International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 15 August 2023 **Revised:** 16 September 2023 **Accepted:** 7 October 2023*

Navigating the Landscape of Robust and Secure Artificial Intelligence: A Comprehensive Literature Review

Saurabh Suman Choudhuri[1], Jayesh Jhurani[2]

[1]Vice President & Global Head of Digital Modalities, SAP America Inc.
Email id: s.choudhuri@sap.com[1]; [1]IEEE id: 99962111.
[2]IT Manager, ServiceTitan, Inc. **Email id: jjhurani@servicetitan.com[2]**

Abstract
Addressing the multidimensional nature of Artificial Intelligence assurance, this thorough survey is dedicated to elaborating on various aspects of ensuring the reliability and safety of computerized systems. It steers through the turbulent seas of model enervates, unmodelled phenomena, and security menaces to give an elaborate lit review. The review touches upon the boisterous ways of addressing these intricate mitigation strategies for model errors used in the past, the challenges of under-specification with modern ML models, and how understanding uncertainty is crucial. In addition, it evaluates the AI system's security basis, the emerging Adversary Machine Learning field, and its processes necessary for testing and evaluation of weaker adversarial case studies. The review of literature also looks upon the situation of DoD context, how the terrain surrounding developmental and operational testing is altering with all these shifts in culture that must be implemented if not to implement robust but secure AI implementation.

Introduction

The only issue of concern is the resilience of Artificial Intelligence (AI) systems and several challenges which include difficulty in model accuracy, tasks we are unable to model, and security threats. Correcting errors regarding the model, that is, imperfections, such as problems related to optimization mechanisms and criteria of regularization and inference algorithmics, is one of the areas addressed by contemporary literature. Despite the expectations, the under-specification problem has significant hindrances in modern machine learning, especially deep learning, facilitating hidden biases in prediction and leading to undesired failures once deployed. The review further gives special focus to calibrated uncertainty measures that are central in steering the accrued intricacies occasioned by the introduced dataset shifts. Security issues developing in AI systems are analyzed focusing on the proven software security formation, as well as the nascent adversarial machine learning domain. The literature review reveals how AI systems emerge as domain-level solutions that are embedded intrinsically in the defense function, thereby illustrating the developmental and operational testing landscape for the Department of Defense (DoD).

Robustness of AI Components and Systems

Robustness in AI systems represents a complex examination, interrogating the layers of mistakes arising from model errors and unmodeled states. This paper captures the current understanding of those challenges while highlighting the researcher's significant detail given on the controversies proposed to be in place and gaps that require research.

Addressing Model Errors

Interestingly, there is a huge literature that stresses the need to consider model errors in considering steps to increase the automation resilience or AI system [1]. A vast range of techniques related to such eventually used include robust optimization, regularization, risk-sensitive objective functions, and robust inference algorithms that have shown much coverage. Although some of these approaches have shown effectiveness in controlled environments, a significant gap remains ever since even a few tools translate these theoretical breakthroughs into everyday effective implementation.

Underspecification in Modern ML Systems

As further discovered specification is becoming a massive hurdle in developing ML systems, more pronouncedly those that use deep learning, to achieve robustness. This phenomenon has been identified as under-specification due to

***Article Received:** 15 August 2023 **Revised:** 16 September 2023 **Accepted:** 7 October 2023*

the algorithm's ability to solve optimization problems when training deep neural networks, resulting in various solutions with very similar average performance [2]. This is in line with the models that emerge owing to this phenomenon; such models harbor a hidden bias and underlined fault, thus making their eventual implementation filled with inconsistencies. Model selection in deep learning is the gaping hole of an explorable AI as described by many researchers who practice XAI (**Figure 1**). Lack of seeing the potential consequences of human actions that can be revealed through understanding model decisions has given significant room for blame and faults to enter advocacy and government policy making. The resulting xenophobia from such assumed complicity presents a social challenge if not addressed.

