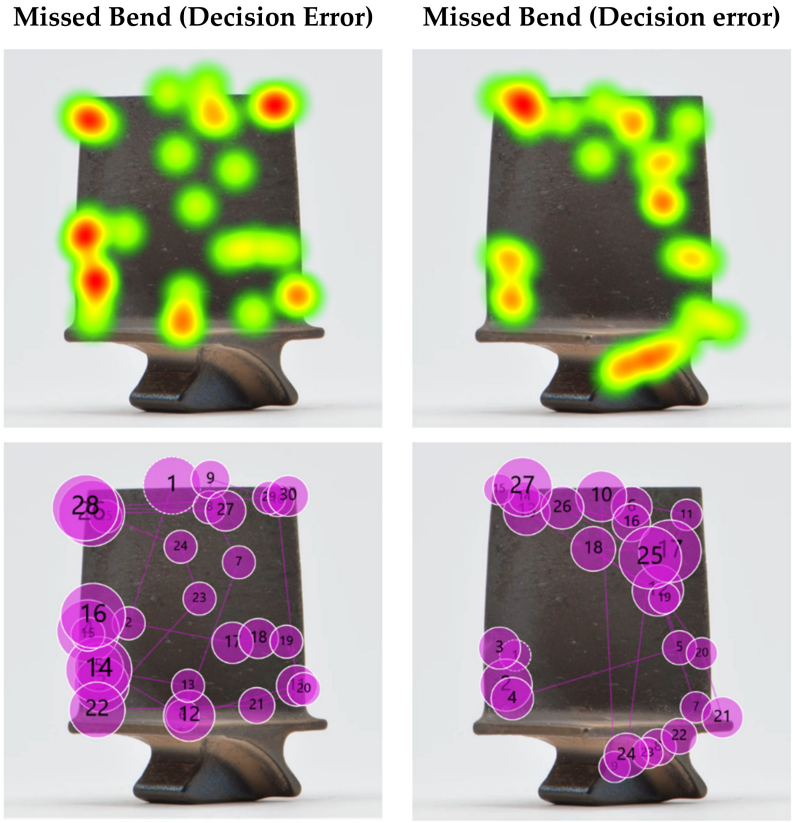

Figure 24. Two participants missed the defect from the most favourable blade perspective.

6. Discussion

6.1. Summary of Research Findings and Comparison with Other Studies

In this study the effect of different influence factors on the inspection performance was evaluated. Statistical analysis was used for screening of the significant effects, followed by semi-quantitative assessment of the eye tracking data to gain additional insights into the inspection process. The tested hypotheses are summarised in Table 8 along with a brief overview of the results. The findings are further discussed and compared to other studies in the field.

6.1.1. Inspection Performance

Overall, an inspection accuracy of 75.5% across all influence factors was achieved. This is comparable to the inspection performance reported for other maintenance activities, ranging from 53% to 77% [9,13,34]. Similar accuracies were measured in the manufacturing industry and range from 45% to 76% [6–8,35,36]. Interestingly, the performance seems to settle around the 80% mark, independent of the industry, inspecting part, and defect type (manufacturing or operational). This suggests that there may be a natural limit to human performance.

The results show that the inspection rates in borescope inspections are significantly lower than for piece-part inspection with 63.8% and 82.6%, respectively. This was consistent for all three groups of expertise. Since it is the first study that quantifies the borescope inspection performance, no comparison to the literature could be made.

Table 8. Research hypotheses and findings.

Hypotheses		Findings
H1	The inspection type affects the inspection performance.	Accepted. Piece-part inspection showed a higher inspection rate than borescope inspection.
H2	The defect type affects the inspection and classification accuracy.	Accepted. Nicks, tears, and tip curls had the highest detection rates, while blockage, cracks, and airfoil dents were the categories more often missed. The classification accuracy was highest for coating loss, nicks, blockage, and tip curl. Airfoil dents, burns, and tears were most difficult to classify.
H3	The defect severity affects the inspection performance.	Accepted. The inspection accuracy decreased with decreasing severity and the critical threshold was identified between severity level S1 and S2.
H4	The blade perspective affects the inspection performance.	Accepted. Perspective P3 lead to the highest accuracy, while P2, P4, and P8 also achieved good results. The worst perspective was P5.
H5	The background colour affects the inspection performance.	Rejected. The background colour had no impact on the inspection performance.

Assessment of the defect types revealed that some defects are more difficult to detect than others. Airfoil dents (19.0%), cracks (11.0%), and blockage (8.0%) showed the lowest detection rates, while nicks (100.0%), tears (95.5%), and tip curls (89.0%) had the highest. This supports previous research whereby salient defects such as tears had a high detection rate and surface defects (e.g., airfoil dents) were often missed [37]. This was consistent for different inspection methods including visual and visual-tactile inspections [37]. Megaw and Richardson [4] pointed out that knowing the critical defects and their appearances can lead not only to better search strategies, but also to higher inspection performances. This corresponds to the comments made by some participants of the present study (mainly non-inspecting staff), who stated that a defect list with a description and sample photograph would have been helpful to better detect and classify defects.

The classification accuracy was also dependent on the defect type. While airfoil dents (5.3%), burns (38.4%), and tears (44.9%) showed the lowest performance, coating loss (98.1%), nicks (90.0%), and blockage (87.5%) were most accurately classified. Across all defect types an average classification accuracy of 62.3% was achieved, which is higher than the 39.1% reported in previous work [37]. This could be due to the current study having a larger research population and sample size with a bigger variety of defect types with a more distinct manifestation. Examination of the classification results identified specific confusion between defect types. The biggest confusion occurred between dents and nicks, with 45.7% of dents being classified as nicks. This is in alignment with previous findings [37] where tears were often confused with breakage. In the present study 21.9% of tears were confused with breakage, and 21.3% with cracks. The present findings show that bends are the most common defect type that that is confused with tip curl in 11.5% of the cases and reflects previous findings [37]. None of the other defect types and their misclassification distribution were previously analysed.

A study by Spencer [16] found that if an area had two defects, e.g., in the case where one defect propagated to another one, the inspectors would name only one or the other. In the present study we found that multiple participants understood the propagation well and verbally expressed the propagation of the defect. Since they were asked to select only

one defect type, they would naturally choose the worst one as it validates the removal of the part from service.

