Sudip Mittal – Promotion Dossier

Table of Contents

Application Letter

MISSISSIPPI STATE UNIVERSITY

Bagley College of Engineering
Dept. of Computer Science & Engineering

Date: Aug 13, 2024

Dear Dr. Shahram Rahimi,

I am writing to formally apply for promotion to the position of Associate Professor in the Department of Computer Science and Engineering (CSE) at Mississippi State University (MSU). Over the past three years, I have had the privilege of serving as an Assistant Professor in this esteemed department, during which time I have made significant contributions to research, teaching, and service in the field of cybersecurity and AI.

My research has focused on developing the next generation of AI-assisted cyber defense systems to protect various organizations and individuals, with funding from the National Science Foundation, the National Institute of Health, and the United States Department of Defense. My work has been recognized both nationally and internationally through numerous peer-reviewed publications in competitive conferences and journals. While at MSU, I have published at leading peer-reviewed conference venues that include ACM SACMAT, SECRYPT, IEEE/IFIP DSN, IEEE SSCI, IEEE TrustCom, ACM ACSAC, IEEE Big Data, IEEE CIC, IEEE TPS, and journal venues that include IEEE Transactions on Services Computing (TSC), IEEE Access, IEEE Internet Computing, and The Cyber Defense Review.

Teaching is a passion that I have pursued with great dedication. I have developed and taught a range of undergraduate and graduate courses, including CSE 8173 Advanced Cyber operations, CSE 4773/6773 Intro to Cyber Operations, CSE 4293/6293 AI for Cybersecurity, CSE 4990/6990 Special Topics - SCADA Security, CSE 4990/6990 Special Topics - Research Methods in Cybersecurity, where I have consistently received positive feedback from students. Beyond the classroom, I have had the honor of mentoring several graduate students, many of whom have gone on to successful careers in academia and industry. My teaching philosophy focuses on continuously fostering curiosity, which I believe has had a long-term impact on my students' education.

In addition to my research and teaching responsibilities, I have actively contributed to the department and the broader academic community. I have served on several NSF review panels, as well as on program committees of multiple cybersecurity and artificial intelligence conferences. At MSU, I am 1 of 2 faculty points of contact for our National Center of Academic Excellence in Cybersecurity (NCAE-C) managing all 3 NSA CAE designations i.e. CAE in Cyber Defense (CAE-CD), CAE in Cyber Research (CAE-R), and CAE in Cyber Operations (CAE-CO).

Looking ahead, I am excited about the opportunities for further growth and contribution as an Associate Professor. I plan to expand my research, continue to innovate in my teaching, and take on greater leadership roles within the department. I am confident that my experience, achievements, and commitment make me a strong candidate for promotion.

Thank you for considering my application. I am eager to continue my journey at MSU and contribute to the ongoing success and reputation of the Department of Computer Science and Engineering. I look forward to the opportunity to discuss my application with you further.

Sudip Mittal, PhD
Assistant Professor
Dept. of Computer Science & Engineering
Mississippi State University

Promotion and Tenure Form

Revised 07/25/2023

Mississippi State University
Application for Promotion and/or Tenure

Please check response(s) in both columns	
TENURE:	**PROMOTION:**
☐ Mandatory tenure decision	☐ Promotion to Instructor II
	☐ Promotion to Instructor III
☒ Not applicable (early promotion or professional track position or already possess tenure)	☒ Promotion to Associate Professor
	☐ Promotion to Full Professor
	☐ Not applicable (only tenure decision)

Faculty members eligible for consideration for promotion or tenure must provide the department head or appropriate official with all pertinent available information by **October 1**. The department head or other appropriate official has the responsibility to assist the faculty member in preparing for tenure or promotion review.

Materials to be provided in the applicant's dossier include:

*1. Cover letter from the candidate requesting promotion and/or tenure.
*2. Completed University Promotion and Tenure application form (this cover page and attached pages) with appropriate responses and associated documentation. This must include a summary sheet of teaching evaluations.
*3. Complete up-to-date vita.
*4. Copy of the initial offer letter and, if necessary, an additional letter detailing significant changes.
*5. Letters from external reviewers (to be added by the department head). The department head should include a sample letter sent to external reviewers and biographical information about reviewers as appropriate.
6. All materials required by the academic unit's procedural guidelines.
7. All supporting documentation desired by the candidate.

*Only these items will be reviewed routinely above the college level. Items 1-7 must go to dept. head and dept. committee. Deans, college committees, and the Provost require items 1-5 but may also request items 6 and 7. Department heads and deans can use their discretion in sending forward any important information included in items 6 and 7. All department head, dean, and committee recommendations should be included in the package to the Provost.

Note: Please refer to the Faculty Handbook for information pertaining to the Promotion & Tenure process.
To apply and be considered for tenure requires that you be a citizen of the United States or be a permanent resident or have begun the permanent residence process (verification required) in order to be eligible for permanent employment in this country.

Are you a citizen or permanent resident of the United States: ☒ Yes ☐ No

If No, have you applied for permanent residency: ☐ Yes ☐ No (Date process initiated (if Yes): _____)

Name of Applicant: **Sudip Mittal**

Present rank: **Assistant Professor** Date of appointment at current rank: **August 16, 2021**

College/School: **Bagley College of Engineering** Department: **Computer Science and Engineering**

Department Head **Shahram Rahimi**

Preferred Mailing Address (Include City and Zip Code): **103 Eudora Welty Dr. APT O-10, Starkville, MS, 39759**

Initial rank at MSU with date of appointment: **Assistant Professor**

Tenure track date of appointment: **August 16, 2021** Years of transferred service (if applicable): **1**

Advanced Degrees with Dates: **PhD in CS 2019, Mtech in CSE 2014, BTech in CSE 2013**.

Salary Funding (%): E&G: _____ MSU Research Unit: _____ Extension: _____ Other: _____

All other information contained in the attached application is correct to the best of my knowledge.

Date: **August 12, 2024** Signed:

Faculty Member

I. **Current Fall semester responsibilities**:

A. Current instruction

	Course number	Title	Credit hours	Number of students
1. Undergraduate:				
2. Graduate:				

3: Advisees: Undergraduate _**0**__ Master's/Specialist __**3**__ Doctoral __**8**__ Postdoctoral __**0**__

B. Current or on-going research/creative/performance activities

My broad research interests are in the areas of cybersecurity and artificial intelligence. I aim to develop the next generation of cyber defense systems that help protect various organizations and people.The use of AI/ML to solve cybersecurity problems has been gaining more traction within industry and academia. This data-driven automation will enable security systems to identify and respond to cyber threats in real time. I work in the following sub-areas:

1) AI for Malware Detection: I believe that in the near future, current signature-based detection methods will become obsolete. All malwares will be polymorphic. Defensive strategies that are based on behavioral analysis need to be developed. One such defensive strategy could be the development of AI based solutions. Such systems will need robust knowledge representations that include various malware behavioral details as well as mitigating instructions that can be leveraged by an AI based anti-virus system deployed on an enterprise network. I imagine these AI based solutions exchanging malware representations with each other, analogous to the current signature sharing schemes. I wish to work on the development of these malware representations and AI based anti-viruses. I have already done some preliminary work on deep learning systems that help detect malware on the cloud infrastructure. A datacenter infected with malware can cause data loss and/or major disruptions to service for its users. These systems will serve as the basis for future AI assisted malware analysis techniques. Another focus here is to make these Intrusion Detection Systems more Explainable (X-IDS). We have been focusing on creating state of the art X-IDS systems.
2) Cybersecurity Analyst Augmentation Systems: In modern enterprises, security analysts monitor threats in a security operations center (SoC) by watchstanding, akin to a lookout on a ship watching the environs for danger. Screens typically show warnings and alerts from individual products and detectors that the enterprise has installed. Watchstanding permits a highly trained security analyst to look at all the disparate pieces of information, and see if they 'click together' to form some pattern which might indicate an attack. The detection efficacy of a security analyst depends on her operational and strategic knowledge about current security landscape and the associated intelligence. I have created various cyber security informatics systems -- Cyber-All-Intel, CyberTwitter. The goal is to add value and augment an analyst's understanding of threats and vulnerabilities. These systems take as input textual Open Source Intelligence (OSINT), then represents this information using a hybrid knowledge representation scheme that combines knowledge graph and vector space embeddings. These systems also proactively try to improve the underlying cybersecurity knowledge. Other extensions that were driven by security analyst demand, were to understand and represent multi-lingual threats address the issue of supply chain attacks on software development, and more recently combining information from different sources with actual malware behavior data collected from a sandbox environment.
3) Autonomous Intrusion Response: When it comes to the use of AI for cyber defense, there are significant gaps and research opportunities that exist in the field of autonomous Intrusion Response. Current batch of Intrusion Response Systems (IRSs) offer trivial response capabilities, usually based on a static mapping between an identified attack and a pre-defined response. Such methodologies exhibit evident limitations mainly related to scalability and a lack of generalizability. I envision a fully capable IRS that can automatically compute and identify a response to an ongoing attack, exploiting additional knowledge about the attacker behavior and of the underlying system that it protects. This involves developing novel AI techniques to compute the optimal or near-optimal countermeasures that a cyber defense agent can take to stop or mitigate an ongoing attack. The problem is further exacerbated when the protected systems exhibit a non-stationary/dynamic behavior, and therefore needs an IRS with the ability to automatically adapt to these changes while dynamically predicting a near optimal response to an intrusion. We have developed the first open-source licensed software prototype that implements a dynamic IRS, named irs-partition, it uses Deep Q-Networks (DQN), Reinforcement Learning (RL), and transfer learning to cope with the non-stationary/dynamic behavior of complex computer systems.

C. Current service/administrative assignments

1. Public service and off-campus professional service activities (non-assessment activities, such as guest lectures and presentations, external committee/board memberships, business/industry/stakeholder advisement, etc. with dates, organizations, & places):

:

Conference & Workshop Organization

I have regularly organized conferences and workshops in cybersecurity and AI:

- ACM CODASPY 2021 & 2022, Panel Co-Chair, Organizing Committee.
- Workshop on Big Data for Cybersecurity (BigCyber 2018 @ BigData 2018, BigCyber 2019 @ BigData 2019, BigCyber 2020 @ BigData 2020, BigCyber 2021 @ BigData 2021, BigCyber 2022 @ BigData 2022, BigCyber 2023 @ BigData 2023)
- KDD Workshop on Knowledge-infused Learning, Co-located with 29TH ACM SIGKDD 2023.
- ACM Workshop on Secure and Trustworthy Cyber-Physical Systems (SaT-CPS) 2021 (Co-located with ACM CODASPY 2021)
- Workshop on Smart Farming, Precision Agriculture, and Supply Chain (SmartFarm-2020 @ BigData 2020)
- Workshop on Analytics for Security in Cyber Physical System (ASCPS - 2019 @ ICDCN 2019)

2. Professional association service, as offices held, etc.:

Program Committee Member/ Journal Reviewer

I have refereed and evaluated computer science, cybersecurity, and artificial intelligence research conducted by others at the following journals and conference venues:

- IEEE Big Data 2021, 2022, 2023
- IEEE ICMLA 2021, 2022, 2023
- ACM CODASPY 2020, 2021, 2022, 2023, 2024
- ACM ACSAC 2022, 2023
- IEEE SMARTCOMP 2020, 2021
- IEEE ICTAI 2020, 2021, 2022, 2023
- NAACL 2020, 2021, 2022, 2023
- EMNLP 2019, 2020, 2021, 2022
- AAAI 2019, 2020, 2021, 2022, 2023
- ACL - IJCNLP 2019, 2020, 2021, 2022, 2023
- IEEE CLOUD 2018
- ACM ICDCN 2019
- AACL 2019, 2020
- SQUEET 2019
- ICSNC 2019
- PLOS ONE
- IEEE ACCESS
- IEEE Transactions on Services Computing (TSC)
- ACM Transactions on Privacy and Security (TOPS)
- IEEE Transactions on Information Forensics and Security (TIFS)
- ACM Transactions on Internet Technology (TOIT)
- IEEE Transactions on Dependable and Secure Computing (TDSC)
- Elsevier Computers & Security
- Elsevier Sustainable Computing
- Springer Nature
- Nature Machine Intelligence

3. University and departmental committee and administrative accomplishments:

At MSU, I have been heavily involved in maintaining our 3 NSA CAE designations. MSU is 1 of 16 universities that has the prestigious CAE-CO designation and is 1 of 10 universities in USA that have all 3 CAE designations.

I have served on the following departmental committees:

- Cybersecurity curriculum committee (2021- Present)
- NSA POC for CAE CD/CO/R Designations (2021 – Present)
- Hiring Committee (2022, 2023)
- ABET Assessment Committee (2022 – 2023)
- CSE Facilities Committee (2022 – Present)

D. <u>Other</u>

N/A

II. **Activities since last promotion (or initial appointment for tenure):**

A. <u>Teaching</u>

1. Evidence of quality of instruction, both credit and non-credit (check items submitted):

(The faculty member should provide material describing their teaching activities and documentation supporting effectiveness. This material must include a summary statement of student survey responses and may include any of the following or any other items deemed appropriate:

☐ peer evaluations (internal or external),
☒ course syllabi and exams,
☐ non-credit education program plans with assessment,
☐ non-credit education program outcomes and impacts,
☐ student input in the form of letters, emails, faculty nominations, etc.,
☐ recordings of teaching sessions, graduate student theses and dissertations, and other materials demonstrating teaching effectiveness.)

Course	Distance	Course Title	Semester	Evaluation Score
CSE 8673	No	Machine Learning	Fall 2021	4/4
CSE 8673	Yes	Machine Learning	Fall 2021	3.5/4
CSE 4990	No	Special Topics - AI for Cyber Security	Spring 2022	4/4
CSE 6990	No	Special Topics - AI for Cyber Security	Spring 2022	4/4
CSE 6990	Yes	Special Topics - AI for Cyber Security	Spring 2022	4/4
CSE 4990	No	Special Topics - SCADA Security	Spring 2022	3.5/4
CSE 6990	No	Special Topics - SCADA Security	Spring 2022	4/4
CSE 4773	No	Introduction to Cyber Operations	Fall 2022	4/4
CSE 4773	Yes	Introduction to Cyber Operations	Fall 2022	4/4
CSE 6773	Yes	Introduction to Cyber Operations	Fall 2022	3.5/4
CSE 6293	No	AI for Cybersecurity	Spring 2023	4/4
CSE 6293	Yes	AI for Cybersecurity	Spring 2023	3/4
CSE 4990	Yes	Special Topics - Research Methods in Cybersecurity	Fall 2023	4/4
CSE 6990	Yes	Special Topics - Research Methods in Cybersecurity	Fall 2023	3/4
CSE 4773	Yes	Introduction to Cyber Operations	Winter 2023	4/4
CSE 8713	Yes	Advanced Cyber Operations	Spring 2024	4/4

Table 1: Student evaluation scores based on anonymous feedback.

2.

Number of Students Supervised	**Major Professor**	**Minor Professor**
Undergraduate Students	**6**	
Undergraduate Research		
Clinical Interns & Residents		
Master's Students	7	
Specialist Students		
Doctoral Students	10	
Postdoctoral Students		

3. Courses initiated or innovations instituted:

During my time at MSU, I have designed and taught multiple new courses in the computer science sub-fields of artificial intelligence and cybersecurity. I have created a new permanent course: *CSE 4293/6293 AI for Cybersecurity*.

As part of my involvement in NSA CAE activities, I have regularly taken part in national educational and research efforts representing MSU:

1) In Spring 2022, I was the participating MSU faculty, in an NSA CAE effort to create subject matter experts in the field of Supervisory Control and Data Acquisition (SCADA) Security. This involves teaching students critical infrastructure defense techniques. Students learned how to protect dams, water treatment plants, nuclear reactors, etc. from cyber attacks. At MSU this was taught as *CSE 4990/6990 Special Topics - SCADA Security*. Other universities involved included: UAH, Texas A&M, U of South Alabama, UWF.
2) In Fall 2023, I created a research course taught as *CSE 4990/6990 Special Topics - Research Methods in Cybersecurity*. This course was part of the national CAE INSURE+E program. Other universities involved included: Texas A&M, UTD, UCSB, Iowa State, etc. This national research effort provides academic institutions working with government agencies, national labs, and FFRDCs a means to offer opportunities to their students on national cybersecurity research priorities.

4. Non-credit educational programs initiated or instituted (documented, non-credit instruction/teaching with student assessment, such as certification programs, short courses, workshops, in-service trainings, workshops, etc.):

- *As part of the National Security Agency (NSA) Center for Academic Excellence in Cyber Defense (CAE-CD) managed the Bachelor of Computer Science & Engineering + Information Assurance Certificate at MSU.*
  Participated in the validation of the program of study, the assurance that program learning objectives were met, the monitoring of program resources, assessment, course administration, student admissions, advising, and evaluation.

- *As part of the National Security Agency (NSA) Center for Academic Excellence in Cyber Defense (CAE-CO) managed the Masters in Cyber Security and Operations program at MSU.*
  Involved in validation of the program of study, the assurance that program learning objectives were met, the monitoring of program resources, assessment, course administration, and student research/thesis advising.

5. Other (academic advisement may be described here or as service):

Student Advising & Mentorship

Doctoral Students Graduated
- Jesse Ables (MSU, 2023, first job: Assistant Professor, Department of Computer Science, University of South Alabama)

Masters Students Graduated
- Lander, Teddy (MSU)
- Whitman, Joshua (MSU)
- Martin Duclos (MSU)
- William Anderson (MSU)
- Eric Kudjoe Fiah (MSU)
- Rickie James Cashwell (UNCW)
- Shaik Barakhat Aziz (UNCW)
- Deepthi Thalagundamatada (UNCW)

Current Doctoral Students at MSU
- Ivan Fernandez
- Morgan Reece
- Keith Strandell
- Damodar Panigrahi
- Trisha Chakraborty (Committee Co-Chair)
- Shaswata Mitra
- Martin Duclos
- William Anderson

Current Masters Students at MSU

- Wilson Patterson
- Kaneesha Moore

B. Research, creative endeavor, or performances

1. Publications, performances or creative activities:
(For books, indicate date of publication and publisher; for articles, indicate refereed journals; for art shows, indicate judged competition; for musical shows, attach copies of programs; for reports, indicate those done for in-house use.)

Journals

1. Subash Neupane, Shaswata Mitra, Ivan A Fernandez, Swayamjit Saha, Sudip Mittal, Jingdao Chen, Nisha Pillai, Shahram Rahimi. Security Considerations in AI-Robotics: A Survey of Current Methods, Challenges, and Opportunities (IEEE Access 2024).
2. Kevin Gao, Andrew Haverly, Sudip Mittal, Jiming Wu, and Jingdao Chen. AI Ethics: A Bibliometric Analysis, Critical Issues, and Key Gaps. International Journal of Business Analytics (IJBAN) 11, no. 1 (2024): 1-19.
3. Portia Pusey, Maanak Gupta, Sudip Mittal, & Mahmoud Abdelsalam. (2024, February). An Analysis of Prerequisites for Artificial Intelligence/Machine Learning-Assisted Malware Analysis Learning Modules. In Journal of The Colloquium for Information Systems Security Education (Vol. 11, No. 1, pp. 5-5).
4. Aritran Piplai, Anantaa Kotal, Seyedreza Mohseni, Manas Gaur, Sudip Mittal, Anupam Joshi. Knowledge-Enhanced Neurosymbolic Artificial Intelligence for Cybersecurity and Privacy. (IEEE Internet Computing 2023).
5. Keith Strandell, and Sudip Mittal. Risks to Zero Trust in a Federated Mission Partner Environment. (The Cyber Defense Review (CDR), Army Cyber Institute (ACI) at West Point, Summer 2023).
6. Subash Neupane, Jesse Ables, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, Maria Seale. Explainable Intrusion Detection (X-IDS): A Survey of Current Methods, Challenges, and Opportunities. (IEEE Access 2022).
7. Trisha Chakraborty, Shaswata Mitra, Sudip Mittal, Maxwell Young. AI_Adaptive_POW: An AI assisted Proof Of Work (POW) framework for DDoS defense. (Elsevier Software Impacts, 2022).
8. Valeria Cardellini, Emiliano Casalicchio, Stefano Iannucci, Matteo Lucantonio, Sudip Mittal, Damodar Panigrahi, Andrea Silvi. An Intrusion Response System utilizing Deep Q-Networks and System Partitions. (Elsevier SoftwareX, 2022).
9. Ketki Joshi, Karuna Joshi, and Sudip Mittal. Semantically Rich Framework to Automate Cyber Insurance Services. (IEEE Transactions of Service Computing, 2021)
10. Manoj Vanajakumari, Sudip Mittal, Geoff Stoker, Ulku Clark, Kasey Miller. Towards a Leader-Driven Supply Chain Cybersecurity Framework. (Journal of Information Systems Applied Research 2021)
11. Sudip Mittal, Secure V2V and V2I Technologies for the Next-Generation Intelligent Transportation Systems in IEEE Computer, vol. 54, no. 02, pp. 4-6, 2021.
12. Aritran Piplai, Sudip Mittal, Anupam Joshi, Tim Finin, James Holt, and Richard Zak. Creating a Cybersecurity Knowledge Graph from Malware After Action Reports. (IEEE Access 2020)
13. Maanak Gupta, Mahmoud Abdelsalam, Sajad Khorsandroo, and Sudip Mittal. Security and Privacy in Smart Farming: Challenges and Opportunities. (IEEE Access 2020)
14. Sai Sree Laya Chukkapalli, Sudip Mittal, Maanak Gupta, Mahmoud Abdelsalam, Anupam Joshi, Ravi Sandhu, and Karuna Joshi. Ontologies and Artificial Intelligence Systems for the Cooperative Smart Farming Ecosystem. (IEEE Access 2020)

Book Chapters

1. Andrew McDole, Mahmoud Abdelsalam, Maanak Gupta, Sudip Mittal, Mamoun Alazab. Deep Learning Techniques for Behavioural Malware Analysis in Cloud IaaS. (Book Chapter - Springer, Malware Analysis using Artificial Intelligence and Deep Learning (MAAIDL 2020))

Conference Publications

1. Morgan Reece, Teddy Lander Jr, Sudip Mittal, Nidhi Rastogi, Josiah Dykstra, Andy Sampson. Defending Multi-Cloud Applications Against Man-in-the-Middle Attacks (ACM Symposium on Access Control Models and Technologies (ACM SACMAT 2024)).
2. Martin Duclos, Ivan A. Fernandez, Kaneesha Moore, Sudip Mittal, and Edward Zieglar. Utilizing Large Language Models to Translate RFC Protocol Specifications to CPSA Definitions. Hot Topics in the Science of Security (HoTSoS 2024).
3. Manas Gaur, Efthymia Tsamoura, Sarath Sreedharan, Sudip Mittal. KiL 2023: 3rd International Workshop on Knowledge-infused Learning (29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 2023)).

4. Shaswata Mitra, Stephen A Torri, Sudip Mittal. Survey of Malware Analysis through Control Flow Graph using Machine Learning (22nd IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom-2023)).
5. Morgan Reece, Teddy Lander Jr, Sudip Mittal, Nidhi Rastogi, Josiah Dykstra, Andy Sampson. Emergent (In) Security of Multi-Cloud Environments (ACM Annual Computer Security Applications Conference (ACM ACSAC 2023)).
6. Nicholas Cummins, Brad Killen, Somayeh Bakhtiari Ramezani, Shahram Rahimi, Maria Seale, Sudip Mittal. A Comparative Study of Continual, Lifelong, and Online Supervised Learning Libraries. (The International FLAIRS Conference 2023).
7. Trisha Chakraborty, Shaswata Mitra, Sudip Mittal. CAPoW: Context-Aware AI-Assisted Proof of Work based DDoS Defense. (20th International Conference on Security and Cryptography (SECRYPT 2023)).
8. Wilson Patterson, Ivan Fernandez, Subash Neupane, Milan Parmar, Sudip Mittal, and Shahram Rahimi. A White-Box Adversarial Attack Against a Digital Twin. (ACM Annual Computer Security Applications Conference (ACM ACSAC 2022)).
9. Subash Neupane, Ivan A. Fernandez, Wilson Patterson, Sudip Mittal, and Shahram Rahimi. A Temporal Anomaly Detection System for Vehicles utilizing Functional Working Groups and Sensor Channels. (IEEE International Conference on Collaboration and Internet Computing 2022 (IEEE CIC 2022)).
10. Chuyen Nguyen, Caleb Morgan, and Sudip Mittal. CTI4AI: Threat Intelligence Generation and Sharing after Red Teaming AI Models. (ACM Conference on Computer and Communications Security (ACM CCS 2022)).
11. Jesse Ables, Thomas Kirby, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, Maria Seale. Creating an Explainable Intrusion Detection System (X-IDS) using Self Organizing Maps. (IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2022)).
12. Trisha Chakraborty, Shaswata Mitra, Sudip Mittal, Maxwell Young. A Policy Driven AI-Assisted PoW Framework. (52nd IEEE/IFIP International Conference on Dependable Systems and Networks (IEEE/IFIP DSN 2022)).
13. Murat Kantarcioglu, Barbara Carminati, Sagar Samtani, Sudip Mittal, Maanak Gupta. Enforcement of Laws and Privacy Preferences in Modern Computing Systems. (ACM Conference on Data and Application Security and Privacy (ACM CODASPY 2022)).
14. Elisa Bertino, Ravi Sandhu, Bhavani Thuraisingham, Indrakshi Ray, Wenjia Li, Maanak Gupta, and Sudip Mittal. Security and Privacy for Emerging IoT and CPS Domains. (ACM Conference on Data and Application Security and Privacy (ACM CODASPY 2022)).
15. Sai Sree Laya Chukkapalli, Priyanka Ranade, Sudip Mittal, and Anupam Joshi. A Privacy Preserving Anomaly Detection Framework for Cooperative Smart Farming Ecosystem. (IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (IEEE TPS 2021)).
16. Shaswata Mitra, Aritran Piplai, Sudip Mittal, Anupam Joshi. Combating Fake Cyber Threat Intelligence using Provenance in Cybersecurity Knowledge Graphs. (IEEE Big Data 2021).
17. Sai Sree Laya Chukkapalli, Nisha Pillai, Sudip Mittal, Anupam Joshi. Cyber-Physical System Security Surveillance using Knowledge Graph based Digital Twins - A Smart Farming Usecase. (IEEE Intelligence and Security Informatics (IEEE ISI) 2021).
18. Caitlin Moroney, Evan Crothers, Sudip Mittal, Anupam Joshi, Tulay Adali, Christine Mallinson, Nathalie Japkowicz and Zois Boukouvalas. The Case for Latent Variable vs Deep Learning Methods in Misinformation Detection: An Application to COVID-19. (24th International Conference on Discovery Science 2021 (DS 2021))
19. Priyanka Ranade, Aritran Piplai, Sudip Mittal, Anupam Joshi, Tim Finin. Generating Fake Cyber Threat Intelligence Using Transformer-Based Models. (International Joint Conference on Neural Network 2021 (IJCNN 2021))
20. Sai Sree Laya Chukkapalli, Shaik Barakhat Aziz, Nouran Alotaibi, Sudip Mittal, Maanak Gupta, Mahmoud Abdelsalam. An Ontology driven Attribute Based Access Control for Smart Fisheries Ecosystem. (ACM Workshop on Secure and Trustworthy Cyber-Physical Systems (ACM SaT-CPS 2021))
21. Ambareen Siraj, Nigamanth Sridhar, Drew Hamilton, Latifur Khan, Siddharth Kaza, Maanak Gupta, Sudip Mittal. Is there a Security Mindset and Can it be Taught?. (ACM CODASPY 2021).
22. Elisa Bertino, Murat Kantarcioglu, Cuneyt Gurcan Akcora, Sagar Samtani, Sudip Mittal, Maanak Gupta. AI for Security and Security for AI. (ACM CODASPY 2021).
23. Nitu Kedarmal Choudhary, Sai Sree Laya Chukkapalli, Sudip Mittal, Maanak Gupta, Mahmoud Abdelsalam, Anupam Joshi. YieldPredict: A Crop Yield Prediction Framework for Smart Farms. (IEEE Big Data 2020).
24. Manoj Vanajakumari, Sudip Mittal, Geoffery Stoker, Ulku Clark. Enhancing Supply Chain Security through Leader driven Cybersecurity Efforts. (Proceedings of the Conference on Information Systems Applied Research 2020).
25. Aritran Piplai, Priyanka Ranade, Anantaa Kotal, Sudip Mittal, Sandeep Nair Narayanan, Anupam Joshi. Using Knowledge Graphs and Reinforcement Learning for Malware Analysis. (IEEE Big Data 2020).
26. Sina Sontowski, Maanak Gupta, Sai Sree Laya Chukkapalli, Mahmoud Abdelsalam, Sudip Mittal, Anupam Joshi, Ravi Sandhu. Cyber Attacks on Smart Farming Infrastructure. (6th IEEE International Conference on Collaboration and Internet Computing (IEEE CIC 2020)).
27. Aritran Piplai, Sudip Mittal, Mahmoud Abdelsalam, Maanak Gupta, Anupam Joshi, Tim Finin. Fusing Knowledge Representations for Malware Threat Intelligence and Behavioral Data. (IEEE Intelligence and Security Informatics (IEEE ISI) 2020).
28. Matthew Stills, Priyanka Ranade, Sudip Mittal. Cybersecurity Threat Intelligence Augmentation and Embedding Improvement - A Healthcare Usecase (IEEE Intelligence and Security Informatics (IEEE ISI) 2020).

