VALIDATED AND DEPLOYABLE AI/ML FOR NDT DATA DIAGNOSTICS

BY ERIC LINDGREN

While artificial intelligence/machine learning (AI/ML) methods have shown promise for the analysis of image and signal data, applications using nondestructive testing (NDT) for managing the safety of systems must meet a high level of quantified capability. Engineering decisions require technique validation with statistical bounds on performance to enable integration into critical analyses, such as life management and risk analysis. The Air Force Research Laboratory (AFRL) has pursued several projects to apply a hybrid approach that integrates AI/ML methods with heuristic and model-based algorithms to assist inspectors in accomplishing complex NDT evaluations. Three such examples are described in this article, including a method that was validated through a probability of detection (POD) study and deployed by the Department of the Air Force (DAF) in 2004 (Lindgren et al. 2005). Key lessons learned include the importance of considering the wide variability present in NDT applications upfront and maintaining a critical role for human inspectors to ensure NDT data quality and address outlier indications.

Introduction

There is a growing increase in interest and attention in AI/ML, which are statistical methods for data analysis. The promise of AI/ML is to use statistical methods to self-extract attributes in the data, such as relationships and/or trends in data that are not as quickly and reliably made through typical human observation. The DAF has embraced the use of these tools for applications where it can accelerate decision-making in representative campaigns, as shown in Figure 1. The objective defined for one of these efforts is summarized as: "The Air Force aims to harness and wield the most optimal forms of artificial intelligence to accomplish all mission-sets of the service with greater speed and accuracy" (USAF n.d.).

With the potential to secure more NDT data through the transformation to fully digital instruments connected as envisioned by the Internet of Things (IoT) and NDE 4.0, there is an increased interest to use AI/ML methods as the diagnostic tool to determine if a flaw is present in NDT data. Justification for the use of AI/ML includes improved accuracy, improved reliability, and faster disposition time by decreasing or eliminating dependence on human interpretation and analysis of NDT data. The initial focus for the use of AI/ML addresses the detection of flaw indications, although there is exploration in the use of AI/ML to provide additional information on characterizing the size and location of discontinuities.

When considering the applicability of AI/ML for flaw detection, it is important to recall that these technical approaches are based on statistical methods, namely regression or classification of data. The concept includes the use of multiple statistical methods in parallel combined with multiple layers of analysis to extract statistical trends in the data to enable decisions that are not readily detectable through more classical methods. These multidimensional data analysis methods frequently are called neural networks. These approaches can either be trained using data with known ground truths called supervised AL/ML, or be allowed to form the statistical relationships without training data, called unsupervised AI/ML. As these methods rely on

US AIR FORCE GRAPHIC BY TRAVIS BURCHAM

Figure 1. The Department of the Air Force artificial intelligence/machine learning campaign illustration.

data, critical attributes of the data must be considered for their use. This includes the amount of available data, the accuracy of the data, and noise present in the data.

The intent of this article is to discuss some of the challenges of using AI/ML exclusively for the analysis of NDT data through a representative outcome when considering noise and data quantity. The approach being used by the researchers at the AFRL to enhance manual interpretation of NDT data is discussed, and several representative examples that integrate attributes of AI/ML into diagnostic capability are presented. The intent is to highlight the capabilities and opportunities within the NDT community to facilitate and accelerate the analysis of NDT data.

AI/ML Requirements for Engineering Decisions

The detection of flaws using NDT capabilities is an engineering decision that requires a statistical metric of capability to ensure the safety of systems. In aviation, the capability is frequently validated by a POD study that follows the guidance provided in MIL-HDBK-1823A (US DOD 2009). To make these types of assessments possible, it is necessary to have metrics on the data that include such factors as quantity, quality, and fidelity, which includes such relatively simple factors as signal-to-noise ratios (SNRs). The outcome of a POD study that follows the guidelines of MIL-HDBK-1823A will be appropriate statistical metrics for risk calculations, ensuring the safety of systems. In the DAF, this is part of the Aircraft Structural Integrity Program (ASIP) (US DOD 2016) and the Propulsion Systems Integrity Program (PSIP) (US DOD 2008).

