Master's Thesis of Engineering

On the Effects of Image-based Algorithmic Suggestions to Bearing Fault Inspectors

August 2019

Graduate School of
Convergence Science & Technology
Seoul National University

Won Shin

Abstract

On the Effects of Image-based Algorithmic Suggestions to Bearing Fault Inspectors

Won Shin
School of Convergence Science & Technology
The Graduate School
Seoul National University

With the wide adoption of predictive maintenance in the past decade, visual inspection using imaging devices has become more important in machinery operation and maintenance (O&M). A variety of imaging devices like endoscope and thermal imaging camera are used to check faults in areas invisible or inaccessible to human eyes. These inspections are now becoming the center of general maintenance work and are increasingly being performed by general maintenance engineers who have limited experience in fault diagnosis. However, human-oriented visual inspection has several major shortcomings including variability of accuracy depending on inspector's expertise level and inconsistency of decisions among the inspectors.

To address these shortcomings, a plethora of research has been undertaken in computer vision and inspection automation. Recently, several researchers applied deep learning-based solutions to machine fault diagnosis problems and achieved promising performance. A recent survey indicates that predictive maintenance is expected to be one of the first fields where AI-based technologies, such as deep learning, will be put into practice. Accordingly, there is a great interest in understanding how effective the AI-based technologies will be if applied in actual industrial settings. But, previous studies have focused on the isolated performance of deep learning-based solutions and have not measured the full effect of the actual users utilizing them. In the field of medicine, however, computer-assisted diagnostics have already been put into practical use since the late 1990's, and active research has been conducted on the effect on radiologist performance.

To the best of our knowledge, this study is the first published research in the field of predictive maintenance that measured the effect of algorithmic suggestion to the human user. We developed a deep learning-based algorithmic suggestion system for bearing fault detection using image data from 138 wind turbines. Then, we performed a user experiment to measure the effect on the technical inspectors with varying expertise level. Thirty-four generalists and twenty specialists participated in the experiment.

The results showed that the algorithmic suggestion system had a statistically significant impact on improving the inspector's specificity and time efficiency. In the case of cognitive load, it reduced for the generalists but slightly increased for the specialists. Both groups found the algorithmic suggestion system useful for their task and displayed willingness to reuse it in the future.

Also, we have found that the effect varied depending on the level of expertise. The generalists showed greater improvements than the specialists in specificity and time efficiency. As a result, with the algorithmic suggestion system, the generalists improved to a level that was not statistically different from the unaided specialists.

The result revealed that each group made different types of mistakes in image classification. The algorithmic suggestion system was able to help in the weak areas of each group and contribute to the correct classification of images in the task.

Keyword: algorithmic suggestion, AI assistance, deep learning algorithm, predictive maintenance, image-based fault detection, computer-aided diagnostics, bearing fault, endoscope inspection

Student Number: 2015-26031

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

For machinery operation and maintenance, it is critical to identify any fault at its early stage before it progresses to a critical stage and cause consequential damage. With the wide adoption of predictive maintenance, visual inspection of machinery to identify machine faults early in advance has become increasingly important in routine maintenance. A variety of visual imaging devices like endoscope and thermal imaging camera are used to check faults in areas invisible or inaccessible to human eyes.

Although visual data collection technology has advanced significantly, human inspectors still play predominant roles in analyzing images for fault diagnosis. Human-oriented inspection, however, has several shortcomings including variability of accuracy depending on an inspector's level of expertise, the inconsistency of diagnosis among inspectors, and cost. Moreover, the volume of image data that a diagnostic specialist needs to process has vastly increased while the specialist resource remains finite.

To address these shortcomings, a plethora of research has been undertaken in the area of computer vision and inspection automation[1–3]. Recently, several researchers applied deep learning solutions to machine fault diagnosis problems and achieved promising performance[4–6].

While a deep learning-based diagnosis solution may not be able to perform a fully automated inspection on its own, employing the solution as an algorithmic suggestion to a human user can make meaningful contribution in fault diagnosis. But, previous studies have focused on the isolated performance of deep learning-based solutions and have not measured the full effect of the actual users utilizing them. In the field of medicine, however, computer-assisted diagnostics have already been put into practical use since the late 1990's, and active research has been conducted on the effect on radiologist performance.

The effect of algorithmic suggestion system on inspectors should be extensively investigated for it to make a meaningful impact on the actual field of the industry. What difference would the algorithmic suggestion system make to the inspectors' diagnostic performance? Will the Inspector be willing to use the algorithm suggestion system? Whether there would be any differences for generalists and specialist inspectors?

Therefore, we extend the previous research by including a measure of the effect on the actual inspectors. At first, a deep learning-based algorithmic suggestion system for bearing fault detection task is developed using 3,000 endoscope images from 138 wind turbine gearboxes. Secondly, a user experiment is performed to measure the effect of algorithmic suggestion system to the performance and perception of specialist and generalist engineer groups. A total of 54 inspectors with an average of seven years of experience participated in the experiment.

The foremost aim of this study is to evaluate whether algorithmic suggestion system improves the diagnostic performance of the inspector and how the improvement differs between generalist and specialist groups. The second objective is to assess the perception of generalist and specialist inspectors after using the algorithmic suggestion system on their task.

Chapter 2. Related works

Statistical measures of diagnostic performance

Specificity and sensitivity are the most commonly used statistical measures for the diagnostic system according to survey research by Gonçalves[7]. Specificity measures the proportion of actual normality (negatives) that are correctly identified as normal. It is also called true negative rate. High specificity means that actual normality is rarely mistaken as abnormal, so false positives are few. Sensitivity measures the proportion of actual abnormality (positives) that are correctly identified as abnormal. It is also called true positive rate or recall. High sensitivity means that actual abnormality is rarely overlooked, and hence false negatives are few. The terms "positive" and "negative" refer to the presence or absence of an abnormality condition, which is the condition of interest.

Shortcomings of human-oriented inspection

Visual inspection is the primary means of evaluating the health conditions of machinery as well as humans and civil structures. This method comes with many shortcomings in diagnostic performance, and they are well studied in the fields of machine vision[8,9] and medical imaging[10].

First, accuracy is highly dependent on the inspector's expertise level. In the experiment by Sickles E et al., the rate of incorrect interpretation by general radiologists was about 30% higher than specialist radiologists in mammographic readings[11].

Second, it is difficult to ensure consistency of diagnostic decision among multiple inspectors. This was well studied by the experiment by Varun Gulshan et al[12]. Seven US certified Ophthalmologists were asked to grade the diabetic retinopathy in retinal fundus photographs into severity level one to five. Surprisingly, consistency among ophthalmologists was very poor. Out of 683 cases, only 20% case showed full agreement among the group of ophthalmologists.

Computer-aided methods for fault detection for industrial application

To address the limitation of human-oriented visual inspection, researchers applied computer vision methods to detect a fault. In the past, computer vision studies focused on various image processing techniques (IPTs). Despite their achievements, the inherent limitation in this approach is that it requires a feature to be developed; not only a time-consuming endeavor but also the feature is only applicable for a particular fault type.

The introduction of a deep learning solution (Table 1) addressed the limitation of IPTs. Damage sensitive features are automatically determined during the training process, and multiple fault types can be detected as long as additional training data is provided. Miaoyiquan et al. applied Convolution Neural Network (CNN) to a fault detection of power cables and achieved 81 to 87 percent accuracy[13]. Park J applied CNN to surface faults of various texture surfaces like wood, fabric, and metal and achieved 98 percent accuracy.[14] Cha et al.[5] applied CNN for assessing concrete cracks and achieved an equally high accuracy of 98 percent. Ren et al. proposed Faster Region-based CNN (R-CNN) architecture that automated the training of object detector, which had to be manually selected before. It can also provide a real-time object detection[15]. Cha et al. applied the faster R-CNN to detect five types of structural surface damages on bridges and achieved a mean average precision of 97 percent[16].

Despite the impressive improvement in the reported accuracy, the deep learning solution is trained and tested for a limited number of failure modes that are usually fewer than five classes. However, the actual number of failure modes that need to be covered to automate inspection is often significantly larger. For example, the ISO 15243 standard classifies bearing failure modes to 16 classes[17]. Since some of the failures rarely occur, it is practically impossible to collect enough training data to cover every class. The likelihood of fully-automated inspection is low except in some cases such as a quality check at automated manufacturing lines where there is a limited set of possible failure modes, and sufficient data can be collected for every failure mode.

The deep learning-based solution may not be able to perform fully automated inspection on its own, however, employing the solution as an algorithmic suggestion to human user can make meaningful contribution in fault detection. Such algorithmic suggestion system makes diagnostic suggestions for human inspector when the inspector makes a final diagnostic decision. In this aided decision-making process, measuring algorithmic suggestion performance in isolation is not sufficient to truly understand its effect. It is crucial to measure change in the inspector's performance when algorithmic suggestion is provided. To the best of our knowledge, no published work has evaluated the effect of algorithmic suggestion to the inspectors.