29. Andrew McDole, Mahmoud Abdelsalam, Maanak Gupta, Sudip Mittal. Analyzing Deep Learning Based Behavioural Malware Detection Techniques in Cloud Infrastructure as a Service (International Conference on Cloud Computing (CLOUD) 2020).
30. Sai Sree Laya Chukkapalli, Aritran Piplai, Sudip Mittal, Maanak Gupta, Anupam Joshi. A Smart-Farming Ontology for Attribute Based Access Control. (6th IEEE International Conference on Big Data Security on Cloud (IEEE BigDataSecurity 2020)).
31. Nitika Khurana, Sudip Mittal, Aritran Piplai and Anupam Joshi. Preventing Poisoning Attacks on Artificial Intelligence based Threat Intelligence Systems. (2019) (IEEE Machine Learning For Signal Processing, (IEEE MLSP) 2019).
32. Ketki Joshi, Karuna Joshi, Sudip Mittal. A Semantic Approach for Automating Knowledge in Policies of Cyber Insurance Services. (14th IEEE International Conference on Web Services (IEEE ICWS), 2019).
33. Aditya Pingle, Aritran Piplai, Sudip Mittal, Anupam Joshi, James Holt, and Richard Zak. RelExt: Relation Extraction using Deep Learning approaches for Cybersecurity Knowledge Graph Improvement. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (IEEE/ACM ASONAM 19), August 27--30, 2019, Vancouver, BC, Canada.
34. Sowmya Ramapatruni, Sandeep Nair, Sudip Mittal, Anupam Joshi, Karuna Joshi. Anomaly Detection Models for Smart Home Security. (IEEE BigDataSecurity 2019).
35. Priyanka Ranade, Sudip Mittal, Anupam Joshi, and Karuna Joshi. Using Deep Neural Networks to Translate Multi-lingual Threat Intelligence. (IEEE Intelligence and Security Informatics (IEEE ISI) 2018).
36. Lorenzo Niel, Sudip Mittal, Anupam Joshi. Mining Threat Intelligence about Open-Source Projects and Libraries from Repository Issues and Bug Reports. (IEEE Intelligence and Security Informatics (IEEE ISI) 2018).
37. Priyanka Ranade, Sudip Mittal, Anupam Joshi, and Karuna Joshi Understanding Multi-lingual Threat Intelligence for AI based Cyber-defense Systems (IEEE International Conference on Technologies for Homeland Security (IEEE HST), 2018)
38. Vishal Rathod, Sandeep Nair Narayanan, Sudip Mittal, and Anupam Joshi. Semantically Rich, Context Aware Access Control for Openstack. (2018 IEEE 4th International Conference on Collaboration and Internet Computing (IEEE CIC 2018)).
39. Maithilee Joshi, Sudip Mittal, Karuna Pande Joshi, and Tim Finin.Semantically Rich Oblivious Access Control for Cloud Storage (IEEE International Conference on Edge Computing (IEEE EDGE) 2017.
40. Sudip Mittal, Aditi Gupta, Karuna Pande Joshi, Claudia Pearce, and Anupam Joshi.A Question Answering System for Management of Cloud Service Level Agreements (IEEE International Conference on Cloud Computing (IEEE CLOUD) 2017)
41. Agniva Banerjee, Raka Dalal, Sudip Mittal, Karuna Pande Joshi. Generating Digital Twin models using Knowledge Graphs for Industrial Production Lines (9th International ACM Web Science Conference, Industrial Knowledge Graphs (ACM WEBSCI) (2017))
42. Karuna P Joshi, Aditi Gupta, Sudip Mittal, Anupam Joshi, Tim Finin and Claudia Pearce. Semantic Approach to Automating Management of Data Privacy Policies for Cloud Consumers (IEEE Big Data 2016)
43. Karuna P. Joshi, Aditi Gupta, Sudip Mittal, Claudia Pearce, Anupam Joshi and Tim Finin. ALDA : Cognitive Assistant for Legal Document Analytics (AAAI Fall Symposium (2016))
44. Sudip Mittal, Aditi Gupta, Karuna Pande Joshi, Claudia Pearce, and Anupam Joshi. A Semantic Framework for Automated Analysis of Cloud Service Level Agreements (2016).
45. Sandeep Nair, Sudip Mittal, Anupam Joshi. Using Semantic Technologies to Mine Vehicular Context for Security. (37th IEEE Sarnoff Symposium (2016))
46. Sudip Mittal, Prajit Kumar Das, Varish Mulwad, Anupam Joshi, and Tim Finin. Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (IEEE/ACM ASONAM), pp. 860-867. IEEE Press, 2016.
47. Sandeep Nair Narayanan, Sudip Mittal, and Anupam Joshi. OBD SecureAlert: An Anomaly Detection System for Vehicles. (IEEE SmartSYS 2016).
48. Aditi Gupta, Sudip Mittal, Karuna Pande Joshi, Claudia Pearce, and Anupam Joshi. Streamlining Management of Multiple Cloud Services. (IEEE International Conference on Cloud Computing (IEEE CLOUD) 2016).
49. Sandeep Nair Narayanan, Sudip Mittal, and Anupam Joshi. Using Data Analytics to Detect Anomalous States in Vehicles. (2016).
50. Sudip Mittal, Karuna Pande Joshi, Claudia Pearce, and Anupam Joshi. Automatic Extraction of Metrics from SLAs for Cloud Service Management. In IEEE International Conference on Cloud Engineering (IEEE IC2E), 2016.
51. Sudip Mittal, Karuna Pande Joshi, Claudia Pearce, and Anupam Joshi. Parallelizing Natural Language Techniques for Knowledge Extraction from Cloud Service Level Agreements. In 2015 IEEE International Conference on Big Data, (IEEE Big Data 2015).
52. Sudip Mittal, Neha Gupta, Prateek Dewan, and Ponnurangam Kumaraguru. Pinned it! A Large Scale Study of the Pinterest Network. In Proceedings of the IKDD Conference on Data Sciences, pp. 1-10. ACM, 2014.

2. Professional papers read; indicate whether invited, refereed, or volunteered.
Cite organization, date, and title:

1. Defending Multi-Cloud Applications Against Man-in-the-Middle Attacks, ACM Symposium on Access Control Models and Technologies (ACM SACMAT 2024), May 2024, Refereed.
2. Utilizing Large Language Models to Translate RFC Protocol Specifications to CPSA Definitions. Hot Topics in the Science of Security (HoTSoS 2024), March 2024, Refereed.
3. Emergent (In) Security of Multi-Cloud Environments, National Security Agency, CAE-R Research Symposium, October 2023, Invited.
4. Combating Fake Cyber Threat Intelligence using Provenance in Cybersecurity Knowledge Graphs. (IEEE Big Data 2021), December 2021, Refereed.

3. Grants for research or study:

Proposals submitted since last promotion and total dollar amount: $ 13 million

1. National Science Foundation (NSF), Collaborative Research: RET Site: Collaborative Research in Cybersecurity Intellligence (CRSI), Andy Perkins, Sudip Mittal, $403,732.00, Not funded.

2. National Science Foundation (NSF), NRI: Crop Anomaly Detection and management with Continually Learning Robot Swarms, Jingdao Chen, Cindy Bethel, Sudip Mittal, $1,199,997.00, Not funded.

3. National Science Foundation (NSF), Collaborative Research: SaTC: CORE:Medium:Fully Multivariate and Interpretable Data Fusion for Fair Misinformation Detection During High Impact Events, Sudip Mittal, $174,405.00, Not funded.

4. National Science Foundation (NSF), Theme 1: Bringing Machine Learning, Knowledge Representation & Reasoning, and Agent-based Systems to Cybersecurity, Sudip Mittal, $624,132.00, Not funded.

5. United States Department of Transportation (DOT), Southeast Transportation Regional Institute for Equity and Advanced Mobility (STREAM), Jingdao Chen, Sudip Mittal, $200,000.00, Not funded.

6. National Science Foundation (NSF), Understanding the social, cognitive, and learning aspects of using digital co-creation and collaborative education platform: Teaching and Learning in the Metaverse (TLM), $841,178.00, Not funded.

7. United States Department of Defense (DOD), VICEORY Capacity Building, George Trawick, Sudip Mittal, Shelly Hollis, $1,990,672.00, Not funded.

8. National Science Foundation (NSF), SaTC: EDU: Project Nimbus- Securing Multi-Cloud Machine Learning Pipelines Against Cyberattack, Sudip Mittal, $164,795.00, Pending.

Proposals funded (cite source, title of project, role [PI, etc.], $ amount, dates): 11 million

1. National Institutes of Health (NIH) - STTR: Phase 1: MyDocSaid: AI assisted Patient Journey utilizing Knowledge Graphs and Large Language Models. Sudip Mittal (PI), ($103,500, July 2024 - July 2025)

2. United States Department of Defense (DOD) - VICEROY for the NCAE-C Southeast Region. - Virtual Institute for Cyber and Electromagnetic Spectrum Research and Employ (VICEROY for the NCAE-C Southeast Region). George Trawick (PI), Sudip Mittal (co-PI), Shelly Hollis (co-PI), ($550,000 - Year 4, July 2024 - July 2025)

3. U.S. Department of Defense (DOD), DoD Cybersecurity Academy Program. George Trawick (PI), Sudip Mittal (co-PI) ($186,000, Aug 2024 - July 2025).

4. National Science Foundation - CyberCorps: Scholarship for Service at Mississippi State University (Renewal). Andy Perkins (PI), Sudip Mittal (Co-PI), Reed Mosher (Co-PI), ($4.2 Million, Aug 2023 - July 2028).

5. National Science Foundation - SaTC: EDU: Inculcate a culture of preparedness against AI security threats to pervasive robotic systems. Sudip Mittal (PI), Jingdao Chen (co-PI), ($400,000, Aug 2023 - July 2026).

6. “RASPET, UAV Security” Sponsored by Department of Homeland Security, Federal. Awarded, Grant. Trawick, G., Torri, S. A. (Co-Principal), Mittal, S. (Co-Principal), (Approx $100,000 August, 2023 - July 2024).

7. U.S. Department of Defense (DOD), DoD CySP Program. George Trawick (PI), Sudip Mittal (co-PI) ($76,540, Aug 2023 - July 2024).

8. U.S. Department of Defense (DOD), US Army Corps of Engineers (USACE); En- gineer Research and Development Center (ERDC) - Adoptable Predictive Maintenance Using Transformer Neural Networks (TNN). Shahram Rahimi (PI), Sudip Mittal (co-PI) (approx $500,000, Start Date - Jan 2023 - Jan 2025).

9. U.S. Department of Defense (DOD), US Army Corps of Engineers (USACE); En- gineer Research and Development Center (ERDC) - Explainable Predictive Maintenance (XPM): An Explainable AI (XAI) Technology for Predictive Maintenance. Shahram Rahimi (PI), Sudip Mittal (co-PI) (approx $500,000, Start Date - Jan 2023 - Jan 2025).

10. National Science Foundation - CyberCorps: Scholarship for Service - A Continuation Program at Mississippi State University. PI - Andy Perkins, Co-PI - Sudip Mittal, Reed Mosher. ($3,112,393.00, end date - Aug 2025).

11. U.S. Department of Defense (DOD), US Army Corps of Engineers (USACE); Engineer Research and Development Center (ERDC) - Enhanced Network Cybersecurity Research - A Self Organizing Maps and Danger Theory Hybrid for Intrusion Detection, Reed Mosher (PI) Shahram Rahimi (co-PI), Sudip Mittal (co-PI). (approx $1,000,000, March 2022 - Feb 2025).

12. NSF, NCyTE - NCAE-C Southeast - Airforce JROTC Summer Program , Grant. Trawick, G., Torri, S. A. (Co-Principal), Mittal, S. (Co-Principal), ($116,500, April 2022 - Aug 2022)

13. National Science Foundation - Collaborative Research: SaTC: EDU: Artificial Intelligence Assisted Malware Analysis (#2133190, $105,790, Aug 2020 - July 2023)

14. National Science Foundation - REU Supplement to Award 2133190. “Explainable Malware Intrusion Detection based on Artificial Immune Systems”, ($16,000, May 2022 - Dec 2022).

4. Other:

N/A

C. <u>Service</u>

1. Public service, non-assessment activities such as guest lectures and presentations, external committee/board memberships, business/industry/stakeholder advisement, etc. (with dates, organizations, places):

NSF Panel Service
I served on 3 NSF review panels for the Secure and Trustworthy Cyberspace (SaTC) Program. 2 times for the SaTC CORE subprogram (2021. 2022) and 1 time for the SaTC EDU subprogram (2023).

2. Professional association service (offices held, journals edited, etc.):

IEEE Mississippi Executive Committee, Professional Activities Chair (2022)

3. University service (committees, administrative accomplishments, etc.):

I have served on the following departmental committees:

- Cybersecurity curriculum committee (2021- Present)
- NSA POC for CAE CD/CO/R Designations (2021 – Present)
- Hiring Committee (2022, 2023)
- ABET Assessment Committee (2022 – 2023)
- CSE Facilities Committee (2022 – Present)

4. Other (academic advisement may be described here or as teaching):

High School Student Research Mentorship
In my time here at MSU I have been part of the MSU and the Mississippi School for Mathematics and Science (MSMS) Research Experience Program, where 2 MSMS students spent a semester in my lab. I hosted the following 2 MSMS students:

- Sephora Poteau (Spring 2023)
- Samar Rosas (Spring 2024)

III. Awards and distinctions (title, date, organization):

N/A

IV. Memberships in learned and professional societies. Society, dates of membership, and offices held:

- Institute of Electrical and Electronics Engineers, IEEE. (2015 - Present)
- Association for Computing Machinery, ACM, International. (2016 - Present)

V. Previous academic ranks, institutions, and dates:

- Assistant Professor, Department of Computer Science, University of North Carolina Wilmington, August 2019-August 2021

VI. Non-academic positions held prior to appointment at MSU:

N/A

VII. Summary listing of all required and supporting documentation (items 6 and 7 on the cover of the application form). This listing should be less than one page in length.

Table of Contents

Application Letter

Promotion and Tenure Form

Initial Offer Letter

Letter regarding significant changes in duties: Appointment as Associate Director of Predictive Analytics and Technology Integration (PATENT) Laboratory

Curriculum Vitae

Teaching Accomplishments

Research Accomplishments

Service Accomplishments

Selected Publications

Department Head's Recommendation for Promotion or Tenure
(Cite the following information and sign.)

1. Name of candidate:_____ Present rank:_____

2. Recommended for promotion to the rank of:_____
   (Or not recommended):_____

3. Recommended for tenure: Yes/No/NA

   Assessment and evaluation by department head: strong points that warrant promotion should be listed, with documentation wherever possible; stress such items as teaching and advising of students, research accomplishments, and university and community service. Please avoid platitudes or general, subjective opinions. It would be useful, too, to comment upon the quality of personal relationships of the candidates with peers, superiors, and any who may report to them, as well as upon their professional performance. Finally, consider the candidate in relation to what you picture as the ideal candidate for this recommended position rather than in relation to other members of your department. In situations where "demonstrated excellence" is required, please provide various supporting evidence such as peer evaluations, reviews of publications, letters of commendation, student survey responses, or any other relevant measures of excellence. Attach relevant departmental committee recommendations.

Date_____ Signed:_____
Department Head

****

Dean's Recommendation for Promotion or Tenure
(Cite the following information and sign.)

1. Name of candidate:_____ Present rank:_____

2. Recommended for promotion to the rank of:_____
   (Or not recommended):_____

3. Recommended for tenure: Yes/No/NA

   Recommendation: Use materials provided by the candidate and department head, as appropriate, but please indicate your evaluation of the candidate's performance to date and prospects for the future. Avoid general, subjective opinions, stress obvious strong points, and indicate where further development may be expected. Attach relevant college/school committee recommendations.

Date_____ Signed:_____
Dean

Curriculum Vitae

SUDIP MITTAL

mittal@cse.msstate.edu ⋄ www.sudipmittal.com
Assistant Professor, Dept. of Computer Science & Engineering, Mississippi State University
665 George Perry Street, 300 Butler Hall, Box 9637, Mississippi State, MS 39762

RESEARCH INTERESTS

Cybersecurity, Artificial Intelligence, Cyber-Physical Systems

CURRENT APPOINTMENTS

Assistant Professor (tenure-track) *August 2021 - Present*
Department of Computer Science & Engineering
Bagley College of Engineering, Mississippi State University

Associate Director *January 2023 - Present*
Predictive Analytics and TEchnology iNTegration (PATENT) Laboratory
Mississippi State University

EDUCATION

Ph.D. in Computer Science, University of Maryland Baltimore County *2014-2019*

M.Tech. in Computer Science & Engineering, IIIT Delhi *2013-2014*

B.Tech. in Computer Science & Engineering, IIIT Delhi *2009-2013*

EXTERNAL FUNDING (CHRONOLOGICAL, LATEST FIRST)

1. **National Institutes of Health** (NIH) - STTR: Phase 1: MyDocSaid: AI assisted Patient Journey utilizing Knowledge Graphs and Large Language Models. Sudip Mittal (PI), ($103,500, July 2024 - July 2025)
2. **Griffiss Institute - VICEROY for the NCAE-C Southeast Region.** - Virtual Institute for Cyber and Electromagnetic Spectrum Research and Employ (VICEROY for the NCAE-C Southeast Region). George Trawick (PI), Sudip Mittal (co-PI), Shelly Hollis (co-PI), ($550,000 - Year 4, July 2024 - July 2025)
3. **U.S. Department of Defense (DOD)**, DoD Cybersecurity Academy Program. George Trawick (PI), Sudip Mittal (co-PI) ($186,000, Aug 2024 - July 2025).
4. **National Science Foundation** - CyberCorps: Scholarship for Service at Mississippi State University (Renewal). Andy Perkins (PI), Sudip Mittal (Co-PI), Reed Mosher (Co-PI), ($4.2 Million, Aug 2023 - July 2028).
5. **National Science Foundation** - SaTC: EDU: Inculcate a culture of preparedness against AI security threats to pervasive robotic systems. Sudip Mittal (PI), Jingdao Chen (co-PI), ($400,000, Aug 2023 - July 2026).
6. "**RASPET, UAV Security**" Sponsored by Department of Homeland Security, Federal. Awarded, Grant. Trawick, G., Torri, S. A. (Co-Principal), Mittal, S. (Co-Principal), (Approx $100,000 August, 2023 - July 2024).
7. **U.S. Department of Defense (DOD)**, **DoD CySP Program**. George Trawick (PI), Sudip Mittal (co-PI) ($76,540, Aug 2023 - July 2024).

8. **U.S. Department of Defense (DOD), US Army Corps of Engineers (USACE); Engineer Research and Development Center (ERDC)** - Adoptable Predictive Maintenance Using Transformer Neural Networks (TNN). Shahram Rahimi (PI), Sudip Mittal (co-PI) (approx $500,000, Start Date - Jan 2023 - Jan 2025).

9. **U.S. Department of Defense (DOD), US Army Corps of Engineers (USACE); Engineer Research and Development Center (ERDC)** - Explainable Predictive Maintenance (XPM): An Explainable AI (XAI) Technology for Predictive Maintenance. Shahram Rahimi (PI), Sudip Mittal (co-PI) (approx $500,000, Start Date - Jan 2023 - Jan 2025).

10. **National Science Foundation** - CyberCorps: Scholarship for Service - A Continuation Program at Mississippi State University. PI - Andy Perkins, Co-PI - Sudip Mittal, Reed Mosher. ($3,112,393.00, end date - Aug 2024, link).

11. **U.S. Department of Defense (DOD), US Army Corps of Engineers (USACE); Engineer Research and Development Center (ERDC)** - Enhanced Network Cybersecurity Research - A Self Organizing Maps and Danger Theory Hybrid for Intrusion Detection, Reed Mosher (PI) Shahram Rahimi (co-PI), Sudip Mittal (co-PI). (approx $1,000,000, March 2022 - Feb 2025).

12. **NSF, NCyTE** - NCAE-C Southeast - Airforce JROTC Summer Program ($116,500, April 2022 - Aug 2022)

13. **National Science Foundation** - Collaborative Research: SaTC: EDU: Artificial Intelligence Assisted Malware Analysis (#2133190, $105,790, Aug 2020 - July 2023, UNCW link, MSState link).

14. **National Science Foundation** - REU Supplement to Award 2133190. "Explainable Malware Intrusion Detection based on Artificial Immune Systems", ($16,000, May 2022 - Dec 2022).

15. **Defense Intelligence Agency**, Oak Ridge National Laboratory (ORNL, Department of Energy) - UNCW, Development of Advanced Solutions for Triage and Visualization of Large Volumes of Images and Videos. UNCW PI - Karl Ricanek, Sudip Mittal, Toni Pence (UNCW portion, $451,301, Dec 2019 - Dec 2021)

16. **U.S. Department of Defense (DOD)**, Cyber-Seahawk (C-Hawk) **DoD CySP Program**. UNCW PI - Ulku Clark, Co-PI - Minoo Modaresnezhad, William Wetherill, Greg Vandergriff, Sudip Mittal, Ronald Vetter, Laurie Patterson, Geoffrey Stoker, Manoj Vanajakumari ($196,782, 09/15/2020 - 12/14/2021)

PUBLICATIONS

h-index = 27, i10-index = 47, updated list on Google Scholar

Journals

1. Cummins, Logan, Alex Sommers, Somayeh Bakhtiari Ramezani, Sudip Mittal, Joseph Jabour, Maria Seale, and Shahram Rahimi. "Explainable Predictive Maintenance: A Survey of Current Methods, Challenges and Opportunities." (IEEE Access 2024).

2. Gao, Di Kevin, Andrew Haverly, Sudip Mittal, Jiming Wu, and Jingdao Chen. "AI Ethics: A Bibliometric Analysis, Critical Issues, and Key Gaps." International Journal of Business Analytics (IJBAN) 11, no. 1 (2024): 1-19.

3. Subash Neupane, Shaswata Mitra, Ivan A Fernandez, Swayamjit Saha, Sudip Mittal, Jingdao Chen, Nisha Pillai, Shahram Rahimi. "Security Considerations in AI-Robotics: A Survey of Current Methods, Challenges, and Opportunities" (IEEE Access 2024).

4. Pusey, P., Gupta, M., Mittal, S., & Abdelsalam, M. (2024, February). An Analysis of Prerequisites for Artificial Intelligence/Machine Learning-Assisted Malware Analysis Learning Modules. In Journal of The Colloquium for Information Systems Security Education (Vol. 11, No. 1, pp. 5-5).
5. Aritran Piplai, Anantaa Kotal, Seyedreza Mohseni, Manas Gaur, Sudip Mittal, Anupam Joshi. "Knowledge-Enhanced Neurosymbolic Artificial Intelligence for Cybersecurity and Privacy". (IEEE Internet Computing 2023).
6. Keith Strandell, and Sudip Mittal. "Risks to Zero Trust in a Federated Mission Partner Environment." (The Cyber Defense Review (CDR), Army Cyber Institute (ACI) at West Point, Summer 2023).
7. Subash Neupane, Jesse Ables, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, Maria Seale. "Explainable Intrusion Detection (X-IDS): A Survey of Current Methods, Challenges, and Opportunities". (IEEE Access 2022).
8. Trisha Chakraborty, Shaswata Mitra, Sudip Mittal, Maxwell Young. "AI_Adaptive_POW: An AI assisted Proof Of Work (POW) framework for DDoS defense". (Elsevier Software Impacts, 2022).
9. Valeria Cardellini, Emiliano Casalicchio, Stefano Iannucci, Matteo Lucantonio, Sudip Mittal, Damodar Panigrahi, Andrea Silvi. "An Intrusion Response System utilizing Deep Q-Networks and System Partitions". (Elsevier SoftwareX, 2022).
10. Ketki Joshi, Karuna Joshi, and Sudip Mittal. "Semantically Rich Framework to Automate Cyber Insurance Services". (IEEE Transactions of Service Computing, 2021)
11. Manoj Vanajakumari, Sudip Mittal, Geoff Stoker, Ulku Clark, Kasey Miller. "Towards a Leader-Driven Supply Chain Cybersecurity Framework". (Journal of Information Systems Applied Research 2021)
12. Sudip Mittal, "Secure V2V and V2I Technologies for the Next-Generation Intelligent Transportation Systems" in IEEE Computer, vol. 54, no. 02, pp. 4-6, 2021.
13. Aritran Piplai, Sudip Mittal, Anupam Joshi, Tim Finin, James Holt, and Richard Zak. "Creating a Cybersecurity Knowledge Graph from Malware After Action Reports". (IEEE Access 2020)
14. Maanak Gupta, Mahmoud Abdelsalam, Sajad Khorsandroo, and Sudip Mittal. "Security and Privacy in Smart Farming: Challenges and Opportunities". (IEEE Access 2020)
15. Sai Sree Laya Chukkapalli, Sudip Mittal, Maanak Gupta, Mahmoud Abdelsalam, Anupam Joshi, Ravi Sandhu, and Karuna Joshi. "Ontologies and Artificial Intelligence Systems for the Cooperative Smart Farming Ecosystem". (IEEE Access 2020)

Book chapters

16. Andrew McDole, Mahmoud Abdelsalam, Maanak Gupta, Sudip Mittal, Mamoun Alazab. "Deep Learning Techniques for Behavioural Malware Analysis in Cloud IaaS". (Book Chapter - Springer, Malware Analysis using Artificial Intelligence and Deep Learning (MAAIDL 2020))

Conferences

17. Morgan Reece, Teddy Lander Jr, Sudip Mittal, Nidhi Rastogi, Josiah Dykstra, Andy Sampson. "Defending Multi-Cloud Applications Against Man-in-the-Middle Attacks" (ACM Symposium on Access Control Models and Technologies (ACM SACMAT 2024)).

18. Duclos, Martin, Ivan A. Fernandez, Kaneesha Moore, Sudip Mittal, and Edward Zieglar. "Utilizing Large Language Models to Translate RFC Protocol Specifications to CPSA Definitions." Hot Topics in the Science of Security (HoTSoS 2024).

19. Gao, Di Kevin, Andrew Haverly, Sudip Mittal, and Jingdao Chen. "A Bibliometric View of AI Ethics Development." In 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pp. 1-5. IEEE, 2023.

20. Manas Gaur, Efthymia Tsamoura, Sarath Sreedharan, Sudip Mittal. "KiL 2023: 3rd International Workshop on Knowledge-infused Learning" (29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 2023)).

21. Shaswata Mitra, Stephen A Torri, Sudip Mittal. "Survey of Malware Analysis through Control Flow Graph using Machine Learning" (22nd IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom-2023)).

22. Morgan Reece, Teddy Lander Jr, Sudip Mittal, Nidhi Rastogi, Josiah Dykstra, Andy Sampson. "Emergent (In) Security of Multi-Cloud Environments" (ACM Annual Computer Security Applications Conference (ACM ACSAC 2023)).

23. Nicholas Cummins, Brad Killen, Somayeh Bakhtiari Ramezani, Shahram Rahimi, Maria Seale, Sudip Mittal. "A Comparative Study of Continual, Lifelong, and Online Supervised Learning Libraries". (The International FLAIRS Conference 2023).

24. Trisha Chakraborty, Shaswata Mitra, Sudip Mittal. "CAPoW: Context-Aware AI-Assisted Proof of Work based DDoS Defense". (20th International Conference on Security and Cryptography (SECRYPT 2023)).

25. Wilson Patterson, Ivan Fernandez, Subash Neupane, Milan Parmar, Sudip Mittal, and Shahram Rahimi. "A White-Box Adversarial Attack Against a Digital Twin." (ACM Annual Computer Security Applications Conference (ACM ACSAC 2022)).

26. Neupane, Subash, Ivan A. Fernandez, Wilson Patterson, Sudip Mittal, and Shahram Rahimi. "A Temporal Anomaly Detection System for Vehicles utilizing Functional Working Groups and Sensor Channels." (IEEE International Conference on Collaboration and Internet Computing 2022 (IEEE CIC 2022)).

27. Nguyen, Chuyen, Caleb Morgan, and Sudip Mittal. "CTI4AI: Threat Intelligence Generation and Sharing after Red Teaming AI Models." (ACM Conference on Computer and Communications Security (ACM CCS 2022)).

28. Jesse Ables, Thomas Kirby, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, Maria Seale. "Creating an Explainable Intrusion Detection System (X-IDS) using Self Organizing Maps". (IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2022)).

29. Trisha Chakraborty, Shaswata Mitra, Sudip Mittal, Maxwell Young. "A Policy Driven AI-Assisted PoW Framework". (52nd IEEE/IFIP International Conference on Dependable Systems and Networks (IEEE/IFIP DSN 2022)).

30. Murat Kantarcioglu, Barbara Carminati, Sagar Samtani, Sudip Mittal, Maanak Gupta. "Enforcement of Laws and Privacy Preferences in Modern Computing Systems". (ACM Conference on Data and Application Security and Privacy (ACM CODASPY 2022)).

31. Elisa Bertino, Ravi Sandhu, Bhavani Thuraisingham, Indrakshi Ray, Wenjia Li, Maanak Gupta, and Sudip Mittal. "Security and Privacy for Emerging IoT and CPS Domains". (ACM Conference on Data and Application Security and Privacy (ACM CODASPY 2022)).

32. Sai Sree Laya Chukkapalli, Priyanka Ranade, Sudip Mittal, and Anupam Joshi. "A Privacy Preserving Anomaly Detection Framework for Cooperative Smart Farming Ecosystem". (IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (IEEE TPS 2021)) .

33. Shaswata Mitra, Aritran Piplai, Sudip Mittal, Anupam Joshi. "Combating Fake Cyber Threat Intelligence using Provenance in Cybersecurity Knowledge Graphs". (IEEE Big Data 2021).

34. Sai Sree Laya Chukkapalli, Nisha Pillai, Sudip Mittal, Anupam Joshi. "Cyber-Physical System Security Surveillance using Knowledge Graph based Digital Twins - A Smart Farming Usecase". (IEEE Intelligence and Security Informatics (IEEE ISI) 2021).

35. Caitlin Moroney, Evan Crothers, Sudip Mittal, Anupam Joshi, Tulay Adali, Christine Mallinson, Nathalie Japkowicz and Zois Boukouvalas. "The Case for Latent Variable vs Deep Learning Methods in Misinformation Detection: An Application to COVID-19". (24th International Conference on Discovery Science 2021 (DS 2021))

36. Priyanka Ranade, Aritran Piplai, Sudip Mittal, Anupam Joshi, Tim Finin. "Generating Fake Cyber Threat Intelligence Using Transformer-Based Models". (International Joint Conference on Neural Network 2021 (IJCNN 2021))

37. Sai Sree Laya Chukkapalli, Shaik Barakhat Aziz, Nouran Alotaibi, Sudip Mittal, Maanak Gupta, Mahmoud Abdelsalam. "An Ontology driven Attribute Based Access Control for Smart Fisheries Ecosystem". (ACM Workshop on Secure and Trustworthy Cyber-Physical Systems (ACM SaT-CPS 2021))

38. Ambareen Siraj, Nigamanth Sridhar, Drew Hamilton, Latifur Khan, Siddharth Kaza, Maanak Gupta, Sudip Mittal. "Is there a Security Mindset and Can it be Taught?". (ACM CODASPY 2021).

39. Elisa Bertino, Murat Kantarcioglu, Cuneyt Gurcan Akcora, Sagar Samtani, Sudip Mittal, Maanak Gupta. "AI for Security and Security for AI". (ACM CODASPY 2021).

40. Nitu Kedarmal Choudhary, Sai Sree Laya Chukkapalli, Sudip Mittal, Maanak Gupta, Mahmoud Abdelsalam, Anupam Joshi. "YieldPredict: A Crop Yield Prediction Framework for Smart Farms". (IEEE Big Data 2020).

41. Manoj Vanajakumari, Sudip Mittal, Geoffery Stoker, Ulku Clark. "Enhancing Supply Chain Security through Leader driven Cybersecurity Efforts". (Proceedings of the Conference on Information Systems Applied Research 2020).

42. Aritran Piplai, Priyanka Ranade, Anantaa Kotal, Sudip Mittal, Sandeep Nair Narayanan, Anupam Joshi. "Using Knowledge Graphs and Reinforcement Learning for Malware Analysis". (IEEE Big Data 2020).

43. Sina Sontowski, Maanak Gupta, Sai Sree Laya Chukkapalli, Mahmoud Abdelsalam, Sudip Mittal, Anupam Joshi, Ravi Sandhu. "Cyber Attacks on Smart Farming Infrastructure". (6th IEEE International Conference on Collaboration and Internet Computing (IEEE CIC 2020)).

44. Aritran Piplai, Sudip Mittal, Mahmoud Abdelsalam, Maanak Gupta, Anupam Joshi, Tim Finin. "Fusing Knowledge Representations for Malware Threat Intelligence and Behavioral Data". (IEEE Intelligence and Security Informatics (IEEE ISI) 2020).

45. Matthew Stills, Priyanka Ranade, Sudip Mittal. "Cybersecurity Threat Intelligence Augmentation and Embedding Improvement - A Healthcare Usecase" (IEEE Intelligence and

Security Informatics (IEEE ISI) 2020).

46. Andrew McDole, Mahmoud Abdelsalam, Maanak Gupta, Sudip Mittal. "Analyzing Deep Learning Based Behavioural Malware Detection Techniques in Cloud Infrastructure as a Service" (International Conference on Cloud Computing (CLOUD) 2020).

47. Sai Sree Laya Chukkapalli, Aritran Piplai, Sudip Mittal, Maanak Gupta, Anupam Joshi. "A Smart-Farming Ontology for Attribute Based Access Control". (6th IEEE International Conference on Big Data Security on Cloud (IEEE BigDataSecurity 2020)).

48. Nitika Khurana, Sudip Mittal, Aritran Piplai and Anupam Joshi. "Preventing Poisoning Attacks on Artificial Intelligence based Threat Intelligence Systems". (2019) (IEEE Machine Learning For Signal Processing, (IEEE MLSP) 2019).