Similar to POD studies, the same factors of the data affect the use of AI/ML. These factors become more critical as a function of the risk to a system if a flaw is not detected during an inspection cycle. Therefore, detailed understanding of the data being used is important to enable proper use of the AI/ML algorithms when using them to extract information from this data. Recent work has illustrated the impact of data quantity and SNR on the ability of a supervised neural network–based classifier (Lindgren 2022). The study used a synthetic dataset and introduced Gaussian noise at different percent levels while varying the number of data points used to train the AI/ML algorithm. The neural network used for this study was a multilayered perceptron with four layers and 50 layers in each hidden layer. The results of this evaluation are shown in Figure 2. The plot illustrates the log of the mean square error of the neural network as a function of SNR for varying the number of data points in each dataset. The SNR varies from an infinite value to one that is poor of only 10 to 1. The number of data points in each dataset varies from 50 up to 14 000. The outcomes are presented in standard box plots showing the interquartile region (IQR) and whiskers based on the 1.5 IQR value, and the outliers are indicated by red indices for each set of numbered data points.

It is clear from this study that the improved SNR and larger datasets result in a lower value for the mean squared error. This outcome is intuitively anticipated as it is expected that more data with higher fidelity will result in improved model outcomes. However, this example highlights some of the challenges of using AI/ML for NDT data analysis. Even with the highest level of SNR, using smaller datasets for training will produce outliers that are considerably deviant for the mean values. When considering the impact on safety of systems, these outliers are the equivalent of a large, missed

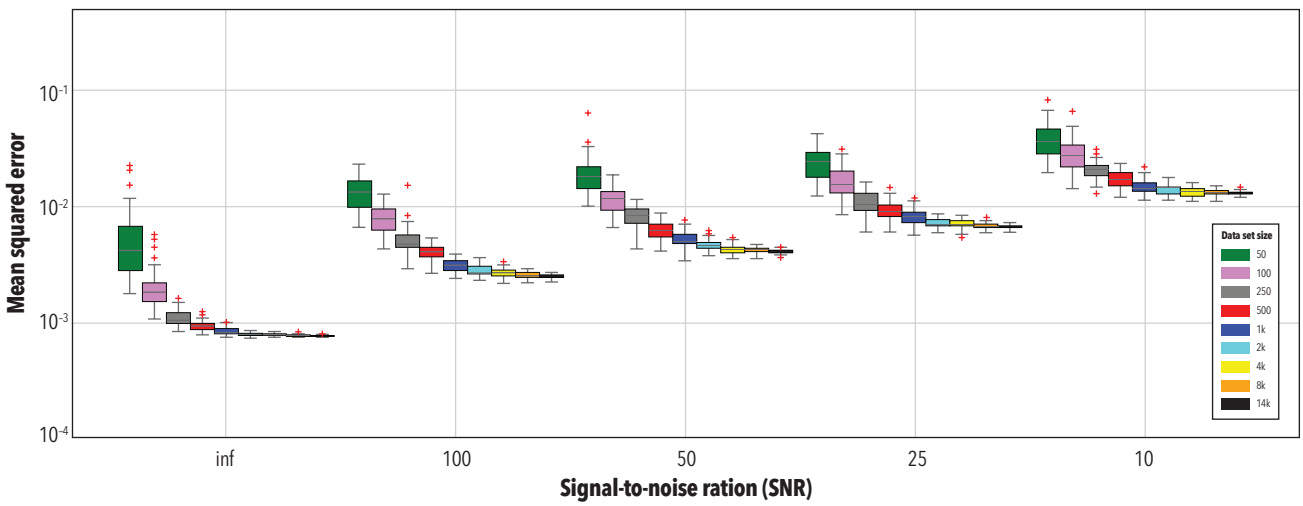

Figure 2. Multilayer perceptron results illustrating mean square error as a function of data quantity and signal-to-noise ratio.

flaw that could lead to an increased risk of a catastrophic outcome. It is important to recall that it is not the smallest flaw that can be detected, but the largest flaw that could be missed that impacts the safety of a system. This is especially true in aviation where single-load path structures are expected to have an extraordinarily low risk of failure when risk is managed by damage tolerance (US DOD 2016).