Table 1. Recent researches applying deep learning algorithms to the fault detection problem

Year, Author	**Title**	**Summary**
2016, Park J, et al	Machine learning-based imaging system for surface fault inspection	Accuracy 98% (Manpower inspection 98.5%) 4,000 images for 4 classes
2017, Miaoyiquan	Fault Detection for Power Line Based on Convolution	Accuracy 81.2-87.3% 510 images for 3 classes

W., et al	Neural Network	
2018, Cha et al	Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks	Accuracy 97% 332 images for 5 classes

Computer-aided diagnostics in the medical field

In the medical field, Computer-aided diagnostics (CAD) was introduced to the clinic in 1998 (Van Ginneken et al.[18]). Many studies have been undertaken to assess the effect of CAD on radiologist performance.

Several studies have shown that CAD effectively improves radiologists' performance. Matthew et al. compared the performance of radiologists with and without CAD assist and showed the CAD assist resulted in a significantly increased sensitivity (90.4% from 81.4%) at the cost of a small increase in the recall rate[19]. Jamie G Cooper performed a meta-analysis of ten studies and demonstrated an overall mean percentage improvement of 17.25% in the clinical diagnostic accuracy with the use of CAD systems[20]. In the experiment by Freer and Ulissey[21], a radiologist first analyzed the mammogram without CAD assistance and then the radiologist reviewed the suggestion from CAD and rendered a final interpretation; thus, changes in the reading were attributed to CAD. In total 12,860 mammograms were used, and the result showed a 19.5% increase in the cancer detection rate. A similar result was published by Birdwell et al. [22], Ko et al.[23], and Helvie et al.[24].

V.A. Fisichella et al. studied the effect of CAD on radiologists with different expertise levels[25]. The result showed that with CAD assist, difference in the performance between experienced and inexperienced readers for lesions greater than or equal to six millimeters was eliminated, suggesting a potentially beneficial role of CAD as a means to reduce inter-observer variability resulting from the expertise level.

As the diagnostic accuracy of CAD is not perfect, it also has a potential negative impact arising from inaccurate diagnoses performed by CAD. Two types of incorrect suggestions from CAD are mainly studied: a false positive (FP) suggestion and a false negative (FN) suggestion. The large number of FPs is one of the fundamental limitations of the CAD system. Several researches studied effects of the false positives on human performance. In the experiment by Fisichella et al., the false positive rates increased significantly for inexperienced readers with CAD[25]. As a result, there was no increase in the overall diagnostic performance measured by FROC analysis despite improvement in the sensitivity. However, in the subgroup analysis of larger (>6mm) lesions, which is less difficult to detect than small-sized lesions, the large number of false positive CAD marks did not increase the FP rate of the inexperienced readers. This is in line with the findings by Mani et al. About 30 FP CAD marks were shown to the radiologists, but it did not lower the specificity[26]. This can be interpreted that the effect of FP CAD marks can vary depending on the task complexity and user's ability to dismiss them.

Eugenio Alberdi et al. studied the effects of false negative output of the CAD system on human decision-making in mammography[27]. The average sensitivity of readers with CAD was significantly lower than the average sensitivity of readers without CAD. He interpreted that the absence of the detection marker on a mammogram may have caused readers to be complacent and pay less attention to the case. As a result, anomalies which they would not have missed without CAD were overlooked.

Method of diagnostic result representation for user

For effective collaboration between the algorithmic suggestion system and user, it is essential to study how data should be provided in the algorithmic suggestion system. Again, there have been several studies in the medical field that are applicable to the machine fault diagnosis problem. Ca et al. surveyed 220 medical professionals to measure user evaluations of the medical information system, and one of the most important factors in meeting users' needs was the method of data representation[28].

The most common visualization in the commercially available CAD system is a marker on the area of potential disease. With the advancement of 3D imaging, some CAD systems also provide 3D markers. The marked area is rendered in the 3D image to assist a user to accurately and intuitively understand the location of disease.

Another useful but less common visualization method is temporal subtraction, which detects any changes over time. Shiraishi et al. developed a CAD scheme for the detection of interval changes in successive whole-body bone scans by use of a temporal-subtraction image[29]. Some CAD systems provided image enhancement to expose the potential disease area more clearly to the user than the original image.

Some of the CAD systems provide a probability value of the suggestion it is making. Feng Li et al. performed user experiments and observed synergistic effects on observer interpretation when probability values were provided[30]. The radiologists were able to correctly maintain their own opinions on obvious cases, whereas computer output assisted in improving their decisions in the majority of subtle cases.

User's perception of new technology to aid their performance

How technology affects user's performance is theorized by Goodhue and Thompson in the task-technology fit model (1995). Task-technology fit (TTF) is "the degree to which technology assists an individual in performing his or her portfolio of tasks"[31]. The fit is maximized when the decision support minimizes user's effort to complete the task. Therefore, the performance and utilization will increase if the technology has a good fit with the task.

A good task-technology fit is expected to decrease user's cognitive load (Zigurs, 1998), since "a good fit is the most efficient way of dealing with the cognitive load of a given task"[32]. Cognitive load refers to the effort being used in the working memory (Sweller, 1998)[33].

The good fit positively influences perceived usefulness and reuse intention of a system. Perceived usefulness is defined as "the degree to which a person believes that using the system would enhance his or her job performance." (Davis, 1989)[34]. Reuse intention is user's behavioral intention to use the system in a similar situation in the future (Venkatesh, 2000)[35]. Reuse intention is also strongly associated with the user's perceived quality of information provided by the system and trust in the system (Nicolaou and Mcknight, 2006)[36]

Chapter 3. Research Questions

This study aims to measure the effect of algorithmic suggestion on the performance and perception of technical inspectors. Also, we want to evaluate how the effect differs by the level of expertise. To that end, the following research questions are formulated:

RQ 1. How does algorithmic suggestion affect the performance of specialist and generalist inspectors?

RQ 2. What are the perceived usefulness, reuse intention, and cognitive load of algorithmic suggestion to inspectors?

Chapter 4. Method

4.1 Development of algorithmic suggestion system based on a deep-learning approach

For this study, we developed an algorithmic suggestion system that detects bearing faults in endoscope images based on a deep learning solution. The system assisted human inspector's task by providing a suggestion for bearing fault detection in a treatment condition.

Bearing endoscope image dataset for deep learning algorithm training

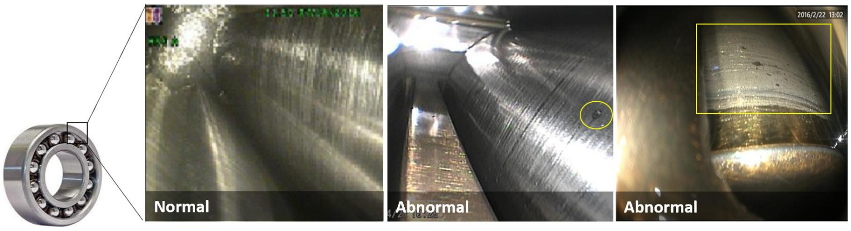


Figure 1. Rolling element bearing (far left), endoscope image of the rolling element in a normal condition (left), and two images of an abnormal bearing condition (right)

Figure 1 shows a rolling element bearing and example endoscope images in a normal and abnormal condition. For ease of understanding, fault areas are highlighted with the yellow box in this figure. Note that such marking of fault areas is not included in dataset used for the algorithmic suggestion system development.

A total of 3,073 endoscope images of rolling-element bearing were collected from in-situ inspections of 138 wind turbine gearboxes and main bearings. All inspections were carried out as an in-situ inspection of the wind turbines in operation, and thus endoscopy was used. Due to the nature of the endoscopy, the position and orientation of the bearings in images vary. Also, inspected wind turbines are of diverse models and manufacturers, and their operating age ranges from new to more than ten years. Various models of endoscopes are used for the inspection. For this reason, the images contained in the dataset varied in brightness, resolution, and image quality.

The data preprocessing was done as follows. First, we checked the quality of the images that were not suitable for the training of algorithmic suggestion system. Images were removed from the datasets if (1) the rolling bearing elements take up less than 20% of the image; (2) the focus of the photo is missed or fuzzy; and (3) the surface is not visible enough due to excessive oil or grease. As a result, a total of 772 images were removed. The same criteria are used for the manual analysis of human inspectors.