During the study, the researchers noted that several participants described defects with their own terminology, e.g., they were calling a nick a 'notch'. Moreover, defects tended to be classified based on their visual appearance rather than on contextual knowledge or by taking into account the potential root causes. Hence inconsistency was found in the use of defect terminology. Many participants were ignorant, causing inability to discriminate the defects. Appropriate training and a standardised defect taxonomy might offer potential to improve the classification accuracy, see also [25].

Defects of severity level S1 (72.0%) were more difficult to detect than severity level S2 (92.8%) and S3 (99.0%). Previous studies [14–18] reported the detection performance in the form of probability of detection (PoD) curves, which cannot be compared to our findings. However, it is generally accepted that with increasing defect size, and thus severity, the likelihood of detecting the defect increases [15].

The blade perspective had a significant effect on the inspection performance with the best perspective being P3 (87.5%) followed by P2, P8, and P4 with 81.5%, 81.0%, and 80.5%, respectively. The worst perspectives were P5, P1, and P7 with 51.0%, 66.5%, and 69.0%, respectively. Megaw and Richardson [4] analysed the perspective factor in visual inspection of electrical connectors. Those authors determined the best perspective based on eye tracking parameters, namely fixation times and number of fixations, while we ranked the different perspectives based on the inspection accuracy. Several studies [14,15] assessed the effect the tilting angle of flat composite panels has on the detection rate of surface dents. The main difference to our work is that we analysed the perspective based on an incremental rotation of the part as opposed to tilting. Moreover, in [15] the angle between light source and panel surface changed with changing tilting angle, thus a combination of angle and lighting effect was assessed. While the results are not comparable to our study, it might provide an opportunity for future research, i.e., to assess the effect of illumination including light source angle and distance, colour temperature, and luminous flux. An interesting finding is that the identified favourable perspectives of the present study are also beneficial for automated inspection systems such as the one in [32]. Both the human and software apply (computer) vision for the visual search and inspection task. We propose that the perspective highlights a variety of edge and surface defects and thereby contributes to visual perception and recognition.

Surprisingly, the background colour had no significant effect on the inspection accuracy. This is contrary to other studies [4,14,15]. Waite [38] found that surface damages are easier to find on green painted composite panels. In the present study, the green background colour showed the highest positive-rate (TP & FP) and was significantly higher compared to the white background, $F(3, 796) = 2.315$, $p < 0.05$. However, due to the high false-positive rate, the inspection accuracy was overall the lowest. Hence, it can be said that a green inspection background is beneficial for detecting any irregularities that require further investigation, while it is less favourable for making a serviceability decision as it leads to a high rejection rate of serviceable (non-damaged) blades. Another interesting finding of the present study was that there seems to be a personal preference towards one colour or another, but without a clear tendency towards a specific one for the research population as a whole.

This study generally found no correlation between expertise and inspection performance (accuracy). The only difference measured was in severity level S1, i.e., blades with very small defects, which accounts for 8.75% of the research sample. In this specific case, inspectors performed best (80.2%), followed by assembly operators (73.2%), with engineers performing the worst (61.6%). The general observation of the present work is supported by several studies [13,16,35,39]. The S1 finding is also consistent with other research [15,40]. This shows that the general literature is ambiguous about the effect of the expertise. The lowest severity level S1 was perceived as most challenging according to participants' comments. Thus, the significance of expertise in this case might be explained

by the task difficulty, i.e., the more challenging the inspection, the more important the previous experience and contextual knowledge.

The statistical analysis revealed that none of the demographic factors had a significant effect on the inspection performance (with expertise being the only exception in the severity sub-study, as discussed previously). This is in line with the earlier literature [8,13,15] that found that demographic variables such as work experience in the industry, previous experience in inspection, education, certification, and visual acuity did not affect the inspection performance. It is possible to conclude that there are other personal factors that need to be considered, such as visual perception capabilities that go beyond the sole visual acuity, e.g., search strategies or cognitive feature recognition. Furthermore, there might be a knowledge component affecting the serviceability decision and thus performance. Since the years of work experience in the industry seem to have no effect, the knowledge might be individually developed and could have been influenced by previous training.

6.1.2. Inspection Approach

Assessment of the eye tracking data in the form of heat maps and gaze plots showed the complexity of borescope inspection compared to piece-part inspection in the form of longer inspection times and the larger number of fixations. This is in alignment with the general literature, which agrees that an increase in eye tracking parameters indicate more complex tasks and a higher cognitive workload [40–47].

The eye tracking recordings further showed significant differences between the different groups of expertise, which was in accordance with previous work [13,35,39,48]. A review of the gaze plots revealed that inspectors made fewer fixations than engineers and assembly operators. Moreover, it was apparent that bench inspectors in piece-part inspection and borescope inspectors in borescope inspection, respectively, applied a systematic search strategy with clear focus on the edges. Engineers and assembly operators, in contrast, inspected the blade in an unstructured way. Their gaze plots showed widely spread fixations across the stimuli and multiple revisits of the same areas. This could indicate a level of 'technical anxiety'.

There is a relationship between the number of fixations and inspection time, i.e., more fixations mean additional dwell time for each of those and more 'travelling' (time) for the participants' eyes between the fixations [4,13]. Therefore, it was not surprising that inspectors performed the fastest, while engineers and assembly operators required more time. Comparison of the findings with those of other studies confirms that experts have a clear search strategy with fewer fixations and thus shorter times compared to novices [13,16,35,39].

Consistent with the literature [8,13,16], this research found that participants looked at salient features first and most often. If a blade had more than one defect, participants tended to detect the most obvious one, while less salient defects were missed. The results further support the idea of an underlying mental model introduced in [13], i.e., previous work experience and contextual knowledge influences the search focus and inspection approach. This was observed in several ways:

Firstly, as previously discussed, inspectors (experts) focused on the blade edges, these being the most critical areas. With experience, they might have developed their individual inspection patterns, e.g., visual circuits with focus on areas where defects usually appear. These findings are consistent with those of Spencer [16], who concluded that experts develop an expectancy of where to find defects. Therein lies the risk that the inspection is performed too rapidly due to high self-efficacy (over confidence) and following a deeply-rooted approach, whereby only suspect areas are inspected, rather than the entire part. This behaviour might explain why in the present study the most common inspection error made by inspectors was a search error, i.e., the defective area was not inspected, possibly because the participant did not expect a defect at this location. This could further explain why their inspection performance was not significantly better than any of the other groups.