This data sensitivity study demonstrates two critical issues that need to be considered when applying AI/ML algorithms to NDT data. The first is the number of data points required to achieve improved performance of AI/ML methods. Large training sets of actual flaws are hard to generate due to the time and cost of preparing such samples. A common complaint of POD studies that follow the guidance of MIL-HDBK-1823A is the high cost to prepare samples with characterized flaws. The minimum number of flaws for a versus a-hat (i.e., flaw size versus magnitude of the signal response from the measurement system) assessments is 40 and for hit-miss assessments is 60. Large datasets of flaw responses in NDT data are difficult to find from service since the engineering response to the detection of a growing number of flaws is either to modify or replace the structural element of concern before a large population of flaws is present. An option that has been pursued includes the use of simulation to generate the required datasets for training. However, the challenge is to create simulations that are representative of the flaws found in actual structures. This approach would require a validation process with a good amount of empirical data covering the wide range of test conditions expected from an engineering perspective.

The second issue is the ability to address outliers and nuances in data that can be indicators of flaws. The concern is the tendency of statistical methods to ignore such features when using large datasets. Unless the attributes of the outlier and nuance change in data are included in sufficient large quantities in training, the approach would tend to dismiss such features in the data, which could result in missed flaws. Conversely, if the AI/ML is sensitive to outliers, then the concern becomes that a large number of false calls could decrease the value of implementing the AI/ML algorithm.

Thus, the lessons learned from the analysis of representative data includes the need to have the right data for training, including multiple flaws that are independent from each other. It is extremely important to recall that resampling the same data is not acceptable unless proper statistical methods to address correlated data are included in the analysis. Similarly, it is not acceptable to test AI/ML methods using the same data that was used for training. Another aspect is to ensure factors that can affect the statistical analysis of data (such as SNR) are included in the training datasets. In addition, if simulation data is used in training, it must be from validated models that capture all the anticipated variances found in the NDT data for the inspection. Lastly, the desired precision and accuracy of the diagnostics to be performed by AI/ML must be defined to ensure the amount of available data is sufficient to meet these objectives. This last consideration is especially true if unsupervised methods are being considered.

Challenges for AI/ML in NDT

As indicated by the sensitivity studies in the previous section, a significant challenge for the use of AI/ML in NDT data is to capture the effect of all the factors that can influence the capability to detect the flaws of interest. Figure 3 is a representation of these factors that the author has used extensively to illustrate the additional challenges when migrating from a laboratory to an operational environment. The three general classes of challenges can be summarized as equipment variability, structural complexity and variability, plus flaw complexity and variability. In addition, these parameters can change as a function of the life of a system, which increases the capability validation difficulty of the NDT system when integrated into system life management.

Equipment variability is the easiest of the three sources of variability to address from a research and development perspective. The variability in equipment settings can be defined and managed, but the unknown that frequently needs to be quantified is sensor variability and its impact on the diagnostics of flaws. Common NDT procedures address this with calibration processes, which alleviate many of these

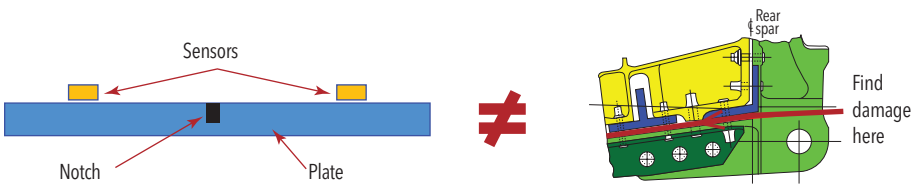

Figure 3. Representative increase in challenges when migrating from a laboratory environment (left) to an operational environment (right).

concerns. However, small changes in sensor configuration, such as coil tilt within eddy current sensors or slight depolarization of well-used ultrasonic transducers, can influence the flaw detection response.