Second, the ground truth label was obtained by a majority vote method. Most of the bearing image data used in this experiment had no additional information such as vibration analysis or disassembly inspection results that could ultimately confirm the condition of the bearing; so the diagnosis had to be based on the inspector's subjective judgment of the image. As a result, there was a problem of variability in which the accuracy of diagnosis could vary depending on the level of expertise of the inspector, which is a common problem in the in-situ endoscopy. To address the issue, we referred to Gulshan and his colleagues' study (2016) [12] and constructed a rigorous labeling process based on a majority vote of a group of examiners. Following the process, images were labeled by three specialist inspectors, and the majority vote was taken to determine a ground truth label. Before starting the labeling process, the specialist inspectors were provided with a technical standard including the criteria of abnormality. There are several criteria used in the industry depending on the purpose of the inspection, which often causes inconsistency among inspectors. We used ISO 15243[17] as a basis for the technical standards. Abnormality is defined as a fault that affects performance or durability and general wear acceptable to its operating age or cosmetic anomalies is not considered as the abnormality. The labeled dataset has 2,301 images consisting of 886 abnormal bearing images and 1,415 normal.

Third, the data set is divided into training and test datasets, containing of 2,101 and 200 images, respectively. The training dataset is a set of examples used during the training phase of an artificial neural network model. Weights parameters of connections between neurons in the model are fitted using the training dataset. The test dataset is a dataset used during the evaluation period of the trained model. The training dataset does not overlap with the training dataset but has the same probability distribution as the training dataset. Therefore, it provides an unbiased evaluation of a developed model. Data augmentation was applied to the training data set to improve the robustness of the training and reduce the imbalance between normal and abnormal data; rotation by 90, 180, 270 degrees and left-right flip was applied to images.

Characteristic of the algorithmic suggestion system for bearing fault detection

An algorithmic suggestion system for bearing fault detection was developed using a CNN. The input was an image with 150×150×3-pixel (height x width x channel (RGB)) resolutions. The deep CNN architecture had four convolutional layers and two fully connected layers. The batch normalization and dropout layers were also applied to improve model performance. The SoftMax layer predicted whether each input data was normal or abnormal.

The system performance for the test set is provided in Table 2. Specificity and sensitivity are 0.91 and 0.89, respectively.

Table 2. The algorithmic suggestion system performance in isolation

No. of images = 220	Predicted: Abnormal	Predicted: Normal
Actual: Abnormal	55	7
Actual: Normal	13	135

4.2 Experiment

An experimental study was designed to investigate the effect of the algorithmic suggestion on specialist and generalist engineers for the bearing fault detection task using endoscope images. The participants performed a classification task for bearing endoscopy images. The experiment was conducted using a mixed design with two factors – the presence of an algorithmic suggestion and expertise level of participants. The following section describes participants, task, experimental procedures, and performance metrics.

Participants

The participants consist of experienced engineers in the wind energy industry. Since bearing failure is a prominent issue that can cause sudden failure and prolonged downtime of a wind turbine, preventing bearing failure by early detection is a matter of significant interest in the industry. A total of 54 engineers with two levels of expertise in diagnostics participated in this study.

The specialists group (n=20) consists of engineers who perform the bearing fault diagnosis as a part of their routine tasks and have more than three years of experience. In the generalists group (n=34) are engineers and technicians who have a basic understanding of machinery and bearing fault symptoms but have limited experience in the bearing fault diagnosis task.

Table 3 summarizes the demographic information of the two groups. On average, the subjects have more than 6 years of experience in the related areas. Specialists are mostly inspection engineers or engineering consultants. Generalists are mostly engaged in the operation and maintenance of wind farms and typically perform various maintenance tasks of wind turbine equipment.

Table 3. Participants demography

		Generalist	Specialist
Number of participants		34	20
Gender – males (% of males)		34 (100%)	19 (95%)
Average work experience in a related area		6.38 years	8.15 years
Major	Mechanical engineering	33	17
	Electrical, electronic engineering	14	11
	Others	7	6
Specialty area	Inspection and failure analysis		18
	Bearing specialist		2
	Operation and maintenance	31	
	Gearbox design and analysis	3	
Nationality	Korea	25	5
	US		7
	Japan	5	
	UK		4
	India		3
	Others	4	1

Task

The participants were asked to analyze 17 endoscope images, each coming from different bearings, and perform a binary classification task of deciding if the bearing health condition is normal or abnormal. This is a simple but typical first step for image-based fault diagnosis.

Experiment procedure

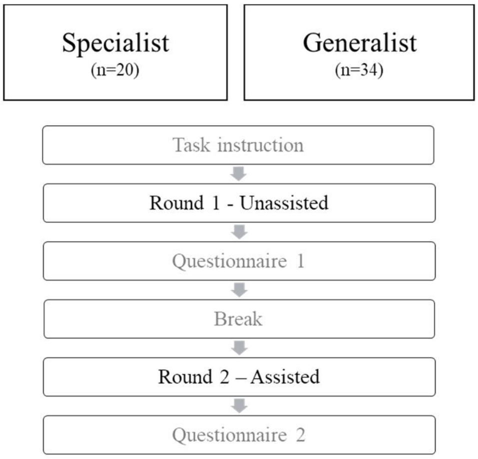
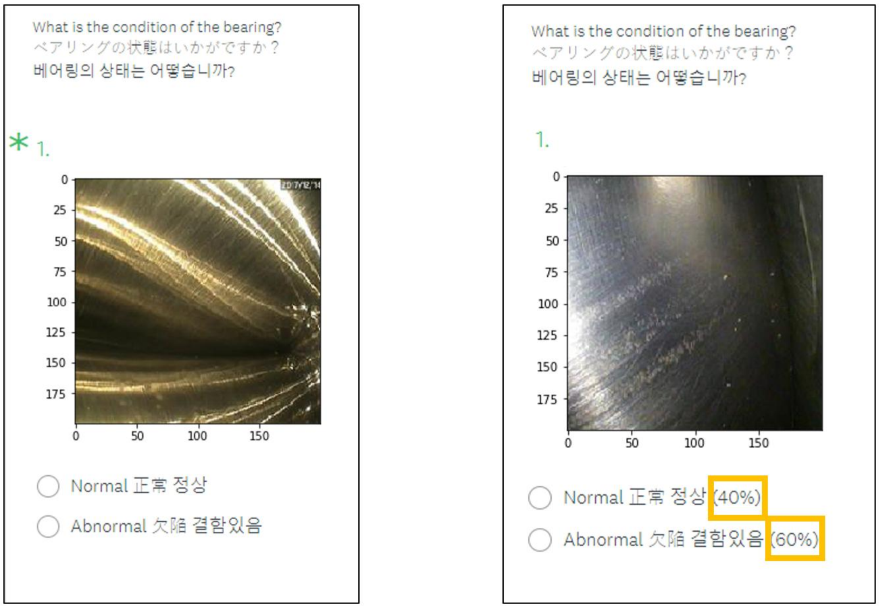


Figure 2. Experiment procedure

The experiment procedure is summarized in Figure 2. Using a repeated measures design, the participants performed the task in two rounds: first without algorithmic suggestion and the second with the algorithmic suggestion. In the second round, algorithmic suggestion is provided in the form of probability values for each decision, as shown in Figure 3 (right).



What is the condition of the bearing?
ベアリングの状態はいかがですか？
베어링의 상태는 어떻습니까?

* 1.

○ Normal 正常 정상
○ Abnormal 欠陥 결함있음

What is the condition of the bearing?
ベアリングの状態はいかがですか？
베어링의 상태는 어떻습니까?

1.

○ Normal 正常 정상 (40%)
○ Abnormal 欠陥 결함있음 (60%)

Figure 3. Example of user interface for unassisted (left, Round 1) and assisted task (right, Round 2)

The experiment procedure is as follows. The participants first received instructions about the task, user interface, and criteria to determine normal and abnormal conditions. Then, they were asked to perform the task without a algorithmic suggestion (Round 1). An example of the user interface for the unassisted task is shown in the left of Figure 3. After the task, participants were asked to evaluate the perceived cognitive load of the task. After the first round, adequate break time was provided before starting the second round for preventing the effects of fatigue.

In Round 2, the participants were given additional instruction about the algorithmic suggestion. The standalone performance of the algorithmic suggestion was explained to allow participants to set the appropriate level of trust. Then, the participants performed the task with the algorithmic suggestion.

After the task was completed, they were asked to evaluate their cognitive load again. The participants' perception of reuse intention and usefulness of the algorithmic suggestion were measured. In addition to the questionnaire, some of the participants voluntarily provided various user feedback.

To control the learning effects due to the sequence of the task, two different sets of images (A, B) were used. Half of the participants who were randomly selected in each group were provided with images from Set A in the first round and Set B in the second round. The other half of the participants were given images from the B set in the first round and A set in the second round.