Secondly, if a defect was detected that justified the removal of the part from service, then the search was aborted and the next blade was inspected. Working in an environment under time pressure imposes this behaviour further. Industry practitioners understand the time constraints and aim to work efficiently. Hence, searching for additional defects on a blade that has already been identified as unserviceable would be a non-value adding activity and considered as 'waste' according to lean principles.

Furthermore, inspectors were able to distinguish defects from other conditions such as deposits or reflections of the borescope light. This could be seen in their eye tracking recordings, whereby the gaze plots showed no fixations in areas with conditions, but rather a clear focus (fixations with long dwell times) on the defect.

An interesting finding was that borescope experts inspected anticipatorily, i.e., when a defect was found on the foremost blade (edge), they drew an imaginary trajectory of the foreign object that caused that defect and searched for any damage on the subsequent blades along that trajectory. Figure 25 shows the defect markings of a borescope inspector (red circles) and the trajectory (blue line, added later for better understanding). This search behaviour was only observed for borescope inspectors and could have been consciously or subconsciously undertaken.

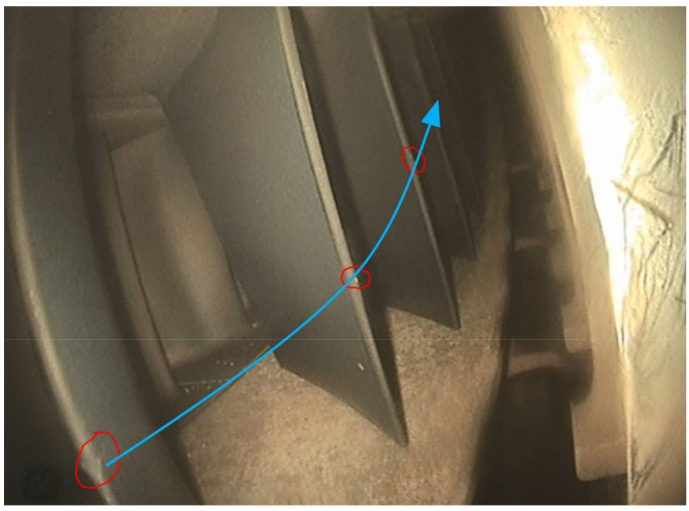

Figure 25. Inspection results of a borescope inspector (red circles) and foreign object trajectory (indicated blue line).

The results show that eye tracking can be used to identify the occurring inspection error of each individual operator and thus areas of improvement. For example, search errors are visible in the gaze plots in the form of missing fixations in the suspect area. The visual locus does not overlap with the defective area. If participants fail to pause their search to look at an irregularity, a recognition error occurred. The gaze plot shows short fixations below the recognition threshold. It appears that recognition errors often occur in situations where people get distracted by other, more prominent features like reflections, deposits or other defects. When the defect was detected but misjudged as being an acceptable condition, then a decision error was made. This is related to the classification ability and understanding of the quality systems.

An interesting finding was that each expertise group tended to make a different inspection error. Inspectors were prone to search errors, while engineers did not recognise anomalies and continued their search without pausing. Assembly operators struggled to differentiate between defects and conditions, indicating the lack of contextual knowledge and experience in inspection. Moreover, this study found a correlation between the inspection time and type of error. Both findings were not previously reported in the literature and add to the understanding of visual inspection.

6.2. Towards a Revised Visual Inspection Framework Including Inspection Errors

The existing framework for visual inspection entails five functions: Initiate, Access, Search, Decision, and Response [34,49]. This concept may need to be reconsidered in the light of the present findings and previous research [13], taking into account the eye tracking observations.

Firstly, the current results show the importance of the visual locus in the search phase, i.e., the specific search strategy for the preferred retinal locus that the operator selects (perhaps unconsciously) to guide their eye movements. Secondly, the decision component in [34] has been replaced by two processes, namely a recognition and a judgement activity. The recognition comprises ontological knowledge of defect terminology and discrimination ability of irregularities. This is different from the search and the decision phase because search is motoric, i.e., physical movement of the eye, while recognition is cognitive, i.e., the brain processes information and pauses once an alarming feature is found. Thus, the recognition phase is added as an individual step to the inspection framework. Thirdly, the judgement phase requires knowledge of the serviceability conditions and includes an accept-reject decision. If required, a closer inspection might be performed before the defect is confirmed. In the case of borescopy this might involve replaying the video, and re-searching for other visual diagnostic clues. For inspection more generally this might involve looking at the part from other perspectives, tactile inspection, use of magnification, or commitment of the part to another non-destructive test (NDT). Errors in the judgement activity correspond to the 'decision errors' in the above text. Fourthly, the inspection process finishes with the completion of the inspection task, making it a continuous process (see Figure 26). The different process steps of the proposed visual inspection framework are further described in Table 9.

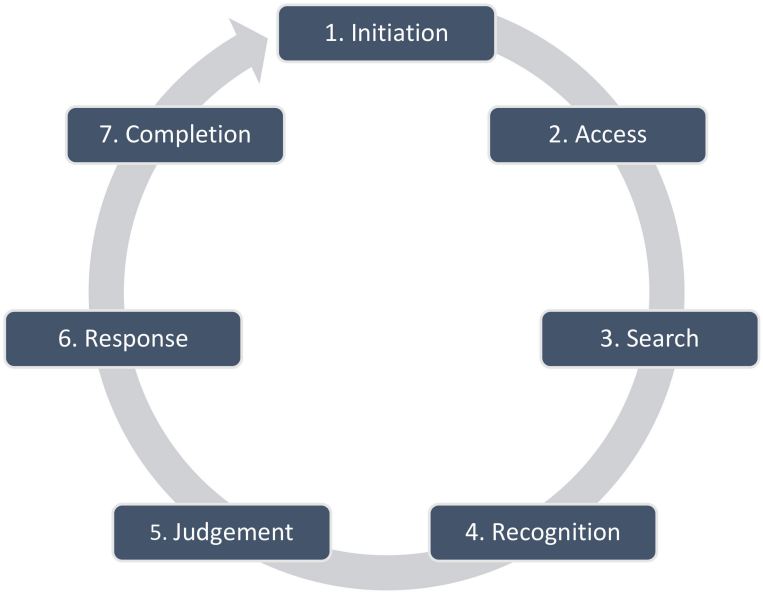

Figure 26. Visual Inspection Framework.