49. Ketki Joshi, Karuna Joshi, Sudip Mittal. "A Semantic Approach for Automating Knowledge in Policies of Cyber Insurance Services". (14th IEEE International Conference on Web Services (IEEE ICWS), 2019).

50. Aditya Pingle, Aritran Piplai, Sudip Mittal, Anupam Joshi, James Holt, and Richard Zak. "RelExt: Relation Extraction using Deep Learning approaches for Cybersecurity Knowledge Graph Improvement." IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (IEEE/ACM ASONAM '19), August 27–30, 2019, Vancouver, BC, Canada.

51. Sowmya Ramapatruni, Sandeep Nair, Sudip Mittal, Anupam Joshi, Karuna Joshi. "Anomaly Detection Models for Smart Home Security". (IEEE BigDataSecurity 2019).

52. Priyanka Ranade, Sudip Mittal, Anupam Joshi, and Karuna Joshi. "Using Deep Neural Networks to Translate Multi-lingual Threat Intelligence". (IEEE Intelligence and Security Informatics (IEEE ISI) 2018).

53. Lorenzo Niel, Sudip Mittal, Anupam Joshi. "Mining Threat Intelligence about Open-Source Projects and Libraries from Repository Issues and Bug Reports". (IEEE Intelligence and Security Informatics (IEEE ISI) 2018).

54. Priyanka Ranade, Sudip Mittal, Anupam Joshi, and Karuna Joshi "Understanding Multilingual Threat Intelligence for AI based Cyber-defense Systems" (IEEE International Conference on Technologies for Homeland Security (IEEE HST), 2018)

55. Vishal Rathod, Sandeep Nair Narayanan, Sudip Mittal, and Anupam Joshi. "Semantically Rich, Context Aware Access Control for Openstack". (2018 IEEE 4th International Conference on Collaboration and Internet Computing (IEEE CIC 2018)).

56. Maithilee Joshi, Sudip Mittal, Karuna Pande Joshi, and Tim Finin."Semantically Rich Oblivious Access Control for Cloud Storage" (IEEE International Conference on Edge Computing (IEEE EDGE) 2017.

57. Sudip Mittal, Aditi Gupta, Karuna Pande Joshi, Claudia Pearce, and Anupam Joshi."A Question Answering System for Management of Cloud Service Level Agreements" (IEEE International Conference on Cloud Computing (IEEE CLOUD) 2017)

58. Agniva Banerjee, Raka Dalal, Sudip Mittal, Karuna Pande Joshi. "Generating Digital Twin models using Knowledge Graphs for Industrial Production Lines" (9th International ACM Web Science Conference, Industrial Knowledge Graphs (ACM WEBSCI) (2017))

59. Karuna P Joshi, Aditi Gupta, Sudip Mittal, Anupam Joshi, Tim Finin and Claudia Pearce. "Semantic Approach to Automating Management of Data Privacy Policies for Cloud Consumers" (IEEE Big Data 2016)

60. Karuna P. Joshi, Aditi Gupta, Sudip Mittal, Claudia Pearce, Anupam Joshi and Tim Finin. "ALDA : Cognitive Assistant for Legal Document Analytics" (AAAI Fall Symposium (2016))

61. Sudip Mittal, Aditi Gupta, Karuna Pande Joshi, Claudia Pearce, and Anupam Joshi. "A Semantic Framework for Automated Analysis of Cloud Service Level Agreements" (2016).

62. Sandeep Nair, Sudip Mittal, Anupam Joshi. "Using Semantic Technologies to Mine Vehicular Context for Security". (37th IEEE Sarnoff Symposium (2016))

63. Sudip Mittal, Prajit Kumar Das, Varish Mulwad, Anupam Joshi, and Tim Finin. "Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities." In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (IEEE/ACM ASONAM), pp. 860-867. IEEE Press, 2016.

64. Sandeep Nair Narayanan, Sudip Mittal, and Anupam Joshi. "OBD SecureAlert: An Anomaly Detection System for Vehicles". (IEEE SmartSYS 2016).

65. Aditi Gupta, Sudip Mittal, Karuna Pande Joshi, Claudia Pearce, and Anupam Joshi. "Streamlining Management of Multiple Cloud Services". (IEEE International Conference on Cloud Computing (IEEE CLOUD) 2016).

66. Sandeep Nair Narayanan, Sudip Mittal, and Anupam Joshi. "Using Data Analytics to Detect Anomalous States in Vehicles". (2016).

67. Sudip Mittal, Karuna Pande Joshi, Claudia Pearce, and Anupam Joshi. "Automatic Extraction of Metrics from SLAs for Cloud Service Management". In IEEE International Conference on Cloud Engineering (IEEE IC2E), 2016.

68. Sudip Mittal, Karuna Pande Joshi, Claudia Pearce, and Anupam Joshi. "Parallelizing Natural Language Techniques for Knowledge Extraction from Cloud Service Level Agreements." In 2015 IEEE International Conference on Big Data, (IEEE Big Data 2015).

69. Sudip Mittal, Neha Gupta, Prateek Dewan, and Ponnurangam Kumaraguru. "Pinned it! A Large Scale Study of the Pinterest Network." In Proceedings of the IKDD Conference on Data Sciences, pp. 1-10. ACM, 2014.

70. Sudip Mittal, Neha Gupta, and Ponnurangam Kumaraguru. "A Pinteresting World", Security and Privacy Symposium - 2013.

71. Sudip Mittal, Shwetank Kumar Saha, and Daksha Yadav. "Privacy and security in Open Government Data." Research Showcase - 2012 (IIIT - Delhi)(Poster) (Best Poster Award: Research Track).

Thesis & Dissertations

72. Sudip Mittal. "Knowledge for Cyber Threat Intelligence." (2019) (Ph.D. Dissertation, Advisor: Dr. Anupam Joshi, UMBC.)

73. Sudip Mittal. "Broker Bots: Analyzing automated activity during High Impact Events on Twitter." (2014) (M.Tech Thesis, Advisor: Dr. Ponnurangam Kumaraguru, IIIT-Delhi.)

Technical Reports

74. Sudip Mittal, Jingdao Chen. "AI Security Threats against Pervasive Robotic Systems: A Course for Next Generation Cybersecurity Workforce". (Technical Report).

75. Maanak Gupta, Mahmoud Abdelsalam, Sudip Mittal. "Preparing Next Generation AI assisted Cybersecurity Warriors". (2021).

76. Maanak Gupta, Sudip Mittal, Mahmoud Abdelsalam. "AI assisted Malware Analysis: A Course for Next Generation Cybersecurity Workforce" (2020).

77. Maanak Gupta, Mahmoud Abdelsalam, Sudip Mittal. "Enabling and Enforcing Social Distancing Measures using Smart City and ITS Infrastructures: A COVID-19 Use Case" (2020).

78. Sudip Mittal, Anupam Joshi, and Tim Finin. "Thinking Fast and Slow! Combining Knowledge Graphs and Vector Spaces." (2017).

79. Sudip Mittal, Neha Gupta, Prateek Dewan, and Ponnurangam Kumaraguru. "The pin-bang theory: Discovering the pinterest world." (2013).

Under Submission

80. Khatib, H. S. A., Neupane, S., Manchukonda, H. K., Golilarz, N. A., Mittal, S., Amirlatifi, A., & Rahimi, S. (2024). Patient-Centric Knowledge Graphs: A Survey of Current Methods, Challenges, and Applications. arXiv preprint arXiv:2402.12608.

81. Mitra, Shaswata, Subash Neupane, Trisha Chakraborty, Sudip Mittal, Aritran Piplai, Manas Gaur, and Shahram Rahimi. "LOCALINTEL: Generating Organizational Threat Intelligence from Global and Local Cyber Knowledge." arXiv preprint arXiv:2401.10036 (2024).

82. Ables, Jesse, Nathaniel Childers, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, and Maria Seale. "Eclectic Rule Extraction for Explainability of Deep Neural Network based Intrusion Detection Systems." arXiv preprint arXiv:2401.10207 (2024).

83. Morgan Reece, Theodore Edward Lander Jr, Matthew Stoffolano, Andy Sampson, Josiah Dykstra, Sudip Mittal, Nidhi Rastogi. "Systemic Risk and Vulnerability Analysis of Multi-cloud Environments". (Submitted under review).

84. Anderson, William, Kaneesha Moore, Jesse Ables, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, and Maria Seale. "Designing an Artificial Immune System inspired Intrusion Detection System." (Submitted under review).

85. Reece, Morgan, and Sudip Mittal. "Self-Sovereign Identity in a World of Authentication: Architecture and Domain Usecases." (Submitted under review).

86. Charles Moore, Shaswata Mitra, Nisha Pillai, Marc Moore, Sudip Mittal, Cindy Bethel, Jingdao Chen. "URA*: Uncertainty-aware Path Planning using Image-based Aerial-to-Ground Traversability Estimation for Off-road Environments" (Submitted under review).

87. Subash Neupane, Ivan A Fernandez, Sudip Mittal, Shahram Rahimi. "Impacts and Risk of Generative AI Technology on Cyber Defense". (Submitted under review).

PAST EXPERIENCE

Assistant Professor (Tenure Track) *August 2021 - Present*
Department of Computer Science & Engineering, Mississippi State University

Assistant Professor (Tenure Track) *August 2019 - July 2021*
Department of Computer Science, University of North Carolina Wilmington

Visiting Lecturer *August 2018 - May 2019*
Department of Computer Science & Electrical Engineering, UMBC

Accelerated Cognitive Cybersecurity Lab *Dec 2016 - May 2019*
Graduate Research Assistant
Advisor: Prof. Anupam Joshi, Prof. Tim Finin

Ebiquity Research Group & CHMPR *Sept 2014 - May 2019*
Graduate Research Assistant
Advisor: Prof. Anupam Joshi

TEACHING EXPERIENCE

- Advanced Cyber Operations (MSU CSE 8713) *Spring 2024*
- Intro to Cyber Operations (MSU CSE 4773/6773) *Fall 2022*
- AI for Cybersecurity (MSU CSE 4293/6293) *Spring 2022, Spring 2023*
- Machine Learning (MSU CSE 8673) *Fall 2021*
- Foundations of Cybersecurity (UNCW CYBR 201) *Fall 2020*
- Cyber Operations (UNCW CSC 324/592) *Fall 2020*
- Computer Security (UNCW CSC 424/592) *Spring 2020, Spring 2021*
- Computer Networks (UNCW CSC 344) *Fall 2019, Spring 2020, Fall 2020, Spring 2021*
- Computer Organization (UNCW CSC 242) *Fall 2019*
- Computer Security (UMBC CMSC 426) *Spring 2019*
- Data Science (UMBC CMSC 491) *Spring 2019*
- Artificial Intelligence (UMBC CMSC 471, 2 sections) *Fall 2018*

ADVISING

Ph.D. Students

Committee Chair

1. Jesse Ables (MSSTATE CSE, 2021 - 2023), Graduated Fall 2023.
2. Trisha Chakraborty (MSSTATE CSE, 2021 - 2024), (Chair: Dr. Maxwell Young), Graduated Spring 2024.
3. Ivan Fernandez (MSSTATE CSE, 2021 - Present)
4. Morgan Reece (MSSTATE CSE, 2021 - Present)
5. Keith Strandell (MSSTATE CSE, 2021 - Present)
6. Damodar Panigrahi (MSSTATE CSE, 2021 - Present)
7. Shaswata Mitra (MSSTATE CSE, 2021 - Present)
8. Martin Duclos (MSSTATE CSE, 2023 - Present)
9. Derek True (MSSTATE CSE, 2023 - Present)
10. William Anderson (MSSTATE CSE, 2023 - Present, *SFS Scholar*)

Masters Students

Committee Chair

1. Ashlyn Martin (MSSTATE CSE, Spring 2025, *SFS Scholar*)
2. Joshua Whitman (MSSTATE CSE, Spring 2024, *SFS Scholar*)
3. Teddy Lander (MSSTATE CSE, Spring 2024, *SFS Scholar*)
4. Clarke Bowen (MSSTATE CSE, Spring 2024, *SFS Scholar*)
5. Martin Duclos (MSSTATE CSE, Graduated Fall 2023)
6. William Anderson (MSSTATE CSE, Graduated Fall 2023, *SFS Scholar*)
7. Eric Kudjoe Fiah (MSSTATE CSE, Graduated Spring 2022)
8. Rickie James Cashwell (UNCW, Graduated Spring 2021)
9. Shaik Barakhat Aziz (UNCW, Graduated Fall 2020)
10. Deepthi Thalagundamatada (UNCW, Graduated Spring 2020)

Undergraduate Students

1. Yasmin Chambers (MSSTATE, Fall 2023 - Present, *SFS Scholar*)
2. Anna Booth (MSSTATE, Fall 2023 - Present, *SFS Scholar*)

3. Joshua Bean (MSSTATE, Fall 2023 - Present, *SFS Scholar*)
4. Gavin Seiler (MSSTATE, Fall 2023 - Present, *SFS Scholar*)
5. Kaneesha Moore (MSSTATE, Summer 2022 - Present, *REU Student, SFS Scholar*)
6. Matthew Morgan (MSSTATE, Summer 2022 - Present, *SFS Scholar*)
7. Wilson Patterson (MSSTATE, Summer 2022 - Present, *REU Student, SFS Scholar*)
8. Daniel Tanner (MSSTATE, Summer 2022 - Fall 2023, *SFS Scholar*)
9. La'andrea Gates (MSSTATE, Fall 2021 - Spring 2022, *SFS Scholar*)
10. Kyanie Waters (MSSTATE, Fall 2021 - Spring 2022, *SFS Scholar*)
11. Ritik Khanna (IIIT Delhi, Main advisor - Dr. Raghava Mutharaju. Fall 2021-Spring 2022.)
12. Anmol Kumar (IIIT Delhi, Main advisor - Dr. Raghava Mutharaju. Fall 2021-Spring 2022.)
13. Christina Newkirk Fairbanks (UNCW, Spring 2020)
14. Matthew Sills (UNCW, Spring 2020 - Spring 2021)
15. Leon Neverov (UNCW, Spring 2021)
16. Meredith Metcalf (UNCW, Spring 2020)
17. Payton Weatherspoon (UNCW, Spring 2020)
18. Nouran Alotaibi (UNCW, Spring 2020)
19. Alexander J. Cossifos (UNCW, First Year Research Experience (FYRE), Spring 2020)
20. Ariana Gabrielle Curtis (UNCW, First Year Research Experience (FYRE), Spring 2021)

High School Student

1. Samar Rosas, Spring 2024, Mississippi School for Mathematics and Science.
2. Sephora Poteau, Fall 2022, Mississippi School for Mathematics and Science.

PROFESSIONAL ACTIVITIES & SERVICE

Professional Association Memberships

- Institute of Electrical and Electronics Engineers (IEEE)
- Association for Computing Machinery (ACM)
- IEEE Computer Society Technical Committee on Security and Privacy
- IEEE Computer Society Technical Committee on Semantic Computing
- DHS Industrial Control Systems Joint Working Group (ICSJWG)

Panel Review

- NSF Panel Reviewer 2021, 2022, 2023

Conference & Workshop Organization

- ACM CODASPY 2021 & 2022, Panel Co-Chair, Organizing Committee.
- Workshop on Big Data for Cybersecurity (BigCyber 2018 @ BigData 2018, BigCyber 2019 @ BigData 2019, BigCyber 2020 @ BigData 2020, BigCyber 2021 @ BigData 2021, BigCyber 2022 @ BigData 2022, BigCyber 2023 @ BigData 2023)
- KDD Workshop on Knowledge-infused Learning, Co-located with 29TH ACM SIGKDD 2023.
- ACM Workshop on Secure and Trustworthy Cyber-Physical Systems (SaT-CPS) 2021 (Co-located with ACM CODASPY 2021)
- Workshop on Smart Farming, Precision Agriculture, and Supply Chain (SmartFarm-2020 @ BigData 2020)
- Workshop on Analytics for Security in Cyber Physical System (ASCPS - 2019 @ ICDCN 2019)
- International Students' Workshop on Smart Computing (SmartStudents - 2019 @ SmartCOMP 2019)

Program Committee Member/Reviewer

- IEEE Big Data 2021, 2022, 2023
- IEEE ICMLA 2021, 2022, 2023

- ACM CODASPY 2020, 2021, 2022, 2023, 2024
- ACM ACSAC 2022, 2023
- IEEE SMARTCOMP 2020, 2021
- IEEE ICTAI 2020, 2021, 2022, 2023
- NAACL 2020, 2021, 2022, 2023
- EMNLP 2019, 2020, 2021, 2022
- AAAI 2019, 2020, 2021, 2022, 2023
- ACL - IJCNLP 2019, 2020, 2021, 2022, 2023
- IEEE CLOUD 2018
- ACM ICDCN 2019
- AACL 2019, 2020
- SQUEET 2019
- ICSNC 2019
- PLOS ONE
- IEEE ACCESS
- IEEE Transactions on Services Computing (TSC)
- ACM Transactions on Privacy and Security (TOPS)
- IEEE Transactions on Information Forensics and Security (TIFS)
- ACM Transactions on Internet Technology (TOIT)
- IEEE Transactions on Dependable and Secure Computing (TDSC)
- Elsevier Computers & Security
- Elsevier Sustainable Computing
- Springer Nature
- Nature Machine Intelligence

University Service

At Mississippi State University

- NSA CAE POC/APOC for CAE-CD, CAE-CO, CAE-R.
- NSA CAE-R Redesignation 2024. In-process.
- NSA CAE-CO Re-designation 2022. Successful, designation expires 2029.
- NSA CAE-CD Re-designation 2021 - 2022. Successful, designation expires 2028.
- ABET Assessment Committee, 2022 - Present.
- Cybersecurity Studies Committee, 2021 - Present.
- CSE Facilities Committee, 2022 - Present.
- Faculty Search Committee, 2023.

At University of North Carolina Wilmington

- Faculty Search Committee, Security line, 2021
- Faculty Advisor, UNCW Center for Cyber Defense Education (NSA/DHS CAE-CDE) (2019 - 2021)
- Faculty Advisor for CS students part of UNCW Cyber Defense Club.
- Cybersecurity Curriculum Committee, UNCW (2020 - 2021)
- Computer Science Assessment Committee, UNCW (2019 - 2021).
- MS Data Science Advisory Committee, UNCW (2019 - 2021).
- Program Coordinator CSC 131/242/342/344/360/434, UNCW (2019 - 2021).

PRESS

- Researchers show how platforms can scrub COVID conspiracies, election lies and other misinformation link (Jan 28, 2022) (Published in Readme.Security magazine)
- Misinformation risk and the security community. (June 21, 2021) (Published in the Cyberwire)

- Machine Learning Can Use Tweets to Spot Critical Security Flaws. (March 07, 2019) (Published in the WIRED magazine)
- Russians hack home internet connections - here's how to protect yourself. (May 04, 2018)(`https://lat.ms/2JQsAPM`) (Published in: Business Insider, Los Angeles Times, etc.)

Citizenship: India, US Permanent Resident.

Teaching Accomplishments

Contents

Highlights

- Taught multiple courses in cybersecurity and AI with consistently strong overall teaching evaluations.
- Took an active part in broad cybersecurity departmental educational efforts by managing MSU's 3 CAE designations.
- Developed multiple new courses and regularly represents MSU in national educational and research initiatives.

In this section of my dossier, I demonstrate my dedication to teaching and mentoring students. Specifically, I offer an explanation of my teaching philosophy, which includes examples of student feedback and overall scores from teaching evaluations. I also provide an outline of my efforts to design and instruct innovative courses at Mississippi State University, as well as a more detailed discussion of my responsibilities as a mentor and advisor.

Teaching Philosophy

My teaching philosophy is rooted in pedagogical techniques that aim to continuously foster curiosity. I have seen others and myself, work harder to find answers when we are curious about them. Being curious about the evolution of computing technology, gaps in the current state of the art systems, and possible future innovations, has inspired my career as an academic. I aim to inculcate the same passion for curiosity in my students. A classroom is a great opportunity to help students explore their curiosities by working on different concepts, resulting in their growth as scholars.

During my tenure as a faculty member, I have been fortunate to work with a diverse pool of students, teachers, mentors, and researchers. This variety of experiences has been helpful to me in understanding the needs and challenges of students as well as in improving my teaching.

I find teaching and advising rewarding because it is fulfilling to see my students work, understand challenging material, and mature with time. The primary classroom instructional strategies that I follow are explained below:

1. *Balancing Theory and Application*: Deep theoretical comprehension of fundamental ideas is necessary for computer science, as is their practical application to real-world systems. I convey to students that theory is a toolkit they may use and continue to grow throughout their schooling and in their future careers. I stress the value of building a solid scientific basis that can be used to solve any type of computer science problem, including creating optimized programs, creating novel

algorithms, and designing new systems. Their homework assignments, projects, and in-class examples serve as illustrations of the above. My objective is to ensure that students receive a traditional education in computer science while simultaneously teaching them the modern skills necessary to pursue their ideal vocations. I do this by balancing academic knowledge of the content with practical exercises.

2. *Active Classroom Environment*: Through my teaching experience, I have realized that the best way to engage students in class is to offer them active discussion scopes. I especially enjoy addressing questions during these discussions since students frequently present interesting perspectives. My own grasp of these topics has improved as a result of responding to these questions and comprehending their perspectives. Alongside traditional approaches of asking conversation stimulating questions and discussing hypothetical scenarios, I also often arrange demos in class. I discovered that students showed more attention when I gave demos of AI or cybersecurity concepts in class. In fact, the demonstrations sparked stimulating discussions that helped them better understand the fundamental ideas. Once they see the likes of deep neural networks, decision trees, public key infrastructure, malware detonation, etc. in action, they can co-relate it with their theoretical understanding. These interactions enable me as a teacher to assess whether students are grasping the material and helps me make sure that they are paying attention. Before tackling more complicated subjects, I can better retrace and course-correct if I can determine where there are comprehension gaps.

3. *Integrating Real-World Research and Engineering Problems*: In order for students to comprehend the larger significance of the technologies they actively study about, it is crucial to present social consequences and big-picture concepts to them in class. Since they can understand the long-term importance of the subject matter, I discovered that addressing real-world research and engineering challenges energizes students and encourages them to stay interested in the course as the semester unfolds. This strategy is useful for attracting students to university-based research. If a student shows interest in the research issues raised, I will try to direct them to the necessary sources, encourage them to submit an application, and set up a meeting with the relevant professors.

4. *Continuous Teaching Refinement and Development*: In order to adjust to specific student demands, shifting contexts, and developing technology, teaching as a discipline requires ongoing learning and strategy improvement. I draw upon current research and academic evidence to improve my classroom learning settings in addition to using student feedback to make my courses more understandable and efficient each semester. For example, I attended courses offered to faculty members by the Center for Teaching and Learning at Mississippi State University. These courses on active learning, inclusive pedagogy, concept mapping, online instruction, and flipped learning, have helped me structure my courses better. One such course that has improved my teaching is inclusive pedagogy. The course taught me various techniques to embed inclusivity in my teaching and classroom environments.

Teaching Experience: I started teaching courses as a full time instructor in Fall 2019. I have taught a variety of computer science courses at 3 universities, Mississippi State University (MSU), University of North Carolina Wilmington (UNCW), and University of Maryland Baltimore County:

- Advanced Cyber operations (MSU CSE 8173)
- Intro to Cyber Operations (MSU CSE 4773/6773)
- AI for Cybersecurity (MSU CSE 4293/6293)
- Machine Learning (MSU CSE 8673)
- Foundations of Cybersecurity (UNCW CYBR 201, Freshpersons/Sophomores, University Gen ED)
- Cyber Operations (UNCW CSC 324/592, split level course)
- Computer Security (UNCW CSC 424/592, UMBC CMSC 426, split-level course)
- Computer Networks (UNCW CSC 344, Juniors/Seniors)
- Computer Organization (UNCW CSC 242, Freshpersons/Sophomores)
- Digital Circuits (UNCW CSC 242, Juniors)
- Data Science (UMBC CMSC 491, Juniors/Seniors)
- Artificial Intelligence (UMBC CMSC 471, Juniors/Seniors)

Broad Cybersecurity Education & Management of MSU's NSA CAE designations

As faculty, it is our responsibility to create an environment for learning that benefits students not only in the courses we teach but also in the overall body of information acquired over the course of a degree. At MSU, I am 1 of 2 faculty points of contact for our National Centers of Academic Excellence in Cybersecurity (NCAE-C) designations. Our 2-person team manages all 3 NSA CAE designations, i.e. CAE in Cyber Defense (CAE-CD), CAE in Cyber Research (CAE-R), and CAE in Cyber Operations (CAE-CO). As a result of our efforts, MSU is 1 of only 10 universities in the USA to hold all 3 designations. The outcome of our work resulted in a successful CAE-CD re-designation until 2028. Additionally, we recently completed our CAE-CO re-designation till 2029. We are currently in process of renewing our CAE-R designation.

On my part, this entails helping manage various cybersecurity degrees through the Cybersecurity Curriculum Committee. To keep these degree programs in compliance with CAE designations, we regularly update and modify various courses that are part of the following degree programs:

1. *CAE in Cyber Operations (CAE-CO) designated Master of Science (M.S) in Cyber Security & Operations (MS-CYSO)*. Both of our undergraduate programs of study (BS-CSE and BS-CYSO) are aligned with our MS-CYSO program, creating the opportunity for students to earn their Master's degree in 5 years (4+1 program). This allows our students majoring in CSE, along with cybersecurity, to enroll in a 4+1 program with the CAE-CO designated Program of Study.

2. *CAE in Cyber Defense (CAE-CD) designated Information Assurance Certificate (IAC)* that students can complete with their B.S. degrees in Computer Science (BS-CSE), Software Engineering, Electrical Engineering, Computer Engineering, Cybersecurity, or Information Systems.

3. *B.S. in Cybersecurity*, CAE-CO program of study validation and designation application will be filed in 2025/2026.

New Courses & Curriculum Development

During my time at MSU, I have designed and taught multiple new courses in the computer science sub-fields of artificial intelligence and cybersecurity. Keeping up with new developments in these fields and the changing requirements of NSA CAE designations, I assisted with the addition of the following new courses:

- CSE 4293/6293 AI for Cybersecurity
- CSE 4783/6783 Cloud Computing and Security (Primary Lead Dr. Charan Gudla)

As part of my involvement in NSA CAE activities, I have regularly taken part in national educational and research efforts representing MSU:

- In Spring 2022, I was the participating MSU faculty, in an NSA CAE effort to create subject matter experts in the field of Supervisory Control and Data Acquisition (SCADA) Security. This involves teaching students critical infrastructure defense techniques. Students learned how to protect dams, water treatment plants, nuclear reactors, etc. from cyber attacks. At MSU this was taught as *CSE 4990/6990 Special Topics - SCADA Security*. Other universities involved included: UAH, Texas A&M, U of South Alabama, UWF.
- In Fall 2023, I created a research course taught as *CSE 4990/6990 Special Topics - Research Methods in Cybersecurity*. This course was part of the national CAE INSURE+E program. Other universities involved included: Texas A&M, UTD, UCSB, Iowa State, etc. This national research effort provides academic institutions working with government agencies, national labs, and FFRDCs a means to offer opportunities to their students on national cybersecurity research priorities.

Student Teaching Evaluations

At the end of each semester, students may provide anonymous feedback in the form of specific scored questions, Table 1 sumarizes my numerical scores. For all my courses, I have received excellent student course evaluations. Some anonymous student comments include:

- The professor explained the subject matter very well and the class was more participatory. The project assignment helped students to get a better knowledge of the subject.

Course	Distance	Course Title	Semester	Evaluation Score
CSE 8673	No	Machine Learning	Fall 2021	4/4
CSE 8673	Yes	Machine Learning	Fall 2021	3.5/4
CSE 4990	No	Special Topics - AI for Cyber Security	Spring 2022	4/4
CSE 6990	No	Special Topics - AI for Cyber Security	Spring 2022	4/4
CSE 6990	Yes	Special Topics - AI for Cyber Security	Spring 2022	4/4
CSE 4990	No	Special Topics - SCADA Security	Spring 2022	3.5/4
CSE 6990	No	Special Topics - SCADA Security	Spring 2022	4/4
CSE 4773	No	Introduction to Cyber Operations	Fall 2022	4/4
CSE 4773	Yes	Introduction to Cyber Operations	Fall 2022	4/4
CSE 6773	Yes	Introduction to Cyber Operations	Fall 2022	3.5/4
CSE 6293	No	AI for Cybersecurity	Spring 2023	4/4
CSE 6293	Yes	AI for Cybersecurity	Spring 2023	3/4
CSE 4990	Yes	Special Topics - Research Methods in Cybersecurity	Fall 2023	4/4
CSE 6990	Yes	Special Topics - Research Methods in Cybersecurity	Fall 2023	3/4
CSE 4773	Yes	Introduction to Cyber Operations	Winter 2023	4/4
CSE 8713	Yes	Advanced Cyber Operations	Spring 2024	4/4

Table 1: Student evaluation scores based on anonymous feedback.

- The slides and other materials did a great job of explaining the subjects.
- In class lecture was effective, and examples and diagrams shown were beneficial to understanding the material. Assignments were great and tied in well with the current topics discussed. overall good instruction from the professor.
- The class's curriculum and schedule was very good, and the overarching topics were very well chosen. Going from AI for cyber security to cyber security for AI was very interesting.

Sample Syllabi

Below, I have provided the syllabi for two courses that I designed and taught. The first is CSE 4293/6293 AI for Cybersecurity, and the second is CSE4773/6773 Introduction to Cyber Operations.

MISSISSIPPI STATE UNIVERSITY
Dept. of Computer Science & Engineering

Course Title – "Artificial Intelligence for Cybersecurity"
CSE – 4293/6293
Section –
Credit hours - 3

Instructor – Dr. Sudip Mittal
Email – mittal@cse.msstate.edu
Phone – 325-7449
Preferred method of contact – Email
You will receive a response within 24 hours.
Office – Rice Hall 103
Office Hours –

Location –
(University operates in the CST time zone)
Scheduled Time –
Method of Delivery –

Course description
The use of Artificial Intelligence (AI) and Machine Learning (ML) to solve cybersecurity problems has been gaining attraction within industry and academia, in part as a response to widespread malware attacks on critical systems, such as cloud infrastructures, government offices or hospitals, and the vast amounts of data they generate. AI- and ML-assisted cybersecurity offers data-driven automation that could enable security systems to identify and respond to cyber threats in real time.

Catalog description
Three hours lecture. Prerequisite: CSE 4633 with a grade of C or better. The use of artificial intelligence and machine learning to solve cybersecurity problems, including advanced topics in applying these techniques to real-world datasets to learn about Cyber Threat Intelligence (CTI), malware analysis, and classification.

Learning objectives
1. Understand AI applications in Cybersecurity, Threat Intelligence, Malware Analysis.
2. Describe threats to AI models.
3. Understand advanced AI research topics and case studies such as adversarial learning and advanced threat detection.