Flaw-to-flaw variability can have a much greater impact on the NDT response. Previous studies have illustrated that the same size flaw can vary in amplitude response from an eddy current inspection by over 20% of a full screen height reading (Forsyth et al. 2015). Similar results can occur in ultrasonic testing as well as other NDT techniques. For ultrasound, fatigue crack morphology and tortuosity can affect a response. Local stress considerations from a fit-up of assemblies and changes due to use can vary crack closure, which, in turn, affects the magnitude of the ultrasonic signal. The variability can be addressed in simulation provided all the attributes of the flaw that affect detection are included in the simulation studies. This includes their interaction, which can become a very large study, especially when considering engineering level validation of the simulation.

While flaw-to-flaw variability can be broadly categorized as a function of the type of flaw, structural variability can become much more challenging in the analysis of NDT data. This is largely due to the extensive range of structures evaluated by NDT, which includes power generation, infrastructure, and transportation, the latter which can be segmented into ground, aviation, and space categories. In addition, other considerations include the materials being used, including metals, polymers, ceramics, and composites; the manufacturing process being used, for example, automation, partial automation, or hand assembly; plus, the assembly process used to join components, such as welding, fastening, and bonding. With all these parameters, it becomes very clear why NDT is the ultimate multidisciplinary engineering domain!

A significant challenge is how to evaluate the effect all these parameters, both individually and through important interactions, have on the NDT response. Consider the simple fastened joint between two metal surfaces, where up to 22 factors addressing equipment, flaws, and structure need to be included in a sensitivity study (Lindgren et al. 2007). Structural considerations include such things as composition of each layer; the possibility of shims and their composition; assembly quality, such as fastener hole tilt or skewed fasteners; and fit-up stresses as a function of what type of fastener is used and how it is installed. In addition, how these factors change as a function of time due to maintenance, repairs, modifications, and even use need to be included.

Using AI/ML techniques for these applications can become very daunting when considering all the parameters that need to be addressed to make diagnostic decisions using automated processes. This includes how the statistical processes adjust to account for changes that occur as a function of time. In addition, how these affect the diagnostic capability of the NDT data must be validated to enable their use in system risk and life management. Therefore, the proper capturing of these factors in statistically representative methods presents itself as a significant challenge, but also a significant research and development opportunity.

DAF Approach to AI/ML for NDT Data

AFRL has been leading the development of algorithms to assist in the diagnostics of NDT data, including one of the first implementations for an aviation NDT application (Lindgren et al. 2005). Attributes that have made this approach successful include the use of multiple approaches to develop algorithms for the diagnostic capability combined with the approach that the algorithms will not replace all human interpretation of NDT data. The algorithms are used as a capability to facilitate and guide the interpretation to make the workload on an inspector easier and focused on the critical elements of the diagnostic process that do not easily lend themselves for automation. AFRL has called this approach intelligence augmentation (IA), but an alternative term being used in the scientific community is collaborative intelligence (CI) (Epstein 2015). This reflects how software tools and capabilities can be used to assist in the analysis of NDT data, which AFRL has named assisted data analysis (ADA).

ADA algorithms combine multiple approaches to provide an optimized method to facilitate NDT diagnostics. These algorithms can be grouped into three general categories. The first uses heuristic-based methods that incorporate “rules of the road” that closely mimic the procedures by which inspectors interpret data. The second is a model-based inversion algorithm that uses simulation to represent the measurement response and iteratively solve for the unknown flaw or material state in the presence of variability. The third uses AI/ML methods trained using NDT data and as much diagnostics information as possible from available datasets. Frequently, the amount of well understood NDT data is much smaller than what would be required for robust AI/ML analysis, and likely requires supplementation from simulated data or transfer learning.

Successful application of ADA has frequently included at least two of these approaches into an integrated diagnostic algorithm for the specific

NDT application being addressed. This includes the use of test data to ensure the intent of the application is being met and that the available data meets the needs of the application before a comprehensive validation study is accomplished. The output of the ADA diagnostic is not the final disposition of an indication. Depending on the application, the output enables inspectors to focus their attention on portions of the inspection data that have features of possible indications by screening data with no attributes of a possible flaw. Alternatively, the output can be used to provide guidance on the nature of an indication so the proper disposition process can be rapidly identified and implemented, minimizing the time a system is in the inspection stage of a maintenance process. The key attribute of this approach is the human inspector remains in the loop. The inspector functions to ensure data quality, data fidelity, and can review any ADA outputs to make the final determination regarding an indication.