Evaluation

Table 4. Performance metrics

Variable	Method
Specificity	$\frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$
Sensitivity	$\frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$
Time efficiency	$\frac{\text{Number of images in the task}}{\text{Time to complete the task (minutes)}}$

The participants' performance of the task and their perception of the algorithmic suggestion were evaluated by metrics in Tables 4 and 5. Specificity, sensitivity, and time efficiency were used for measuring user performance. Time efficiency measures the average number of images analyzed per minute. Time is a necessary constraint for industrial tasks (E. D. Megaw. 1979)[8] as it is directly related to cost.

In terms of perception, we used a questionnaire to measure the participants' cognitive load (adapted from Hart and Staveland, 1988), behavior reuse intention and perceived usefulness (adapted from Davis et al., 1989; Venkatesh et al., 2003). All the items had a 5-point Likert scale (strongly disagree = 1; strongly agree = 5). The participants responded to the questionnaire without knowing the result of the diagnostic task; they did not know if the algorithmic suggestion improved their performance or not. The example items for each metric are provided in

Table 5 with its reliability (Cronbach's alpha).

Table 5. Perception metric

Variable	Example questions	Cronbach's alpha
Cognitive load	This task required much mental and perceptual activity to complete	0.75 - 0.76
Behavioral intention to use	If I had a software that provides algorithmic suggestion the next time I do the bearing inspection, I predict I would use it	0.77
Perceived Usefulness	Using the algorithmic suggestion system would enhance my effectiveness in my job (e.g., better accuracy and reliability of inspections reporting)	0.95

To assess the algorithmic suggestion system's effects on the inspector's performance, we first conducted a descriptive analysis. Then, the mixed design of two-way ANOVA analyses coupled with post-hoc analysis was employed to assess the effect of the algorithmic suggestion (unassisted vs. assisted) and inspector's expertise level (generalist vs. specialists) on three performance metrics: specificity, sensitivity and time efficiency. To further analyze the characteristics of the images on which the algorithmic suggestion enhanced the performance of each group, a case analysis is performed by calculating the percentage of correct classification per image for each group.

To investigate the effects of the system on the inspector's perception, we conducted descriptive analysis and ANOVA with post-hoc analysis for cognitive load. Also, descriptive analysis was performed for the two other perception metrics: behavioral intention to use and perceived usefulness.

Chapter 5. Results

5.1 Effect on performance

Specificity

The summary of specificity result is provided in Table 6. Panel A shows descriptive statistics of the four condition groups. Between the unassisted test and assisted test results, specificity improved for both the generalists and the specialists. Between the generalists and the specialists, specificity is greater for the specialists than for generalists at both tests.

Panel B presents the ANOVA result. Level of expertise is a significant factor impacting specificity (F(1,52)=12.94, p=0.001). More importantly, the result shows that the impact of the algorithmic suggestion (F(1,52)=9.30, p=0.004) is as significant as the expertise level. The interaction between the level of expertise and the algorithmic suggestion is found to be significant (p=0.076), which indicates that the impact of algorithmic suggestion is dependent on the inspector's level of expertise.

Panel C provides the result of the post-hoc analysis. The unassisted test result highlights significant underperformance of the generalists compared to the specialists (C3). The generalists had an average specificity 24% lower and standard deviation 85% higher than those of the specialists. At assisted test, the algorithmic suggestion significantly improved their specificity (C2) (p=0.001). As a result, if we compare 'generalist, assisted' and 'specialist, unassisted' (C5), the generalist' underperformance was effectively mitigated (p=1.00).

For the specialists, the algorithmic suggestion improved specificity although the difference was not significant (C1) (p=1.00). When both groups had the algorithmic suggestion, the difference was not significant but it was larger than C5. As expected, when specialists had the algorithmic suggestion while generalists did not, the specificity difference was significant (C6).

Table 6. Specificity summary

Panel A. Descriptive statistics

Level of expertise Algorithmic suggestion		Generalist (n=34)	Specialist (n=20)
Unassisted	Mean (std. dev)	0.65 (0.24)	0.85 (0.13)
Assisted	Mean (std. dev)	0.81 (0.16)	0.89 (0.13)

Panel B. ANOVA results

Effect	DF	F	p-value	
Level of expertise (generalist vs. specialist)	1	12.94	0.001	***
Algorithmic suggestion (unassisted vs. assisted)	1	9.30	0.004	**
Level of expertise * algorithmic suggestion	1	3.27	0.076	.

Panel C. Post-hoc analysis

Pairwise comparison	Dif.	t	p-value	
C1. Specialist: assisted - unassisted	0.04	0.78	1.000	
C2. Generalist: assisted - unassisted	0.16	3.99	0.001	**
C3. Unassisted: specialist - generalist	0.20	3.95	0.001	***
C4. Assisted: specialist - generalist	0.09	1.71	0.546	
C5. Specialist, unassisted - generalist, assisted	0.04	0.94	1.000	
C6. Specialist, assisted - generalist, unassisted	0.24	4.71	<.0001	***

P value adjustment: Bonferroni method

Sensitivity

The summary of sensitivity result is presented in Table 7. Panel A shows both generalists and specialists groups displayed equally high values of the average sensitivity at unassisted test. In the assisted test, sensitivity improved albeit in a small magnitude. More importantly, the standard deviation of the sensitivity was reduced for both groups.

Panel B shows that neither level of expertise nor the algorithmic suggestion had any statistically significant impact on the sensitivity. Because both groups' sensitivity was already high in the unassisted test, there was only small room for improvement that the algorithmic suggestion could realize. High sensitivity and low specificity is a common characteristic of the screening task of binary classification[37]. Since the main effect was not significant, a post-hoc analysis was not carried out for sensitivity.

Table 7, Sensitivity

Panel A. Descriptive statistics

Algorithmic suggestion \ Level of expertise		Generalist (n=34)	Specialist (n=20)
Unassisted	Mean (std. dev)	0.93 (0.13)	0.93 (0.11)
Assisted	Mean (std. dev)	0.95 (0.09)	0.96 (0.07)

Panel B. ANOVA results

Effect	DF	F	p-value
Level of expertise (generalist vs. specialist)	1	0.05	0.824
Algorithmic suggestion (unassisted vs. assisted)	1	1.55	0.218
Level of expertise * algorithmic suggestion	1	0.19	0.662

Case analysis

In the case analysis, we noted that each group made different types of mistakes in image classification. The algorithmic suggestion was able to help in the weak areas of each group and contribute to the correct classification of images in the task.

Figure 4**오류! 참조 원본을 찾을 수 없습니다.** shows the percentage of correct classification for each image by groups in the unassisted and assisted tests. Each data point corresponds to the individual images used in the test. Jitter was applied to the many overlapping points. The chart is divided into four quadrants using 80% as a threshold for good and poor performance. Given the imperfection in the algorithmic suggestion's diagnostic capabilities, it expectedly made incorrect prediction for some images, and such incorrectness is analyzed separately later in this section.

On the images in section A (Figure 4(a)), both specialists and generalists groups showed good performance. Reviews of the images in this section reveal that it was very obvious and easy to determine whether they were normal or abnormal. One of the images presented in Figure 5(a) shows apparent abnormality called abrasive wear. There were 17 images in this section. When the algorithmic suggestion is provided, 16 of them remained in A' (Figure 4(B)), and one moved to B’ at assisted test.

Section B displays images on which the specialists group performed well while the generalists did poorly. These images contained various artifacts like oil bubble, light reflection, and various wear patterns that can be mistaken as an abnormality but are actually normal. Figure 5(b) shows one of the images in this area. Although the ground truth label of this image is normal, the reflection of light on the bearing's metal surface is usually mistaken as an abnormality. In the unassisted test, only 53% of the generalists correctly classified this image. In the assisted test, it improved to 88%. In section B, there were nine images in the unassisted test. Eight moved to A' and one moved to B' in the assisted test.

Both groups showed poor performance on the Section D images. Figure 5 (c) shows one of the images in this area. The ground truth label of this image is abnormal. Since the severity of abrasive wear present in this image is at the subtle boundary of the normal and abnormal condition, correct classification depends on which criterion is selected for the evaluation. At unassisted test, only 40% of the generalist answered it correctly. It improved to 90% when assisted. In Section D, there were four images in the unassisted test. Three moved to A' and one moved B' at assisted test.

There were no images belonging to Section C.