Topics
1. Cyber Threat Intelligence (CTI) and Analysis (3 contact hours)
2. Malware attack stages (3 contact hours)

3. CTI standards (3 contact hours)
4. Malware knowledge representation (3 contact hours)
5. CTI sharing (3 contact hours)
6. Malware data collection and feature identification (3 contact hours)
7. AI-assisted malware detection (3 contact hours)
8. Malware classification and Attribution (3 contact hours)
9. Advanced research topics (21 contact hours)
9.1. Adversarial Machine Learning (15 contact hours)
9.1.1. Model Evasion (3 contact hours)
9.1.2. Function Extraction (3 contact hours)
9.1.3. Model Poisoning (3 contact hours)
9.1.4. Model Inversion (3 contact hours)
9.1.5. Traditional attacks (3 contact hour)
9.2. Advanced Persistent Threat (APT) detection (6 contact hours)

Grading Scheme

	Component	Weightage
1	Assignments (Approx. 6 programming Assignments)	30%
2	Midsemester Exam	15%
3	Final Exam	15%
4	Paper Reading & Review (Technical Summary & Critique)	15%
5	Course Project	25%
6	Survey Paper (**Graduate Students Only**, for grad students the total possible points will be normalized from 110 to 100)	10%

Midterm Exam Data and Time –
Endterm Exam Date and Time –

Grading Scale

Percentage	Letter Grade
90.0 - 100	A
80.0 - 89.9	B
70.0 - 79.9	C
60.0 - 69.9	D
00.0 - 59.9	F

Attendance policy

The course does not use attendance as part of a student's grade, however students are expected to attend all class meetings. Please refer to Academic Operating Policy 12.09, regarding attendance expectations and accommodations.

Distance Learning Education

The distance learning section will be conducted synchronously, with students will participating via MSU's WebEx infrastructure. All assignments and deliverables will be submitted using CANVAS. Interactions between the distance learning students and the instructor will happen regularly during the class as well as WebEx enabled office hours. Distance learning students will be encouraged to interact with peers via some group assignments and the group project. Distance Learning section requires a WebEx enabled computer with a headphone/speaker.

Late Submission Policy

Assignments must be submitted by 11:59pm on the due dates. A penalty of 25% will be deducted from your score for the first 6-hour period if your submission is late, and 50% for the first 24-hour period. A penalty of 75% will be deducted from your score for >= 48-hour period. Assignments and reports submitted more than 72 hours late will not be assessed and will score as a zero (0). Weekend days will be counted. Penalties will not be assessed for late submissions in the case of unforeseen university-approved excused absences as outlined in MSU AOP 12.09, but the instructor must be contacted and work submitted as soon as possible.

Examinations and Assignments

All assignments must have your name, student ID and course name/ number. Examinations will heavily emphasize the **conceptual understanding** of the material. No make-up exams will be given for any other reasons than those approved by the university (serious illness, medical emergency, etc. as described in MSU AOP 12.09). The exam format will be a mixture of multiple-choice questions and short/long answer questions. If you do not think that your test was graded appropriately, you need to send a valid written explanation for the requested change. This must be done within three days from the date the test was returned to you.

Continuity of Instruction

In the event that face-to-face classes are suspended due to extenuating circumstances, such as weather, the instructor will continue instruction in a manner that best supports the course content and student engagement. In this event, all instructors will notify students of the change via their university email address (the official vehicle for communication with students). At that time, they will provide details about how instruction and communication will continue, how academic integrity will be ensured, and what students may expect during the time that face-to-face classes are suspended. If a student becomes unable to continue class participation due to extenuating circumstances, (e.g., health and safety, loss of power, etc.) the student should contact their instructor and advisor for guidance. For additional guidance, please refer to Academic Operating Policy 12.09.

Disability Resource Center

Mississippi State University is committed to providing equitable access to learning opportunities for all students. The Disability Resource Center (01 Montgomery Hall) collaborates with students who have disabilities to arrange reasonable accommodations. If you have, or think you may have, a disability, please contact drc@saffairs.msstate.edu or 662-325-3335 to arrange a confidential discussion regarding equitable access and reasonable accommodations. Disabilities may include, but are not limited to, conditions related to mental health, chronic health, attention, learning, autism, brain injury, vision, hearing, mobility, speech, or intellectual disabilities. In the case of short-term disabilities (e.g., broken arm), students and instructors can often work to minimize barriers. If additional assistance is needed, please contact the Disability Resource Center.

Title IX

MSU is committed to complying with Title IX, a federal law that prohibits discrimination, including violence and harassment, based on sex. This means that MSU's educational programs and activities must be free from sex discrimination, sexual harassment, and other forms of sexual misconduct. If you or someone you know has experienced sex discrimination, sexual violence and/or harassment by any member of the University community, you are encouraged to report the conduct to MSU's Director of Title IX/EEO Programs at 325-8124 or by e-mail to titleix@msstate.edu. Additional resources are available at Dean of Students Sexual Misconduct and Sexual Assault.

University Safety Statement

Mississippi State University values the safety of all campus community members. Students are encouraged to register for Maroon Alert texts and to download the Everbridge App. Visit the Personal Information section in Banner on your MyState portal to register. To report suspicious activity or to request a courtesy escort via Safe Walk, call University Police at 662-325-2121, or in case of an emergency, call 911. For more information regarding safety and to view available training including helpful videos, visit ready.msstate.edu

Student Honor Code

Mississippi State has an approved Honor Code that applies to all students. The code is as follows: "As a Mississippi State University student, I will conduct myself with honor and integrity at all times. I will not lie, cheat, or steal, nor will I accept the actions of those who do." Upon accepting admission to Mississippi State University, a student immediately assumes a commitment to uphold the Honor Code, to accept responsibility for learning, and to follow the philosophy and rules of the Honor Code. Students will be required to state their commitment on examinations, research papers, and other academic work. Ignorance of the rules does not exclude any member of the MSU community from the requirements or the processes of the Honor Code. For additional information, please visit the Honor Code Policy.

CSE 4773/6773 - Introduction Cyber Operations

Faculty: **Dr. Sudip Mittal**

Office: **Room 329 Butler Hall**

Office Hours: By appointment (email)

Email: mittal@cse.msstate.edu

Class Time & Dates:

This course is an online only, asynchronous course. This course will be taught entirely online using Canvas Learning Management System. It is the responsibility of the student to obtain the lecture materials presented and be prepared for any assessments that may be given.

Course Overview

This course is designed to develop the students' knowledge of cyberspace operations concepts and methodologies. Graduates should be able to assist in the analysis, synthesis, and evaluation of management, engineering, and operational approaches to solve complex problems within cyberspace, defensive and offensive.

Learning Outcomes

1. Students will gain an appreciation for the technologies that will enable the study of cyber operations, both defensive and offensive.
2. Students will understand the fullest set of capabilities that exist for defensive and offensive cyber operations that can be discussed in an unclassified environment.
3. Students will gain an appreciation for the complex rules and limitations for cyber operations.

Student Outcomes:

Students will have a sound understanding of the technologies and methods utilized to defend systems and networks. They will be able to describe, evaluate, and operate a defensive network architecture employing multiple layers of protection using technologies appropriate to meet mission security goals.

Assessments

1. Podcast Reports……20%
2. Assignments……60%
3. Final Exam ……..10%
4. 3 page summaries……..10%
    - National Cyber Strategy-2018
    - Executive Order on Improving the Nation's Cybersecurity 2021
5. Topic Paper (Grad Students ONLY)……20% (total score normalized to 100)

Course Topics

Course Topics

Topic	Contact Hours
Introduction to cyberspace operations	
Building systems for cyber operations	3
Operational awareness	3
Review of network technologies	3
Defending operational networks	
Current enterprise network architecture requirements	2
Information assurance, cyber defense, and enabling technologies	2
Fundamental principles of cyber defense	3
Basic cyber defense models	3
Conducting risk assessments	3
Vulnerability assessment of enterprise networks	3
Intrusion detection technologies	2
Attacking operational networks	
Legal limitations	3
Domain name service	3
Disrupt, deny, or degrade operations	3
Deceive, neutralize, or destroy operations	3
Conducting offensive cyber space operations	3
Reporting and documenting offensive cyberspace operations	3

Note – The Mississippi State University Syllabus contains all policies and procedures that are applicable to every course on campus and online. The policies in the University Syllabus describe the official policies of the University and will take precedence over those found elsewhere. It is the student's responsibility to read and be familiar with every policy. The University Syllabus may be accessed at any time on the Provost website under Faculty and Student Resources and at - https://www.provost.msstate.edu/faculty-student-resources/university-syllabus

Research Accomplishments

Contents

Highlights

- I have served as a PI/Co-PI on multiple grants from NSF, NIH, and Department of Defense, totaling $11 million approx at MSU CSE.
- While at MSU, I have published at leading peer-reviewed conference venues that include ACM SACMAT, SECRYPT, IEEE/IFIP DSN, IEEE SSCI, IEEE TrustCom, ACM ACSAC, IEEE Big Data, IEEE CIC, IEEE TPS, and journal venues that include IEEE Transactions on Services Computing (TSC), IEEE Access, IEEE Internet Computing, The Cyber Defense Review, and Elsevier SoftwareX. Overall, according to google scholar I have a h-index of 27 and an $i-10$ index of 47.
- I worked on interesting cybersecurity and AI topics with researchers from the university, national labs, and government agencies.

My broad research interests are in the areas of cybersecurity and artificial intelligence. I aim to develop the next generation of cyber defense systems that help protect various organizations and people. A brief description of my main research efforts follows. The efforts described below have been funded primarily through the below listed grants.

External Funding

- National Institutes of Health (NIH) - STTR: Phase 1: MyDocSaid: AI assisted Patient Journey utilizing Knowledge Graphs and Large Language Models. Sudip Mittal (PI), ($103,500, July 2024 - July 2025)
- Griffiss Institute - VICEROY for the NCAE-C Southeast Region. - Virtual Institute for Cyber and Electromagnetic Spectrum Research and Employ (VICEROY for the NCAE-C Southeast Region). George Trawick (PI), Sudip Mittal (co-PI), Shelly Hollis (co-PI), ($550,000 - Year 4, July 2024 - July 2025)
- National Science Foundation - SaTC: EDU: Inculcate a culture of preparedness against AI security threats to pervasive robotic systems. Sudip Mittal (PI), Jingdao Chen (co-PI), ($400,000, Aug 2023 - July 2026).
- National Science Foundation - CyberCorps: Scholarship for Service at Mississippi State University (Renewal). Andy Perkins (PI), Sudip Mittal (Co-PI), Reed Mosher (Co-PI), ($4.2 Million, Aug 2023 - July 2028).
- National Science Foundation - Collaborative Research: SaTC: EDU: Artificial Intelligence Assisted Malware Analysis. Sudip Mittal (PI) (#2133190, $105,790, Aug 2020 - July 2023).

- National Science Foundation - REU Supplement to Award 2133190. "Explainable Malware Intrusion Detection based on Artificial Immune Systems". Sudip Mittal (PI) ($16,000, May 2022 - Dec 2022).
- U.S. Department of Defense (DOD), US Army Corps of Engineers (USACE); Engineer Research and Development Center (ERDC) - Enhanced Network Cybersecurity Research - A Self Organizing Maps and Danger Theory Hybrid for Intrusion Detection, Shahram Rahimi (PI), Sudip Mittal (co-PI) (approx. 1 million USD 2022-2025).
- National Science Foundation - CyberCorps: Scholarship for Service - A Continuation Program at Mississippi State University. PI - Andy Perkins, Co-PI - Sudip Mittal, Reed Mosher. ($3,112,393.00, end date - Aug 2025)
- U.S. Department of Defense (DOD), US Army Corps of Engineers (USACE); Engineer Research and Development Center (ERDC) - Adoptable Predictive Maintenance Using Transformer Neural Networks (TNN). Shahram Rahimi (PI), Sudip Mittal (co-PI) (approx. 1/2 million USD 2022-2025).
- U.S. Department of Defense (DOD), US Army Corps of Engineers (USACE); Enigineer Research and Development Center (ERDC) - Explainable Predictive Maintenance (XPM): An Explainable AI (XAI) Technology for Predictive Maintenance. Shahram Rahimi (PI), Sudip Mittal (co-PI) (approx. 1/2 million USD 2022-2025).
- U.S. Department of Defense (DOD), DoD Cybersecurity Academy Program. George Trawick (PI), Sudip Mittal (co-PI) ($189,000, Aug 2024 - July 2025).
- U.S. Department of Defense (DOD), DoD CySP Program. George Trawick (PI), Sudip Mittal (co-PI) ($76,540, Aug 2023 - July 2024).

External Collaborative Research Efforts

Other than setting up internal MSU research collaborations, I have set up numerous external collaborative efforts. I base my research and mentoring activities on the principles of *synergistic collaborations*. I have been successful in establishing collaborative grants and research activities with several internal and external colleagues. Currently, I have active collaboration with various faculties at Rochester Institute of Technology (RIT), University of Texas at San Antonio (UTSA), American University, Tennessee Tech University, North Carolina A&T State University, Augusta University, Clark Atlanta University, etc. On the other hand, I regularly work with researchers at Argonne National Laboratory, Army Corps of Engineers (USACE), National Security Agency (NSA), etc.

Summary of Research Efforts in AI and Cybersecurity

The use of AI/ML to solve cybersecurity problems has been gaining more traction within industry and academia. This data-driven automation will enable security systems to identify and respond to cyber threats in real time. I work in the following sub-areas:

AI for Malware Detection: I believe that in the near future, current signature-based detection methods will become obsolete. All malwares will be polymorphic. Defensive strategies that are based on behavioral analysis need to be developed. One such defensive strategy could be the development of AI based solutions. Such systems will need robust knowledge representations that include various malware behavioral details as well as mitigating instructions that can be leveraged by an AI based anti-virus system deployed on an enterprise network. I imagine these AI based solutions exchanging malware representations with each other, analogous to the current signature sharing schemes. I wish to work on the development of these malware representations and AI based anti-viruses. I have already done some preliminary work on deep learning systems that help detect malware on the cloud infrastructure. A datacenter infected with malware can cause data loss and/or major disruptions to service for its users. These systems will serve as the basis for future AI assisted malware analysis techniques. Another focus here is to make these Intrusion Detection Systems more Explainable (X-IDS). We have been focusing on creating state of the art X-IDS systems.

MSU PhD Students Involved: Jesse Ables, William Anderson, Trisha Chakraborty
Some Publications at MSU:

- Neupane, Subash, et al. "Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities." IEEE Access 10 (2022): 112392-112415.
- Ables, Jesse, et al. "Creating an explainable intrusion detection system using self organizing maps." 2022 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2022.
- Mitra, Shaswata, et al. "Survey of malware analysis through control flow graph using machine learning." 2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE, 2023.
- Chakraborty, Trisha, et al. "CAPoW: Context-Aware AI-Assisted Proof of Work Based DDoS Defense." International Conference on Security and Cryptography (SECRYPT 2023).
- Pusey, Portia, et al. "An Analysis of Prerequisites for Artificial Intelligence/Machine Learning-Assisted Malware Analysis Learning Modules." Journal of The Colloquium for Information Systems Security Education. Vol. 11. No. 1. 2024.

Cybersecurity Analyst Augmentation Systems: In modern enterprises, security analysts monitor threats in a security operations center (SoC) by watchstanding, akin to a lookout on a ship watching the environs for danger. Screens typically show warnings and alerts from individual products and detectors that the enterprise has installed. Watchstanding permits a highly trained security analyst to look at all the disparate pieces of information, and see if they 'click together' to form some pattern which might indicate an attack. The detection efficacy of a security analyst depends on her operational and strategic knowledge about

current security landscape and the associated intelligence. I have created various cyber security informatics systems – Cyber-All-Intel, CyberTwitter. The goal is to add value and augment an analyst's understanding of threats and vulnerabilities. These systems take as input textual Open Source Intelligence (OSINT), then represents this information using a hybrid knowledge representation scheme that combines knowledge graph and vector space embeddings. These systems also proactively try to improve the underlying cybersecurity knowledge. Other extensions that were driven by security analyst demand, were to understand and represent multi-lingual threats address the issue of supply chain attacks on software development, and more recently combining information from different sources with actual malware behavior data collected from a sandbox environment.

MSU PhD Students Involved: Shaswata Mitra, Martin Duclos, Morgan Reece
Some Publications at MSU:

- Mitra, Shaswata, et al. "Localintel: Generating organizational threat intelligence from global and local cyber knowledge." Under Submission.
- Reece, Morgan, et al. "Defending Multi-Cloud Applications Against Man-in-the-Middle Attacks." Proceedings of the 29th ACM Symposium on Access Control Models and Technologies. 2024.
- Reece, Morgan, et al. "Emergent (In) Security of Multi-Cloud Environment." Poster at the 2024 ACM Annual Computer Security Applications Conference (ACM ACSAC 2024)
- Mitra, Shaswata, et al. "Combating fake cyber threat intelligence using provenance in cybersecurity knowledge graphs." 2021 IEEE International Conference on Big Data (Big Data). IEEE, 2021.
- Duclos, Martin, et al. "Utilizing Large Language Models to Translate RFC Protocol Specifications to CPSA Definitions." Hot Topics in the Science of Security (HoTSoS 2024).

Autonomous Intrusion Response: When it comes to the use of AI for cyber defense, there are significant gaps and research opportunities that exist in the field of autonomous Intrusion Response. Current batch of Intrusion Response Systems (IRSs) offer trivial response capabilities, usually based on a static mapping between an identified attack and a predefined response. Such methodologies exhibit evident limitations mainly related to scalability and a lack of generalizability. I envision a fully capable IRS that can automatically compute and identify a response to an ongoing attack, exploiting additional knowledge about the attacker behavior and of the underlying system that it protects. This involves developing novel AI techniques to compute the optimal or near-optimal countermeasures that a cyber defense agent can take to stop or mitigate an ongoing attack. The problem is further exacerbated when the protected systems exhibit a non-stationary/dynamic behavior, and therefore needs an IRS with the ability to automatically adapt to these changes while dynamically predicting a near optimal response to an intrusion. We have developed

the first open-source licensed software prototype that implements a dynamic IRS, named irs-partition, it uses Deep Q-Networks (DQN), Reinforcement Learning (RL), and transfer learning to cope with the non-stationary/dynamic behavior of complex computer systems.

MSU PhD Students Involved: Damodar Panigrahi
Some Publications at MSU:

- Panigrahi, Damodar, et al. "REGARD: Rules of EngaGement for Automated cybeR Defense to aid in Intrusion Response." Under Submission.
- Cardellini, Valeria, et al. "irs-partition: An Intrusion Response System utilizing Deep Q-Networks and system partitions." SoftwareX 19 (2022): 101120.

Summary of Research Efforts in Security of AI enabled Cyber Physical Systems (CPS)

My research in this area has spanned numerous emerging 'smart' domains, such as robotics, smart cities, campuses, digital twins, communities, homes, industrial IoT infrastructure, etc. At MSU, the focus has been on creating Adoptable Predictive Maintenance Using Transformer Neural Networks and eXplainable Predictive Maintenance (XPM) AI system which will be able to augment current big data analytics, and advanced sensing and control technologies by including explainable diagnostic analysis, descriptive PdM models, and an easily understandable chain of reasoning from the sensed data, through the system's knowledge and inference, to the resulting predictions. In a recent NSF funded effort, we are also looking at the security of AI enabled robotics systems.

MSU PhD Students Involved: Ivan Fernandez, Subash Neupane, Alexander Sommers, Logan Cummins.
Some Publications at MSU:

- Cummins, Logan, et al. "Explainable predictive maintenance: a survey of current methods, challenges and opportunities." IEEE Access (2024).
- Fernandez, Ivan A., et al. "A Survey on Privacy Attacks Against Digital Twin Systems in AI-Robotics." IEEE International Conference on Collaboration and Internet Computing (IEEE CIC 2024).
- Sommers, Alexander, et al. "Generating Synthetic Time Series Data for Cyber-Physical Systems." 2024 IEEE 10th World Forum on Internet of Things (IEEE WF-IoT 2024).
- Neupane, Subash, et al. "Security Considerations in AI-Robotics: A Survey of Current Methods, Challenges, and Opportunities." IEEE Access (2024).

Service Accomplishments

Contents

Highlights

- I served as a reviewer for a number of conferences and journals.
- I have mentored a large number of students, from the high school level to PhD students.
- I have served on multiple departmental cybersecurity communities and have managed MSU's 3 CAE designations.

I am proud to have provided a significant amount of service, both in the broader academic community and within MSU. Below, I summarize some of this service; more details are provided in my CV.

NSF Panel Service

I served on 3 NSF review panels for the Secure and Trustworthy Cyberspace (SaTC) Program. 2 times for the SaTC CORE subprogram (2021. 2022) and 1 time for the SaTC EDU subprogram (2023).

Workshop Organization

I have organized workshops that follow the theme: AI for Security and Security for AI.

- 3rd International Workshop on Knowledge-infused Learning @ 29TH ACM SIGKDD `https://aiisc.ai/kiml2023/index.html`.

  Workshop Report: Manas Gaur, Efthymia Tsamoura, Sarath Sreedharan, and Sudip Mittal. 2023. KiL 2023: 3rd International Workshop on Knowledge-infused Learning. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23). Association for Computing Machinery, New York, NY, USA, 5857–5858. `https://doi.org/10.1145/3580305.3599199`

- Workshop on Big Data for Cybersecurity (BigCyber 2016 - 2023) held at IEEE International Conference on Big Data (Big Data 2016 - 2023) `https://bigcyber.umbc.edu/program-committee/`.

Each workshop was a single-day event where speakers shared research findings relevant to the wider community of AI and cybersecurity researchers. For the BigCyber workshop, me and my collaborators have made it into an annual event focusing on the aspects of data science, AI, and cybersecurity. The 7th BigCyber is scheduled to be held in December 2024 at IEEE Big Data 2024.

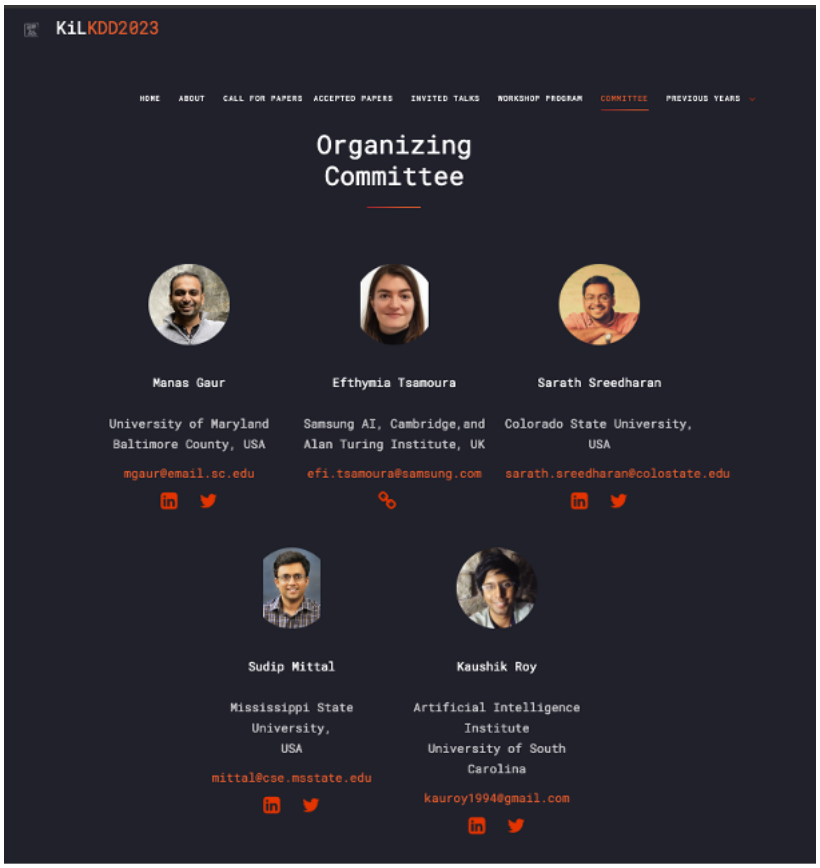


Figure 1: Organizing committee for 3rd International Workshop on Knowledge-infused Learning.

Figure 2: Organizing committee for Workshop on Big Data for Cybersecurity 2023 (BigCyber 2023).

High School Student Research Mentorship

In my time here at MSU I have been part of the MSU and the Mississippi School for Mathematics and Science (MSMS) Research Experience Program, where 2 MSMS students spent a semester in my lab. I hosted the following 2 MSMS students:

- Sephora Poteau (Spring 2023)
- Samar Rosas (Spring 2024)

In Spring of 2024 Samar Rosas authored a full conference paper: 'ReVisE: Emulated Visual Outfit Generation from User Reviews using Generative-AI'. Samar Rosas, Subash Neupane, Shaswata Mitra, and Sudip Mittal. Presented at 33rd SEDE Conference: Software & Data Engineering (SEDE), October 21-22, 2024, San Diego, California, USA.

Broader PC member/Reviewer Service

I have been a PC member/reviewer for a broad range of Cybersecurity and AI conferences and journals. These include:

- IEEE Big Data 2021, 2022, 2023
- IEEE ICMLA 2021, 2022, 2023
- ACM CODASPY 2020, 2021, 2022, 2023, 2024
- ACM ACSAC 2022, 2023
- IEEE SMARTCOMP 2020, 2021
- IEEE ICTAI 2020, 2021, 2022, 2023
- NAACL 2020, 2021, 2022, 2023
- EMNLP 2019, 2020, 2021, 2022
- AAAI 2019, 2020, 2021, 2022, 2023
- ACL - IJCNLP 2019, 2020, 2021, 2022, 2023
- IEEE CLOUD 2018
- ACM ICDCN 2019
- AACL 2019, 2020
- SQUEET 2019
- ICSNC 2019

- PLOS ONE
- IEEE ACCESS
- IEEE Transactions on Services Computing (TSC)
- ACM Transactions on Privacy and Security (TOPS)
- IEEE Transactions on Information Forensics and Security (TIFS)
- ACM Transactions on Internet Technology (TOIT)
- IEEE Transactions on Dependable and Secure Computing (TDSC)
- Elsevier Computers & Security

Departmental Committee Service

I have served on the following departmental committees:

- Cybersecurity curriculum committee (2021- Present)
- NSA POC for CAE CD/CO/R Designations (2021 – Present)
- Hiring Committee (2022, 2023)
- ABET Assessment Committee (2022 – 2023)
- CSE Facilities Committee (2022 – Present)

Student Advising

I have had the opportunity to work with the following students on research topics in cybersecurity, AI, and cyber-physical systems.

Doctoral Students Graduated:

1. Jesse Ables (December 2023, First Position: Assistant Professor, Computer Science, University of South Alabama.)
2. Trisha Chakraborty (May 2024, Chair: Maxwell Young, Committee Co-Chair: Sudip Mittal, First Position: Amazon R&D., USA.)

Masters Students Graduated:

1. Joshua Whitman (May 2024)
2. Teddy Lander (May 2024)

3. Martin Duclos (December 2023)
4. William Anderson (December 2023)
5. Eric Kudjoe Fiah (May 2023)

Current Doctoral Students at MSSTATE as Committee Chair & Major Professor:

1. Morgan Reece (expected graduation: Spring 2025)
2. Keith Strandell (expected graduation: Spring 2025)
3. Damodar Panigrahi (expected graduation: Spring 2025)
4. Ivan Fernandez (expected graduation: Fall 2025)
5. Shaswata Mitra (expected graduation: Spring 2026)
6. Martin Duclos (expected graduation: Spring 2026)
7. William Anderson (expected graduation: Spring 2026)
8. Derek True (expected graduation: Spring 2027)

Conference Organization

In 2022, I was part of the organizing committee for ACM Conference on Data and Application Security and Privacy (CODASPY). URL: `https://www.codaspy.org/2022/organizing-committee.html`

Invited Academic Participant to NSA CAE-R Special Topic Workshop on Generative AI Tools for Cybersecurity

I was invited to help shape the discourse on the topic of GenAI for Cybersecurity:

Special Topic Workshop on Generative AI Tools for Cybersecurity*

Moraine Valley Community College
Burr Ridge, Illinois

September 21, 2023

* Supported by the CAE CoP-R as part of the Research Symposium

Organizers: Dr. Benjamin Blakely, Argonne National Laboratory
Mr. Neil Fendley, Johns Hopkins Applied Physics Laboratory
Dr. Bradford Kline, National Security Agency
Dr. Agnes Chan, Northeastern University
Dr. Susanne Swetzel, Stevens Institute of Technology

Academic Participants

Dr. Dipankar Dasgupta
Department of Computer Science, University of Memphis
https://cs.memphis.edu/~dasgupta/
Dr. Dasgupta's pioneering research spans across computational intelligence during last 30 years, including various applications of *AI for security and security of AI*. He made significant contributions to build digital immunity and Generative AI to design survivable systems, and has +300 publications with over 20,500 citations as per Google Scholar.

Prof. Yingfei Dong
Dept. of Electrical and Computer Engineering, University of Hawaii
https://sites.google.com/hawaii.edu/yingfeidong/home
Yingfei works on IoT and control network S&P, unmanned aerial system (UAS) security, S&P in networked systems and applications. Today, generative AI tools post serious new threats to networked systems, such that many existing S&P methods may be easily compromised or disrupted. New research in this direction is urgently needed to address these emerging issues

Sanjay Goel
Department of Information Security and Digital Forensics, State University of New York at Albany
http://www.albany.edu/~goel
Sanjay Goal is actively working on understanding the threats to AI systems and auditability of AI systems, including: 1) misuse by hackers and criminals for gathering intelligence and facilitating attacks and perpetrating fraud, 2) attack vectors including, data poisoning, malformed input, and model tampering; 3) and various societal issues.

Berk Gulmezoglu
Department of Electrical and Computer Engineering, Iowa State University
https://www.ece.iastate.edu/bgulmez/
Berk Gulmezoglu works in hardware security, including computer architecture and side-channel attacks. On the defensive side, we develop new techniques relying on GenAI techniques to limit the side-channel leakage at the software and hardware level. We are currently developing new vulnerability analysis tools using ChatGPT and other GenAI models, which can detect leakages in cryptographic libraries.

Latifur Khan
Department of Computer Science, University of Texas at Dallas
www.utdallas.edu/~lkhan
We have integrated, extended, and applied our NSF-funded efforts to extract data from text about political and social conflict using generative AI (large language model (LLM)). Our research has led to the development of ConfliBERT, a domain-specific LLM publicly available at Hugging Face, trained on an expert-curated corpus of 33.7 GB about conflict and political violence.

Sudip Mittal
Department of Computer Science & Engineering, Mississippi State University
www.sudipmittal.com
Sudip Mittal's current research focus is to understand how GenAI can be used by adversaries to compromise organizational security.

Dinh C. Nguyen
Department of Electrical and Computer Engineering, The University of Alabama in Huntsville
https://sites.google.com/view/dinh-chi-nguyen/home
My research interests focus on network security with two main research areas, namely wireless security with federated machine learning and network authentication with blockchain.

Matthew Wright
Department of Cybersecurity, Rochester Institute of Technology
https://sites.google.com/site/matthewkwright/
Matt Wright is an expert in applying deep learning and usability to cybersecurity problems like detecting deepfakes, traffic analysis, and malware classification. He is spending a lot of time reading about generative AI, helping his Department faculty get ready for students using generative AI, and preparing a new course for Spring 2024 on "Generative AI in Cybersecurity."

Management of MSU's NSA Center of Academic Excellence (CAE) Designations in Cyber Defense, Cyber Operations, and Research

At MSU, I have been heavily involved in maintaining our 3 NSA CAE designations. MSU is 1 of 16 universities that has the prestigious CAE-CO designation and is 1 of 10 universities in USA that have all 3 CAE designations (see below).

- Undergraduate Education
- Graduate Education
- Computer Science
- Software Engineering
- Cyber Security
- Accelerated B.S./M.S. Program
- Student Organizations
- Academic Calendar
- Final Exam Schedule
- Speaker Seminar Series

Center of Academic Excellence Designations

The **National Security Agency (NSA)** and the **Department of Homeland Security (DHS)** jointly sponsor the **National Centers of Academic Excellence in Cyber Defense (CAE-CD)** program. The goal of the program is to reduce vulnerability in our national information infrastructure by promoting higher education and research in cyber defense and producing professionals with cyber defense expertise. NSA's CAE in **Cyber Operations (CAE-CO)** program supports the President's **National Initiative for Cybersecurity Education (NICE)**: Building a Digital Nation, and furthers the goal to broaden the pool of skilled workers capable of supporting a cyber-secure nation.