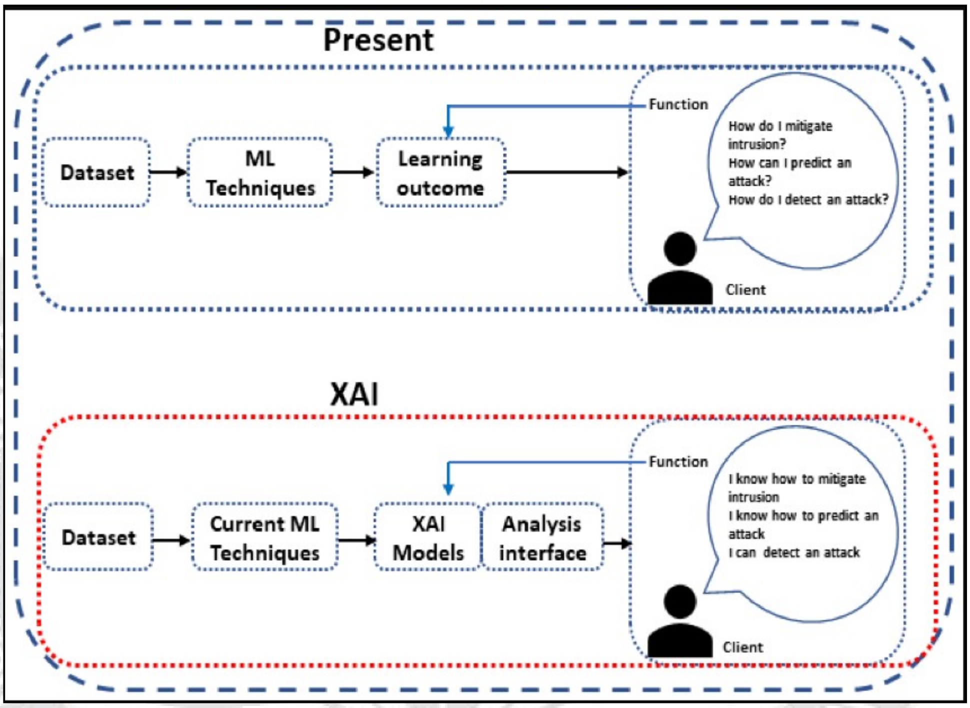


Figure 1. The concept of XAI [3].

Understanding Uncertainty in ML Models

Understanding and trusting ML model uncertainty means having a high level of finality to the structure of modern ML algorithms. The phenomenon, referred to as dataset shift and defined [1], consists of a distortion of the training data distribution due to changes made during the operational period. Calibration methods are noted in the literature as essential techniques that aid in setting uncertainty scores so that they accurately capture the likelihood of prediction accuracy. Effectively calibrated uncertainty measures allow the design, integration, and monitoring of policies to office prudence in the operation of AI systems.

Challenges and Opportunities in Robust AI

The roadblocks accompanying the development of Robust AI are opportunities to call to mind that they do not present obstacles but call for rapid attention. However, there are problems involving the determination of robustness criteria and testing strategies in the AI system life cycle despite frequent mentions throughout the literature. From items such as model evaluation, deployment, and continuous monitoring throughout operational stages, the literature illuminates these interim challenges that are critical [2], [4]. Three directions are identified as promising strategies including “building robustness in” through smart design and customization and utilizing algorithms with robustness features. And while the literature notes the need for additional study and coordinated unification of testing procedures.

Tools and Practices for Measuring Robustness

The reviewed literature stresses the need for creating methods, frameworks, and practices that would allow measuring AI component resiliency as well as system-level resiliency. The abovementioned individuals must use better tools, of which the following are recognized as necessarily proper for AI engineers, product managers, designers, software engineers, systems engineers, and operators. These tools serve as the centerpiece of activities in developing,

Article Received: *15 August 2023* ***Revised:*** *16 September 2023* ***Accepted:*** *7 October 2023*

engineering, and running AI-embodied features with certainty as well as confidence.