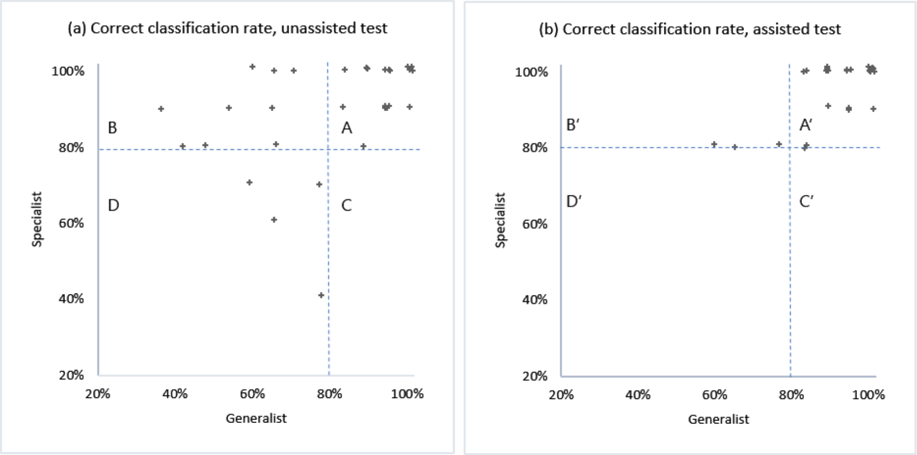


Figure 4. Correct classification rate per image by groups (correctly predicted by algorithmic suggestion)

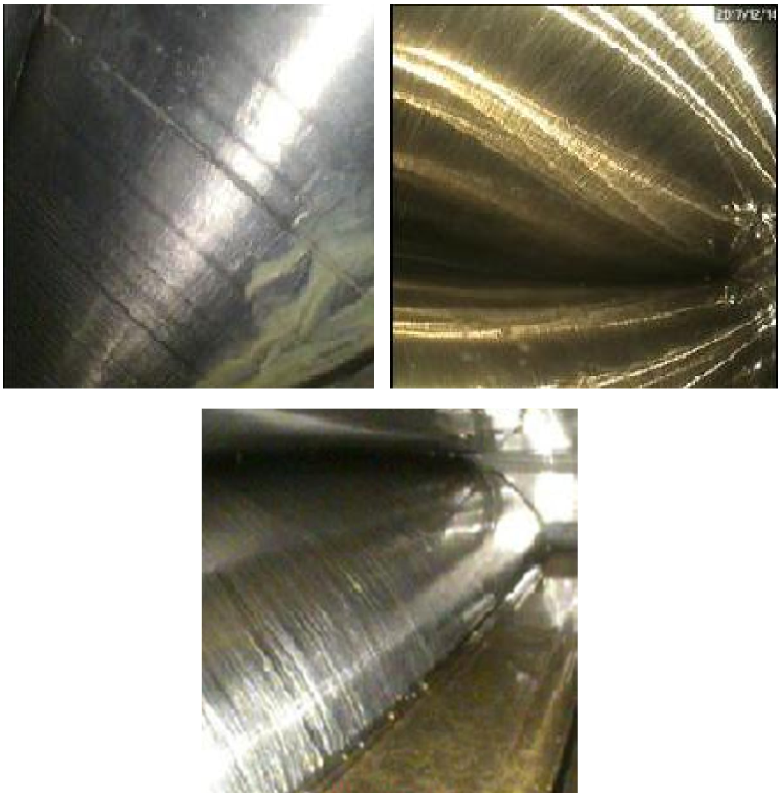


Figure 5. Example images of each section (correctly predicted by algorithmic suggestion): (a) Section A (top left), (b) Section B (top right), (c) Section D (bottom)

As mentioned above, the algorithmic suggestion made a false prediction for a few cases. There were four images that the algorithmic suggestion classified incorrectly. Figure 6 shows that the correct classification rate was reduced for all of them. Figure 7 shows an example of two cases where the algorithmic suggestion provided false predictions. Figure 7 (a) is the case where the negative effect of the false prediction was small. The image clearly shows a failure mode called pitting. The negative effect of the false prediction was large in Figure 7 (b). This image was in Section B at the unassisted test but moved to Section D' at the assisted test. This is the boundary case that can be either normal or abnormal depending on the criteria selected. In both cases, the amount of training data corresponding to the associated failure mode was very small.

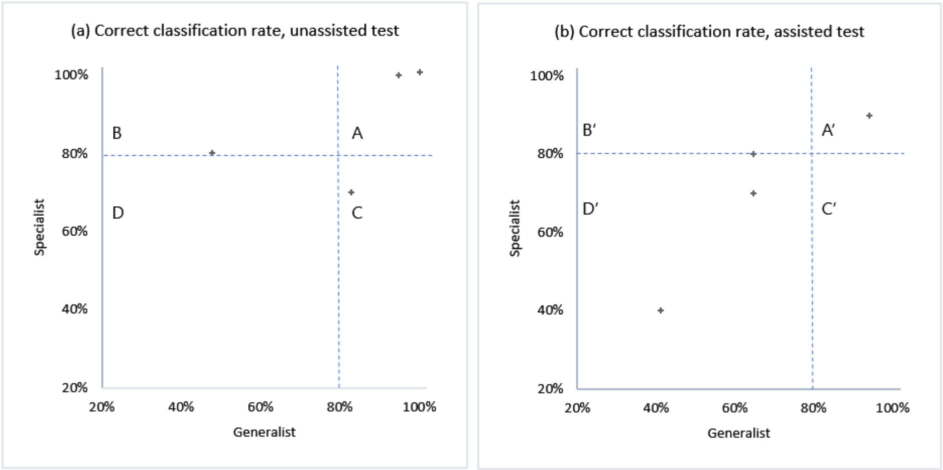


Figure 6. Correct classification rate per image by group (Incorrectly predicted by algorithmic suggestion)

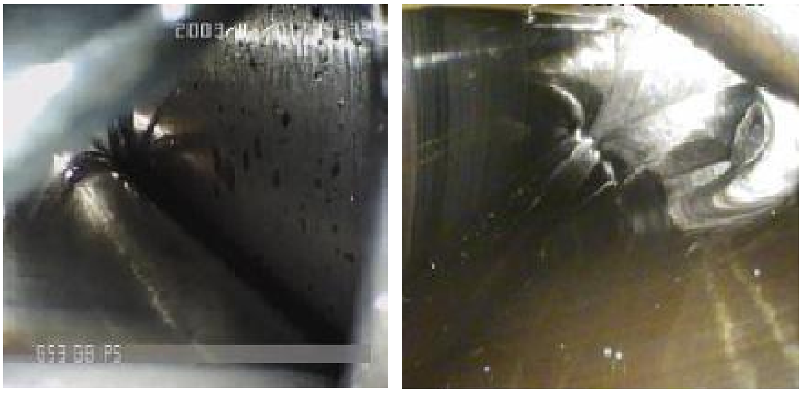


Figure 7. Example images of each section (Incorrectly predicted by algorithmic

suggestion): (a) Section A-A’ (left), (b) Section B-D’ (right)

Time efficiency

The summary of the time efficiency result is provided in Table 8. Panel A result shows that time efficiency increased by 25.3% (from 5.88 to 7.38) for the generalists and 6.4% (from 6.28 to 6.68) for the specialists.

Panel B shows that the algorithmic suggestion was a significant factor for time efficiency (p=0.023) whereas the level of expertise was not. The interaction between the level of expertise and the algorithmic suggestion was marginally significant.

Panel C shows significant improvement for generalists (C2) (p=0.035) but not for specialists (C1) (p=1.00).

Table 8. Time efficiency

Panel A. Descriptive statistics

Level of expertise Algorithmic suggestion		Generalist (n=29)	Specialist (n=20)
Unassisted	Mean (std. dev)	5.88 (2.36)	6.28 (2.87)
Assisted	Mean (std. dev)	7.38 (3.09)	6.68 (3.16)

Panel B. ANOVA results

Effect	DF	F	p-value	
Level of expertise (generalist vs. specialist)	1	0.04	0.839	
Algorithmic suggestion (unassisted vs. assisted)	1	5.52	0.023	*
Level of expertise * algorithmic suggestion	1	1.82	0.184	

Panel C. Post hoc analysis

Pairwise comparison	Dif.	t	p-value	
C1. Specialist: assisted - unassisted	0.40	0.65	1.000	
C2. Generalist: assisted - unassisted	1.49	2.89	0.035	*
C3. Unassisted: specialist - generalist	0.39	0.48	1.000	
C4. Assisted: specialist - generalist	-0.69	-0.83	1.000	

C5. Specialist, unassisted - generalist, assisted	-1.09	-1.32	1.000
C6. Specialist, assisted - generalist, unassisted	0.80	0.96	1.000

P value adjustment: Bonferroni method

5.2 Effect on perception

Cognitive load

Cognitive load result summary is provided in Table 9. Panel A shows descriptive statistics of the four condition groups. At unassisted task, the cognitive load was higher for generalists than for specialists. At assisted task, cognitive load is reduced by 7.0% (from 3.97 to 3.69) for the generalists, whereas it increased by 2.4% (from 3.29 to 3.37) for the specialists.