Representative DAF Successes

The following represents several examples developed by AFRL and transitioned to the DAF. The ADA capabilities are presented as a function of increasing complexity from the perspective of combining the three technical approaches outlined in the previous section. However, this order should not be considered a listing of increasing complexity as each application had its unique degrees of complexity and used different approaches to tailor to the need and to the desired outcome of the inspection.

A representative application that emphasizes the use of heuristics occurs in the manufacturing of aerospace composite structures, especially primary load carrying structures such as wing and fuselage skins. These parts require 100% ultrasonic inspection to detect delaminations and porosity where common rejection criteria are for delaminations greater than 6.35 mm (0.25 in.) in diameter or porosity that exceeds 2%. When considering the large areas to be inspected at manufacturing (note: this is not a requirement once a system is fielded), a bottleneck in the production flow can occur with the large volume of data to be assessed by inspectors. To minimize this bottleneck, a heuristic-based algorithm was developed to closely mimic the steps taken by an inspector to review data collected from these inspections (Aldrin et al. 2016).

The ADA algorithm leverages the available A-scan and B-scan data that accompanies the C-scan data. Multiple steps are taken in each of the three data representations to determine if an indication has features associated with delaminations that exceed the reject criteria. The representative result is shown in Figure 4 where C-scan features are identified as suspected defects and others are identified as benign. Though both may appear similar in the C-scan, attributes of the front wall, back wall, and volumetric gating can be used to distinguish between acceptable and rejectable features. The rejectable features are highlighted to the trained inspector who makes the final determination regarding the indication. With this approach, inspection processes have been greatly accelerated, though exact metrics are not available for publication.

Another representative case study includes the use of both simulation and heuristics to identify defects and discriminate between types of defects. The specific application is for rotating turbine engine components evaluated by an automated inspection system that can provide highly registered data. Using a combination of model-based assessments and heuristic analysis methods, the response from data with varying probe conditions can be evaluated and provide guidance on what features are from suspected indications and what are due to the probe variability (Aldrin et al. 2019b). A representative illustration of this approach is the experimental response from a subsurface nonmetallic inclusion in the presence of probe variation

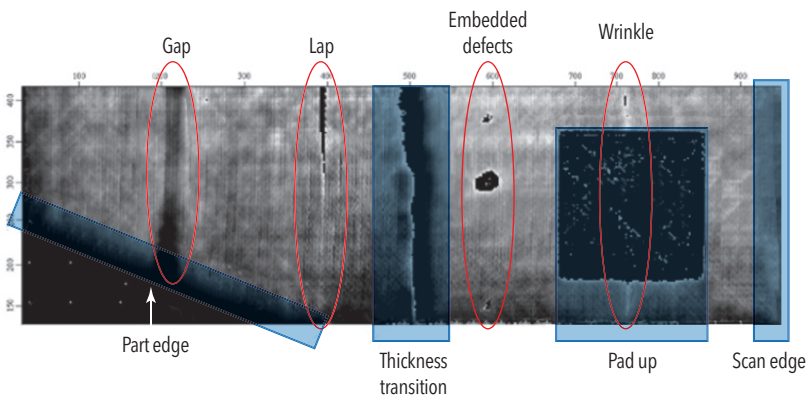

Figure 4. Ultrasonic C-scan of a composite test article indicating regions identified by the assisted data analysis algorithms as potential defects.