Security Challenges in Modern AI Systems

However, one critical aspect that needs to be addressed while building AI systems in the modern technological landscape is security. This study provides an integrated look at the complex sophistication of the security risks in modern AI systems, specifically focusing on securing against intentional perversion and involuntary failure.

Foundations in Software Security

AI is a system of amalgamation of some software and data being part of the bigger world namely, it comes under systems included in the roofs designated to software cyber-physical systems. In consideration of this, AI engineers should resort to previously known information and best practices originating from software security domains. The following measures are innovative to fill this gap: integrating MITRE's ATT&CK (**Figure 2**) framework for securing ML systems in production and building on prioritized security features focused on AI.

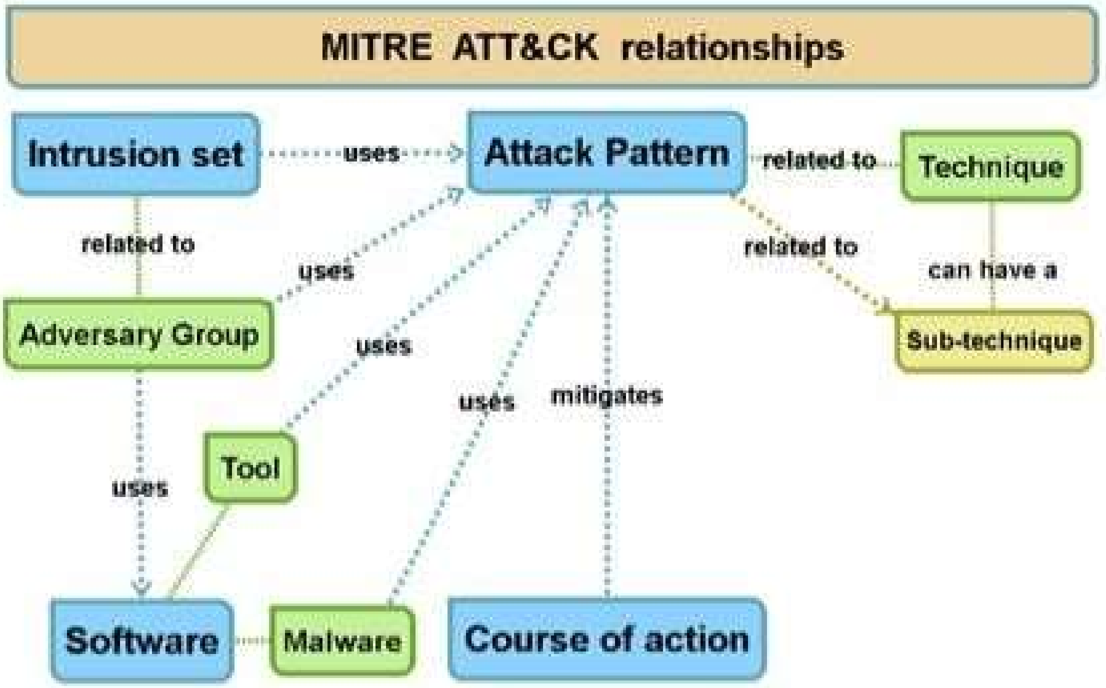


Figure 2. ATT & CK Model relationships [5].

Adversarial Machine Learning: Taxonomy and Strategies

Advances in modern ML algorithms, with notable ones being deep learning have unveiled contemporary attack routes hence the popularity of adversarial machine learning. Scientists who research in this field want to understand machine learning models and their underlying threats as well as ways by which we can protect them from being attacked [6], [7]. Taxonomy categorizes attacks into three areas: Learning something incorrect, performing an action of the wrong sort, and even revealing something that is correct but should unveiled, not made available nor exposed. Misrepresenting or corrupting operational data, attackers exploit models by providing malicious examples to take unexpected and unfavorable responses. Additionally, the attackers can take advantage of elements that allow them to target ML models deployed in production.