Panel B shows that expertise level is a statistically significant factor impacting inspector's cognitive load (p=0.043), which is no surprise as the specialists are more familiar with the task and require less effort than generalists. Interestingly, the algorithmic suggestion did not affect cognitive load (p=0.416) for the specialists. The interaction between expertise level and the algorithmic suggestion is marginally significant even though the change in each group shows the opposite.

Panel C presents that cognitive load is significantly different between specialists and generalists at unassisted task (C3) (p=0.084).

Table 9. Cognitive load

Panel A. Descriptive statistics

Level of expertise Algorithmic suggestion		Generalist (n=29)	Specialist (n=19)
Unassisted	Mean (std. dev)	3.97(0.90)	3.29(0.85)
Assisted	Mean (std. dev)	3.69(1.00)	3.37(0.81)

Panel B. ANOVA results

Effect	DF	F	p-value

Level of expertise (generalist vs. specialist)	1	4.34	0.043	*
Algorithmic suggestion (unassisted vs. assisted)	1	0.67	0.416	
Level of expertise * algorithmic suggestion	1	2.18	0.146	

Panel C. Post hoc analysis

Pairwise comparison	Dif.	t	p-value	
C1. Specialist: assisted - unassisted	0.08	0.42	1.000	
C2. Generalist: assisted - unassisted	-0.28	-1.83	0.446	
C3. Unassisted: specialist - generalist	-0.68	-2.52	0.084	*
C4. Assisted: specialist - generalist	-0.32	-1.20	1.000	
C5. Specialist, unassisted - generalist, assisted	-0.40	-1.49	0.839	
C6. Specialist, assisted - generalist, unassisted	-0.60	-2.23	0.175	

P value adjustment: Bonferroni method

Behavioral reuse intention and usefulness

Table 10 presents a summary of behavioral reuse intention and usefulness. They were measured once after the assisted task, unlike other metrics. The participants did not know if their performance improved with the algorithmic suggestion because task results were not shared with participants. Hence, the results are based on their perception of the diagnostic assistance provided during the assisted task. Both groups show an equally positive perception.

Answer choices had a 5-point Likert scale (strongly disagree = 1; strongly agree = 5). The generalists perceives the algorithmic suggestion as useful (*M*=3.95, *SD*=0.89) and show a willingness to reuse it (*M*=3.79, *SD*=0.89). Specialist's result is similar: useful (*M*=3.93, *SD*=0.64) and intention for reuse (*M*=3.96, *SD*=0.46).

Table 10. Behavioral reuse intention and usefulness

Descriptive statistics

		Generalist (n=33)	Specialist (n=20)
Behavioral reuse intention	Mean (std. dev)	3.95 (0.89)	3.93 (0.64)

Usefulness	Mean (std. dev)	3.79 (0.89)	3.97 (0.46)

Chapter 6. Discussion

We assessed the effect of the algorithmic suggestion to inspectors performing bearing endoscope classification task. The results showed that the algorithmic suggestion has potential to significantly improve the performance of inspectors. Algorithmic suggestion improved specificity, sensitivity, and time efficiency. Specificity and time efficiency showed greater improvement for the generalists than the specialists. In the case of cognitive load, it reduced for the generalists but increased for the specialists. Both groups had a positive response to the usefulness and the reuse intention of the algorithmic suggestion system.

Specificity and sensitivity

In terms of specificity (Table 6), the algorithmic suggestion effectively improved it while the improvement was greater for the generalists than the specialists. Initially, the unassisted test result showed a concerning level of underperformance by the generalists compared to the specialists (Table 6, Panel C-C3). When the algorithmic suggestion was provided to the generalists, the specificity gap between groups was effectively removed. In addition, the standard deviation among the generalists is greatly reduced.

This result implies that the algorithmic suggestion is a useful tool for reducing variability in diagnostic accuracy attributed by the inspector's expertise level. A similar improvement is reported in the experiment of generalist radiologists' performance using CAD[25]. Generalist with algorithmic suggestion can take on parts of a specialist's inspection task without sacrificing diagnostic quality. Furthermore, the cost is reduced, and efficiency is improved as a specialist inspector is a scarce and costly resource. If this change can free up a requirement for specialists, their role may be shifted to new areas[3]. For example, a specialist's time can be better used at producing data sets for the training of algorithmic suggestion while generalist inspectors conduct the actual inspection at improved performance with the help of algorithmic suggestion. Such discussion is actively underway in other industries as well, including legal services[38] and medical diagnostics[39].

In the case of the specialists, the algorithmic suggestion slightly improved specificity with a statistically insignificant impact. It should be noted that the specialists' specificity in the unassisted test was already high leaving small room for any further improvement.

In terms of sensitivity, the algorithmic suggestion also improved it for both groups. Although the algorithmic suggestion on its own had lower sensitivity than an unaided specialist, it was still effective in improving specialists' performance. The reduction in standard deviation shows that algorithmic suggestion is effective in reducing variability within the group, which is one of the major shortcoming of human-oriented inspection.

Case analysis

The case analysis sheds light on how the algorithmic suggestion improved specificity and sensitivity. In the unassisted test, both groups had low specificity and high sensitivity (Table 6 and Table 7) indicating that they avoided false negatives at the expense of false positives. Inspectors tend to classify an uncertain case as abnormal rather than normal, which is a reasonable choice given the higher cost of false negatives than that of false positives. However, as in the result of the generalists, if the number of false positives increases beyond a certain level, the overall value of the inspection is greatly reduced. The trust in the diagnostics is undermined substantially, and there is an increase in the extra cost for following up on a false positive[40]. Therefore, it is crucial to reduce false positives while maintaining true positives. However, the challenge for a generalist is that there is a large variety of visual artifacts leading to a high rate of false positives, as shown in Figure 6. Accurate diagnosis of these cases requires training and experience that are time-consuming and costly.

Therefore, it is important to provide support to the generalists in reducing false positives, and the result proves that the algorithmic suggestion was effective in this aspect. Training an algorithmic suggestion system to effectively reduce false positives is an achievable goal if sufficient training data is available. Although false positives come in a large variety, they frequently occur in any inspection and colleting sufficient data is feasible. In the medical field, CAD specifically designed to reduce false positives were developed for a similar reason[41].

On the other hand, the specialists were not good at the subtle boundary cases (Figure 4). Even when the evaluation criteria are given, it is difficult for multiple inspectors to consistently apply the criteria. This was well studied in the experiment by Varun Gulshan et al[12].

The result suggests the potential benefits of the algorithmic suggestion in improving consistency. Because the training data was labeled through a rigorous process ensuring consistency, the algorithmic suggestion could make a consistent prediction following the criteria. Consistency is more important in the severity rating task than the binary classification task used in this study. Abnormal cases are classified into multiple severity scales, and the progression of the failure is identified based on the severity scale trend over the years. For this trend data to be valid, the consistency of diagnosis should be ensured.

Negative effects of algorithmic suggestion were also observed. When algorithmic suggestion made a false prediction, the percentage of correct classification was reduced. The reduction was much less when the fault in the image was visually obvious to the inspectors, since the inspector was not misled but maintained their correct decision. This result is in line with the previous findings by Fisichella [25] and Mani[26], which showed that the negative effect of false prediction varies depending on the task complexity and user's ability to dismiss them. Measures to minimize negative effects from incorrect predictions with algorithmic suggestion system are important research topics in the future.

Time efficiency

The algorithmic suggestion improved time efficiency for both the generalists (25.3%) and specialists (6.4%), suggesting that it played a positive role in helping inspectors analyze images quicker. The participants' feedback mentioned that "The algorithmic suggestion helped weed out obviously normal and abnormal cases quickly," which hints to that the time saving is mostly for obvious cases than subtle ones. The use of algorithmic suggestion can help free up human resources to focus on high-level tasks. The overall improvement is in line with the previous studies[42]. The improved time efficiency has a direct implication on the cost-saving. As the visual inspection method using an imaging device is more widely adopted as a routine maintenance task, the amount of image data to be processed is rapidly increasing. Hence, the benefit of cost-saving by algorithmic suggestion can be significant.

Cognitive load

In the case of cognitive load, there was a marginally significant interaction between expertise level and cognitive load. (Table 9, Panel B). For generalists, the algorithmic suggestion reduced average cognitive load by 7.0% (p=0.446). In the screening task, the inspectors processed a large number of images for many hours, which caused mental fatigue to the operators, an increase in the number of mistakes and a decrease in the working efficiency. Therefore, reduction in cognitive load is an important benefit. Since the task in the experiment involved a relatively small number of images, the effect of reduced cognitive load can be more significant in actual tasks where the volume of images processed are far greater than those employed here.