Table 9. Description of the Visual Inspection Framework exemplary for borescope inspection.

Process Step (Hazard)	Process Description	Potential Error (Top Event) and Possible Causes (Threats)	Tasks in the Example of Borescope Inspection (Barriers)
1. Initiation	Inspection workplace setup and part preparation. Provision and setup of required inspection tools. Comprehend standard working procedure (SOP).	Initiation error: • Incorrect part preparation • Inadequate, missing, or non-compliance with procedures • Inadequate setup knowledge • Incorrect, inoperative, or non-calibrated tools • Inadequate inspection environment	Pre-wash engine, select appropriate borescope tip and correct camera settings, read and comply with engine manual and standard work procedures.
2. Access	Locate inspection area and gain access to location. Ensure best possible position for reliable part inspection.	Access error: • Incorrect part accessed • Incorrect part presentation (distance, angle, or lighting) • Part or surrounding components damaged during access	Remove borescope hole plug, insert borescope, and manoeuvre borescope into appropriate location for blade inspection.
3. Search	Comprehensively scan the part and systematically search for any irregularity. Ensure an adequate search strategy covering the entire stimulus is used.	Search error: • Inappropriate search strategy • Inexperience staff • Human factors, e.g., fatigue • Operational time pressure	Start video recording, initiate engine rotation, and search for any damages on the blade.
4. Recognition	Process visual information and perceive indications of possible anomalies. Recognise that the part differs from its ideal condition and discriminate against other possibilities.	Recognition error: • Indication missed • False memory of ideal part • Distracted by other (more salient) features • Human factors, e.g., fatigue	Pause engine rotation if indication is found. Inspect finding in detail.
5. Judgement	Classify indication as condition or defect. Determine defect type and compare finding to corresponding limits in standard. Decide whether finding is within limits (acceptable) or outside limits (reject).	Decision or judgement error: • Irregularity forgotten before decision is made • Misclassification of indication • Misjudgement or incorrect measurement of indication size and location • Incorrect comparison to standard	Decide whether the finding is acceptable or needs to be repaired, i.e., whether an engine tear-down is required.
6. Response	Record and report findings. Complete inspection documentation.	Response error: • Findings not reported • Incorrect documentation • Task not signed off	Take a picture (snapshot) of the paused borescope video, report finding (classify defect and location). Repeat from step 3 until all blades are inspected.
7. Completion	Remove equipment from inspection area and return to storage for next use. Release part to next process in accordance with inspection outcomes (e.g., repair).	Completion error: • Misinterpretation of the reported findings • Incorrect action performed • Incomplete action	Remove borescope and re-install borescope plug. Repeat from step 2 until all stages have been inspected. Continue maintenance procedure, e.g., engine tear down and repair.

The inspection framework can be used as an operational guide and can be represented in two ways: (a) as a table, or (b) graphically as a bowtie diagram. The two representations are complimentary and could be used for different purposes, i.e., to express different

information. While the table can be used as a checklist to manage the inspection task, the bowtie diagram allows representing barriers and escalation factors in a concise way that is easy to communicate and to understand, which may have benefits for training. Moreover, the bowtie diagram allows for the colour-coding of the barrier to show their effectiveness, and using the 6M framework to further explore other contributing factors [22].

A bowtie diagram was drawn for the first process step (Initiation) to exemplify the concept, see Figure 27. The process step forms the hazard of the bowtie diagram and the top event is the inspection error that can occur in this step. The threats are any inherent risks in the process that can cause an inspection error, e.g., incorrect part preparation or incorrect equipment. Any procedure that is part of the process can act as a barrier, e.g., compressor wash or selecting an appropriate borescope tip. Moreover, preceding procedures, such as appropriate training, can be included as barriers. If a task is performed incorrectly and the barriers fail, an error occurs. It can cascade through the subsequent processes, causing an incorrect serviceability decision (e.g., missing a defect), ultimately affecting part reliability and operational safety. The development of a bowtie representation of the visual inspection framework is not the key objective of this paper, hence only the 'Initiation' phase is shown here. There is an opportunity for further research.