Mitigation Strategies and Trade-offs

Combating the adversarial AI involves embracing intricate coordinating principles as well as embedded compromises. As is the tact of such arm-lifting decision makers in this scenario, defenders – system builders and operators are required to make decisions as they find themselves between 'dirty' or somewhat shady tradeoffs in information advantage afforded to attacker versus defender [8]–[10]. Interestingly, the latter study does not make an explicit connection to its critics but also clearly reveals the complicated tradeoff

***Article Received:** 15 August 2023 **Revised:** 16 September 2023 **Accepted:** 7 October 2023*

between do right thing policy enforcement, learning, and disclosure. It is a common paradox that such models are created to "do the right things"; they may, however, eventually become more deserving of information unveiling.

Research and Development Imperatives

Since AI is becoming hostile, the surrounding area makes it continuously necessary to research and invent new ideas. The main areas of focus include accounting for the relations between several defense policies, taking the level of information availability by both attackers and defenders into consideration, as well as budget limitation management. Furthermore, there is a significant need for certain products that enable the builders of AI systems to understand what it need security-wise.

Expanding Security Coordination and Red Teaming

There are two major fields of opportunity for upgrading AI systems in terms of their security against developing threats. First, there is an indication to fully exploit the research findings of security vulnerability coordination as a way of accommodating AI technologies for connecting with new vulnerabilities. With AI increasingly being used in real-world systems, developing strategies to understand and resolve its inevitable peculiar security implications constitutes one of the most pressing issues [11], [12]. Red teaming presents a second area of benefit: improving red teaming capabilities is an effective approach. A traditional practical part of optimizing security in software systems, red teaming can be operationalized as an inspective tool to evaluate the aspects of the security panorama in AI-lion environments.

Processes and Tools for Testing, Evaluating, and Analyzing AI Systems

Considering the popularity of new trends such as AI systems development and deployment, consideration for robustness and security have been given a lot of attention. This detailed study covers crucial processes and tools that need to be performed for testing, verification, and other procedures for AI systems evaluation. Seen from an AI engineering point of view, this discussion highlights the need for dedicated tools, methods design patterns, and artificial intelligence standards that help the application projection of a full-scale solution responsible building and operation.

AI Engineering Landscape

For a full and accurate understanding of AI systems' level of strength in robustness and security, one needs to look closely at the technical, algorithmic, as well as mathematical constructs upon which such interaction is structured. Though innovative instruments are not uncommon, the highly specific nature of AI systems and particularly ML applications requires an entirely different set of tools and processes. Different from conventional system software engineering, AI focuses on issues that are usually larger in their profile, less resolved but formulated rather vague with more inherent complexity of input and result spaces [3], [12], [13]. Although some traditional software engineering tools are useful in providing support, they fail in solving AI problems completely. This also shows that such tools are particularly required, and innovative ones developed for the slight difference of AI development.

Challenges in Existing Testing Tools

Traditional testing tools, mostly designed for conventional software development, often have limitations when used to test AI and ML algorithms. Large problem spaces, fuzzy objectives due to understandings of end states or emergent behavior in smart systems, and complex mapping of inputs to outputs, all require greater refinements to which traditional testing methods do not cater Sometimes, a clear gap is observable that leads to the development of completely new verticals. Responding to the nuances of the AI systems requires going beyond the ordinary and using tools fine-tuned for addressing nuances of AI development.

Incorporation into Modern Software Development

Seamlessly performing the operations necessitates AI system tools integrated into present-day software development processes. In parallel with traditional software engineering ideas, AI engineers need to have instruments such as those designed for software reverse engineer rings, static and dynamic code analysis, and fuzz testing [5], [8]. On the other hand, AI is unique in its methodological requirements which involve additional innovative approaches that can be oriented into making standard testing different in certain ways. The use of AI-specific tactics that are part of contemporary action trajectories enables coherent incorporation into the system compliance with the overall purpose of achieving soundness and security.