On the other hand, the algorithmic suggestion increased the average cognitive load of the specialists' case by 2.4% (p=1.000), although the difference was statistically insignificant. According to the task-technology fit[32] theory, cognitive load is an important indicator of how well a technology fits to a particular user and task. One plausible explanation for the opposite trend is information overload. In the experiment conducted by Irina et al. (2014) [43], when the decision support is provided, the cognitive load is increased if the user is already familiar with the task because the additional information can act as information overload. Another explanation is the trust effect. When the specialists used the algorithmic suggestion for the first time, they would have made additional cognitive efforts to evaluate the accuracy of the algorithmic suggestion while carrying out the task at hand to determine his or her level of trust in the system. Additional research is required for more accurate measurement of the effect on the cognitive load.

Perceived usefulness and reuse intention

In terms of perceived usefulness and reuse intention, both groups responded equally positively. This result indicates that both groups are willing to use the algorithmic suggestion if provided. For the generalist group, this is in line with the prediction of the task-technology fit model. A good task-technology fit[32] can be assumed as the cognitive load was reduced. In this case, positive usefulness and reuse intention are expected according to the model. However, the result for the specialists is not aligned with the model's prediction. The good fit cannot be assumed because the cognitive load was slightly increased, but usefulness and reuse intention were as positive as those of the generalists.

Limitations and future works

The present study has some limitations. Firstly, trust effect was not accounted for in the experiment. Time efficiency improvement and cognitive load reduction are affected by the trust level that the inspector has on the algorithmic suggestion. One participant in the specialist group mentioned in his feedback that "it can be quicker next time since I now know the accuracy level of the algorithmic suggestion" implying that the time efficiency improvement and cognitive load reduction can be larger after the user's trust level on the algorithmic suggestion is settled. A previous study [44] showed that the trust level changes over time while the user interacts with the system, and it takes a while for the trust level to stabilize. In order to ensure that the levels of trust and confidence are stabilized, participants should be provided with sufficient experience of using the algorithmic suggestion before the experiment.

Secondly, the task in the experiment was among the simplest in the area of fault diagnostic tasks that specialists routinely perform. To assess the effect of algorithmic suggestion on specialists more accurately, more complex tasks like severity scale rating and failure mode classification should be used for the experiment. In particular, the algorithmic suggestion is expected to be most effective for severity scale rating task for which the consistent application of criteria is critically important.

Thirdly, the specificity and sensitivity of the algorithmic suggestion used in this study can be further improved. Because the aim of this study is not achieving the highest possible specificity and sensitivity, we chose a well-known deep-learning model that offers reasonably good performance and is not difficult to train. If we used a more complicated model optimized for fault detection, such as the Faster R-CNN, specificity and sensitivity might have improved further. However, even with the improved performance, the findings of this study are still expected to be valid unless 100% specificity and sensitivity can be achieved for every failure mode. There are often ten or more failure modes per component and collecting sufficient training data for each mode can be cumbersome. Some failure modes occur only rarely. In some cases, failure modes develop too quickly to catastrophic failures leaving little to no time to collect data. Therefore, once the specificity and sensitivity reach a certain level, it is expected that they are limited by insufficiency of training data rather than a selection of a model. Rather, in such a case, there are cases in which a user can correctly classify as shown in the example of Figure 7(a). Therefore, with the sophistication of algorithmic suggestions, collaboration with humans can become more and more important to overcome limitations imposed by data.

Chapter 7. Conclusion

To the best of our knowledge, this study is the first published research in the field of predictive maintenance that measured the effect of algorithmic suggestion to human users. This study extended previous studies to include measurement of the differences in specificity and sensitivity when a user is assisted with algorithmic suggestion. Also, evaluation metrics were expanded to include time efficiency, cognitive load and user's perception. Since the effectiveness may vary depending on the level of expertise of the user, the experiment participants included generalist and specialist inspectors. We reviewed a wide range of previous studies in the medical field that measured the effect of computer-aided diagnostic (CAD) systems on radiologists.

We developed a deep learning-based algorithmic suggestion for bearing fault detection using image data from 138 wind turbines. Then, we performed a user experiment to measure the effect on technical inspectors with varying expertise levels. Thirty-four generalists and twenty specialists participated in the experiment.

The results showed that algorithmic suggestion had a statistically significant impact on improving the inspectors' specificity and time efficiency. In the case of cognitive load, it reduced for the generalists but slightly increased for the specialists. Both groups had a positive response to the usefulness and reuse intention of the algorithmic suggestion.

We have found that the effectiveness varied depending on the level of expertise. Specificity and time efficiency showed greater improvements for the generalists than for the specialists. As a result, when the generalists were assisted with the algorithmic suggestion, their specificity improved to a level that was not statistically insignificant from the unaided specialists. When specialists had algorithmic suggestion, average specificity and sensitivity improved slightly although those of algorithmic suggestion on its own was similar to that of an unaided specialist.

Types of problems where algorithmic suggestion was helpful differed for the generalists and the specialists. In the case of the generalists, algorithmic suggestion reduced false positives that were caused by artifacts. For the specialists, algorithmic suggestion improved the consistency of diagnostic decisions on subtle boundary cases.

Negative effects of algorithmic suggestion were also observed. When algorithmic suggestion made a false prediction, the average correct classification rates were reduced.

The results have the following implications. First, significant cost savings are expected as inspectors assisted by algorithmic suggestion can perform more accurate diagnostics in less time. Second, changes may occur in the roles of the generalists and specialists in the predictive maintenance field. The generalists with algorithmic suggestion can take on parts of the specialist inspection task. The specialist’s time can be better used at producing high quality data sets that are integral in the development of algorithmic suggestion.

Bibliography

1. Chauhan, V. & Surgenor, B. A Comparative Study of Machine Vision Based Methods for Fault Detection in an Automated Assembly Machine. Procedia Manuf. 1, 416–428 (2015).

2. Bhuvanesh, A. & Ratnam, M. M. Automatic detection of stamping defects in leadframes using machine vision: Overcoming translational and rotational misalignment. Int. J. Adv. Manuf. Technol. 32, 1201–1210 (2007).

3. Benke, K., Benke, G., Benke, K. & Benke, G. Artificial Intelligence and Big Data in Public Health. Int. J. Environ. Res. Public Health 15, 2796 (2018).

4. Mohan, A. & Poobal, S. Crack detection using image processing: A critical review and analysis. Alexandria Eng. J. (2017). doi:10.1016/j.aej.2017.01.020

5. Cha, Y.-J., Choi, W. & Büyüköztürk, O. Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks. Comput. Civ. Infrastruct. Eng. 32, 361–378 (2017).

6. Adhikari, R. S., Moselhi, O. & Bagchi, A. Image-based retrieval of concrete crack properties for bridge inspection. Autom. Constr. 39, 180–194 (2014).

7. Gonçalves, V. M., Delamaro, M. E. & Nunes, F. de L. dos S. A systematic review on the evaluation and characteristics of computer-aided diagnosis systems. Rev. Bras. Eng. Biomédica 30, 355–383 (2014).

8. E.D.Megaw. Factors affecting visual inspection accuracy. Appl. Ergon. 10, 27–32 (1979).

9. A Survey of Automated Visual Inspection. Comput. Vis. Image Underst. 61, 231–262 (1995).

10. Reiner, B. I. & Krupinski, E. The insidious problem of fatigue in medical imaging practice. J. Digit. Imaging 25, 3–6 (2012).

11. Sickles, E. A. et al. Performance Benchmarks for Diagnostic Mammography. Radiology 235, 775–790 (2005).

12. Gulshan, V. et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 316, 2402 (2016).

13. Wang, M., Tong, W. & Liu, S. Fault Detection for Power Line Based on Convolution Neural Network. in Proceedings of the 2017 International Conference on Deep Learning Technologies - ICDLT '17 95–101 (ACM

Press, 2017). doi:10.1145/3094243.3094254

14. Yoshida, H., Näppi, J., MacEneaney, P., Rubin, D. T. & Dachman, A. H. Computer-aided Diagnosis Scheme for Detection of Polyps at CT Colonography. RadioGraphics 22, 963–979 (2002).

15. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.

16. Cha, Y.-J., Choi, W., Suh, G., Mahmoudkhani, S. & Büyüköztürk, O. Autonomous Structural Visual Inspection Using Region-Based Deep Learning for Detecting Multiple Damage Types. Comput. Civ. Infrastruct. Eng. 33, 731–747 (2018).

17. ISO 15243:2017 - Rolling bearings -- Damage and failures -- Terms, characteristics and causes.

18. van Ginneken, B., Schaefer-Prokop, C. M. & Prokop, M. Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology 261, 719–32 (2011).

19. Gromet, M. Comparison of Computer-Aided Detection to Double Reading of Screening Mammograms: Review of 231,221 Mammograms. Am. J. Roentgenol. 190, 854–859 (2008).