See the NICE One-Pager to learn more about the NSA/DHS CAE designations

MSU currently holds the following CAE designations:

- CAE in CD Education (CAE CDE),
- CAE in CD Research (CAE-R),
- CAE in Cyber Operations (CAE-CO)

MSU CAE Cyber Operations Program of Study

Institutions that receive a CAE-C designation have met the rigorous requirements set forth by the sponsor of the program, the National Security Agency (NSA). The NSA awards CAE-C designations to institutions that have committed to producing cybersecurity professionals that will reduce vulnerabilities in our national infrastructure. Upon graduation students who successfully complete all requirements for the NSA validated **MSU CAE Cyber Operations Program of Study** can request a certificate of completion for their records.

CAE Designation Point of Contacts

- Dr. George Trawick: **gtrawick@cse.msstate.edu**
- Dr. Sudip Mittal: **mittal@cse.msstate.edu**

Selected Publications

Contents

Overview

As part of this dossier, I have included five recent publications that represent my cybersecurity and AI research at Mississippi State University. My CV and Google Scholar page `https://scholar.google.com/citations?user=HxIicawAAAAJ` contain a comprehensive list of all my published research.

1. Morgan Reece, Theodore Lander, Sudip Mittal, Nidhi Rastogi, Josiah Dykstra, and Andy Sampson. "Defending Multi-Cloud Applications Against Man-in-the-Middle Attacks." In Proceedings of the 29th ACM Symposium on Access Control Models and Technologies (ACM SACMAT 2024), pp. 47–52, 2024.

Note: The paper exemplifies collaborative work led by PhD student Morgan Reece between MSU, the Rochester Institute of Technology (RIT), and the National Security Agency (NSA). The publication was supported by the NSA and NSF Scholarship for Service (SFS) grant #1565484.

2. Jesse Ables, Thomas Kirby, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, and Maria Seale. "Creating an explainable intrusion detection system using self-organizing maps." In 2022 IEEE Symposium Series on Computational Intelligence (IEEE SSCI), pp. 404–412,. IEEE, 2022.

Note: The paper was funded by a grant from U.S. Army Engineer Research and Development Center (ERDC). It encompasses research aimed at developing explainable intrusion detection systems, which are a top research focus in the field of cybersecurity.

3. Subash Neupane, Jesse Ables, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, and Maria Seale. "Explainable Intrusion Detection Systems (X-IDS): A survey of current methods, challenges, and opportunities." IEEE Access 10 (2022): 112392-112415.

Note: This popular paper is a survey on Explainable Intrusion Detection Systems (X-IDS). It describes how these systems are currently built, as well as research recommendations and opportunities for developing the next generation of X-IDS. The paper was funded by a grant from U.S. Army Engineer Research and Development Center (ERDC). This work served as the basis for PhD dissertation of Jesse Ables and William Anderson.

4. Aritran Piplai, Anantaa Kotal, Seyedreza Mohseni, Manas Gaur, Sudip Mittal, and Anupam Joshi. "Knowledge-enhanced neurosymbolic artificial intelligence for cybersecurity and privacy." IEEE Internet Computing 27, no. 5 (2023): 43–48.

Note: This paper includes a multi-university collaborative vision of how *neurosymbolic AI*, which integrates deep learning and symbolic AI architectures, can be utilized for cybersecurity.

5. Keith Strandell and Sudip Mittal. "Risks to zero trust in a federated mission partner environment." The Cyber Defense Review 8, no. 3 (2023), Army Cyber Institute: 89–98.

Note: This paper, published in the prestigious Cyber Defense Review by the US Army Cyber Institute, highlights the challenges of using zero-trust architectures when federating with non-US mission partners. The work serves as the basis of the PhD dissertation of Keith Strandell.

Selected Publication 1

Defending Multi-Cloud Applications Against Man-in-the-Middle Attacks

Morgan Reece
mlr687@msstate.edu
Mississippi State University
Mississippi State, MS, USA

Theodore Lander
tel127@msstate.edu
Mississippi State University
Mississippi State, MS, USA

Sudip Mittal
mittal@cse.msstate.edu
Mississippi State University
Mississippi State, MS, USA

Nidhi Rastogi
nxrvse@rit.edu
Rochester Institute of Technology
Rochester, NY, USA

Josiah Dykstra
josiahdykstra@acm.org
National Security Agency
Fort Meade, MD, USA

Andy Sampson
agsamps@uwe.nsa.gov
National Security Agency
Fort Meade, MD, USA

ABSTRACT

Multi-cloud applications have become ubiquitous in today's organizations. Multi-cloud applications are being deployed across cloud service provider platforms to deliver services to all aspects of business. With the expansive use of multi-cloud environments, security is at the forefront of concerns when deploying and managing access to multi-cloud applications and the expanded attack surface of these applications. Attackers can exploit vulnerabilities in multi-cloud environments that expose privileged information to inevitable attack.

In this paper we develop a multi-cloud victim web application deployed as component services. These services are deployed on different cloud service providers. Being deployed on the different cloud service providers expands the attack surface of the multi-cloud victim web application. Using the victim multi-cloud application, we demonstrate a man-in-the-middle attack showing the stealing of privileged credentials. Utilizing ParrotOS as the exploitation server, we demonstrate an attack on an application deployed across three cloud service providers: AWS, Azure, and Rackspace. Having successfully attacked the application, we then implement mitigations and verify the protection by attacking the protected application.

CCS CONCEPTS

• **Security and privacy** → **Access control**; *Multi-factor authentication*; *Web protocol security*; *Security protocols.*

KEYWORDS

multi-cloud; man-in-the-middle; ARP poisoning; identity security

ACM Reference Format:
Morgan Reece, Theodore Lander, Sudip Mittal, Nidhi Rastogi, Josiah Dykstra, and Andy Sampson. 2024. Defending Multi-Cloud Applications Against Man-in-the-Middle Attacks. In *Proceedings of the 29th ACM Symposium on Access Control Models and Technologies (SACMAT 2024), May 15–17, 2024, San Antonio, TX, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3649158.3657051

SACMAT 2024, May 15–17, 2024, San Antonio, TX, USA

ACM ISBN 979-8-4007-0491-8/24/05
https://doi.org/10.1145/3649158.3657051

1 INTRODUCTION

The proliferation of cloud-hosted applications continues to increase. While traditional cloud-hosted applications were designed to run as a *single-cloud service*, recently more applications are being deployed to different Cloud Service Providers (CSPs). Called *multi-cloud applications*, these software programs leverage resources and services from multiple cloud providers to fulfill their functionality. For example, organizations leverage *multi-cloud environment*, running applications like human resources on Amazon Web Services (AWS) and IT Service Management on Microsoft Azure [21]. Although this provides interconnectedness, interoperability, and data sharing, the complexity of the multi-cloud environment increases the vulnerability to various threats such as credential theft, privilege escalation, and man-in-the-middle attacks.

In this paper, we present a generic multi-cloud architecture for simulating and testing attacks and their mitigation techniques (see Figure 1). This architecture is implemented in our victim web application. The victim web application is comprised of a web service, an application service, an email service, and a database service, hosted across multiple cloud providers.

Using the victim web application based on the multi-cloud architecture outlined, we execute a Man-In-The-Middle (MITM) attack using ARP Poisoning, which allows the attacker to capture privileged credentials and compromise the database. The purpose of demonstrating the MITM attack is to show the stealing of privileged credentials and how security controls can mitigate vulnerabilities as well as maintain security in a multi-cloud application. Additionally, we show that security controls should encompass strong access control, network security, and encryption of traffic at multiple levels such as at the CSP, the application, and the APIs. It is worth noting that while encryption and access control pose challenges in multi-cloud applications due to differences in implementations by CSPs, using TLS and IAM best practices can significantly improve the security of multi-cloud applications. The main contributions of this paper include:

(1) We develop a multi-cloud-based victim web application deployed as component services on disparate cloud service providers. allowing for the expansion and exposure of the attack surface unique to the multi-cloud environment.
(2) We demonstrate a successful MITM attack that exploits ARP Poisoning in a multi-cloud environment, and bring attention

to the vulnerabilities and relevance of strong access control and network security.

(3) We propose and experimentally analyze a practical mitigation approach for the MITM attack in a multi-cloud environment. We focus on strong access control, network security, and encryption for a holistic security approach.

The rest of the paper is structured as follows. In Section 2, we dive into the background and existing research in multi-cloud attack security and vulnerability mitigation strategies. Next, in Section 3, we discuss the multi-cloud architecture and how it is used to build and deploy a multi-cloud-based victim web application for attack demonstration. In Section 4, we provide the details to execute a MITM attack. In Section 5, we present the mitigation strategies and finally, in Section 6, we summarize our research and suggest future work.

2 BACKGROUND & RELATED WORK

In this section, we first contrast multi-cloud applications with single-cloud applications. We then present the formal architecture of a generic multi-cloud application, APP_{MC}. This architecture serves as a basis for our victim web application (V_{webapp}, Section 3). In addition, we include a comprehensive examination of security issues in multi-cloud applications.

2.1 Single vs. Multi-Cloud Applications

A multi-cloud application is a software program that executes its functions by utilizing the resources and services of multiple cloud providers. Multi-cloud applications offer a range of advantages compared to single-cloud applications. These include cost optimization, improved performance and scalability, enhanced availability and resiliency, and the ability to avoid vendor lock-in [7]. By selecting the most economically viable services for various components of the application, multi-cloud can yield substantial cost reductions in comparison to a single-vendor approach. Different cloud providers demonstrate proficiency in distinct domains. Multi-cloud deployments enable the utilization of the unique advantages of each platform, resulting in enhanced performance. Deploying the application across multiple cloud providers mitigates dependence on a single platform, thereby improving resilience to failures and reducing the likelihood of downtime. With multi-cloud, the application is not constrained to the ecosystem of a single vendor. This affords greater autonomy in selecting services that align with system requirements and in negotiating favorable pricing [8].

Conversely, managing and optimizing a multi-cloud application can be complex, requiring skilled personnel and specialized tools [16]. Ensuring data security and access control across multiple cloud platforms with differing security standards necessitates careful implementation and execution [12]. Multiple contracts, billing systems, and support channels are often part of multi-cloud, which makes operational management difficult and time-consuming.

2.2 Multi-Cloud Application Architecture

A generic multi-cloud application APP_{MC}, shown in Figure 1, has multiple component services $\{s_1, s_2, s_3, ..., s_k\} \epsilon S$, hosted on different cloud platforms $\{c_1, c_2, c_3, ..., c_n\} \epsilon C$. Typically, a many-to-many map exists between the sets of cloud providers C and component services S.

The component services S, hosted by the cloud providers C, communicate with each other using Application Programming Interfaces (APIs) over the Internet. These multiple communication paths between component services can be represented as $\{p_{user}, p_{1,1}, p_{1,2}, ..., p_{i,j}, ..., p_{k,k}\} \epsilon P$, where $p_{i,j}$ is the communication path between component services s_i and s_j. p_{user} is the connection utilized by the user to communicate with APP_{MC}. Generally, multiple communication paths are programmed into APP_{MC} by its developers based on the multi-cloud application software requirements.

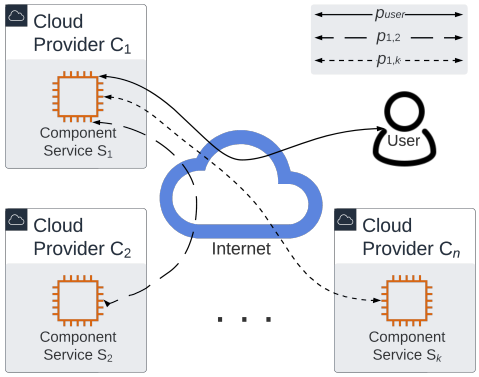


Figure 1: Multi-Cloud Architecture The implementation of an application across a multi-cloud environment can have its k number of component services, $\{s_1, s_2, s_3, ..., s_k\} \epsilon S$, distributed across n number of cloud providers, $\{c_1, c_2, c_3, ..., c_n\} \epsilon C$. Communication paths between component services are represented as $\{p_{1,1}, p_{1,2}, ..., p_{i,j}, ..., p_{k,k}\} \epsilon P$, where $p_{i,j}$ is the communication path between component services s_i and s_j. The user communicates with component service 1 over path p_{user}.

In practice, most of the deployed multi-cloud applications have three distinct types of component services. A *web service* manages user interaction through a website or a mobile application. Multiple *application services* process computational logic. *Data repositories and database services* store APP_{MC} data for manipulation and retrieval. The following subsection addresses the vulnerabilities that arise in this architecture as a result of its design decisions.

2.3 Security Concerns in Multi-Cloud Applications

In Section 2.1, we presented differences between single and multi-cloud applications. The architectural and implementation differences identified do not mitigate vulnerabilities; rather, they contribute to the vulnerabilities. Multi-cloud applications inherit all the security issues found in single-cloud environments [5]. In our efforts to find supporting literature for our research, we discovered that there is little research on multi-cloud risk and vulnerability analysis, let alone multi-cloud exploitation.

Reece et al. present a multi-cloud risk and vulnerability analysis model that integrates well-known industry standard frameworks

to produce a holistic risk model [21]. The current fragmented approach leads to gaps in vulnerability identification and mitigation, leaving security holes within the multi-cloud environment. The paper addresses the need for this integrated approach by using STRIDE, DREAD, and the MITRE ATT&CK frameworks to identify, qualify, and mitigate risks in the multi-cloud environment. The six attack vector categories that are identified and analyzed are cloud architecture, APIs, authentication, automation, management differences, and cybersecurity legislation.

Afolaranmi et al. present a method for enhancing the security of a multi-cloud environment by utilizing a security evaluation framework. The presence of buffer overflow and cross-site scripting attacks is emphasized, as they can have a significant impact on multi-cloud environments. To address this issue, the developed framework incorporates essential components, including operational and architectural perspectives [3].

Lingle et al. in their paper provide an overview of Security as a Service (SECaaS) [13]. The preliminary examination reveals the existing fragmented service offerings, such as logging as a service, IAMaaS, SOCaaS, DLP, and IPS/IDS services, which are included by each cloud service provider in their cloud offerings.

The subsequent stage of the analysis involves introducing a novel third-party SECaaS (Security as a Service) that utilizes innovative cloud security techniques.

IAM and its associated policies dictate the authorization process for individuals seeking access to cloud resources. Additionally, they represent a primary focus for potential attackers. This risk can be exacerbated by multi-cloud strategies, which increase complexity and make it more difficult to enforce a uniform IAM policy across multiple cloud environments [2]. Attackers can exploit weak IAM policies or poorly managed identities to gain unauthorized access to resources and data. For instance, a common attack is identity spoofing, where an attacker impersonates a legitimate user to bypass access control measures [15]. In multi-cloud environments, the risks are intensified due to the presence of individual Identity and Access Management (IAM) systems for each cloud platform, which can complicate the coordination of their administration.

Vulnerabilities in a multi-cloud environment stem from differences in how CSPs implement common technologies, management control schemes, and communication between component services deployed in difference CSPs [21]. MITM attacks target exploiting the expanded attack surface that is created when component services are deployed in different CSPs. The MITM attack can be characterized as a 'fishing' expedition where the attacker hopes to find privileged information in the traffic that they have captured while running the attack. In this paper, we exploit weaknesses in the APP_{MC} architecture. We next present our victim web application, upon which we execute a MITM attack.

3 MULTI-CLOUD BASED VICTIM WEB APPLICATION

To demonstrate the MITM attack in Section 4, we first describe a multi-cloud victim web application (V_{webapp}). V_{webapp} implements the multi-cloud architecture (Section 2.2). The component services of V_{webapp} serve distinct functions demonstrative of a multi-cloud application: *web service and application service*, *email service*, and *database service*. V_{webapp} setup leverages three cloud providers: Microsoft Azure, Amazon Web Services (AWS), and Rackspace (Figure 2).

V_{webapp} web service is a python-flask application that provides a user interface. It runs on a Linux server deployed in the Azure cloud [17]. The web service renders the web pages, takes the input from the user, and sends the data to the V_{webapp} application service for processing. The V_{webapp} application service written in Python receives user data from the web service. Once data is received, the application service executes the requested operation on the data. V_{webapp} application service communicates with two other services; database service and email service. V_{webapp} uses the Internet for users to connect and the communication needs of the different services hosted by the different cloud providers. User authentication is provided through username/password credentials communicated to V_{webapp} over HTTPS.

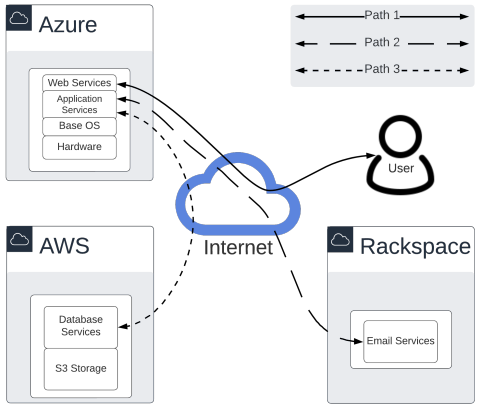


Figure 2: Multi-Cloud based Victim WebApp Utilizing three cloud service providers hosting four services. Communication between services takes place over the Internet via APIs.

The application service leverages APIs to communicate with the database, and the email services hosted on AWS and Rackspace respectively. The application service utilizes username and password credentials to authenticate via the database service API. The connection to the database service uses the standard MySQL port and API over the Internet. We use MySQL as our database service hosted on AWS [6].

The database service utilizes the AWS S3 storage system for its block storage of tables and data. The V_{webapp} application service communicates with the email service over the Internet by sending email messages using the Simple Mail Transport Protocol (SMTP) [11] and receiving email messages using the Internet Message Access Protocol (IMAP) [1] API. The email service is hosted by Rackspace [20], and connections to the email service are authenticated using credentials unique to the user sending or receiving the email.

V_{webapp} experimentation environment allows us to demonstrate attacks on the Internet-exposed interface between the application service and the database (see Section 4). Our V_{webapp} implementation, typical of multi-cloud deployment environments, provides us

with a representative attack environment. These systemic vulnerabilities are rooted in data-sharing requirements in multi-cloud environments [21]. When information is exchanged among component services that are hosted on distinct cloud providers, trust is required among the various component services of the victim web application. Through access control, the confidentiality, availability, and integrity of the information are validated, thereby establishing trust between the different component services. Trust is a prerequisite for information exchange among component services of V_{webapp} that are hosted on distinct cloud providers [18]. By implementing access control that verify the information's availability, confidentiality, and integrity, trust can be established among component services.

The interconnections and implementation details of the V_{webapp} deployment are also shown in Figure 2. The Internet-exposed data communications paths that are susceptible to attacks have been defined below:

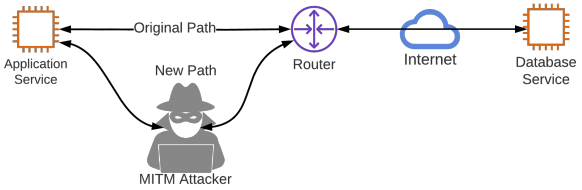


Figure 3: Attack Diagram: The MITM attack removes the original path between the application service and the router and replaces it with a path through the attacker.

- **Path 1 - User to/from Web Service**: The V_{webapp} web services run in Azure and receive the user input over the Internet (Figure 2). This is a common vector for many types of attacks, including DDoS, traffic sniffing, credential stealing, and cross-site scripting.
- **Path 2 - Application Service to/from Database Service**: The application service, which is also deployed on the Azure cloud server communicates with the AWS database service over the Internet through an API identified as *Path 2* in Figure 2. This path is susceptible to attacks like MITM attacks, substitution attacks, and privilege elevation.
- **Path 3 - Application Service to/from Email Service**: The application service utilizes *Path 3* to communicate email operation commands to the email service. Utilizing the information in the database service, the application service configures the authentication with the email service to allow the user to send, read, and manage their emails. *Path 3* utilizes the SMTP (Simple Mail Transfer Protocol) and the IMAP (Internet Message Access Protocol) API to communicate the email operations requested by the user through the user interface. A common attack on *Path 3* is packet sniffing, which includes credential stealing and other traffic interception or injection attacks.

4 MAN IN THE MIDDLE ATTACK DEMONSTRATION

A MITM attack was successfully executed on the V_{webapp} utilizing ARP Poisoning. The attack on the application service was launched from the attack server, which is located on the same subnet as the application service. The attack was executed on the path between the application service, hosted on Azure, and the database, hosted on AWS, along *Path 2* as seen in Figure 2. We used Ettercap [19] application running on a ParrotOS [14] Linux attack server to execute the ARP Poisoning attack and capture the re-routed network traffic with Wireshark. Ettercap sends ARP messages that replace the router's MAC address with the attack server's MAC address in the applications service's ARP table.

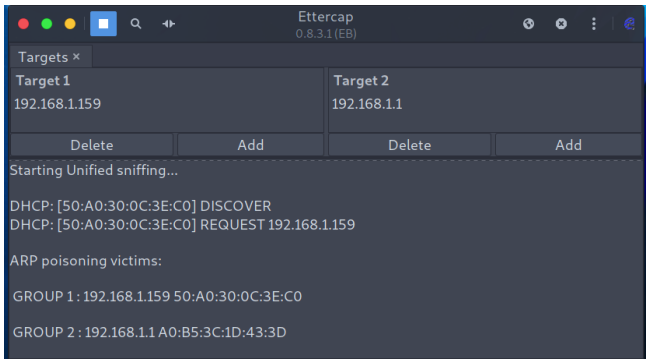


Figure 4: Ettercap ARP Poisoning: Using Ettercap, the attacker can send ARP commands to the application service and replace the MAC address of the router with the MAC address of the attack server.

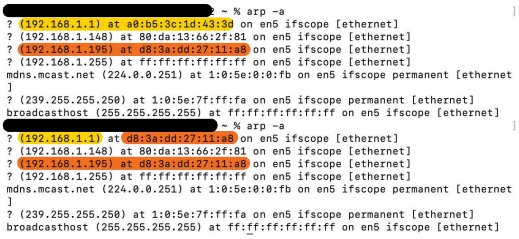


Figure 5: Attacker ARP Poisoning: Utilizes ARP protocol to update the ARP entry for the network router in the ARP table of the application service.

A MITM attack is a cyber-attack where an attacker intercepts and potentially alters the communication between two unsuspecting parties. In a multi-cloud environment, an attacker executes a MITM attack by inserting an attack server between two component services along their communication path (see Figure 3). The attacker utilizes ARP [10] as the initial attack vector. ARP is a network protocol that communicates Media Access Control (MAC) addresses of each device on a subnet to every other device on the subnet to allow proper delivery of Ethernet packets throughout that subnet. The MAC address is added to each Ethernet packet of network communication to ensure proper delivery of the packet. The ARP table on a network device, such as a server, contains the IP Address and associated MAC address for each device on the local subnet. ARP is a foundational protocol to Ethernet and has no inherent security to prevent attacks such as ARP Poisoning. ARP

Poisoning is the corruption of a service's ARP table. Using Ettercap on ParrotOS, the ARP Poisoning attack was executed on the application service by replacing the MAC address of the router with the MAC address of the attack server (see Figure 4). With the MAC address replaced, the attack server now receives all the traffic from the application service that is leaving the subnet, which includes communication to the database service (see Figure 5).

The attacker will do three things with the traffic from the application service: First, capture and store the traffic. Second, search the captured traffic for user credentials and other privileged information. Third, forward the traffic onto the router. Forwarding the traffic keeps the communication between the application service and database service active, and therefore prevents either service from realizing that there is a MITM attack occurring.

The execution of the MITM attack attack on the V_{webapp} follows these steps. The attacker compromises the V_{webapp} network by gaining control of a system on the same subnet. This is not covered in our research and is assumed that the attacker has compromised a system already on the network. The attacker installs the Ettercap application with the ARP Poisoning functionality. The attacker runs Ettercap and sets up the ARP Poisoning attack. The application service is set as Target 1 and the router for the subnet is set as Target 2. The ARP Poisoning attack is then initiated. The Ettercap application sends an ARP update to the application service with its own MAC address to be associated with the IP address of the router. In Figure 5, the "arp -a" command is run before and after the ARP Poisoning attack has begun. The first set shows the IP address of the router, 192.168.1.1, which has a different MAC address than the attack server, 192.168.1.195. The second set of outputs from the "arp -a" command shows that the router and the attack server have the same MAC address. With the ARP table poisoned, the application service traffic meant for the router is instead going to the attack server. Shown in Figure 3, the ARP Poisoning changes the communication path with the router from the "Original Path" to the "New Path".

With the communication path now going through the MITM Attacker, they can capture and inspect/analyze the packets from the application service. In the analysis of the application's network traffic, we can find the admin credentials for the database service (see Figure 6). The admin credentials to the database service allow our attacker to read all data in the database. After successfully retrieving data from the database, the attacker can then use the extracted information to gain unauthorized access to the email accounts stored in the database. The compromise of the email accounts is beyond the scope of our MITM attack, thus we did not exploit any of the users' email accounts.

A successful MITM attack leads to the interception of privileged information by the attacker. The compromise of sensitive information enables the attacker to exploit it for the purpose of advancing the attack or to trade or disclose the pilfered data.

5 MITIGATION AGAINST MITM ATTACKS

Mitigating MITM attacks in multi-cloud environments centers around strong access control and network security. Attackers very often deploy an ARP attack as the initial step to a MITM attack since, as mentioned in Section 4, ARP has no inherent security, and is therefore susceptible to ARP Poisoning with limited direct mitigation available.

Preventing ARP Poisoning can be done through a couple of techniques. Static ARP tables would prevent the attacker from being able to change the ARP table of the application service. However, this limits the dynamic nature of the subnet, and in large environments would require a large amount of manual effort by network engineers. Dynamic ARP Inspection, found on higher-end switches verifies all ARP updates, and discards updates that look malicious. This advanced feature is becoming more common on small business switches but requires a higher level of network administration effort and expertise.

If the techniques to prevent an ARP attack cannot be implemented, then we can deploy mitigations that limit the impact of a successful attack. Isolating each service to a subnet would help mitigate an ARP attack. The network would have to be configured to have one device per subnet, therefore ARP messages and updates/changes would not be sent to other devices. Also, with the isolation of each device, it would be easy to detect if an attacker was able to add a device to a subnet which would be removed from the network.

The mitigation against the MITM attack deployed in our V_{webapp} is encryption, where we encrypt all the network traffic to and from the application service. Several challenges come with deploying encryption in a multi-cloud environment. APP_{MC} are deployed across multiple CSPs. Each CSP has its own method of implementing and managing encryption. Strict coordination efforts must be put in place to ensure that the communication path between services is secure. The best practice noted in Scott et al. [23], is to always use TLS. Our V_{webapp} mitigation implemented TLSv1.2. TLS must be implemented on both sides of the communication path and should use the highest common version that each service can support. Our database service supported TLSv1.3, the latest TLS version. However, our application service only supported up to TLSv1.2, therefore, TLSv1.2 was implemented. To encrypt the path from the application service to the database service we enabled encryption AWS for our database instance. The counterpart in the application service is to include the SSL/TLS certificate bundle (us-east-2-bundle.pem) from AWS. The two sides of this configuration set up TLS encryption of the communication path between the application service and the database service. With encryption enabled, when our attacker

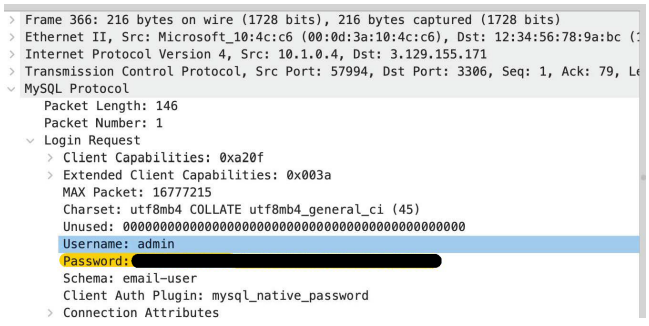


Figure 6: Attacker stolen credentials: Using Wireshark, the attacker is able to capture and inspect the network traffic to steal the user/administrator credentials.

captures the network traffic, they are unable to decipher the data that is being transferred between the component services.

Other mitigations noted in Scott et al. [23] target access control, such as implementing SSO and secure API key management. Utilization of a Single Sign-On (SSO) technology enables strong access control. SSO offloads the authentication of a user to a third-party identity provider that has implemented advanced authentication methods like Multi-Factor Authentication (MFA). Access control and IAM management policy differences between the CSPs create vulnerabilities that can be exploited by attackers, which were explained in Section 2.3. Without a third-party identity provider, we were not able to implement SSO. Implementation of API keys and their secure management is another mitigation that supports strong access control in the APP_{MC}. Because our V_{webapp} is integrating one application service and one database service, the increased security gained through the use of an API key over username/password combination is limited and therefore out of scope for our experimentation.

A significant development in access control is Self Sovereign Identity (SSI) [4, 9]. Authentication through SSI is dependent on the trustworthiness of a third-party verifier and the verification of the user's identity. Integrating SSI in the multi-cloud environment would enhance supervision of authentication across all component services and empower users with greater control over privileged information [22].

6 CONCLUSION & FUTURE WORK

The nature of a multi-cloud application being spread across multiple cloud providers opens it up to attacks. The challenge in multi-cloud architectures lies in securely integrating and accessing the distributed component services of the victim web application (V_{webapp}) discussed in our research. We conducted experiments that involved launching targeted attacks against V_{webapp}. These attacks were specifically focused on exploiting the limited or weak access controls in place during the inter-service communication process. The executed MITM attack showcased the attacker's capability to intercept and acquire the user's information in the absence of sufficient mitigations. The presence of limited mitigations increases the vulnerability of the V_{webapp}, potentially leading to more severe and damaging attacks. We also discussed the benefits of incorporating robust access controls, such as sophisticated user authentication and network security technology, to minimize the vulnerability to unauthorized access by potential attackers. Furthermore, we have described the efficacy of implementing mitigations in reducing the vulnerabilities inherent in such an environment. Additional research is necessary in this domain as a result of the emergence of new attacks carried out by the attackers. There is a need to investigate and understand how these new attacks are being deployed and leveraged by attackers; and for the development of targeted security strategies specific to these newly developed attacks.

ACKNOWLEDGMENTS

This research was supported by NSA H98230-21-1-0317, and National Science Foundation (NSF) grant #1565484.