Integration into DevOps and MLOps Pipelines

For AI development and deployment, it is necessary to integrate AI tools into DevOps or Machine Learning Operations (MLOps) pipelines. It is this integration that makes the process easier and simpler - by enabling Continuous Integration and Continuous Delivery (CI/CD) (**Figure 3**) along the way. Integrating CM into the CI/CD framework requires the promotion of continuous monitoring and security enhancement. This ongoing monitoring ensures that the resilience and security of AI systems are dynamic measures and not a steady solution as it guides throughout the system lifecycle [4], [7], [8]. One of the roles that Continuous Monitoring plays is to facilitate real-time assessment for consistent identification of potential weaknesses with necessary and incremental implementations such as controls, mitigations, model retraining even systems redesign based on the actual performance of the running systems.

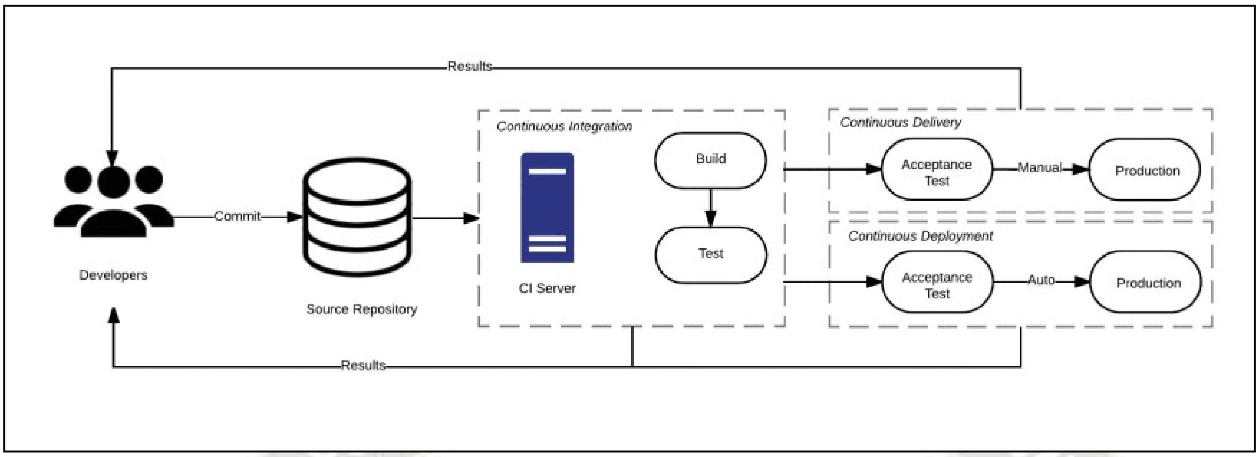


Figure 3. Relationship Between Continuous Integration, Delivery, And Deployment [14].

Foundations of Robust and Secure AI

Apart from their inherent value, these properties support mission success and develop other similar characteristics like safety, availability, credibility, deliverability, and conformity. Strong and stable systems have a significant role in fulfilling policy dependencies such as privacy, equity, and morality. The highly volatile DoD necessitates a complete paradigm shift in the developmental test and evaluation (DT&E) and operational test and evaluation (OT&E) processes to incorporate AI within business-as-usual strategies.

Evolution of DT&E and OT&E

AI systems cannot be evaluated using traditional platforms, and the entire process of DoD's acquisition must involve processes such as DT&E and OT&E. However, this evolution needs to involve careful deliberation on the generation of system testing requirements procurement of such and cost-related issues concerned with continuous monitoring [6], [10], [12]. AI was included in OT&E through the recent workshop that was organized by the University of Maryland's ARLIS, where it exposed the needs and challenges it introduced to carry out this process. Interestingly, the workshop highlighted the gap between what can be right easily measured at measured and what materially influences operations.