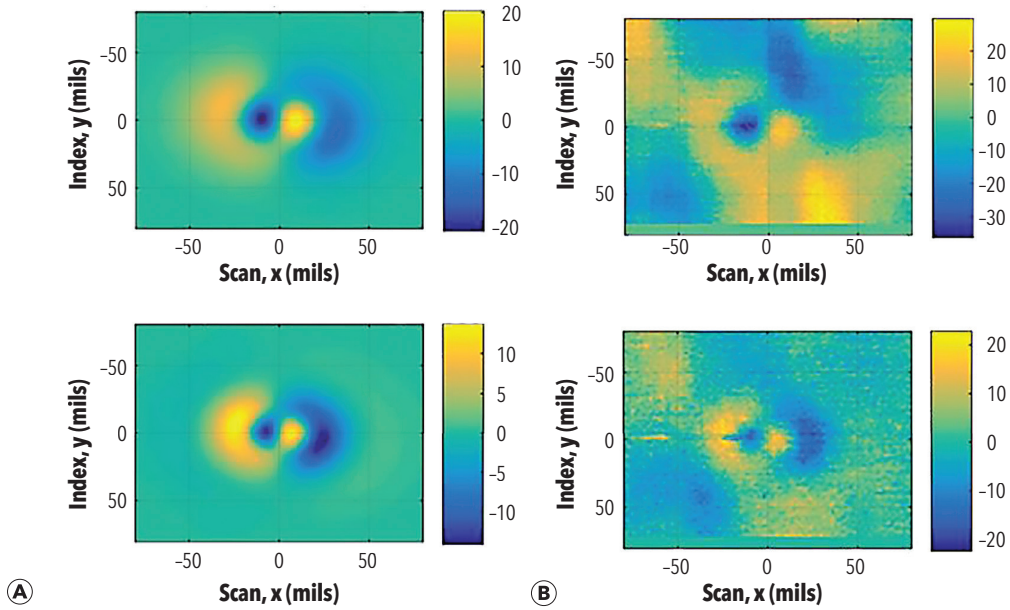

Figure 5. Model fit (a) to experimental data (b) for the vertical (top) and horizontal (bottom) representations of eddy current scans from a sub-surface nonmetallic particle with differences due to probe and material variability.

and material noise. Final results from iterative comparison and adjustments of the model data, being compared in impedance planes, are shown in Figure 5 and highlight the ability to evaluate the buried nonmetallic inclusion size and depth. Additional steps in the development process resulted in the ADA algorithms providing guidance to the inspectors when features in the data indicate when a fatigue crack is emanating from a nonmetallic inclusion. The ADA being developed for this application is in its final stages of refinement before it will be evaluated by a formal validation process.

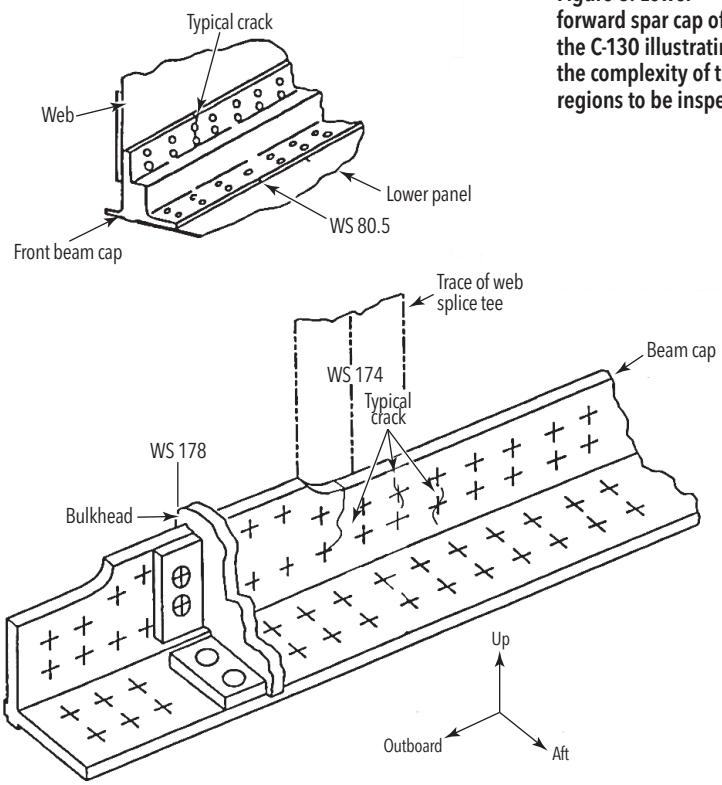

Figure 6. Lower forward spar cap of the C-130 illustrating the complexity of the regions to be inspected.