20. Cooper, J. G., West, R. M., Clamp, S. E. & Hassan, T. B. Does computer-aided clinical decision support improve the management of acute abdominal pain? A systematic review. Emerg. Med. J. 28, 553–7 (2011).

21. Freer, T. W. & Ulissey, M. J. Screening Mammography with Computer-aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center. Radiology 220, 781–786 (2001).

22. Birdwell, R. L., Bandodkar, P. & Ikeda, D. M. Abbreviations: BI-RADS Breast Imaging Reporting and Data System CAD computer-aided detection PPV positive predictive value. Radiology 236, 451–457 (2005).

23. Ko, J. M., Nicholas, M. J., Mendel, J. B. & Slanetz, P. J. Prospective Assessment of Computer-Aided Detection in Interpretation of Screening Mammography. Am. J. Roentgenol. 187, 1483–1491 (2006).

24. Helvie, M. A. et al. Sensitivity of Noncommercial Computer-aided Detection System for Mammographic Breast Cancer Detection: Pilot Clinical Trial. Radiology 231, 208–214 (2004).

25. Computer-aided detection (CAD) as a second reader using perspective filet view at CT colonography: effect on performance of inexperienced readers.

Clin. Radiol. 64, 972–982 (2009).

26. Mani, A. et al. Computed tomography colonography: feasibility of computer-aided polyp detection in a &quot;first reader&quot; paradigm. J. Comput. Assist. Tomogr. 28, 318–26

27. Alberdi, E., Povyakalo, A., Strigini, L. & Ayton, P. Effects of Incorrect Computer-aided Detection (CAD) Output on Human Decision-making in Mammography 1 Computer Assisted Radiology and Surgery. doi:10.1016/j.acra.2004.05.012

28. Kilmon, C., Fagan, M., Pandey, V. & Belt, T. Using the task technology fit models as a diagnostic tool for electronic medical records systems evaluation. Issues Inf. Syst. 9, 196–204 (2008).

29. Shiraishi, J., Li, Q., Appelbaum, D., Pu, Y. & Doi, K. Development of a computer-aided diagnostic scheme for detection of interval changes in successive whole-body bone scans. Med. Phys. 34, 25–36 (2006).

30. Aoyama, M. et al. Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images. Med. Phys. 30, 387–394 (2003).

31. Dale, L. & Ronald, L. Task-technology fit and individual performance. MIS Q. 19, 213–236 (1995).

32. Zigurs, I. & Buckland, B. K. A Theory of Task/Technology Fit and Group Support Systems Effectiveness. MIS Q. 22, 313 (1998).

33. Sweller, J. Cognitive Load During Problem Solving: Effects on Learning. Cogn. Sci. 12, 257–285 (1988).

34. Venkatesh, V. et al. USER ACCEPTANCE OF INFORMATION TECHNOLOGY: TOWARD A UNIFIED VIEW 1. User Acceptance of IT MIS Quarterly 27, (2003).

35. Venkatesh, V. Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model. Information Systems Research 11, (2000).

36. Andreas, I., Nicolaou, ; & Mcknight, H. Perceived Information Quality in Data Exchanges: Effects on Risk, Trust, and. 17, (2006).

37. Knop, K., Olejarz, E. & Ulewicz, R. Evaluating and Improving the Effectiveness of Visual Inspection of Products from the Automotive Industry. in 231–243 (Springer, Cham, 2019). doi:10.1007/978-3-030-17269-5_17

38. Reid, M. A Call to Arms: Why and How Lawyers and Law Schools Should Embrace Artificial Intelligence. Univ. Toledo Law Rev. 50, (2018).

39. Jha, S. & Topol, E. J. Adapting to Artificial Intelligence: : Radiologists and Pathologists as Information Specialists. JAMA 316, 2353 (2016).

40. Evaluating Failure Prediction Models for Predictive Maintenance | Machine Learning Blog. Available at: https://blogs.technet.microsoft.com/machinelearning/2016/04/19/evaluating-failure-prediction-models-for-predictive-maintenance/. (Accessed: 21st January 2019)

41. Nagel, R. H., Nishikawa, R. M., Papaioannou, J. & Doi, K. Analysis of methods for reducing false positives in the automated detection of clustered microcalcifications in mammograms. (1998). doi:10.1118/1.598326

42. Burling, D. et al. Virtual colonoscopy: effect of computer-assisted detection (CAD) on radiographer performance. Clin. Radiol. 63, 549–556 (2008).

43. Mălăescu, I. & Sutton, S. G. The effects of decision aid structural restrictiveness on cognitive load, perceived usefulness, and reuse intentions. Int. J. Account. Inf. Syst. 17, 16–36 (2015).

44. Zhang, Z. & Wang, Z. Assessing and assuring trust in E-commerce systems. in CIMCA 2006: International Conference on Computational Intelligence for Modelling, Control and Automation, Jointly with IAWTIC 2006: International Conference on Intelligent Agents Web Technologies ... (2007). doi:10.1109/CIMCA.2006.52

45. Bughin, J. et al. ARTIFICIAL INTELLIGENCE THE NEXT DIGITAL FRONTIER?

국문 초록

이미지기반 알고리즘에 의한 지원이 베어링 고장 진단 검사자에 대해 미치는 효과

서울대학교
융합과학 기술 대학원
신 원

최근에 예지보전 (Predictive maintenance) 개념이 유지보수 영역(O&M)에 폭넓게 도입되면서, 각종 영상 장비를 이용한 검사가 더욱 중요해 지고 있다. 내시경, 열화상 카메라 등이 사람의 눈에 보이지 않거나 접근할 수 없는 영역의 결함을 확인하기 위해 사용된다. 이러한 고장 검사는 이제는 일반 유지보수 업무의 중심이 되어 가고 있으며, 이에 따라 고장 진단 경험이 부족한 일반 정비 기술자들에 의해 수행되는 경우도 늘어나고 있다. 그러나 사람이 행하는 육안 검사는 검사자의 전문성 수준에 따라 정확도가 달라지는 점, 여러명의 검사자가 있을 경우 진단의 일관성이 떨어지는 점 등 몇 가지 주요 문제가 있다.

이러한 단점을 해결하기 위해 컴퓨터 비전 및 검사 자동화에 분야에서 많은 연구가 수행되었다. 최근 여러 연구자들이 기계 결함 진단 문제에 딥러닝 솔루션 (Deep learning solution) 을 적용하여 매우 높은 정확성을 달성하였다. 기존 연구에서는 예지보전 분야가 딥 러닝 등 인공지능 기반 기술이 가장 먼저 실용화될 분야로 예측된 바 있으며. 이에 따라 실제 산업 현장에서 적용할 경우의 효과에 대한 기대가 있다. 그러나 이전의 연구들은 딥러닝 기반 기반 솔루션의 자체적인 진단 정확성에만 초점을 맞추고 있으며, 실제 사용자가 이를 활용할 경우의 효과까지는 측정한 바가 없다. 이에 반해, 의학 분야에서는 이미 90년대 후반부터 컴퓨터 지원 진단(Computer

Aided Diagnostics) 이 실용화되었으며, 영상 의학과 의사의 진단 정확성과 업무 효율성에 미치는 영향에 대한 연구가 활발히 진행되고 있다.

이 연구는 예지 보전 분야에서 알고리즘 기반 진단 지원이 인간 이용자에게 미치는 영향을 측정한 첫 번째 연구다. 138개의 풍력 발전기에서 취득한 내시경 사진 데이터를 사용하여 베어링의 고장 진단을 위한 딥러닝 기반 알고리즘 기반 진단 지원 시스템을 개발했다. 다양한 전문성 수준을 가진 기술자에 대한 영향을 측정하기 위한 사용자 실험을 실시했다. 34명의 일반 기술자와 20명의 고장 진단 전문가들이 이 실험에 참여했다.

실험 결과에 의하면 기술자가 알고리즘 기반 지원을 사용하여 과제를 수행할 경우 특이도과 시간 효율성에 있어서 통계적으로 유의미한 증가가 확인되었다. 인지부하의 경우, 일반 기술자의 경우는 줄었지만 진단 전문가는 소폭 증가했다. 두 그룹 모두 알고리즘 기반 지원의 유용성과 재사용 의도에 대해 긍정적인 반응을 보였다.

효과는 사용자의 전문성 수준에 따라 달라졌다. 일반 기술자 그룹이 진단 전문가 그룹에 비해 더 큰 효과를 보였다. 또한, 알고리즘 기반 지원을 통해서 도움을 받는 영역이 일반 기술자와 진단 전문가가 서로 다름을 확인하였다.

키워드 : AI assistance, 인공지능, 딥러닝, 예지보전, 이미지 기반 고장 진단, 컴퓨터 보조 진단, 베어링 결함, 내시경 검사

학번 : 2015-26031