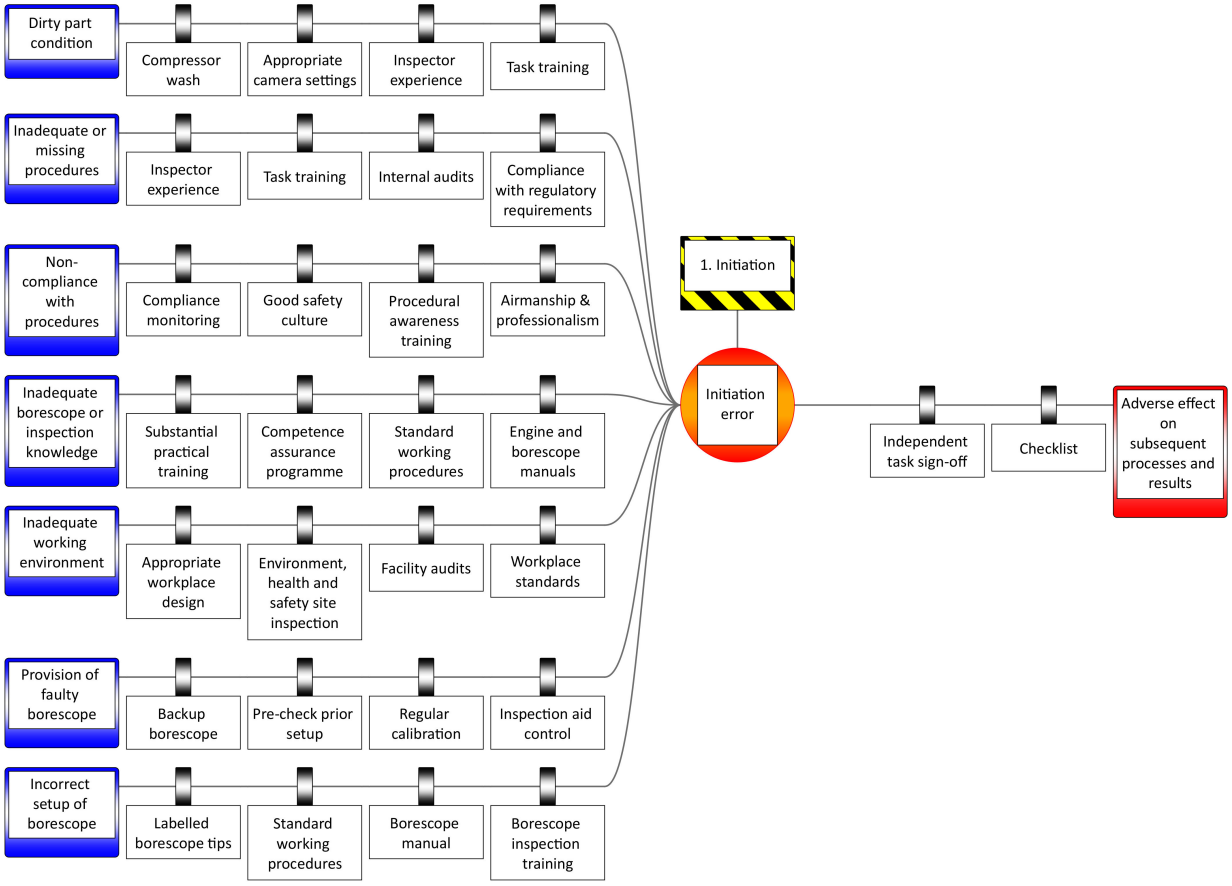

Figure 27. Exemplary Bowtie Diagram for the first step of the inspection process.

In this study there were cases of defects being missed. It should be noted that from an operational perspective, a missed defect is not necessarily a safety issue. This is because the safety consequences are also affected by defect type, size, and severity. For example, a crack might cause a blade to fracture and result in a catastrophic engine failure, whereas the much larger feature of an airfoil dent might merely decrease the fuel efficiency of the engine. Furthermore, the safety outcomes are moderated by regular inspection. From a safety perspective the fundamental objective is that any blade condition that could propagate to a dangerous state during the next engine tour (before the next regular engine shop visit),

should be removed from service before that tour begins. Hence, the minimum inspection requirement is for the operator to ensure that the blade under examination does not have any such conditions. However, this is a complex serviceability decision because of the aforementioned interaction between defect type, size & severity. There are heuristics for this decision, as represented in the engine manuals, but nonetheless a judgement must be made by the operator. It is natural that operators will use the precautionary principle when applying their agency, i.e., err on the side of rejecting a blade rather than risk the adverse consequences to the engine, aircraft, and passenger.

Being a high reliability industry, operators are conditioned by their training and the work culture to make safety-conservative decisions. Nonetheless, missed defects do occur. At a first approximation it can be assumed that this is a genuine slip, lapse or mistake, rather than perverse agency [50]. The eye tracking analysis was useful in identifying where the people gazed relative to the defect, and this informed the development of the revised visual inspection framework. Specifically, we noted that one cause of missed defects was related to the search—recognition—decision error progression (see Figure 23 and related text). We noted a variety of visual search strategies, even within one group of operators, not all of which were effective. In other cases participants looked at a defect, even for some time, but did not classify it as a defect in the end (serviceability decision error). We tentatively suggest that a way to protect against missed defects might be to more deliberately talk about visual search strategies and decision errors when training operators.

It was also apparent from the eye tracking results that borescope inspectors were all using a similar search strategy of examining the edges (see Figure 18). This was generally effective and efficient. However, it should be noted that the serviceability decision that needs to be made by a borescope inspector when examining an engine on the wing is not the same as an operator examining disassembled blades. The borescope inspector needs to determine whether or not the engine needs to be committed to tear down, and defects to the blade edges are the key determinants. This illustrates that it may be possible to develop specific inspection protocols, including visual search strategies, but these would presumably need to be contextualised to each situation.

6.3. Implications for Practitioners

6.3.1. Inspection Environment and Emerging Technologies

The insights gained from this study concerning the effect of different influence factors on the inspection performance could be used to improve the inspection environment and processes, and thus performance. For instance, the favoured perspective could be considered in the engine design (location and orientation of borescope holes). Another approach might be using the most beneficial perspective to standardise the image and video acquisition in piece-part and borescope inspection, respectively. This could be achieved with advanced technologies such as continuum robots [26–28]. Standardisation increases the repeatability and is desirable for automated defect detection software using conventional image processing [32]. It would also benefit artificial intelligence (AI) systems with deep learning (DL) algorithms, as it would require smaller datasets to train the AI, making the training faster and cheaper [51].

6.3.2. Training Implications

This study identified different inspection errors that occurred. Due to the different nature thereof, different training interventions are required. We propose that eye tracking can be used to evaluate any one inspector and identify which types of errors they are most prone to. Subsequently, customised training can be provided to selectively address those errors and improve their performance. For example, if someone shows a high search error rate, then it might be valuable to provide this person with a search strategy, e.g., scanning along the edges of the blade. As previous studies already indicated, there might be an opportunity for eye tracking being used as a training tool by playing the gaze recording of an inspection expert to learn by example.

If a recognition error occurred, staff might need more practice in recognising relevant features that require further assessment. This relates to the person's visual sensitivity, and slowing down their search could potentially allow them to recognise more irregularities. We would tentatively recommend not putting them under time pressure, particularly with regard to novices and new staff.

Decision errors could possibly be mitigated by having a training set of blades with a variety of defects and conditions. It is important to train staff in the difference between a defect and an acceptable condition. Currently, there is no consistent ontological description of defect types within the industry and thus it cannot be assumed that everyone in an organisation understands the difference between the different defect types. We also notice a conservativeness towards part rejection leading to high false positive rates, which shows the importance of operators knowing the criteria for part acceptance and rejection. Ultimately, it is tentatively recommended to work towards a common understanding within an organisation and the industry by applying a standardised defect taxonomy such as the one in [25].

While some factors such as defect type or severity cannot be influenced, it was still worthwhile to analyse them. The insights could be used by maintenance providers to tailor their training endeavours and focus them around the critical and difficult detectable defects and defect locations. The mentioned training implications have all the potential to improve the inspection performance and staff competency, thereby assuring flight safety [52–54].