Pacing Test and Evaluation Practices with Technological Advances

Under the evolution dynamics of modern technologies, there is a need for an agile and proactive test and evaluation community within the DoD. However, this covers an increase in the number of AI testers capable of handling all these complexities caused by leveraging activities involving AI systems and simulators [7], [13]. The cultural burden is associated with creating a belief in risk-taking across the entire set of stakeholders related to AI systems creation and implementation. The proselytizing and prototyping are ingredients fundamentally across domains, AI intelligibility makes a unique set of challenges as some information reasons about whether half-hooked onto a machine setting up the system testing early in program development stages.

The Crucial Role of Rigorous Testing

Despite the common belief that testing as an activity is slow, it is a process of finding defects and redundant features, especially late on during project management functions. Deep iterations of inquiry, learning, construction, and testing are fundamentally critical for the teams accountable for designing and developing AI systems. This method allows for pointing out inconsistencies in the information flow that structures the general program of actions within this system. The consistent assessment of the model's capability to hold up to unanticipated phenomena and tolerate attacks is crucially important. Also, its viability for decision-making is deemed necessary to ensure efficiency.

Interdependence and Experimentation

In any complicated structure, the comprehension of interlocking random elements is vital. Teams must steer by experimentation to 'fingerprint' such interdependencies, on the one hand, and devise contingent plans for unexpected behaviors caused by changes within systems. This throws light on the need for a non-linear perspective analysis that embodies the complex interdependencies within the AI

Article Received: *15 August 2023* ***Revised:*** *16 September 2023* ***Accepted:*** *7 October 2023*

system [4], [10], [11]. In the process of organizing their utilization in high-risk environments such as managing under-control highway systems or power grids, it should ensure the adaptability and safety of AI Systems. Also, the whole issue regarding national security uses must be given even more emphasis because such applications necessarily pose a high level of risk.

Cultural Shift and the Path Forward

As the DoD aims to achieve powerful and resilient AI systems, such an effort should engage a broad-based strategy. This comprises looking at things in different ways, that is, for the known unknowns and the unknown" unknown", as well as a culture of both experimenting and testing. Using AI systems to function in higher-value environments requires serious thought to be put into the threats and weaknesses [4], [8]. With only time, AI systems leveraged throughout the critical infrastructures will form enticing targets and the DoD should remain alert over new risks that could surface.

Conclusion

This review highlights the complex and dense terrain of guaranteeing resilient and safe AI systems. In dealing with the issues relating to model errors, partial specification, and threats, a careful approach is necessary for coming up with intelligent solutions The introduction of AI into the DoD's testing practices and tendencies requires a cultural change, recognizing experimentation as well as proactive testing necessary. Strict testing procedures, persistence in a model's reliability assessment amidst evolving conditions as well as mechanisms of AI systems interdependency are critical for effective achievement. With the rise of AI practices in high-risk environments, it has become necessary for the DoD to proceed with caution in addressing cropping-up risks to ensure resilience as a secure and culturally cautious strategy towards its implementation. To maintain the competitiveness of advantages over emerging cellular and network threats, there must be constant R&D effort underpinning and supporting AI systems that need to meet rigid standards geared towards robustness and security.

References

[1] R. Dzombak and S. Beckman, "Unpacking capabilities underlying design (thinking) process," *Int. J. Eng. Educ.*, vol. 36, no. 2, pp. 574–585, 2020.

[2] S. Patel and K. Mehta, "Systems, design, and entrepreneurial thinking: Comparative frameworks," *Syst. Pract. Action Res.*, vol. 30, pp. 515–533, 2017.

[3] C. I. Nwakanma *et al.*, "Explainable artificial intelligence (xai) for intrusion detection and mitigation in intelligent connected vehicles: A review," *Appl. Sci.*, vol. 13, no. 3, p. 1252, 2023.

[4] J. M. Spring, A. Galyardt, A. D. Householder, and N. VanHoudnos, "On managing vulnerabilities in AI/ML systems," in *New Security Paradigms Workshop 2020*, 2020, pp. 111–126.