The third example combines elements of heuristics, simulations, and large dataset analysis to realize a successful outcome on a very complex inspection. The application addresses the lower forward spar cap on C-130 aircraft (Lindgren et al. 2005), as shown in Figure 6. The approach leverages development at the academic level for both the generation and detection of ultrasonic creeping waves (Nagy et al. 1994), plus the use of algorithms to discern the presence of cracks in a less complex, but still challenging, application (Aldrin et al. 2001). As described in Lindgren et al. (2005), the solution included the use of analytical methods to fully represent the propagation paths within the structure; simulation tools to explore various attributes of the inspection data as it propagates in the structure; plus, the use of advanced processing methods, namely echo dynamics and local correlation functions, to discriminate between responses from potential flaws to those from other geometric reflectors found intermittently in the structure. In addition, over 2000 representative inspection opportunities

from both harvested and mock-up test articles were used to refine the decision-making process for the ADA algorithms.

The inspection process was fully validated by a comprehensive POD study before being deployed. The inspections were accomplished by contractor field teams that would collect the data and ensure it had sufficient quality to be evaluated by the ADA. Suspected indications identified by the ADA were sent to an NDT engineer to make a final determination if the indication was confirmed and needed to be sent to engineering for disposition.

The next generation of ADA will expand the capability of algorithms to facilitate the identification of defects to the capability to characterize the defects in ways that are not available today. While inspectors can use methods to approximate defect size, attributes like fatigue crack depth are especially challenging. However, using a combination of heuristics, simulations, and data-driven analytical methods, the use of ADA to determine the depth of a fatigue crack from a bolt-hole eddy current inspection was shown to have an average accuracy of 8.5% for fastener holes with minimal variability (Aldrin et al. 2019a). The next steps in the development process are to use this integrated approach to address fastener hole variability, such as skew and out-of-round attributes, to provide a crack depth estimated with a statistical bounds on accuracy, to enable rapid disposition of these defects in aerospace structures.

Summary

There is a continued potential for AI/ML methods to enhance data analysis and diagnostics for NDT data. However, there needs to be a realistic approach that includes evaluation of the data quantity, quality, and fidelity. This ensures it has the desired attributes that enable the AI/ML techniques to provide outcomes with sufficient statistical metrics for the results to be used in engineering decisions. In addition, these outcomes require rigorous validation of the diagnostic capability before they can be trusted to help ensure the integrity, or safety, of systems.

A representative example illustrated the challenges in using AI/ML techniques for smaller and noisy datasets, highlighting how this can lead to outliers that would imply potentially missed defects if this approach was used for NDT datasets. Additional challenges exist in data variability from equipment, defects, and structure that impact the amount of quality data required for AI/ML approaches. While data for defects can be augmented by simulations, these must contain all the anticipated variability and complexity of the NDT evaluation technique to represent nuances and outliers that are challenging for AI/ML, but critical for high-accuracy flaw detection.

The challenges of AI/ML when used for NDT data has led AFRL to pursue a hybrid approach that integrates AI/ML with heuristic- and model-based diagnostic algorithms to facilitate and reduce the workload of inspectors while not taking them completely out of the loop. Representative examples for several DAF-related applications have demonstrated the power of combining at least two of these methods to enable complex inspections and diagnostics of NDT data. The ADA algorithms are combined with human analysis to maximize the value of the algorithms by reducing the workload of inspectors so they can focus on the critical data that could be indications of defects being present. Future work includes plans to expand the capabilities of ADA algorithms to characterize defects with statistical metrics of accuracy. Initial development efforts have shown the potential of this capability, which would decrease the disposition time of indications and increase availability of the system to the end user. ME

ACKNOWLEDGMENTS

The author expresses his deep appreciation for the pioneering contributions and collaboration with Dr. John Aldrin of Computational Tools. Mr. David Forsyth of Texas Research Institute – Austin is recognized for his contribution to implementing ADA algorithms for multiple NDT applications. The AI/ML analysis here would not be possible without the work performed by Mr. Tushar Gautam and Drs. Kirby, Hochhalter, and Zhe of the University of Utah.