6.4. Limitations

This study has several limitations. Firstly, the blades were presented as piece-part images on a computer screen rather than handing the physical part to the participant for inspection. Similarly, the borescope images were stills from a borescope video that inspectors would usually see. In both cases, images were used to allow for eye tracking recording and repeatedly measured with consistent parameters such as lighting or blade perspectives. Thus, images provided the best solution and somewhat represent the borescope inspection, which is already a screen-based inspection. However, we acknowledge that eye tracking glasses might be better for piece-part inspection, as it allows for the recording of the eye gaze while holding the actual part, and could be considered for future research. Previous research [37] measured an inspection accuracy of 70.5% when presented with images, and 84.0% when inspecting physical parts. This translates to an improvement of 19.1% over the images. However, the accuracy of image-based piece-part inspection in the present paper was already 82.6%, which is higher than the 70.5% measured in [37]. Thus, the benefit of physical handling might be much lower than 19.1%.

Secondly, there were only two borescope inspectors participating in this study. Thus, it was not possible to include them as an individual group in the statistical analysis. The small number was due to limited staff available at our industry partner with certified borescope experience. Future work could analyse the borescope inspection further, allowing for a sufficiently large research population for statistical evaluation.

Thirdly, safe flight operation is of utmost importance and thus the inspection accuracy was chosen as the main performance measure. The influence factors might also have an effect on the inspection time. However, the inspection time was not analysed in this study due to the nature of eye tracking analysis being a tedious process and could be addressed in the future when eye tracking analysis will be further automated and less laborious. Refer to [13] for a detailed overview of the limitation of eye tracking technology.

We introduced three levels of defect severity to allow for quantification of the inspection risk using the framework provided in [31]. An alternative is to describe the severity as a function of the defect size (similar to a PoD curve) [18]. Still another way of defining severity is to take into account the defect tolerances such as fatigue life [15]. Furthermore, the location of the defect plays a crucial role and is included in the engine limits of the different tolerance zones.

There is a risk that a memory effect might have occurred when the same blade was presented multiple times, e.g., with different background colours or from different perspectives. To minimise this effect, the stimuli were presented in random order and distributed across a large sample set of 120 images. Nonetheless, the possibility of an occurring memory effect cannot be entirely ruled out.

6.5. Future Work

Some future work streams were already discussed previously and are not repeated here. This study did not seek correlation between the various influence factors. Future research could explore those interrelationships. The findings of the present study indicate that colour is not a major variable and therefore we tentatively recommend that if there was a wider study looking at the associations between the variables that the colour variable could be excluded.

There is an interest to study how inspection operations can be enhanced to improve the inspection accuracy and get closer to the 100% mark. Future work could explore several pathways including but not limited to (a) increasing the performance of the human operator by providing better training (as discussed earlier); (b) improving the inspection processes by e.g., introducing a consecutive, independent inspection or developing procedures to counteract human factors; and (c) introducing emerging technologies such as artificial intelligent software for automated defect detection or 3D scanning technologies to complement the human operator.

The performance of the participants and the effect of the influence factors was measured based on the inspection accuracy. However, there are other ways of determining the performance and selecting the best parameter for each influence factor. Weighted statistical analysis could be used to reflect the importance of the defect, e.g., based on the frequency with which the defect type occurs during visual inspection, or based on the associated risk if the defect is missed (i.e., critical defects are weighted more than non-critical ones). This could result in a different outcome when comparing the different groups of expertise.

A quantitative comparison of the eye tracking results of the different participants and expertise groups could potentially be made using areas of interest (AOIs). AOIs are defined regions of a stimulus for which eye tracking data can be specifically extracted. An example for blade inspection is given in Figure 28. After identifying relevant areas, eye tracking metrics such as the number of fixations or the time spent in that area can be extracted and analysed statistically. This is a manual and laborious process and might not be applicable to borescope inspection, where parts of the blade are covered by other blades and the image is highly distorted.

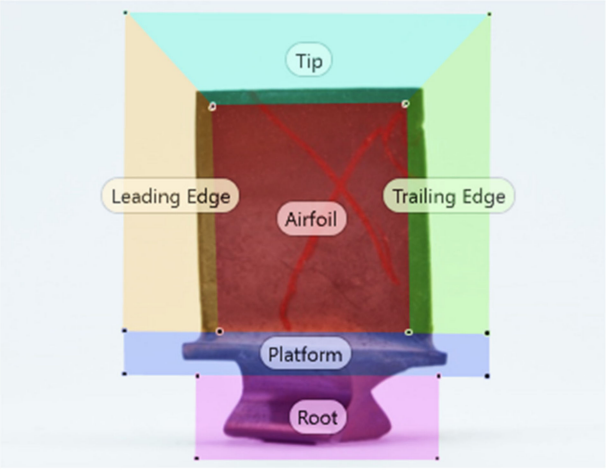

Figure 28. Blade with areas of interest (AOIs).

7. Conclusions

This work makes the following original contributions to the field. Firstly, the effect of different influence factors on the inspection performance was quantitatively and qualitatively assessed. Those factors include the type of inspection, defect type, severity level, blade perspective, and background colour. This complements the previously analysed defect types in [13] and takes into account turbine blade defects such as burns, cracks, coating loss, and cooling hole blockage. Furthermore, the correlations between the influence factors and demographic variables including expertise, education, previous experience in inspection, work experience in the industry, and visual acuity were analysed.

Secondly, eye tracking was applied to further understand the effect of each influence factor on the visual search process. It provided a better understanding of the visual focus and underlying cognitive processes. The different search strategies and inspection errors made by the operator were extracted from the heat maps and gaze plots. Eye tracking has proved to be useful for individual performance assessment and could be beneficial for customised training to improve the inspection performance of the operator.

A third contribution is the suggestion of a revised visual search framework, taking into account the cognitive processes that appeared in the empirical findings. This was applied to borescopy. It was shown that this can in principle be extended into a bowtie framework. The principles appear to be generalisable.

There is a general understanding in the industry that borescope inspection is more challenging than piece-part inspection. The present study is the first work in the literature that quantitatively assessed the operators' performance in borescope inspection and confirmed the hypothesis. This may contribute to a more realistic expectancy of the industry and regulatory authorities regarding the achievable performance of human operators in such an inspection environment. The insights gained might be applicable to other industries using borescopes as inspection aids, including automotive, oil and gas, and power generation.

Overall, the findings contribute in several ways to our understanding of visual inspection and might be applicable to other industries with inspection processes. The insights can be used to improve the inspection environment and to customise training endeavours. This has the potential to increase inspection performance in both reliability and productivity.