REFERENCES

[1] B. Leiba A. Melnikov. 2021. Internet Message Access Protocol, RFC 9051. https://www.ietf.org/rfc/rfc9051.html
[2] Sandesh Achar. 2022. Cloud Computing Security for Multi-Cloud Service Providers: Controls and Techniques in our Modern Threat Landscape. *International Journal of Computer and Systems Engineering* 16, 9 (2022), 379–384.
[3] Samuel Olaiya Afolaranmi, Borja Ramis Ferrer, and Jose Luis Martinez Lastra. 2018. A Framework for Evaluating Security in Multi-Cloud Environments. , 3059–3066 pages. https://doi.org/10.1109/IECON.2018.8591454 ISSN: 2577-1647.
[4] Md. Rayhan Ahmed, A. K. M. Muzahidul Islam, Swakkhar Shatabda, and Salekul Islam. 2022. Blockchain-Based Identity Management System and Self-Sovereign Identity Ecosystem: A Comprehensive Survey. *IEEE Access* 10 (2022), 113436–113481. https://doi.org/10.1109/ACCESS.2022.3216643
[5] Mohammed A. AlZain, Eric Pardede, Ben Soh, and James A. Thom. 2012. Cloud Computing Security: From Single to Multi-clouds. In *2012 45th Hawaii International Conference on System Sciences*. IEEE, New York, NY, USA, 5490–5499. https://doi.org/10.1109/HICSS.2012.153
[6] Amazon. 2024. Amazon Relational Database Service. https://aws.amazon.com/rds/.
[7] M. G. Avram. 2014. Advantages and Challenges of Adopting Cloud Computing from an Enterprise Perspective. *Procedia Technology* 12 (2014). https://doi.org/10.1016/j.protcy.2013.12.525
[8] Ataollah Fatahi Baarzi, George Kesidis, Carlee Joe-Wong, and Mohammad Shahrad. 2021. On Merits and Viability of Multi-Cloud Serverless. In *Proceedings of the ACM Symposium on Cloud Computing* (Seattle, WA, USA) *(SoCC '21)*. Association for Computing Machinery, New York, NY, USA, 600–608. https://doi.org/10.1145/3472883.3487002
[9] Md Sadek Ferdous, Farida Chowdhury, and Madini O. Alassafi. 2019. In Search of Self-Sovereign Identity Leveraging Blockchain Technology. *IEEE Access* 7 (2019), 103059–103079. https://doi.org/10.1109/ACCESS.2019.2931173
[10] Internet Engineering Task Force. 1982. *An Ethernet Address Resolution Protocol: Or Converting Network Protocol Addresses to 48.bit Ethernet Address for Transmission on Ethernet Hardware*. Request for Comments RFC 826. Internet Engineering Task Force. https://doi.org/10.17487/RFC0826 Num Pages: 10.
[11] J. Klensin. 2008. Simple Mail Transfer Protocol, RFC 5321. https://www.rfc-editor.org/rfc/rfc5321.html
[12] P. Ravi Kumar, P. Herbert Raj, and P. Jelciana. 2018. Exploring Data Security Issues and Solutions in Cloud Computing. *Procedia Computer Science* 125 (2018), 691–697. https://doi.org/10.1016/j.procs.2017.12.089
[13] Jason Lingle, Kevin Dickens, Iram Bakhtiar, and James Herford. 2019. Security-as-a-Service in a Multi-Cloud Environment. https://doi.org/10.13140/RG.2.2.27812.12166
[14] Parrot Linux. 2024. Parrot Security. https://parrotlinux.org/. Accessed: 2023-10-05.
[15] Mohammad Masdari and Marzie Jalali. 2016. A survey and taxonomy of DoS attacks in cloud computing. *Security and Communication Networks* 9, 16 (2016).
[16] Lawrence E. Meyer and E. Billionniere. 2021. Upskilling to Meet Cloud Talent Needs. *2021 ASEE Virtual Annual Conference* null (2021), null. https://par.nsf.gov/biblio/10288742
[17] Microsoft. 2024. Azure Cloud. https://azure.microsoft.com/. Accessed: 202-01-12.
[18] NR Paul and D Paul Raj. 2021. Enhanced Trust Based Access Control for Multi-Cloud Environment. *Computers, Materials & Continua* 69, 3 (2021), 3079–3093.
[19] Ettercap Project. 2024. Ettercap. https://www.ettercap-project.org/. Accessed: 2023-11-12.
[20] Rackspace. 2024. Rackspace. https://www.rackspace.com/. Accessed: 2023-12-01.
[21] Morgan Reece, Theodore Edward Lander, Matthew Stoffolano, Andy Sampson, Josiah Dykstra, Sudip Mittal, and Nidhi Rastogi. 2023. Systemic Risk and Vulnerability Analysis of Multi-cloud Environments. arXiv:2306.01862 [cs.CR]
[22] Morgan Reece and Sudip Mittal. 2022. Self-Sovereign Identity in a World of Authentication: Architecture and Domain Usecases. arXiv:2209.11647 [cs.CR]
[23] Sam Scott and Graham Neray. 2021. Best practices for REST API security: Authentication and authorization - Stack Overflow. https://stackoverflow.blog/2021/10/06/best-practices-for-authentication-and-authorization-for-rest-apis/

Selected Publication 2

Creating an Explainable Intrusion Detection System Using Self Organizing Maps

Jesse Ables*, Thomas Kirby*, William Anderson*,
Sudip Mittal*, Shahram Rahimi*, Ioana Banicescu*, and Maria Seale†
* Department of Computer Science & Engineering
Mississippi State University, Mississippi, USA,
(email: {jha92, tmk169, wha41}@msstate.edu, {mittal, rahimi, ioana}@cse.msstate.edu)
† U.S Army Engineer Research and Development Center
Vicksburg, Mississippi, USA, (email: maria.a.seale@erdc.dren.mil)

2022 IEEE Symposium Series on Computational Intelligence (SSCI) | 978-1-6654-8768-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/SSCI51031.2022.10022255

***Abstract*—Modern Artificial Intelligence (AI) enabled Intrusion Detection Systems (IDS) are complex black boxes. This means that a security analyst will have little to no explanation or clarification on why an IDS model made a particular prediction. A potential solution to this problem is to research and develop Explainable Intrusion Detection Systems (X-IDS) based on current capabilities in Explainable Artificial Intelligence (XAI). In this paper, we create a novel X-IDS architecture featuring a Self Organizing Map (SOM) that is capable of producing explanatory visualizations. We leverage SOM's explainability to create both global and local explanations. An analyst can use global explanations to get a general idea of how a particular IDS model computes predictions. Local explanations are generated for individual datapoints to explain why a certain prediction value was computed. Furthermore, our SOM based X-IDS was evaluated on both explanation generation and traditional accuracy tests using the NSL-KDD and the CIC-IDS-2017 datasets. This focus on explainability along with building an accurate IDS sets us apart from other studies.**

I. Introduction

The use of Artificial Intelligence (AI) in cyber-defense solutions, particularly Intrusion Detection Systems (IDS), has been gaining traction to protect against a wide range of cyber attacks. While these AI models have high detection rates, high false positive and false negative rates can dissuade a security analyst from using an AI enabled IDS [1]. These IDS built using AI and deep learning methods are black boxes, meaning a security analyst will have little to no explanations and clarifications on why a model made a particular prediction. With the rise in cyber attacks on critical infrastructure, government organizations, and business networks, there is a pressing need for an explainable, automated detection system that can provide real-time aid to an analyst.

Intrusion Detection Systems are generally utilized as part of a larger cybersecurity defense effort at an organization generally located in a Cyber-Security Operations Center (CSoC). These systems monitor networks and automate attack detection by comparing network activity to the signature of known attacks or by detecting behavior that is anomalous to benign network patterns [2]. Through these methods, a security analyst can use an IDS to detect improper use, unauthorized access, or the abuse of a network. Analysts can then create mitigating strategies to minimize damages and costs of the malicious behavior. The usefulness and cost effectiveness of IDS have therefore been the subject of much research [3], [4].

Previous work in AI enabled IDS has generally focused on improving detection rates while limiting false positives and false negatives. These techniques have been effective at achieving high detection rate, but have failed to provide explanations for their computed predictions. Without the ability to understand how a model reached a decision and which features were relevant to the decision computation, a security analyst will give less credence to these AI enabled IDS. Opaque Deep Learning methods in particular, can be considered as black boxes providing no explanations and feature relevance information, severely limiting adoption in real world cyber-defense scenarios [5].

A potential solution to this problem is to research and develop Explainable Intrusion Detection Systems (X-IDS) based on current capabilities in Explainable Artificial Intelligence (XAI) [6]. The guidelines proposed by the Defense Advanced Research Projects Agency (DARPA) indicate that to be explainable, an AI should explain the reasoning for its decisions, characterize its strengths and weaknesses, and convey a sense of its future behavior [7]. An X-IDS that is transparent in its behavior and decision making process, will empower a security analyst to make better informed actions, understand the feature composition of a prediction, help CSoCs defend from known attacks, and quickly understand zero-day attacks. To address this need, we create an X-IDS using Self Organizing Maps (SOMs).

Data collected from modern networks contain potentially hundreds of different features about the traffic flow, operating systems, network protocols, and other metadata. SOMs work by representing this high dimensional data on a 2-dimensional plane. This also includes maintaining the topographical relationship of the data by grouping similar data [8]. Through this dimensional reduction and various other SOM visualization techniques, a security analyst can view both global and local explanations about a potential attack rather than an opaque prediction generated by a black box model.

As the need for explainable cyber-defense systems increases and to address the lack of XAI research in the field of IDS, the main objective of this paper is to demonstrate the

978-1-6654-8768-9/22/$31.00 ©2022 IEEE

explainability of the SOM based X-IDS rather than creating the most accurate system. Higher accuracy systems can be developed by using complex derivative architectures. However, further research is necessary to make them explainable. Our goal in this paper is to increase trust in IDS and help CSoCs defend from attack through the use of explainable insights. As a secondary focus, we also provide the accuracy scores of our SOM based X-IDS system trained on the NSL-KDD and CIC-IDS-2017 datasets.

Major contributions presented in this paper are -

- A novel X-IDS architecture featuring a SOM, built using DARPA's proposed guidelines for an explainable system. This system is able to produce robust, explanatory visualizations of the SOM model and create accurate IDS predictions.
- A Local and Global explainability analysis using the SOM explainable architecture. The explanation module creates a collection of explainable visualizations that can be used by a security analyst to understand predictions.
- A performative analysis using NSL-KDD and CIC-IDS-2017. The SOM based model is able to achieve accuracies as high as 91% on NSL-KDD and 80% on CIC-IDS-2017 datasets.

The rest of the paper is outlined as follows - In Section II, we discuss some related work on IDS, XAI, and X-IDS. Section III briefly describes the SOM algorithm and how it can be used to achieve explainability. Section IV, outlines our SOM based X-IDS with its architecture presented in Figure 1. Section V lists our experimental results. Finally, the conclusion and future work has been presented in Section VI.

II. Related Work

In this section, we present some related work on Intrusion Detection Systems (IDS), Explainable Artificial Intelligence (XAI), and Explainable Intrusion Detection Systems (X-IDS).

A. Intrusion Detection Systems (IDS)

An intrusion refers to an action that obtains unauthorized access to a network or system [9]. Intrusions can be characterized by a violation of Confidentiality, Integrity, or Availability (CIA). An IDS consists of tools, methods, and resources that help a CSoC protect an organization [10], [11].

IDS can be classified as either a host-based IDS or network-based IDS. Host-based IDS are placed on a host system and monitor host activity, incoming and outgoing network traffic [12]. Network-based IDS are built to survey and protect a network of hosts from intrusion [13]. In addition, IDS can also be categorized into operation-based classes, such as signature, anomaly, and hybrid. Signature-based IDS operate by preventing known attacks from accessing a network. The IDS compares incoming network traffic to a database of known attack signatures. Notably, this method has difficulty in preventing *zero-day* attacks [14]. Anomaly-based IDS look for patterns in incoming traffic to recognize potential threats and leverage complex AI models [15], [16]. A significant drawback of this approach is the the tendency for such systems to categorize legitimate, unseen behavior as anomalous. Hybrid-based IDS incorporate the design philosophy of both signature-based and intrusion-based IDS to improve the detection rate while minimizing false positives [17], [18].

Current AI enabled anomaly-based IDS can be further divided into black box and white box models. White box models are considered *easy to understand* by an expert. This allows the expert to analyze the decision process and understand how the model renders its decision. This "semi-transparent" property allows white box models to be deployed in decision sensitive domains, where auditing the decision process is a requirement. White box models may use regression-based approaches [19], decision trees [20], and SOMs [21], [22]. Black box models, on the other hand, have an opaque decision process. This opaqueness property makes establishing the relationship between inputs and the decision difficult, if not outright impossible. Black box models comprise nearly all the AI enabled state-of-the-art approaches for IDS, as the focus is traditionally on model performance, not explainability. Examples of black box models are Isolation Forest [23], One-Class SVM [24], and Neural Networks [25].

B. Explainable Artificial Intelligence (XAI)

As previously stated, state-of-the-art approaches for IDS, as well as machine learning as a whole, focus on model performance through the lens of model accuracy. This focus on model accuracy has driven the development further away from modeling approaches that are transparent or have methods of explainability. In turn, this creates a separation between model inference and *understanding* model inference, which gives the inability to confirm model fairness, privacy, reliability, causality, and ultimately trust.

The notion of XAI dates back to the 1970s. Moore et al. [26] surveyed works from the 1970s to the 1980s, detailing early methods of explanations. Some early explanations consisted of canned text and code translations, such as the 1974 explainer MYCIN [27]. We can find a more current definition of XAI by DARPA [7]. They define XAI as systems that are able to explain their reasoning to a human user, characterize their strengths and weaknesses, and convey a sense of their future behavior. In turn, the system offers some form of justification for its action, leading to more trust and understanding of the system. The explanations from an XAI system help the user not only in using and maintaining the AI model, but also helping users complete tasks in parallel with the AI system. Tasks can include doctors making medical decisions [27]–[29], credit score decisions [30], detecting counterfeit banknotes [31] or CSoC operators defending a network [7], [32].

C. Explainable Intrusion Detection Systems (X-IDS)

Explainable Intrusion Detection Systems (X-IDS) are still an emerging sub-genre in the field. The need for explainability in IDS is becoming increasingly necessary both to warrant further application in decision sensitive domains, as well as to supplement and empower existing knowledge techniques (e.g. data mining, rule-based development) that black boxes

obfuscate. The users need to be confident in the predictions or recommendations computed by an IDS. Understandable explanations allow users to perform their tasks correctly. The stakeholders of an IDS (e.g. CSoC operators, developers, and investors) are individuals who will be dependent on the performance of the system. CSoC operators will be performing defensive actions based on prediction and explanation results. Developers can use explanations to fortify the model in areas where it is weak. Investors may need explanations to help them in making budgeting decisions for their company.

The current literature consists of many different black box and white box models being used alongside explanation techniques. Common explainer modules for black box models are Local Interpretable Model-agnostic Explanations (LIME) [33], SHapley Additive exPlanations (SHAP) [34], and Layer-wise Relevance Propagation (LRP) [35]. Modern techniques for explaining black box models consist of creating surrogate models that generate explanations either locally or globally. Other methods involve propagating predictions backwards in a Neural Network or decomposing a gradient. More novel approaches have also experimented with making datasets explainable [36] or making GUIs for explainable systems [37].

III. Explainable Self Organizing Maps

In this section, we briefly describe the theoretical and the practical aspects of SOMs and how they can be utilized to detect intrusions and generate explanations.

A. Self Organizing Maps (SOMs)

Self Organizing Maps (SOMs), sometimes referred to as Kohonen Maps [8], [38], Kohonen Self Organizing Maps [39], or Kohonen Networks [40], are a class of unsupervised machine learning algorithms. SOMs are comprised of a network of individual units, each of which has a feature vector of the same size as the dimension of training data. Some implementations also include an *(x,y)* coordinate to allow unit movement in a two-dimensional (2D) space. This 2D space is typically represented as a square or a hexagonal grid, to more easily visualize the represented space.

Training a SOM model, outlined in Algorithm 1, utilizes the following steps: First, the SOM is initialized with n rows and m columns. New nodes are allocated an N element array based on the number of features of a chosen dataset. These values are randomly chosen between 0 and 1. The next phase of the algorithm begins by picking a random training sample. Then, the Best Matching Unit (BMU) is calculated by finding the smallest euclidean distance from the training sample to a SOM unit. After the BMU is found, it and its neighbors are updated using the formula $w_i = w_i - \lambda * (w_i - i_i)$, where w is the set of BMU weights and i is the set of feature values. λ is the learning rate function that considers the current training iteration and distance from the BMU. Lastly, the learning rate, neighborhood radius, and current iteration numbers are updated. The function λ works in a way that it decreases during the course of the training process. This process is done over T epochs.

Algorithm 1 SOM Algorithm where *n* and *m* are the number of columns and rows, *N* is the number of features the dataset uses, *T* is the total number of training epochs, and *W* is the output weights of the SOM

Input: n, m, N, T
Output: W
1: Allocate n * m element array W
2: **for** each node in W **do**
3: Allocate N element array with random values [0,1]
4: **end for**
5: **while** $t < T$ **do**
6: Pick a training sample
7: Find Best Matching Unit using Euclidean Distance
8: Update BMU elements: $w_i = w_i - \lambda * (w_i - i_i)$
9: Update BMU Neighbors
10: Update Learning Rate and Radius
11: $t+ = 1$
12: **end while**
13: **return** W

SOMs have some unique advantages that come with their application. The first is algorithmic simplicity. As shown in Algorithm 1, the brevity of the algorithm helps to maintain the desired properties of algorithmic decomposability and tractability. Additionally, due to its unsupervised nature, SOMs can work on a variety of datasets and applications (e.g. data mining and discovery), not just prediction [41]. By design, SOMs convert high-dimensional data into a lower dimensional representation. This representation can be topologically clustered and explained through visualizations [42]. One challenge that comes with the application of SOMs is the selection of the size parameter, as the size does not dynamically adjust and there is no *best size* heuristic [43]. Finally, another challenge with SOMs is their scalability, both in their time complexity, $O(N^2)$ [44], and space complexity. More methods, such as those in [45], are needed to address these challenges.

B. SOMs and Intrusion Detection

In the past, SOMs have been used to create many IDS. These studies focused on building accurate IDS and did not discuss explainability, unlike our work. Among these approaches, SOMs were used to create both host-based [46] and network-based [42], [47], [48] IDS. The majority of these methods simply trained a SOM based IDS and illustrated mappings between datapoints and the associated BMU. The approaches in [48], [49] use multiple SOMs in conjunction with one another to create a more effective IDS. Only one approach [47] discussed the false positive rate and accuracy of a SOM-based IDS. Their method for prediction involved assigning a label to BMUs based on the training dataset. Using this approach meant that not all SOM units were assigned a label. The authors utilized Gaussian Mixture Modeling (GMM) to make predictions when a testing sample was similar to an unlabeled unit.

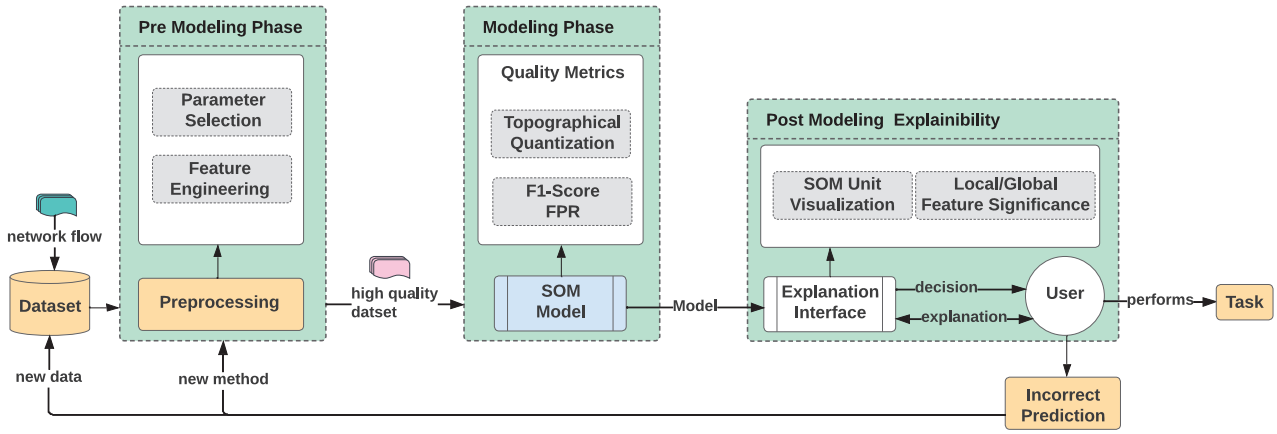


Fig. 1: Architecture for an Explainable Intrusion Detection System (X-IDS) utilizing Self Organizing Maps (SOMs), based on DARPA's recommended architecture for Explainable Artificial Intelligence (XAI) systems [7].

C. SOMs and Explainability

Once trained, SOMs are able to illustrate mappings between datapoints and the associated BMU. As this is generally a 2D representation of the feature space, it can be visually understood by the user. This advantageous SOM property makes them *explainable*. SOM's explainablity can be divided into *Global* and *Local* explanations.

Global explanations are used to give a general idea of how a particular model computes predictions. The U-Matrix (See Section IV-C), is a commonly used technique [50]. This additive metric works by summing the distance to each of a unit's neighbors. A group of low scores represents clusters in the map, while a group of high scores signifies sparseness. The Starburst U-Matrix is an updated variation of the U-Matrix. This updated version helps visualize cluster sizes and locations on the map. Other clustering representations, like K-means clustering, can also aid in visualization. For more fine-grained data, feature heat-maps can be created to visualize SOM feature values and their importance. These techniques provide global explanations.

Local interpretations of data are a more recent development for explaining SOMs. These explanations are generated for individual sample datapoints and are used to explain why a certain prediction value was computed. This allows the user to understand the decision process of the SOM model. The primary use of this method is to explain and visualize feature importance. When making a prediction, a datapoint's features are scored based on how impactful they are to the computed prediction. Wickramasinghe et al. [51] developed an explainable SOM for Cyber-Physical Systems. Their system created both local and global explanations by data-mining a SOM model. The mined information was used to create visualizations including histograms, T-distributed Stochastic Neighbor Embedding (t-SNE) [52], heat maps, U-Matrix, component planes, and U-Map. This variety of visualizations allow the SOM to be explainable not only to domain experts, but also non-domain experts.

IV. X-IDS Architecture

An X-IDS's main goal is to help stakeholders protect their networks and understand various relevant events. The system should act as both a guard and adviser for network security. When an IDS discovers an intrusion, the user should be notified to prevent a possible attack. Explanations generated by the X-IDS should assist CSoC operators in their mission to protect their organization. To help achieve this goal, we propose the novel, proof of concept SOM based X-IDS architecture in Figure IV. The proposed architecture is based on DARPA's recommended architecture for XAI systems [7]. Components of the framework can be changed to suit users' needs. The architecture is abstract enough that methods other than SOMs can be interchanged to create different X-IDS. The architecture consists of three stages: pre modeling, modeling, and post modeling explainability. In the first phase, the model preprocesses raw network data captured into high quality datasets, and selects parameters for the SOM model. Next, the model is trained during the modeling phase. Metrics are recorded to determine the newly trained model's quality. Lastly, in the post modelling phase, the SOM is data-mined to generate explanatory visualizations that allow users to understand how predictions are generated. In the next subsections, we describe each of these phases in detail.

A. Datasets and Pre Modeling Phase

In this work, NSL-KDD [53] and CIC-IDS-2017 [54] were used to test the explainability and effectiveness of our architecture. NSL-KDD was chosen because of its wide use in the literature. It allows our method to be compared to other existing IDS. CIC-IDS-2017 includes more modern attacks and is useful for testing an unbalanced dataset. The datasets are passed through a preprocessing module that normalizes the data. Additionally, the architecture uses Bayesian Probability

of Significance [55] to select features. Any feature significance value over a designer selected threshold is included in the preprocessed dataset. The resulting high quality dataset is used during the modelling phase.

B. Modeling Phase

The modeling phase begins by training the SOM model using the high quality dataset generated during the pre modeling phase. For this paper, we use the POPSOM implementation [56]. Training parameters include total training iterations, learning rate, and SOM size. At the end of the training session, the model will be tested to produce topographical error, quantization error, F1-score, precision, recall, and a confusion matrix. The confusion matrix can be used to determine both the false positive and false negative rate for the model. The quality metrics are used to determine if a model has been sufficiently trained to generate explanations.

Quality Metrics: There have been various metrics and measures proposed to evaluate the quality of a trained SOM. These include quantization error, topographic accuracy, embedding accuracy, and convergence index. Quantization error was used by Kohenen [57], and measures the average distance between nodes and the data points. To measure how much the features of the input space have been preserved in low dimensional output space, a topographic error is used. The topographic error is measured by evaluating how often the BMU and second BMU are next to each other [43], [58]. Map embedding accuracy is similar to quantization error and it measures how similar the distribution of the input data is with respect to that of the SOM units [59]. In order to measure both topographic preservation and distribution similarity between the input and SOM units, the convergence index was proposed to be a measure that linearly combines map embedding accuracy and topographic accuracy [60]. Prediction accuracy metrics are also important to include in an IDS architecture. These metrics include F1-score, false positive rate, and false negative rate. These measurements allow the architecture to be compared to other existing IDS.

C. Post Modeling Explainability

Once the modeling phase has been completed and quality metrics have ensured that the model is a good representation of the data, the model can be used to perform a variety of explainability and visualization tasks. The model itself is a list of SOM units and the weights associated with these units. Visualization tasks include creating local and global explanations, a U-Matrix, and feature heatmaps.

1) Local and Global Explanations: Global and local interpretability can be achieved by examining important features of the trained SOM, and utilizing this information to generate an explanation for a specific data instance classification or cluster classification [61]. Global significance for NSL-KDD is shown in Figure 2b with higher values denoting that a feature has a higher probability of being important. Higher variance features increase the probability that a model will capture the dataset's structure [55]. Through this graph, an analyst can understand which features are important to the overall SOM structure, allowing them to examine predictions at a local level based on globally important features.

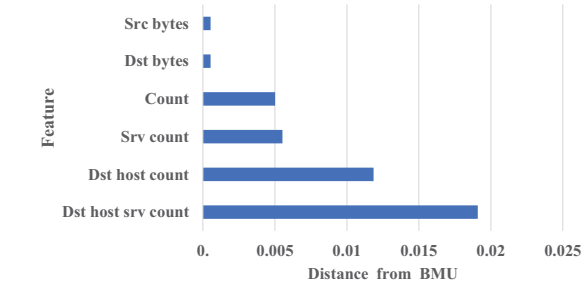
Figure 2a shows the local explanations for a prediction, where each feature on the y-axis has a value representing distance from its respective BMU value (See Section III). In this example, we can see the features with the largest impact on a prediction: duration, dst bytes, and src bytes. These features were the closest to the BMU, and they played a large role in computing the predicted value. Seeing the specific features that influence predictions provides insight about samples labeled as malicious or benign and can further help operators determine the reason of incorrect predictions. These features can also be further investigated with feature value heat maps.

2) Unified Distance Matrix (U-Matrix): The U-Matrix is a visualization of the distances between neighboring SOM units. With distances shown as a color gradient, units far apart will create dark boundaries while areas with similar units will be lighter. This can visually represent the natural clusters of input data. To enhance the standard U-Matrix, the starburst model uses connected component lines of nodes overlaid on the matrix to better represent clusters [62]. For a labeled data set, the user is able to visualize each BMU along with the associated label. Figure 3a shows clear clusters with boundaries separating malicious (1) and benign (0) behavior. Using this information a security analyst can investigate more visualizations and feature importance values to gain an understanding about why certain malicious network activities are being grouped together.

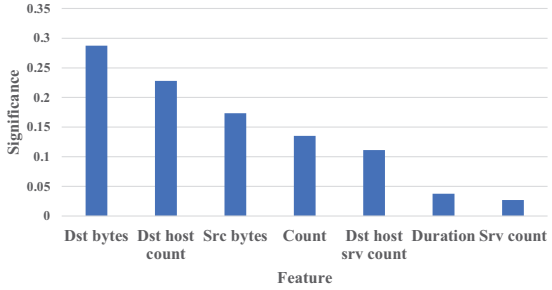
3) Feature Value Heat Map: Heat maps show general trends that a feature has on the entire SOM model. SOM feature values are represented from 0 to 1, and the heat maps denote this with darker and lighter values, respectively. An example feature value heat map can be found in Figure 3c. In this example, the 'dst bytes' features has a cluster of higher values in the bottom-left corner, while the rest of the SOM consists of lower values. Users can use this information to form conclusions about the model. Feature value maps are more powerful when multiple are viewed at a time. In addition, the U-Matrix or K-means clustering charts can then be referenced to make general decisions about the model. The heat maps work well as a fine-grained global explanation that helps users to understand the overall model.

V. Experimental Results and Evaluation

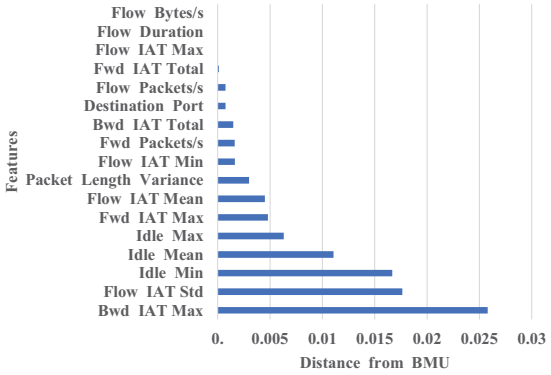
Our SOM based X-IDS was evaluated on both *explanation generation* and *traditional accuracy tests*. Experiments were run using the aforementioned datasets (See Section IV-A), which was used to train two 18x18 SOMs. The training process was completed in 1000 iterations over the dataset. After 1000 iterations, there was no significant improvement in evaluation metrics. In fact, while training on the CIC-IDS-2017 dataset, the SOM model performance began to degrade as a result of over-fitting.



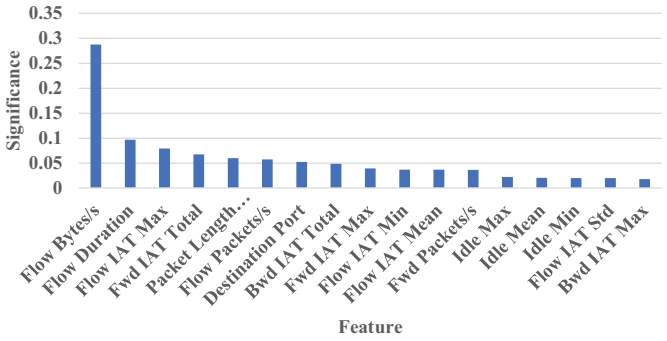
(a) NSL-KDD Local Anomaly Explanation



(b) NSL-KDD Global Feature Significance



(c) CIC-IDS-2017 Local Normal Explanation



(d) CIC-IDS-2017 Global Feature Significance

Fig. 2: These figures show the local and global feature explanations for both the NSL-KDD and CIC-IDS datasets. (a) The local explainability of a prediction is defined by the distance between feature value and BMU. More important features have a lower score than less important features. This figure shows the feature importance for an anomalous sample from the NSL-KDD testing set. (b) Global feature significance is calculated using Bayesian Probability of Significance [55]. Higher values are considered more important than lower values.

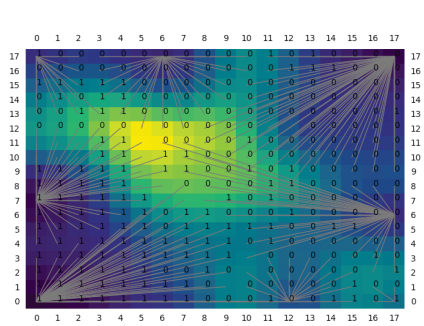
A. Model Explainability

The results for the NSL-KDD dataset can be found in Figures 2a and 2b. The local explanation example shows that the most important features for its prediction were 'Duration', 'Destination (dst) bytes', and 'Source (src) bytes'. The remaining features, 'Service (srv) count', 'Count', and 'Destination (dst) host count' are considered less significant because of their distance from the BMU. Two of the important features coincide with the Global Feature Significance graph. This trend continues when testing on many different test samples. The most important global features are frequently at the forefront for local significance. Similarly to NSL-KDD, CIC-IDS-2017 follows this trend. Many of the top, globally selected features also play a more important role in the local predictions.