AUTHOR

Eric Lindgren: AFRL and ASNT Fellow; eric.lindgren@us.af.mil

CITATION

Materials Evaluation 81 (7): 35–42
https://doi.org/10.32548/2023.me-04364

REFERENCES:

Aldrin, J. C., D. S. Forsyth, and J. T. Welter. 2016. "Design and Demonstration of Automated Defect Analysis Algorithms for Ultrasonic Inspection of Complex Composite Panels with Bonds." *AIP Conference Proceedings* 1706. https://doi.org/10.1063/1.4940591.

Aldrin, J. C., E. A. Lindgren, and D. S. Forsyth. 2019a. "Intelligence augmentation in nondestructive evaluation." *AIP Conference Proceedings* 2102 (1): 020028. https://doi.org/10.1063/1.5099732.

Aldrin, J. C., E. K. Oneida, E. B. Shell, V. Sinha, K. Keller, J. K. Na, A. L. Hutson, H. A. Sabbagh, E. Sabbagh, R. K. Murphy, S. Mazdiyasni, M. R. Cherry, and D. M. Sparkman. 2019b. "Model-based inversion of eddy current data for classification and sizing of planar and volumetric discontinuities." *Vol. 44: Electromagnetic Nondestructive Evaluation XXII* 44:80–85.

Aldrin, J. C., J. D. Achenbach, G. Andrew, C. P'an, R. Grills, R. T. Mullis, F. W. Spencer, and M. Golis. 2001. "Case Study for the Implementation of an Automated Ultrasonic Technique to Detect Fatigue Cracks in Aircraft Weep Holes." *Materials Evaluation* 59 (11): 1313–19.

Epstein, S. L. 2015. "Wanted: Collaborative intelligence." *Artificial Intelligence* 221:36–45. https://doi.org/10.1016/j.artint.2014.12.006.

Forsyth, D.S., J. Ocampo, H. Millwater, and J. Montez. 2015. "Structural Health Monitoring, Risk, and Reliability." Aircraft Structural Integrity Program Conference. available at: https://www.researchgate.net/publication/344886205_Structural_Health_Monitoring_Risk_and_Reliability.

Lindgren, E. A., J. R. Mandeville, M. J. Concordia, T. J. MacInnis, J. J. Abel, J. C. Aldrin, F. Spencer, D. B. Fritz, P. Christiansen, R. T. Mullis, and R. Waldbusser. 2005. "Probability of Detection Results and Deployment of the Inspection of the Vertical Leg of the C-130 Center Wing Beam/Spar Cap." 8th Joint DoD/FAA/NASA Conference on Aging Aircraft.

Lindgren, E. A., J. S. Knopp, J. C. Aldrin, G. J. Steffes, and C. F. Buynak. 2007. "Aging Aircraft NDE: Capabilities, Challenges, and Opportunities." *AIP Conference Proceedings* 894:1731–38. https://doi.org/10.1063/1.2718173.

Lindgren, E.A. 2022. "Intelligence Augmentation for Aviation-based NDE Data." 2022 Aircraft Structural Integrity Program Conference, available at: http://www.arctosmeetings.com/agenda/asip/2022/proceedings/presentations/P23251.pdf and https://doi.org/10.32548/RS.2022.005.

Nagy, P. B., M. Blodgett, and M. Golis. 1994. "Weep hole inspection by circumferential creeping waves." *NDT & E International* 27 (3): 131–42. https://doi.org/10.1016/0963-8695(94)90604-1.

US DOD. 2008. MIL-STD-3024: *Department of Defense Standard Practice, Propulsion System Integrity Program (PSIP)*. US Department of Defense.

US DOD. 2009. MIL-HDBK-1823A: *Department of Defense Handbook, Nondestructive Evaluation System Reliability Assessment*. US Department of Defense.

US DOD. 2016. MIL-STD-1530D: *Department of Defense Standard Practice, Aircraft Structural Integrity Program (ASIP)*. US Department of Defense.

USAF. n.d. "BLUE: THE AI Advantage." US Air Force. https://www.af.mil/News/Photos/igphoto/2002319445/.