Author Contributions: Conceptualisation, J.A. and D.P.; methodology, J.A., D.P. and A.M.; validation, J.A.; formal analysis, J.A.; investigation, J.A.; resources, D.P. and A.M.; data curation, J.A.; writing—original draft preparation, J.A.; writing—review and editing, J.A. and D.P.; visualisation, J.A.; supervision, D.P. and A.M.; project administration, D.P. and A.M.; funding acquisition, D.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research project was funded by the Christchurch Engine Centre (CHCEC), a maintenance, repair and overhaul (MRO) facility based in Christchurch and a joint venture between the Pratt and Whitney (PW) division of Raytheon Technologies Corporation (RTC) and Air New Zealand (ANZ).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Human Ethics Committee of the University of Canterbury (HEC 2020/08/LR-PS approved on the 2 March 2020; HEC 2020/08/LR-PS Amendment 1 approved on the 28 August 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study prior to experiment commencement.

Data Availability Statement: The data are not publicly available due to commercial sensitivity and data privacy.

Acknowledgments: We sincerely thank staff at the Christchurch Engine Centre for their participation in the eye tracking study and contributing to this research. We further thank everyone who supported this research in any form. In particular, we want to thank Ross Riordan and Marcus Wade.

Conflicts of Interest: J.A. was funded by a PhD scholarship through this research project. The authors declare no other conflicts of interest.

References

1. Ackert, S. Engine maintenance concepts for financiers. *Aircr. Monit.* **2011**, *2*, 1–43.
2. Latorella, K.; Prabhu, P. A review of human error in aviation maintenance and inspection. *Int. J. Ind. Ergon.* **2000**, *26*, 133–161. [CrossRef]
3. Nickles, G.; Him, H.; Koenig, S.; Gramopadhye, A.; Melloy, B. *A Descriptive Model of Aircraft Inspection Activities*; Federal Aviation Administration (FAA): Washington, DC, USA, 2019.
4. Megaw, E.D.; Richardson, J. Eye movements and industrial inspection. *Appl. Ergon.* **1979**, *10*, 145–154. [CrossRef]
5. Drury, C.G.; Spencer, F.W.; Schurman, D.L. Measuring Human Detection Performance in Aircraft Visual Inspection. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Albuquerque, NM, USA, 22–26 September 1997; Volume 41, pp. 304–308. [CrossRef]
6. Hayes, A.S. Control of visual inspection. *Ind. Qual. Control* **1950**, *6*, 73–76.
7. Carter, C.W. Quality Control of Visual Characteristics. In *ASQC Convention Transactions*; American Society Quality Control: Milwaukee, WI, USA, 1957; p. 634.
8. See, J.E. Visual Inspection Reliability for Precision Manufactured Parts. *Hum. Factors* **2015**, *57*, 1427–1442. [CrossRef]
9. Leach, J.; Morris, P.E. Cognitive Factors in the Close Visual and Magnetic Particle Inspection of Welds Underwater. *Hum. Factors* **1998**, *40*, 187–197. [CrossRef] [PubMed]
10. Graybeal, B.A.; Phares, B.M.; Rolander, D.D.; Moore, M.; Washer, G. Visual Inspection of Highway Bridges. *J. Nondestruct. Eval.* **2002**, *21*, 67–83. [CrossRef]
11. Heida, J.H. Characterization of inspection performance. In Proceedings of the 12th World Conference on NDT, Amsterdam, The Netherlands, 23–28 April 1989; pp. 1711–1716.
12. Sadasivan, S.; Greenstein, J.S.; Gramopadhye, A.K.; Duchowski, A.T. Use of eye movements as feedforward training for a synthetic aircraft inspection task. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Portland, OR, USA, 2–7 April 2005; pp. 141–149.
13. Aust, J.; Mitrovic, A.; Pons, D. Assessment of the Effect of Cleanliness on the Visual Inspection of Aircraft Engine Blades: An Eye Tracking Study. *Sensors* **2021**, *21*, 6135. [CrossRef]
14. Cook, L. *Visual Inspection Reliability for Composite Aircraft Structures*; Cranfield University: Cranfield, UK, 2009.
15. Baaran, J. *Visual Inspection of Composite Structures*; European Aviation Safety Agency (EASA): Cologne, Germany, 2009.
16. Spencer, F.W. *Visual Inspection Research Project Report on Benchmark Inspections*; Aging Aircraft NDI Validation Center (AANC), Sandia National Labs: Albuquerque, NM, USA, 1996.
17. Erhart, D.; Ostrom, L.T.; Wilhelmsen, C.A. Visual detectibility of dents on a composite aircraft inspection specimen: An initial study. *Int. J. Appl. Aviat. Stud.* **2004**, *4*, 111–122.
18. Chen, X.; Ren, H.; Bil, C. Inspection Intervals Optimization for Aircraft Composite Structures Considering Dent Damage. *J. Aircr.* **2014**, *51*, 303–309. [CrossRef]
19. Annis, C.; Gandossi, L.; Martin, O. Optimal sample size for probability of detection curves. *Nucl. Eng. Des.* **2013**, *262*, 98–105. [CrossRef]
20. See, J.E. Visual inspection: A review of the literature. *Sandia Natl. Lab. Albuq. New Mex.* **2012**. [CrossRef]
21. Megaw, E.D. Factors affecting visual inspection accuracy. *Appl. Ergon.* **1979**, *10*, 27–32. [CrossRef]
22. Aust, J.; Pons, D. A Systematic Methodology for Developing Bowtie in Risk Assessment: Application to Borescope Inspection. *Aerospace* **2020**, *7*, 86. [CrossRef]
23. Aust, J.; Pons, D. Bowtie Methodology for Risk Analysis of Visual Borescope Inspection during Aircraft Engine Maintenance. *Aerospace* **2019**, *6*, 110. [CrossRef]
24. Gant, S.K. Visual assessment of impact damage on painted composite aircraft structures. In Proceedings of the International SAMPE Symposium and Exhibition (Proceedings), Baltimore, MD, USA, 3–7 June 2007; Volume 52.
25. Aust, J.; Pons, D. Taxonomy of Gas Turbine Blade Defects. *Aerospace* **2019**, *6*, 58. [CrossRef]
26. Pernice, K.; Nielsen, J. *How to Conduct Eyetracking Studies*; Nielsen Norman Group: Fremont, CA, USA, 2019.
27. Fan, M.; Shi, S.; Truong, K.N. Practices and Challenges of Using Think-Aloud Protocols in Industry: An International Survey. *J. Usability Stud.* **2020**, *15*, 85–102.
28. Aust, J.; Pons, D. Methodology for Evaluating Risk of Visual Inspection Tasks of Aircraft Engine Blades. *Aerospace* **2021**, *8*, 117. [CrossRef]
29. Aust, J.; Shankland, S.; Pons, D.; Mukundan, R.; Mitrovic, A. Automated Defect Detection and Decision-Support in Gas Turbine Blade Inspection. *Aerospace* **2021**, *8*, 30. [CrossRef]
30. Kundel, H.L.; Nodine, C.F. *Studies of Eye Movements and Visual Search in Radiology*, 1st ed.; Senders, J.W., Fisher, D.F., Monty, R.A., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1978; pp. 317–327.
31. Drury, C.G.; Fox, J.G. The imperfect inspector. *Hum. Reliab. Qual. Control* **1975**, 11–16. [CrossRef]
32. Schoonard, J.W.; Gould, J.D.; Miller, L.A. Studies of visual inspection. *Ergonomics* **1973**, *16*, 365–379. [CrossRef]
33. Jacobson, H.J. A study of inspector accuracy. *Ind. Qual. Control* **1952**, *9*, 16–25.
34. Aust, J.; Mitrovic, A.; Pons, D. Comparison of Visual and Visual–Tactile Inspection of Aircraft Engine Blades. *Aerospace* **2021**, *8*, 313. [CrossRef]
35. Waite, S. Defect Types and Inspection. In Proceedings of the MIL17 Maintenance Workshop, Chicago, IL, USA, 19–21 July 2006.