[5] B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington, and C. B. Thomas, "Mitre att&ck: Design and philosophy," in *Technical report*, The MITRE Corporation, 2018.

[6] J. Helland and N. VanHoudnos, "On the human-recognizability phenomenon of adversarially trained deep image classifiers," *arXiv Prepr. arXiv2101.05219*, 2020.

[7] P. Bajcsy, N. J. Schaub, and M. Majurski, "Designing trojan detectors in neural networks using interactive simulations," *Appl. Sci.*, vol. 11, no. 4, p. 1865, 2021.

[8] K. Tran, W. Neiswanger, J. Yoon, Q. Zhang, E. Xing, and Z. W. Ulissi, "Methods for comparing uncertainty quantifications for material property predictions," *Mach. Learn. Sci. Technol.*, vol. 1, no. 2, p. 25006, 2020.

[9] Y. Ovadia *et al.*, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[10] A. B. Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. fusion*, vol. 58, pp. 82–115, 2020.

[11] A. Tocchetti and M. Brambilla, "The role of human knowledge in explainable AI," *Data*, vol. 7, no. 7, p. 93, 2022.

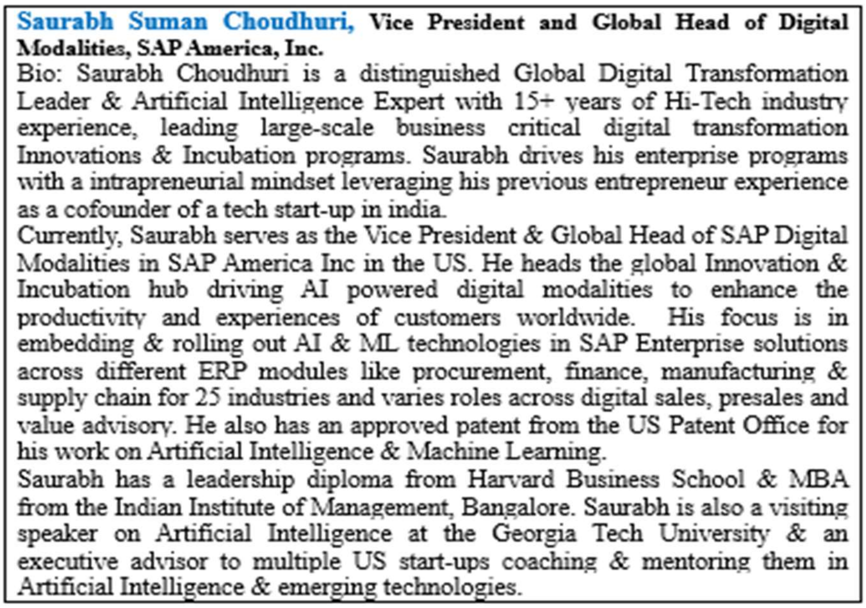
[12] A. D'Amour *et al.*, "Underspecification presents challenges for credibility in modern machine learning," *J. Mach. Learn. Res.*, vol. 23, no. 1, pp. 10237–10297, 2022.

[13] B. Nour, M. Pourzandi, and M. Debbabi, "A survey on threat hunting in enterprise networks," *IEEE Commun. Surv. Tutorials*, 2023.

[14] M. Shahin, M. A. Babar, and L. Zhu, "Continuous integration, delivery and deployment: a systematic review on approaches, tools, challenges and practices," *IEEE access*, vol. 5, pp. 3909–3943, 2017.