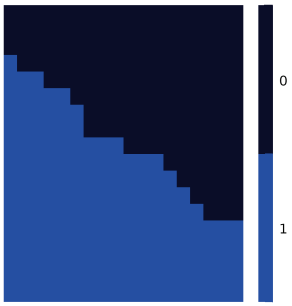
The next set of explainability techniques has been data-mined from the trained SOM. Figures 3a and 3d show the generated Starburst U-Matrix for NSL-KDD and CIC-IDS-2017, respectively. The SOM algorithm was able to separate benign clusters from malicious clusters in the map created from NSL-KDD dataset. The bottom-left corner is primarily malicious samples, while the top-right corner contains mostly benign samples. Additionally, the clusters marked by the starbursts' origins mostly represent one label. On the other hand, the CIC-IDS-2017 map has not separated the labels sufficiently. Most of the labels present in the figure are benign (0) with very few malicious labels (1). CIC-IDS-2017 is an unbalanced dataset, with about 70% of samples being benign and 30% of samples as malicious. This class imbalance causes the SOM to be trained on more benign samples than malicious.

For a simplified label separation, users can visualize a K-means clustering interpretation in Figure 3b. This figure helps to explain which clusters the benign and malicious traffic are grouped in. The NSL-KDD K-means graph is similar to the computed U-matrix. However, the CIC-IDS-2017 K-means cluster graph was unable to form accurate clusters. As mentioned above, there were few units that were labeled malicious (1), and the K-means clustering algorithm chosen was unable to create a meaningful separation.

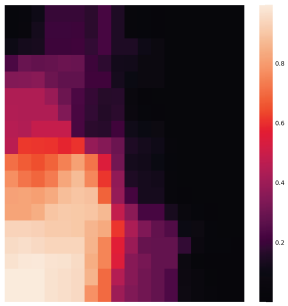
Lastly, the feature value heatmaps are generated for each feature of the dataset. The examples chosen were the most significant features for each dataset: 'destination (dst) bytes' and 'flow bytes/s'. On their own, they can be used to see general training trends for each feature. In Figures 3c and 3e,



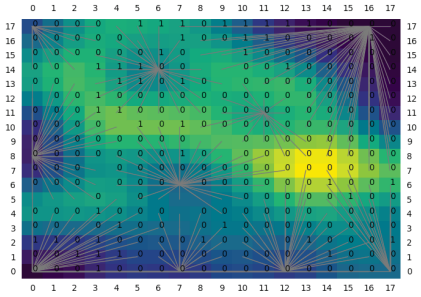
(a) NSL-KDD Starburst U-Matrix



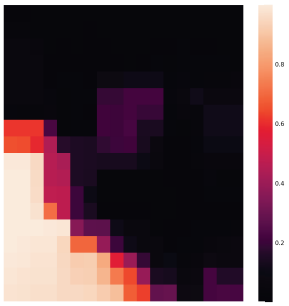
(b) NSL-KDD K-means Clustering Map



(c) Dst byte Feature Map



(d) CIC-IDS-2017 Starburst U-Matrix



(e) Flow bytes/s Feature Map

Fig. 3: (a)(d)The Starburst U-Matrix shows both the most common label for each node and the clusters the SOM has learned. Darker areas represent units that are close Euclidean Distance-wise. Notably, we can see a clear divide between classes on the NSL-KDD dataset as represented in the figure. (b)(e) K-means clustering can be used as a simplified view of where labels appear on the SOM. In this model's iteration, anomalous traffic is mostly grouped on the bottom of the SOM.(c) The feature value heatmap displays the value of a specific feature on each unit in the SOM. Lighter values represent units with values closer to 1, while darker values show values closer to 0. The 'dst byte' example shows that the bottom 'anomalous' cluster values higher values.

we can see that each of these features have higher values in the bottom-left units and lower values elsewhere. Users will be able to build a mental model of the SOM when visualized in conjunction with the features maps. For example, 'destination (dst) byte', 'duration', and 'source (src) byte' all have higher values in the malicious section of the map. One may conclude that when these values are all close to one, the sample is more than likely malicious.

Dataset	F1	Precision	Recall	FPR	FNR
NSL-KDD	91.0%	91.0%	91.4%	9.4%	8.0%
CIC-IDS-2017	80.0%	77.4%	81.8%	22.5%	4.5%

TABLE I: Accuracy Metrics for NSL-KDD and CIC-IDS-2017

Dataset	SOM	Random Forest	DBN
NSL-KDD	91.0%	99.67%	97.5%
CIC-IDS-2017	80.0%	97.1%	94.0%

TABLE II: Accuracy comparison between different algorithms tested on NSL-KDD and CIC-IDS-2017 [63]–[66]

B. Accuracy

When creating an IDS, accuracy is an important metric to consider. Table I shows the accuracy metrics obtained for both the NSL-KDD and CIC-IDS-2017 datasets. The accuracy of the NSL-KDD evaluation can be attributed to the separation of benign and malicious traffic, as mentioned above. The accuracy of CIC-IDS-2017, however, is much lower. The U-matrix shows that not many units are labeled as malicious. Interestingly, only 14% of the units are labeled as malicious, which means that 77.4% of malicious samples are similar to

that small subset of units. Additionally, Table II compares the SOM's accuracy to Random Forest and Deep Belief Networks (DBN). Here we can see that the SOM performs between 7% to 14% worse than the black box models.

The results from the explainability and accuracy experiments show that it is possible to create explainable and relatively accurate SOM based X-IDS. The visualization techniques used can give users an understanding of the model and can empower security analysts to make their own predictions, similar to the model. A 91% F1-score can be attributed to the clear separation the model makes between malicious and benign samples. When compared to more complex algorithms like Random Forest and DBN, SOM performs worse. However, SOMs are far more explainable and easier to understand. We believe that the accuracy and explainability can be further improved with more complex SOM algorithms.

VI. Conclusion and Future Work

In this paper, we created a novel, proof of concept SOM based X-IDS implementation. The implementation was able to produce robust, explainable figures describing the SOM model. Explainability was demonstrated using various forms of visualization including feature significance, U-matrices, and feature heatmaps. Through these, users are able to create their own conclusions about how the model works and makes predictions. Additionally, accuracy was tested using the NSL-KDD and CIC-IDS-2017 datasets. The SOM implementation was able to achieve accuracies of 91% and 80%, respectively. Potential future works will include analysing the explainability and accuracy of Growing SOMs or Growing Hierarchical SOMs. These studies will aim to increase the accuracy of SOMs while simultaneously decreasing false positives and false negatives. In addition, explanations can be improved by surveying security analysts to discover the most useful visualizations and feedback. We will continue to use and improve our architecture to create the state-of-the art in X-IDS.

VII. Acknowledgment

This work by Mississippi State University was financially supported by the U.S. Department of Defense (DoD) High Performance Computing Modernization Program, through the US Army Engineering Research and Develop Center (ERDC) (#W912HZ-21-C0058). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Army ERDC or the U.S. DoD.

References

[1] Alaa Marshan. Artificial intelligence: Explainability, ethical issues and bias. *Annals of Robotics and Automation*, pages 034–037, 08 2021.

[2] Raytheon. Cyber security operations center (csoc), 2017.

[3] Mustapha Belouch, Salah El Hadaj, and Mohamed Idhammad. Performance evaluation of intrusion detection based on machine learning using apache spark. *Procedia Computer Science*, 127:1–6, 2018.

[4] Shelly Xiaonan Wu and Wolfgang Banzhaf. The use of computational intelligence in intrusion detection systems: A review. *Applied soft computing*, 10(1):1–35, 2010.

[5] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[6] Subash Neupane, Jesse Ables, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, and Maria Seale. Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities, 2022.

[7] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40(2):44–58, 2019.

[8] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[9] Dorothy E Denning. An intrusion-detection model. *IEEE Transactions on software engineering*, (2):222–232, 1987.

[10] Rebecca Gurley Bace, Peter Mell, et al. Intrusion detection systems, 2001.

[11] Andrew McDole, Maanak Gupta, Mahmoud Abdelsalam, Sudip Mittal, and Mamoun Alazab. Deep learning techniques for behavioural malware analysis in cloud iaas. In *Malware Analysis using Artificial Intelligence and Deep Learning*. Springer, 2021.

[12] Kopelo Letou, Dhruwajita Devi, and Yumnam Jayanta. Host-based intrusion detection and prevention system (hidps). *International Journal of Computer Applications*, 69:28–33, 05 2013.

[13] Biswanath Mukherjee, Todd L. Heberlein, and Karl N. Levitt. Network intrusion detection. *IEEE Network*, 8:26–41, 1994.

[14] Ashu Sharma and Sanjay Kumar Sahay. Evolution and detection of polymorphic and metamorphic malwares: A survey. *arXiv preprint arXiv:1406.7061*, 2014.

[15] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41:15:1–15:58, 2009.

[16] Andrew McDole, Mahmoud Abdelsalam, Maanak Gupta, and Sudip Mittal. Analyzing cnn based behavioural malware detection techniques on cloud iaas. In *International Conference on Cloud Computing*, pages 64–79. Springer, 2020.

[17] Mateusz Szczepański, Michał Choraś, Marek Pawlicki, and Rafał Kozik. Achieving explainability of intrusion detection system by hybrid oracle-explainer approach. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[18] Guansong Pang, Choubo Ding, Chunhua Shen, and Anton van den Hengel. Explainable deep few-shot anomaly detection with deviation networks. *arXiv preprint arXiv:2108.00462*, 2021.

[19] Basant Subba, Santosh Biswas, and Sushanta Karmakar. Intrusion detection systems using linear discriminant analysis and logistic regression. In *2015 Annual IEEE India Conference (INDICON)*, pages 1–6. IEEE, 2015.

[20] Basim Mahbooba, Mohan Timilsina, Radhya Sahal, and Martin Serrano. Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, 2021, 2021.

[21] Chet Langin, Michael Wainer, and Shahram Rahimi. Annabell island: a 3d color hexagonal som for visual intrusion detection. *Internation Journal of Computer Science and Information Security*, 9(1):1–7, 2011.

[22] Chet Langin, Hongbo Zhou, and Shahram Rahimi. A model to use denied internet traffic to indirectly discover internal network security problems. In *2008 IEEE International Performance, Computing and Communications Conference*, pages 486–490. IEEE, 2008.

[23] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008.

[24] Bernhard Schölkopf, Robert C. Williamson, Alex Smola, John Shawe-Taylor, and John C. Platt. Support vector method for novelty detection. In *NIPS*, 1999.

[25] G.P. Zhang. Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(4):451–462, 2000.

[26] Johanna D Moore and William R Swartout. Explanation in expert systemss: A survey. Technical report, University of Southern California Marina Del Rey Information Sciences Inst, 1988.

[27] Edward Hance Shortliffe. Mycin: a rule-based computer program for advising physicians regarding antimicrobial therapy selection. Technical report, Stanford Univ Calif Dept of Computer Science, 1974.

[28] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[29] Leeanne Lindsay, Sonya Coleman, Dermot Kerr, Brian Taylor, and Anne Moorhead. Explainable artificial intelligence for falls prediction. In *International Conference on Advances in Computing and Data Sciences*, pages 76–84. Springer, 2020.

[30] Ye Eun Chun, Se Bin Kim, Ja Yun Lee, and Ji Hwan Woo. Study on credit rating model using explainable ai. *The Korean Data & Information Science Society*, 32(2):283–295, 2021.

[31] Miseon Han and Jeongtae Kim. Joint banknote recognition and counterfeit detection using explainable artificial intelligence. *Sensors*, 19(16):3607, 2019.

[32] DARPA. Broad agency announcement explainable artificial intelligence (xai). *DARPA-BAA-16-53*, pages 7–8, 2016.

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[34] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[35] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pages 63–71. Springer, 2016.

[36] Sheikh Rabiul Islam, William Eberle, Sheikh K Ghafoor, Ambareen Siraj, and Mike Rogers. Domain knowledge aided explainable artificial intelligence for intrusion detection and response. *arXiv preprint arXiv:1911.09853*, 2019.

[37] Chunyuan Wu, Aijuan Qian, Xiaoju Dong, and Yanling Zhang. Feature-oriented design of visual analytics system for interpretable deep learning based intrusion detection. In *2020 International Symposium on Theoretical Aspects of Software Engineering (TASE)*, pages 73–80. IEEE, 2020.

[38] Erkki Oja and Samuel Kaski. *Kohonen maps*. Elsevier, 1999.

[39] Shyam M Guthikonda. Kohonen self-organizing maps. *Wittenberg University*, 98, 2005.

[40] Teuvo Kohonen and Timo Honkela. Kohonen network. *Scholarpedia*, 2(1):1568, 2007.

[41] Jason Ong and Syed Muhammad Raza Abidi. Data mining using self-organizing kohonen maps: A technique for effective data clustering & visualization. In *IC-AI*, 1999.

[42] VK Pachghare, Parag Kulkarni, and Deven M Nikam. Intrusion detection system using self organizing maps. In *2009 International Conference on Intelligent Agent & Multi-Agent Systems*, pages 1–5. IEEE, 2009.

[43] Gregory Breard. Evaluating self-organizing map quality measures as convergence criteria. 2017.

[44] Dmitri Roussinov and Hsinchun Chen. A scalable self-organizing map algorithm for textual classification: A neural network approach to thesaurus generation. *Communication Cognition and Artificial Intelligence*, 15(1-2):81–111, 1998.

[45] Yao Liu, Jun Sun, Qing Yao, Su Wang, Kai Zheng, and Yan Liu. A scalable heterogeneous parallel som based on mpi/cuda. In *Asian Conference on Machine Learning*, pages 264–279. PMLR, 2018.

[46] Peter Lichodzijewski, A Nur Zincir-Heywood, and Malcolm I Heywood. Host-based intrusion detection using self-organizing maps. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, volume 2, pages 1714–1719. IEEE, 2002.

[47] Emiro De la Hoz, Andrés Ortiz García, Julio Ortega Lopera, Eduardo Miguel De La Hoz Correa, and Fabio Enrique Mendoza Palechor. Implementation of an intrusion detection system based on self organizing map. 2015.

[48] Brandon Craig Rhodes, James A Mahaffey, and James D Cannady. Multiple self-organizing maps for intrusion detection. In *Proceedings of the 23rd national information systems security conference*, pages 16–19. MD Press Baltimore, 2000.

[49] Sahin Albayrak, Christian Scheel, Dragan Milosevic, and Achim Muller. Combining self-organizing map algorithms for robust and scalable intrusion detection. In *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, volume 2, pages 123–130. IEEE, 2005.

[50] Alfred Ultsch and H. Peter Siemon. Kohonen's self organizing feature maps for exploratory data analysis. In *Proceedings of the International Neural Network Conference (INNC-90), Paris, France, July 9–13, 1990*, volume 1, pages 305–308. Kluwer Academic Press, 1990.

[51] Chathurika S. Wickramasinghe, Kasun Amarasinghe, Daniel L. Marino, Craig Rieger, and Milos Manic. Explainable unsupervised machine learning for cyber-physical systems. *IEEE Access*, 9:131824–131843, 2021.

[52] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[53] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A Ghorbani. A detailed analysis of the kdd cup 99 data set. pages 1–6, 2009.

[54] Ranjit Panigrahi and Samarjeet Borah. A detailed analysis of cicids2017 dataset for designing intrusion detection systems. *International Journal of Engineering & Technology*, 7:479–482, 3 2018.

[55] Lutz Hamel and Chris Brown. Bayesian probability approach to feature significance for infrared spectra of bacteria. *Applied Spectroscopy*, 66:48–59, 1 2012.

[56] Li Yuan. *Implementation of self-organizing maps with Python*. University of Rhode Island, 2018.

[57] Teuvo Kohonen. The self-organizing map. *Neurocomputing*, 21:1–6, 1998.

[58] Jouko Lampinen and Erkki Oja. Clustering properties of hierarchical self-organizing maps. *Journal of Mathematical Imaging and Vision*, 2:261–272, 1992.

[59] Lutz Hamel. Som quality measures: An efficient statistical approach. volume 428, pages 49–59. Springer Verlag, 2016.

[60] Self-organizing map convergence. *Int. J. Serv. Sci. Manag. Eng. Technol.*, 9:61–84, 4 2018.

[61] Chathurika S Wickramasinghe, Kasun Amarasinghe, Daniel L Marino, Craig Rieger, and Milos Manic. Explainable unsupervised machine learning for cyber-physical systems. *IEEE Access*, 9:131824–131843, 2021.

[62] Lutz Hamel and Chris Brown. Improved interpretability of the unified distance matrix with connected components. *7th International Conference on Data Mining (DMIN'11)*, 4 2012.

[63] Nabila Farnaaz and MA Jabbar. Random forest modeling for network intrusion detection system. *Procedia Computer Science*, 89:213–217, 2016.

[64] Maxim Nikolaevich Goryunov, Andrey Georgievich Matskevich, and Dmitry Aleksandrovich Rybolovlev. Synthesis of a machine learning model for detecting computer attacks based on the cicids2017 dataset. *Proceedings of the Institute for System Programming of the RAS*, 32(5):81–94, 2020.

[65] Md. Zahangir Alom, VenkataRamesh Bontupalli, and Tarek M. Taha. Intrusion detection using deep belief networks. In *2015 National Aerospace and Electronics Conference (NAECON)*, pages 339–344, 2015.

[66] Othmane Belarbi, Aftab Khan, Pietro Carnelli, and Theodoros Spyridopoulos. An intrusion detection system based on deep belief networks. *arXiv preprint arXiv:2207.02117*, 2022.

Selected Publication 3

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 August 2022, accepted 17 October 2022, date of publication 25 October 2022, date of current version 31 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3216617

SURVEY

Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities

SUBASH NEUPANE[1], JESSE ABLES[1], (Graduate Student Member, IEEE), WILLIAM ANDERSON[1], SUDIP MITTAL[1], (Member, IEEE), SHAHRAM RAHIMI[1], (Member, IEEE), IOANA BANICESCU[1], (Life Senior Member, IEEE), AND MARIA SEALE[2]

[1]Department of Computer Science and Engineering, Mississippi State University, Mississippi, MS 39762, USA
[2]U.S. Army Engineer Research and Development Center, Vicksburg, Mississippi, MS 39180, USA

Corresponding author: Subash Neupane (sn922@msstate.edu)

This work was supported by Mississippi State University through the U.S. Department of Defense (DoD) High Performance Computing Modernization Program through the U.S. Army Engineering Research and Develop Center (ERDC) under Grant W912HZ-21-C0058.

ABSTRACT The application of Artificial Intelligence (AI) and Machine Learning (ML) to cybersecurity challenges has gained traction in industry and academia, partially as a result of widespread malware attacks on critical systems such as cloud infrastructures and government institutions. Intrusion Detection Systems (IDS), using some forms of AI, have received widespread adoption due to their ability to handle vast amounts of data with a high prediction accuracy. These systems are hosted in the organizational Cyber Security Operation Center (CSoC) as a defense tool to monitor and detect malicious network flow that would otherwise impact the Confidentiality, Integrity, and Availability (CIA). CSoC analysts rely on these systems to make decisions about the detected threats. However, IDSs designed using Deep Learning (DL) techniques are often treated as black box models and do not provide a justification for their predictions. This creates a barrier for CSoC analysts, as they are unable to improve their decisions based on the model's predictions. One solution to this problem is to design *explainable IDS* (X-IDS). This survey reviews the state-of-the-art in explainable AI (XAI) for IDS, its current challenges, and discusses how these challenges span to the design of an X-IDS. In particular, we discuss black box and white box approaches comprehensively. We also present the tradeoff between these approaches in terms of their performance and ability to produce explanations. Furthermore, we propose a generic architecture that considers human-in-the-loop which can be used as a guideline when designing an X-IDS. Research recommendations are given from three critical viewpoints: the need to define explainability for IDS, the need to create explanations tailored to various stakeholders, and the need to design metrics to evaluate explanations.

INDEX TERMS Explainable intrusion detection systems, explainable artificial intelligence, machine learning, deep learning, white box, black box, explainability, cybersecurity.

I. INTRODUCTION AND MOTIVATION

The use of Artificial Intelligence (AI) and Machine Learning (ML) to solve cybersecurity problems has been gaining traction within industry and academia, partly as a response to widespread malware attacks on critical systems, such as cloud infrastructures, government institutions, etc. [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. AI- and ML-assisted cybersecurity offers data-driven automation that could enable security systems to identify and respond to cyber threats in real time. Many of these AI-based cyber defense systems are hosted in an organizational Cyber Security Operations Center (CSoC). CSoCs operated by security analysts act as a cybersecurity information hub and a defense base. Here the task is to orchestrate different security systems that are a part of an organization's overall cybersecurity framework,

The associate editor coordinating the review of this manuscript and approving it for publication was Mansoor Ahmed.

many of which have AI components. Examples of these security systems include Security Information and Event Management (SIEM) systems, vulnerability assessment solutions, governance, risk and compliance systems, application and database scanners, Intrusion Detection Systems (IDS), user and entity behavior analytics, Endpoint Detection and Remediation (EDR), etc. Here the security analysts maintain an ''organizational state'', keeping themselves one step ahead of the attackers to prevent potential intrusions [12].

The term ''intrusion detection'' originated in the early 1980s with James Anderson's seminal paper [13]. Dorothy E. Denning [14], following Anderson's work, proposed the first functional IDS in the mid-1980s. An IDS is a software or hardware security system that automates the process of monitoring and analyzing events occurring within a computer system or network for indications of potential security problems before they inflict widespread damage [15], [16].

In general, an intrusion results in a breach of at least one of the principles: *Confidentiality*, *Integrity*, or *Availability* (CIA). These tenets of security are used when protecting modern data infrastructure. They refer to the permissions to access or modify data, the prevention of improper data modification, and the ability to access data. The objective of an IDS is to detect misuse, unauthorized use (outsider without authorization), and abuse (abusing privilege-e.g., insider threat) within an organization, and much research has been done to improve the operational capacity of these IDS [17], [18], and [19].

The literature shows that numerous IDSs have been developed through the application of a variety of techniques from an array of disciplines, including statistical methods, ML techniques, and others [20]. At present, ML and Deep Learning (DL) techniques are widely used to develop IDS because of their ability to attain a high detection rate [16]. This adoption is also attributed to the fact that IDSs based on ML/DL techniques are much more efficient, accurate, and extendable as compared to their counterparts developed using other techniques. The surveys in [1], [21], and [22] primarily focus on intrusion detection techniques based on deep learning. However, *the techniques described in the preceding surveys are deficient in their ability to explain their inference processes and final results*, and they are *frequently treated as a black box by both developers and users* [23]. As a result, there is growing concern about the possibility of bias in these models, which necessitates the requirements for model transparency and post-hoc explainability [24]. Unfortunately, the majority of black box IDS described in the literature are opaque and much is needed to augment transparency.

It is apparent that these opaque/non-transparent models can achieve *impressive prediction accuracies*; however, they lack justification for their predictions. This is due to their nested and non-linear structure, which makes it difficult to identify the precise information in the data that influences their decision-making [25]. Such a *lack of understanding about the inner workings of opaque AI models* or an inability to traverse back from the outputs to the original data raises user trust issues [26]. This black box nature of the models creates problems for several domains in which AI or components of AI are integrated [27]. For example, in the context of an IDS, CSoCs analysts are tasked with the responsibility of analyzing IDS alerts for a variety of purposes, including alert escalation, threat and attack mitigation, intelligence gathering, and forensic analysis among others [28]. The lack of explanation of alerts generated by an IDS creates a barrier for task analysis and subsequently impedes decision-making.

In addition to the issues of *transparency* and *trust* surrounding AI systems, there exists yet another issue referred to as the problem of *decomposability*, specifically for systems built with DL models (See Section IV-A3 for IDS based on the decomposition approach). AI systems that are designed using DL techniques are difficult to interpret due to their inability to be decomposed into intuitive components [29]. The difficulty in interpreting DL models jeopardizes their actual use in production, as the computation behind their decisions are unknown [11]. *Explainable AI* (XAI) seeks to remedy this and other problems.

According to the Defense Advanced Research Projects Agency (DARPA), XAI systems are able to explain their reasoning to a human user, characterize their strengths and weaknesses, and convey a sense of their future behavior [24]. In this sense, by justifying specific decisions, XAI systems aid users in comprehending the model and assisting them in maintaining and effectively using it. On the other hand, transparency about predictions contributes to the development of trust in a system's intended behavior and provides users with confidence that they are performing tasks correctly.

Transparency is an open problem in the field of intrusion detection. Cybersecurity professionals now frequently make decisions based on the recommendations of an AI-enabled IDS. Therefore, the predictions made by the model should be understandable [11]. For instance, when an IDS model is presented with zero-day attacks, the model may misclassify the attacks as normal, resulting in a system breach. Understanding why specific samples are misclassified is the first step toward debugging and diagnosing the system. It is critical to provide detailed explanations for such misclassifications, so as to determine the appropriate course of action to prevent future attacks [6]. Therefore, an IDS should go beyond merely detecting intrusions-i.e., it should provide reasoning for the detected threat. The explanations in the form of correlations of various factors (for example, time of intrusion, type, suspicious network flow) influencing the predicted outcome can assist cybersecurity analysts in quickly analyzing tasks and making decisions [28].

The goal of *XAI in the field of intrusion detection* is to build operator trust and allow for more control of autonomous AI subsystems. Explainable Intrusion Detection Systems (X-IDS) will help build trust in these systems while also aiding CSoC analysts in their task of defending systems.

The major contributions of the paper are as follows:

- We present the state-of-the-art of the XAI approach and discuss the critical issues that surround it, most

importantly, how these issues relate to the intrusion detection domain. We propose a taxonomy based on a literature review to help lay the groundwork for formally defining explainability in intrusion detection.
- A comprehensive survey of the current landscape of X-IDS implementations is presented, with an emphasis on two major approaches: black box and white box. The distinction between the two approaches is discussed in detail, as is the rationale for why the black box approach with post-hoc explainability is more appropriate for intrusion detection tasks.
- We propose a generic explainable architecture with a user-centric approach for designing X-IDS that can accommodate a wide variety of scenarios and applications without adhering to a specific specification or technological solution.
- We discuss the challenges inherent in designing X-IDS and make research recommendations aimed at effectively mitigating these challenges for future researchers interested in developing X-IDS.

The remainder of this paper is organized as follows. Section II provides the background on explainable artificial intelligence (XAI). Section III summarizes our survey and taxonomy. Following that, in Section IV, we review the literature on black box and white box X-IDS approaches. Section V introduces a generic X-IDS architecture that future researchers can use as a guide. Section VI identifies research challenges and makes recommendations to future researchers. Finally, Section VII concludes this survey.

II. EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

The definition of what constitutes an explanation in AI remains an open research question. In the available literature, there are various definitions of the 'explainable AI' (XAI). Lent et al. [30] defines XAI as a system that *"provides an easily understandable chain of reasoning from the user's order, through the system's knowledge and inference, to the resulting behavior"*. The authors of this work used the term XAI to describe their system's ability to explain the behavior of AI-controlled entities. However, the recent definition by DARPA [31] indicates XAI as the intersection of different areas including machine learning, human computer interface, and end user explanation.

XAI has the potential to offer significant benefits to a broad range of domains that rely on artificial intelligence systems. Most importantly, the impact of XAI in the decision-making process for stakeholders in the IDS ecosystem is considerable. IDSs based on deep learning models can be more efficient in detecting malicious traffic with high accuracy. However, a CSoC analyst is still left with a significant task, i.e., to determine why the flow is malicious, and how to best deal with the attack. One solution to this problem is the post-hoc explainability offered by XAI. In [11], for example, the authors demonstrate how the SHapley Additive exPlanation (SHAP) framework [32] can be utilized to gain a deeper understanding of the model characteristics that contribute the most to various types of attacks. Similarly, misclassification is another important issue within the IDS ecosystem. Misclassification of malicious network flow as benign could be catastrophic to an organization (CSoC). For example, the authors in [6], demonstrated how the counterfactual technique can be used to explain misclassification. User trust in IDS predictions is imperative, and providing justification is as important as the prediction itself. For example, in [33], the authors illustrated how user trust can be garnered by providing input feature relevance scores (Layer-wise Relevance Propagation (LRP) method) that indicate the contribution of each input feature to the detection of the intrusion.

Apart from IDS systems, currently, XAI is being used in mission-critical systems and defense [30], [31]. To foster the trust of AI systems in the transportation domain, researchers are proposing explanations systems [34]. Some works based on image processing with explainability is found in [35], [36], [37], and [38]. Transparency regarding decision-making processes is critical in the criminal justice system [27], [39]. Various explainable methods for judicial decision support systems have been proposed by authors in [40], [41], and [42]. Model explainability is essential for gaining trust and acceptance of AI systems in high-stakes areas, such as healthcare, where reliability and safety are critical [43], [44]. Medical anomaly detection [45], healthcare risk prediction system [46], [47], [48], [49], genetics [50], [51], and healthcare image processing [52], [53], [54] are some of the areas that are moving towards adoption of XAI. Another area is finance, such as AI-based credit score decisions [55], [56] and counterfeit banknotes detection [57]. Support for XAI in academia for evaluation tasks are found in [58], [59], and [60]. Lastly, in the entertainment industry XAI for recommender systems is found in the works of [61], [62], and [63].

Arrieta et al. [64] argue that one of the issues that hinders the establishment of common ground for the meaning of the term 'explainability' in the context of AI is the interchangeable misuse of 'interpretability' and 'explainability' in the literature. Interpretability is the ability to explain or convey meaning in human-comprehensible terms [64]. This translates into the ability of a human to understand the model's reasoning without the need for additional explanations [65]. On the other hand, explainability is associated with the concept of explanation as a means of interface between humans and a decision-maker (model) that is both accurate and comprehensible to humans [65]. In this sense, if system users need an explanation as a proxy system to understand the reasoning process, that explanation is precisely represented by the XAI.

A central concept that emerges from all the preceding definitions of the XAI is 'understandability', which is the degree to which a human can comprehend a decision made with respect to a model. However, understandability is tightly coupled with the characteristics of the system's users. For instance, whether or not the explanation made the concept clear or simple to understand is entirely dependent on the audience.

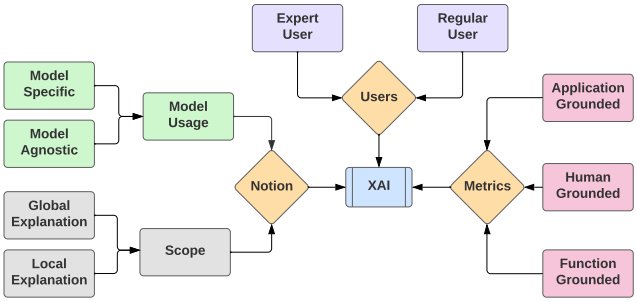


FIGURE 1. A taxonomic approach to explainability definition based on explainability concepts, formalizing explainability tasks from the standpoint of stakeholders, and evaluating explainability techniques. Green represents model dependency, while grey represents the scope of the explanations. Light purple represents various types of stakeholders in the IDS ecosystem. Pink represents techniques to evaluate explanations.

Despite the widespread recognition of the importance of explainability, researchers are struggling to establish universal, objective criteria for developing and validating explanations [66]. This is because XAI is plagued by inherent challenges that need addressing to foster its development. These include (i) achieving consensus on the right notion of model explainability, (ii) identifying and formalizing explainability tasks from the perspectives of various stakeholders, and (iii) designing measures for evaluating explainability techniques [67].

To address these challenges, we propose a taxonomy as depicted in Figure 1 based on our literature review to lay the groundwork for formally developing and validating explanations. In the following subsections, we describe in detail the concepts related to XAI including its notions, its meaning to various stakeholders of the system, and metrics to evaluate explanations.

A. NOTIONS OF EXPLAINABILITY

Several approaches to explanation methods have been proposed by different authors in the pursuit of explaining AI systems. The authors in [65] conducted a survey of black box specific explainability methods and proposed a taxonomy for XAI systems based on four characteristics: (i) the nature of the problem; (ii) the type of explainer used; (iii) the type of black box model processed by the explainer; and (iv) the type of data supported by the black box.