36. Brunyé, T.T.; Carney, P.A.; Allison, K.H.; Shapiro, L.G.; Weaver, D.L.; Elmore, J.G. Eye Movements as an Index of Pathologist Visual Expertise: A Pilot Study. *PLoS ONE* **2014**, *9*, e103447. [CrossRef]
37. Brunyé, T.T.; Mercan, E.; Weaver, D.L.; Elmore, J.G. Accuracy is in the eyes of the pathologist: The visual interpretive process and diagnostic accuracy with digital whole slide images. *J. Biomed. Inf.* **2017**, *66*, 171–179. [CrossRef]
38. Ghanbari, L.; Wang, C.; Jeon, H.W. Industrial Energy Assessment Training Effectiveness Evaluation: An Eye-Tracking Study. *Sensors* **2021**, *21*, 1584. [CrossRef]
39. Brunyé, T.T.; Drew, T.; Weaver, D.L.; Elmore, J.G. A review of eye tracking for understanding and improving diagnostic interpretation. *Cogn. Res. Princ. Implic.* **2019**, *4*, 7. [CrossRef]
40. Chen, S.; Epps, J.; Ruiz, N.; Chen, F. Eye activity as a measure of human mental effort in HCI. In Proceedings of the 16th International Conference on Intelligent User Interfaces, Palo Alto, CA, USA, 13–16 February 2011; pp. 315–318.
41. Debue, N.; van de Leemput, C. What does germane load mean? An empirical contribution to the cognitive load theory. *Front. Psychol.* **2014**, *5*, 1099. [CrossRef]
42. Fabio, R.; Incorpora, C.; Falzone, A.; Caprì, T.; Errante, A.; Carrozza, C.; Mohammadhasani, N. The influence of cognitive load and amount of stimuli on entropy through eye tracking measures. *EAPCogScience* **2015**, *1*, 199–204. Available online: https://hdl.handle.net/11570/3064917 (accessed on 30 June 2021).
43. Hooge, I.T.; Erkelens, C.J. Control of fixation duration in a simple search task. *Percept. Psychophys.* **1996**, *58*, 969–976. [CrossRef]
44. Jacobs, A.M.; O'Regan, J.K. Spatial and/or temporal adjustments of scanning behavior to visibility changes. *Acta Psychol.* **1987**, *65*, 133–146. [CrossRef]
45. Drury, C.G. Exploring search strategies in aircraft inspection. In *Visual Search 2*; Brogan, D., Gale, A., Carr, K., Eds.; Taylor & Francis Ltd.: London, UK, 1993; pp. 101–112.
46. Drury, C.G.; Watson, J. Good Practices in Visual Inspection. Available online: https://www.faa.gov/about/initiatives/maintenance_hf/library/documents/#HumanFactorsMaintenance (accessed on 14 June 2021).
47. Ji, Z.; Pons, D.; Pearse, J. Why do workers take safety risks?—A conceptual model for the motivation underpinning perverse agency. *Safety* **2018**, *4*, 24. [CrossRef]
48. Heilemann, F.; Dadashi, A.; Wicke, K. Eeloscope—Towards a Novel Endoscopic System Enabling Digital Aircraft Fuel Tank Maintenance. *Aerospace* **2021**, *8*, 136. [CrossRef]
49. Wang, M.; Dong, X.; Ba, W.; Mohammad, A.; Axinte, D.; Norton, A. Design, modelling and validation of a novel extra slender continuum robot for in-situ inspection and repair in aeroengine. *Robot. Comput.–Integr. Manuf.* **2021**, *67*, 102054. [CrossRef]
50. Alatorre, D.; Nasser, B.; Rabani, A.; Nagy-Sochacki, A.; Dong, X.; Axinte, D.; Kell, J. Teleoperated, in situ repair of an aeroengine: Overcoming the internet latency hurdle. *IEEE Robot. Autom. Mag.* **2018**, *26*, 10–20. [CrossRef]
51. Folmsbee, J.; Johnson, S.; Liu, X.; Brandwein-Weber, M.; Doyle, S. Fragile neural networks: The importance of image standardization for deep learning in digital pathology. In *Medical Imaging 2019: Digital Pathology*; SPIE—International Society for Optics and Photonics: Bellingham, DC, USA, 2019; Volume 10956.
52. Clemons, A.J. *Training Methods for Visual Inspection Tasks*; Iowa State University: Ames, IA, USA, 2013.
53. Wang, M.-J.; Lin, S.-C.; Drury, C. Training for strategy in visual search. *Int. J. Ind. Ergon.* **1997**, *20*, 101–108. [CrossRef]
54. Khan, R.S.A.; Tien, G.; Atkins, M.S.; Zheng, B.; Panton, O.N.M.; Meneghetti, A.T. Analysis of eye gaze: Do novice surgeons look at the same location as expert surgeons during a laparoscopic operation? *Surg. Endosc.* **2012**, *26*, 3536–3540. [CrossRef]