***Article Received:** 15 August 2023 **Revised:** 16 September 2023 **Accepted:** 7 October 2023*

Author's Profile:-

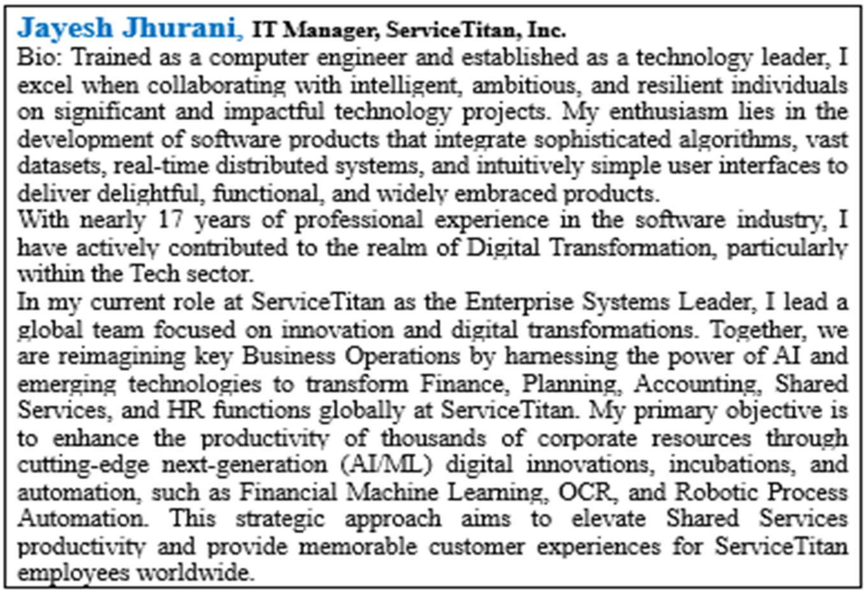


Saurabh Suman Choudhuri, Vice President and Global Head of Digital Modalities, SAP America, Inc.

Bio: Saurabh Choudhuri is a distinguished Global Digital Transformation Leader & Artificial Intelligence Expert with 15+ years of Hi-Tech industry experience, leading large-scale business critical digital transformation Innovations & Incubation programs. Saurabh drives his enterprise programs with a intrapreneurial mindset leveraging his previous entrepreneur experience as a cofounder of a tech start-up in india.

Currently, Saurabh serves as the Vice President & Global Head of SAP Digital Modalities in SAP America Inc in the US. He heads the global Innovation & Incubation hub driving AI powered digital modalities to enhance the productivity and experiences of customers worldwide. His focus is in embedding & rolling out AI & ML technologies in SAP Enterprise solutions across different ERP modules like procurement, finance, manufacturing & supply chain for 25 industries and varies roles across digital sales, presales and value advisory. He also has an approved patent from the US Patent Office for his work on Artificial Intelligence & Machine Learning.

Saurabh has a leadership diploma from Harvard Business School & MBA from the Indian Institute of Management, Bangalore. Saurabh is also a visiting speaker on Artificial Intelligence at the Georgia Tech University & an executive advisor to multiple US start-ups coaching & mentoring them in Artificial Intelligence & emerging technologies.



Jayesh Jhurani, IT Manager, ServiceTitan, Inc.

Bio: Trained as a computer engineer and established as a technology leader, I excel when collaborating with intelligent, ambitious, and resilient individuals on significant and impactful technology projects. My enthusiasm lies in the development of software products that integrate sophisticated algorithms, vast datasets, real-time distributed systems, and intuitively simple user interfaces to deliver delightful, functional, and widely embraced products.

With nearly 17 years of professional experience in the software industry, I have actively contributed to the realm of Digital Transformation, particularly within the Tech sector.

In my current role at ServiceTitan as the Enterprise Systems Leader, I lead a global team focused on innovation and digital transformations. Together, we are reimagining key Business Operations by harnessing the power of AI and emerging technologies to transform Finance, Planning, Accounting, Shared Services, and HR functions globally at ServiceTitan. My primary objective is to enhance the productivity of thousands of corporate resources through cutting-edge next-generation (AI/ML) digital innovations, incubations, and automation, such as Financial Machine Learning, OCR, and Robotic Process Automation. This strategic approach aims to elevate Shared Services productivity and provide memorable customer experiences for ServiceTitan employees worldwide.