In another work [68], the authors presented notions related to the concept of explainability in two clusters. The first cluster refers to attributes of explainability – it contains criteria and characteristics used by scholars in trying to define the construct of explainability. The second cluster refers to the theoretical approaches for structuring explanations. Das and Rad in [69] proposed a taxonomy for categorizing XAI techniques based on explanation scope, algorithm methodology, and usage. Similarly, the authors in [55] surveyed over 180 articles related to explainability and categorized explainability using three criteria: the complexity of interpretability, the scoop of interpretability, and model dependency. The first criterion emphasizes the difficulty of interpreting and explaining complex models, such as those based on deep learning. The second criterion differentiates between local and global explanations, while the third criterion discusses model-specific and model-agnostic explanations. On the other hand, Pantelis et al. in [70] divided explainability methods into four groups based on: the data types used, the scope of explanation, the purpose of explanation, and the model usage.

A common category found in the literature regarding the taxonomy of explainability is the *scope of explainability* and *model dependency*. The following subsections describe these categories in greater detail.

1) LOCAL EXPLAINABILITY

The ability to explain a single prediction or decision is an example of local explainability. This explainability is used to generate a unique explanation or justification of the specific decision made by the model [55].Some of the local explanation methods include the Local Interpretable Model Agnostic Explanation (LIME) [71], the Anchors [72] and the Leave One Covariate Out (LOCO) [73]. LIME was originally proposed by Ribeiro et al. [71] who used a surrogate model to approximate the predictions of the black box model. Rather than training a global surrogate model, LIME uses a local surrogate model to interpret individual predictions.

To explain the behavior of complex models with high precision rules called *Anchors*, representing local, *sufficient conditions* for predictions, the same authors proposed an extension to the LIME method in [72]. Another popular technique for generating local explanation models with local variable importance measures is LOCO [73].

Lundberg et al. [32] proposed a game-theoretic optimal solution based on Shapley values for model explainability referred to as Shapely Additive Explanations (SHAP). SHAP calculates the significance of each feature in each prediction. The authors have demonstrated the equivalence of this model among various local interpretable models including LIME [71], Deep Learning Important FeaTures (DeepLIFT) [74], and LayerWise Relevance Propagation (LRP) [75]. The SHAP value can be computed for any model, not just simple linear models.

2) GLOBAL EXPLAINABILITY

The global explainability of a model makes it easier to follow the reasoning behind all the possible outcomes. These models shed light on the model's decision-making process as a whole, resulting in an understanding of the attributions for a variety of input data [69].

The LIME [71] model was extended with a 'submodular pick algorithm' (SP-LIME) in order to comprehend the model's global correlations. By providing a non-redundant global decision boundary for the machine learning model, LIME provides a global understanding of the model from individual data instances using a submodular pick algorithm.

Concept Activation Vectors (CAVs) proposed by Kim et al. [76] is another global explainability method. This

model can interpret the internal states of a neural network in the human-friendly concept domain. In another work, Yang et al. [77] proposed a novel method, the Global Interpretation via Recursive Partitioning (GIRP), to construct a global interpretation tree based on local explanations for a variety of machine learning models. Other methods of global explanation include an explanation by information extraction [78]. In this study, the authors propose a method of information extraction that is only lightly supervised and provides a global interpretation. They demonstrated that interpretable models can be generated when representation learning is combined with traditional pattern-based bootstrapping.

3) MODEL-SPECIFIC INTERPRETABILITY

The use of model-specific interpretability methods is restricted to a limited number of model classes. With these methods, we are restricted to using only models that provide a specific type of interpretation, which can reduce our options for using more accurate and representative models.

4) MODEL-AGNOSTIC INTERPRETABILITY

Methods that are model agnostic are not tied to any specific type of ML model, and are by definition modular, in the sense that the explanatory module is unrelated to the model for which it generates explanations. Model-agnostic interpretations are used to interpret artificial neural networks (ANNs) and can be local or global. In their survey [69], the authors argue that a significant amount of research in XAI is concentrated on model-agnostic post-hoc explainability algorithms, due to their ease of integration and breadth of application. Based on other reviewed papers, the authors [55] broadly categorize the techniques of model-agnostic interpretability into four types, including visualization, knowledge extraction, influence methods, and example-based explanations.

B. FORMALIZING EXPLAINABILITY TASKS FROM THE USER PERSPECTIVES

To be explainable, a machine learning model must be human-comprehensible. This presents a challenge for the development of XAI because it entails communicating a complex computational process to humans. The interpretable element that serves as the foundation of explanation is highly dependent on the question of *"who"* will receive the explanation. Rosenfeld et al. [79] identified three targets of explanation, including regular user, expert user, and the external entity. According to the authors, an explanation should be specific to user types. For instance, in a legal scenario, the explanation must be made to the expert users, not the regular users. On the other hand, if explanations are geared towards regular users, then the chance of developing trust and acceptance of XAI methods is high. To address the issue of stakeholder-specific explanation requirements, IBM developed an open-source toolkit known as AI Explainability 360 (AIX360) [80].

Adadi et al. [55] emphasize the significance of humans-in-the-loop approach for explainable systems from two perspectives: Human-like explanation and Human-friendly explanation. The first aspect focuses on how to produce explanations that simulate the human cognitive process, while the second aspect is concerned with developing explanations that are centered on humans.

Section V discusses the importance of human-centered design when developing X-IDS systems, and Section VI-B examines the explainability requirements imposed by various stakeholders in the IDS ecosystem.

C. MEASURES FOR EVALUATING EXPLAINABILITY TECHNIQUES

There have been few studies on evaluating explanations and quantifying their relevance despite the growing body of research that produces explainable ML methods. Doshi-Velez and Kim [81] proposed the three classes as evaluation methods for interpretability, including application-grounded, human-grounded, and functionally grounded methods. Application-grounded evaluation is concerned with the impact of the interpretation process's results on the human, domain expert, or end user, in terms of a well-defined task or application. Human-grounded evaluation is concerned with conducting simplified application-grounded evaluation where experiments are run with regular users rather than domain experts. Functionally grounded evaluation does not require human subjects, and rather uses formal, well-defined mathematical definitions of interpretability to determine the method's quality.

On the other hand, in [82], the authors outline three different evaluation criteria of explanations for deep networks, such as processing, representation, and explanation producing. The first criterion includes techniques that simulate data processing to generate insights about the relationships between a model's inputs and outputs. The second criterion describes an approach on how data is represented in networks and explains the representation. The third criterion states that the explanation-producing systems can be evaluated according to how well they match user expectations.

To evaluate local explanation, IBM in their toolkit AIX360 [80] suggested two metrics such as Faithfulness [83] and Monotonicity [84], to quantify the ''goodness'' of a feature-based local explanation. Other types of evaluation criteria found in the body of literature include completeness compared to the original model, ability to detect models with biases, completeness as measured on a substitute task, and human evaluation.

Another significant piece of work that could serve as a benchmark for evaluating explanations is the Florida Institute for Human and Machine Cognition's psychological model of explanation (IHMC) [24]. Section VI-C provides greater detail about this proposed model.

Next, we describe our survey approach and develop a taxonomy for X-IDS grounded in the current literature.

III. SURVEY AND TAXONOMY

The term *intrusion* refers to any unauthorized activity occurring within a network or system. An IDS is a collec-

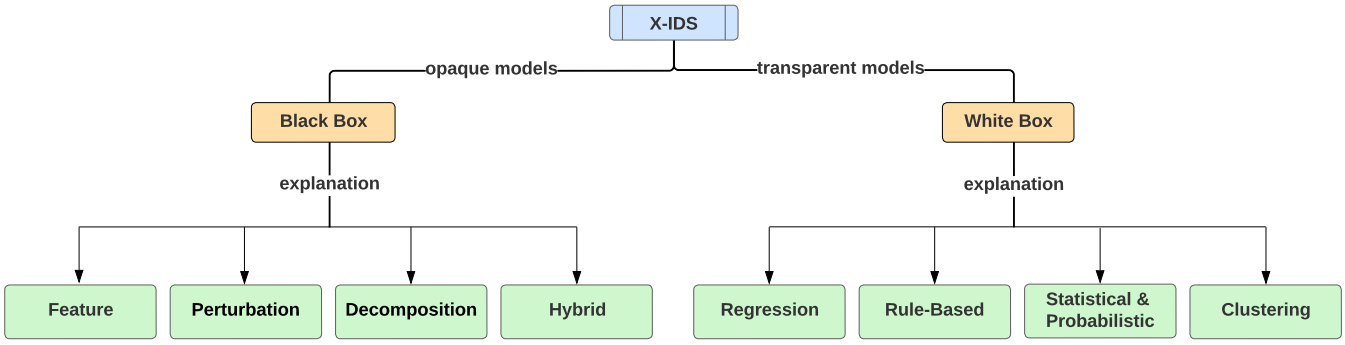


FIGURE 2. An overview of our proposed taxonomy. We categorize X-IDS techniques into two families, white box and black box. White box approaches encompass the techniques of Regression, Rule-Based, Clustering, and Statistical & Probabilistic Methods. Black box approaches encompass Feature, Perturbation, Decomposition, and Hybrid approaches. These approaches define the method of explainability to interpret the model's decision process.

tion of tools, methods, and resources that assist CSoC analysts in identifying, assessing, and reporting intrusions. Intrusion detection is typically a component of protection that surrounds a system and is not a stand-alone protection measure [85]. IDS are classified according to where they look for intrusive behavior: *host-based* or *network-based*. A host-based IDS monitors traffic that is originating from and coming to a specific host. Network-based IDS are strategically positioned in a network to analyze incoming and outgoing communication between network nodes.

IDS are categorized based on three detection techniques: *signature-based*, *anomaly-based*, and *hybrid*. Signature-based IDS monitors network traffic and compares it to a database of known malicious threats' signatures or attributes. However, they are incapable of detecting zero-day attacks, metamorphic threats, or polymorphic threats [86]. On the other hand, anomaly-based IDS look for patterns in data that do not conform to expected behavior [87], allowing them to recognize such threats. However, these detection systems are susceptible to higher false positive rates because they may categorize previously unseen, yet legitimate, system behaviors as anomalies [19]. Hybrid IDS integrate both signature-based and anomaly-based detection methods, which allows for an increased detection rate of known intrusions, the ability to detect unseen intrusions, and reduce false positives.

As previously stated, prior work has focused on XAI from the lens of explainability, qualifying the definitions of *notion*, *users*, and *metrics* (See Section II). This survey follows that direction by creating a taxonomy surrounding current XAI techniques for IDS. The focus is on their relevance and applicability to the domain of intrusion detection, with a particular emphasis on the current hierarchy of families, strengths and weaknesses, and any challenges or assumptions that come with their application. A summary of our taxonomy can be seen in Figure 2. The two primary families of XAI techniques are those of white box models and black box models which greatly affect our survey taxonomy for approaches to X-IDS. The survey of existing systems based on the taxonomy in Figure 2, is available in Section IV. Next, we describe the salient features of white box models and black box models.

A. SALIENT FEATURES OF WHITE BOX TECHNIQUES

White box models provide results that are easy to understand [88]. This *easy to understand* condition is typically defined as an explainable outcome understood by an expert in the field. In practice, this definition is more associated with the popular suite of machine learning models that existed prior to the rise in popularity of neural network based approaches. White box models, while generally not as efficacious as their black box counterparts, bring a layer of transparency that is intrinsic to their decision process. This trait is often preferred, if not a requirement, in domains where the decision system is sensitive or requires a high degree of auditing. These models cover a wide variety of techniques that fall into four distinct families: *Regression*, *Rule-Based*, *Clustering*, and *Statistical & Probabilistic Methods*.

Regression-based approaches comprise the family of regression analysis. These approaches have a well formed background of statistical support and maturity. Therefore, these models are most often employed in the early stages of modeling, in the pipelines of more complex models, and in domains where scrutiny and transparency are of paramount importance. Although not a focal point of comparison for this paper, regression models are highly computationally efficient, allowing for rapid construction, as well as deployment into low-resource systems where detection time is critical, such as IoT edge devices. Regression approaches can be split into Parametric Regression and Non-parametric Regression. The former enforces a constraint on model expectation via a restriction on the parameters of the model, making this modeling approach best for when certain assumptions can be met. The latter enforces no such constraint, which decreases overall interpretability but increases the application to a wider variety of data and assumptions. Popular regression techniques are Linear Regression (LR), Logistic Regression (LoR), various non-linear models, Poisson Regression, Kernel Regression (KR), and Spline Smoothing.

Rule-based approaches leverage a learned set of rules as a means of the model decision process, and thusly, model explainability. Rule based explanations are perhaps the most practical, as they mimic the human decision making process when it comes to defining an anomaly. This process also allows learned rules to then be incorporated into Signature-Based IDS (SIDS), allowing Anomaly-Based IDS (AIDS) to serve as zero-day identifiers and rule miners. Rule-based approaches benefit from the allowance of a very tight definition of rules, known as hard rules or crisp rules, or for a relaxed fuzzy-rule based approach, allowing flexibility and further statistical inference to be rendered on them. A popular approach to modeling for rule-based explanations is the Decision Tree and its many variants.

Statistical & Probabilistic Methods is a broad category for the numerous statistical models of reasoning that exist in the literature. Notably, many of these methods have seen a decline in use as a compliment to the rise in popularity of various black box methods. These less frequently used methods are appropriate for application in specific scenarios or in larger pipelines for multi-stage IDS. Examples of such approaches include moment-based approaches, statistical ensembles, Markov Models, Baysian Networks, and others which are covered more specifically in IV-B3.

Clustering-based approaches use supervised or unsupervised learning to aggregate similar data objects. This *similar* condition is defined by a similarity, or dissimilarity, measure. Traditionally these methods are defined on distance based metrics such as Euclidean, Manhattan, Cosine Measure, Pearson coefficient, and many others. Other attempts to define similarity have had success in the graph representation domain, using graph-based clustering algorithms to accomplish this task. Clustering, due to its ability to be leveraged as an unsupervised learner, still retains a high degree of use due to the importance of data mining for intrusion detection. Examples of popular clustering algorithms are K-Means, Self-Organzing Maps (SOMs), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Agglomerative Clustering, and Spectral Clustering.

B. SALIENT FEATURES OF BLACK BOX TECHNIQUES

Black box models are models where the decision systems are considered opaque [65]. These systems, composing nearly all of the state-of-the-art, are limited due to the lacking ability of model inspection and evaluation. Therefore, if these systems are to be utilized in decision sensitive domains, i.e, those whose applications require safety, privacy, and fairness, some degree of exploration and evaluation of their decision process must be possible. Currently, there exists no singular solution to the black box inspection problem. However, many candidate explanations have emerged, exploring and exploiting various aspects of the machine learning process to create explanations for black box models. These candidate explanations currently fall into four distinct families: *Feature*, *Perturbation*, *Decomposition*, and *Hybrid*.

Feature-based explanations target features as the method of explanation. The goal of feature attribution is to determine how much each feature is responsible for the output prediction. Features were one of the first methods of explainability in black box models due to their impact on model performance and human interpretable relevance. Examples of popular feature based explanations are Partial Dependence Plot (PDP) [89], Accumulated Local Effects (ALE) [90], Individual Conditional Expectations (ICE) [91], H-statistic [92], and SHapley Additive exPlanations (SHAP) [32].

Perturbation-based explanations study changes to the output space with perturbations to the input space. Due to this property, perturbation techniques can be deployed to any general input space, such as tabular data, images, or text. In particular, model sensitivity to feature perturbations has long been regarded as a measure of feature importance. Saliency maps, Randomized Input Sampling for Explanation of Black Box Models (RISE) [93], and Local Interpretable Model-Agnostic Explanations (LIME) [71] are popular perturbation methods.

Decomposition-based explanations decompose the original model prediction. Much like the previous two methods, the goal of decomposition is to allocate a measure of importance to the input space; however, this method does so by the decomposition of a model signal, such as the model's gradients. This is predicated on the assumption that large gradients play a role in shaping explanations. However, gradients are not the only method of decomposition. Many approaches exist, such as Gradient * Input [94], Integrated Gradients (IG) [95], Grad-CAM [96], DeepLIFT [74], Deep Taylor Decomposition (DTD) [97] and Layerwise Relevance Propagation (LRP) [75].

Hybrid-based explanations encapsulate a type of model construction often demonstrated in pipelined machine learning architectures. These models can range from ensembles, a blend of white box and black box approaches working in tandem, to carefully composed IDS pipelines encapsulating many of the best state-of-the-art approaches. Therefore, hybrid approaches present the most variability of explanations, with respect to methodology, location of explanations, and application.

Next, we use the taxonomy showcased in Figure 2, to present a literature survey on approaches to X-IDS.

IV. APPROACHES TO EXPLAINABLE IDS (X-IDS)

As per the survey overview presented in Section III, and the taxonomy showcased in Figure 2, we will now describe in detail the black box and the white box approaches to XAI in intrusion detection systems.

A. BLACK BOX X-IDS MODELS

Guidotti et al. [65], describes a black box predictor as ''a data-mining and machine learning obscure model, whose internals are either unknown to the observer or are known but are uninterpretable by humans''. A black box model is not explainable by itself. Therefore, to make a black box model

explainable, we have to adopt several techniques to extract explanations from the inner logic or the outputs of the model.

To survey the IDS landscape with respect to explainability, we have further divided the literature into different categories of XAI black box models: feature based, perturbations based, decomposition based, and hybrid approaches. These classifications are based upon how explanations are generated. A detailed literature overview is also available in Table 1.

1) FEATURE BASED APPROACHES

One popular scheme for explanations considers the influence features have on prediction. Such schemes are called feature explanations. Existing processes, such as feature engineering and feature selection, are already common in machine learning pipelines. Therefore, it is natural that features emerge as a method of explainability. Several candidate solutions that currently exploit this assumption are Partial Dependence Plot (PDP), Accumulated Local Effects (ALE), H-statistic, and SHAP.

An important generalizable SHAP-based framework is proposed by Wang et al. [11]. Their framework uses both local and global explanations to increase the explainability of the IDS model. The IDS model consists of a binary Neural Network (NN) classifier and a multi-class NN classifier. To generate explanations, both models and predictions are fed to the SHAP module. Local explanations are generated by choosing an attack and randomly selecting 100 of the occurrences. An average Shapely value is calculated, and the SHAP module outputs a confidence score for the prediction. The authors evaluate explainability by using a neptune attack, where a flooding of SYN packets is observed. The explanation results show that the top four features are related to DoS and SYN flood attacks. Using the global explanation produced by the SHAP module, researchers can make inferences about how the model might react during a related attack. However, the model's confidence seems to favor attacks that attempt many network connections (e.g. probe or DoS) over other attacks, such as privilege escalation attacks. The IDS system along with the SHAP explanations are relevant to assist subject matter experts in making security decisions.

In another effort, Islam et al. [98] built a domain knowledge infused explainable IDS framework. Their architecture is composed of two parts: a feature generalizer that uses the CIA principles and an evaluator that compares the black box models using different configurations.

The feature generalizer first maps the top three ranked features to attack types, then maps attack types to the CIA principles. For example, DoS attacks are associated with availability; Heartbleed or PortScan attacks are associated with confidentiality. Using this mapping system, the authors add three new features: C, I, and A. These three features include the aggregate scores of their related features from a data sample. If a feature positively affects a prediction, then it adds to the score; otherwise, it subtracts from the score.

The evaluator, on the other hand, runs four different feature configurations. The first configuration uses the full, preprocessed CICIDS dataset of 78 features. The second is a feature selected dataset of 50 attributes. The final two datasets are domain knowledge based: a 22 feature dataset of domain infused features and a three feature dataset consisting of C, I, and A scoring features.

Tests are run on ANN, SVM, Random Forest (RF), Extra Trees (ET), Gradient Boosting (GB), and Naive Bayes algorithms (NB). The authors outline two types of tests: explainability and generalizability. The first two datasets are used to find a baseline to compare against the authors' novel, domain infused approach. Their findings from initial experimentation show that the RF using the full dataset outperforms all other algorithms with an F1-score of 99.68%. The domain infused and CIA datasets are able to obtain an F1-score of 99.32% and 93.84% on RF and ET algorithms, respectfully. The small difference between the full dataset and the domain infused dataset show that the authors can now create a way to explain predictions without negatively impacting model performance. The authors create another CIA scoring formula that shows how much impact a CIA mapped feature had on the samples prediction. These C, I, and A scores can then be shown to an analyst to explain the prediction. To test their method against unknown attacks, the models are trained on all attacks in the dataset except one. The classifier is tested on a dataset that includes all of the attacks.

The results show that the novel domain infused dataset performs similarly to the full dataset. In one case, the domain infused dataset is able to be used to find an attack that the full dataset configuration could not. The authors have demonstrated that creating an explainable algorithm and dataset can be useful for both accuracy and explanations.

Sarhan et al. proposes another feature based technique [99]. Two feature sets, NetFlow and CICFlowMeter, are evaluated across three datasets. When new IDS datasets are created, they are not necessarily created using the same tools. NetFlow and CICFlowMeter based IDS datasets collect different feature sets. The authors test these different feature sets using Random Forests and Deep Feed Forward algorithms. The results from this experiment show a minor improvement from the NetFlow feature set over the CICFlowMeter set. The most interesting result is the change in false positives between the two feature sets. NetFlow offers a much lower false positive rate than its counterpart in many of the tests. Additionally, NetFlow is slightly faster to make predictions than CICFlowMeter. The authors conclude that NetFlow offers slightly higher quality security features. Explainability is achieved in the form of SHAP. SHAP is used to determine which features are causing this difference in performance. The authors conclude that there are certain features across all datasets that contain more security focused data. However, the most important features vary across datasets. This is attributed to the fact that each dataset has different attacks. The authors work shows the importance of feature selection during dataset creation.

A novel method in [100] uses Auto-Encoders (AE) in combination with SHAP to explain anomalies. Anomalies

TABLE 1. An overview of the existing literature on black-box approaches to intrusion detection systems, with a focus on their scope, contribution, and limitations.

Paper Title	Focus/Objective	Contribution	Limitation
Feature based IDS			
An Explainable Machine Learning Framework for Intrusion Detection Systems [11]	Locally and globally explainable NN using SHAP for IDS.	• Framework that creates both local and global explanations. • First use of SHAP in the field of IDS. • Comparison between one-vs-all classifier and multi-class classifier.	• More intrusion detection datasets should be tested. • SHAP cannot work in real-time. • SHAP needs to be tested on more robust attacks.
Domain Knowledge Aided Explainable Artificial Intelligence for Intrusion Detection and Response [98]	Use CIA principles on data to improve both generalizability and explainability of a model	• Method for the collection and use of Domain Knowledge in IDS • Use CIA principles to aid in explainability • Domain Knowledge increase generalizability	• Domain Knowledge is applied to a specific dataset. New mappings may be needed on new datasets. • More datasets need to be tested.
An Explainable Machine Learning-based Network Intrusion Detection System for Enabling Generalisability in Securing IoT Networks [99]	Explore explainability in IDS by comparing two different IDS feat.	• Evaluate two different Network Intrusion Detection datasets: NetFLow and CICFlowMeter. • Creation of two new datasets in the CICFlowMeter format. • An explainable analysis is performed using SHAP.	• Explanations are only done using SHAP. • No analysis on the performance of the explainer.
Explaining Anomalies Detected by Autoencoders Using SHAP [100]	Use SHAP to create custom explanations for anomalies found with an autoencoder.	• Method for explaining anomalies found by an autoencoder. • Preliminary experiment with real word data and domain experts. • Suggest methods for evaluating explanations.	• Custom explanation lacks any form of visualization to aid the user.
Perturbation based IDS			
A New Explainable Deep Learning Framework for Cyber Threat Discovery in Industrial IoT Networks [101]	Explainable Intrusion Detection in the field of IoT	• Conv-LSTM-based autoencoder for time-series attacks • Detects zero-day attacks • Sliding window technique that increases accuracy of CNN and LSTM model XAI concepts to improve trust	• Tested only on a single dataset • Considers only univariate time-series data
An Adversarial Approach for Explainable AI in Intrusion Detection Systems [6]	Explain models and predictions through an adversarial approach	• Methodology explaining incorrectly classified samples to help improve flaws in the model	• Only tested on DoS attacks from NSL-KDD
Feature-Oriented Design of Visual Analytics System for Interpretable Deep Learning Based Intrusion Detection [102]	A suite of visual tools used to improve explainability of CNNs	• Analysis of Features and Requirements to improve visual analysis of XAI • IDSBoard, a GUI for understanding Deep Learning Intrusion Detection • Demonstrate the effectiveness of visual analytics	• Only tested on a single dataset • Scalability of visual analytics system • Visual analytics system only designed for CNN
Explanation framework for Intrusion Detection [103]	Explaining IDS explanations using a Counterfactual technique.	• Explaining classifications based on feature importance. • Advice on how to change a classification to its desired result. • Outline the decision process so that the user can simulate it themselves.	• Analysis of the counterfactual technique was only run on one type of ML algorithm.
Decomposition/Gradient based IDS			
Toward Explainable Deep Neural Network Based Anomaly Detection. [104]	Initial steps into XAI for DNN Intrusion Detection	• Framework for creating an explainable Deep Network XAI concepts to improve trust	• Experiments are run using only DOS attacks from the NSL-KDD dataset

TABLE 1. *(Continued.)* **An overview of the existing literature on black-box approaches to intrusion detection systems, with a focus on their scope, contribution, and limitations.**

Towards explaining anomalies: A deep Taylor decomposition of one-class models. [105]	Explaining anomalies found by a SVM using Deep Taylor Decomposition.	• A method for 'neuralizing' a one-class SVM to be explained by Deep Taylor Decomposition.	• Experiments solely run using one-class SVM. No comparison to 'real' neural networks.
Hybrid IDS			
Achieving explainability of intrusion detection system by hybrid oracle-explainer approach [106]	Building a hybrid IDS based around 'XAI Desiderata' that does not decrease performance or add vulnerability.	• An explainer module modeled after the 'XAI Desiderata.' • A Hybrid-Oracle explainer Intrusion Detection System.	• Two models need to be effectively trained. • Would benefit from being tested on multiple datasets.
Explainable deep few-shot anomaly detection with deviation networks [107]	An anomaly detection system able of detecting anomalies learned from few anomalous training samples.	• Prior-driven anomaly detection framework. • DevNet, an anomaly detection framework based on Gausian prior, Z-Score-based deviation loss, and multiple instance learning. • A theoretical and empirical analysis of Few-shot anomaly detection.	• Experiments only run using image based datasets with relatively small sample sizes.

are detected using the reconstruction score of the AE. Samples that return a higher reconstruction score are considered anomalous. An explainer module is created with the goal to link the input value of anomalies to their high reconstruction score. Features are split into two sets. The first set contains features that are causing the reconstruction score to be higher, while the second set does the opposite. The authors label these sets 'contributing' and 'offsetting', respectively. Contributing features will have a SHAP score that is negative, and the opposite is true for offsetting. Explanations are presented in the form of a color-coded table where darker values are more important than lighter values. This novel approach to explaining AE can be improved with more iterations of its visualization style and methodology.

In another piece of work, Dang, Q. V. [108] suggests an explainable IDS that uses the eXtreme Gradient Boosting (xgboost) classifier as its base predictor model and makes explanations using PDP plots and the SHAP value. The author uses the CICIDS2017 dataset to train and test the proposed model. The experiment result indicates the proposed classifier has high detection accuracy. However, it requires high computational power. To reduce the computational need, the author utilizes the PDP plots to recursively remove features that cannot be explained without affecting their predictive accuracy.

2) PERTURBATION BASED APPROACHES

Perturbation based approaches make minor modifications to input data to observe changes in output predictions. Their explanations are based on the inclusion, removal, or modification of a feature in a dataset. These approaches are model agnostic (see Section II), therefore, they can be applied to any model.

A representative work by Wu et al. [102] showcases the advantages of this approach. The authors have created a CNN model along with a dashboard user interface (UI) to make the black box deep learning components more explainable. They gather feature requirements for their dashboard from literature. These include: (i) it is important to know the role that individual neurons play in predictions; (ii) multiple models should be tested, and the best parameters should be selected to achieve the best accuracy; (iii) visualization should assist in finding interesting results; (iv) there should be an explanation as to how the model made a decision; (v) we should be able to see the data representation in each layer of the model.

The authors use the NSL-KDD dataset to test their CNN. NSL-KDD is encoded into a 12×12 grayscale image that serves as input. Their model is able to achieve an 80% accuracy. The dashboard UI is able to showcase a variety of visualizations that assists in explainability. The UI includes: a detailed view of each cluster of neurons and the associated feature class, a t-SNE scatterplot of the activation values, a feature map of the convolutional kernel, a feature panel that explains how the model came to a prediction (utilizing LIME and a Saliency chart), a confusion matrix of predicted instances, and a graph for finding input data patterns. The authors demonstrate the advantages of using the dashboard UI by comparing CNNs with fewer layers than their proposed architecture. For example, the last layer in a smaller CNN shows that it is unable to detect one of the attack types (u2r) from the NSL-KDD dataset, while the proposed architecture can detect the attack. The dashboard UI is able to demonstrate that the smaller model may need more layers to be effective.

Khan et al. [101] propose an explainable autoencoder-based detection framework using convolutional and recurrent networks to discover cyber threats in IoT networks. The model is capable of detecting both known and zero-day attacks. It leverages a 2-step, sliding window technique that is used to transform a 1-dimensional (1D) sample into smaller contiguous 2-dimensional (2D) samples. This 2D sample is then fed through a CNN, comprised of a 1D convolutional layer and a 1D max-pooling layer which extracts spatial features. The data is then fed into the auto-encoder based LSTM that extracts temporal features. Finally, the DNN uses

the extracted representation to make predictions. To make the model explainable, the authors use LIME [71] (see Section II-A). The dataset used for experimentation was from a real-world gas pipeline system. It consists of system logs that include packet data used to communicate with the pipeline, along with features such as packet length, pressure setpoint, and PID gain. The authors obtain a 99.35% accuracy using their proposed model. LIME shows that there are five features in the dataset that are primarily responsible for the different predictions.

In another impactful work [6], the authors argue that rather than explaining every prediction, it is possible to create a model that explains misclassifications using a *counterfactual technique*. The goal is to explain adversarial attacks, which aim to confuse models into misclassifying input samples. Using this technique, the authors find weak points in their model and develop strategies to overcome these limitations. When an input sample is classified incorrectly, minimal changes are made to the sample until it is classified correctly. The difference between the original, incorrectly labeled sample and the new, correctly labeled sample are used to explain the occurrence of the misclassification.

NSL-KDD dataset is used to create these models. A linear classifier and a multi-layer perceptron (MLP) are used during testing and the authors achieve an accuracy of 93% and 95%, respectfully. t-SNE is used to visualize the misclassified and corrected samples. The authors technique for minimizing the difference between samples is effective as the projections created by t-SNE are nearly identical. More insight can then be gathered from these projections as they show which features caused the misclassification along with the magnitude of the impact. This method appears to be a good way to communicate why a classification occurred and allows for a user to make the necessary inferences.

Burkart et al. [103] proposes a similar application of counterfactuals on an explainable IDS framework. Here the goal of the system is to answer the question: *Why did X happen and not Y?* The authors aim to create explanations that are *understandable* and *actionable*. By understandable, they mean explaining an instance of classification, and by actionable they mean giving advice for changing the classification. The framework should also allow the users to simulate these changes themselves. The counterfactual technique is used to achieve these goals.

The technique takes a vector x and locates a similar co-ordinate position $x\prime$ in the feature space, that causes a change in the predicted label. $x\prime$ should be a sample that is very similar to x. The authors' method for their explainable technique has 5 phases. In Phase 1, their algorithm finds the first counterfactual point by using an optimization problem. Phase 2 extrapolates that point by finding other points near it that are also opposite of the original vector x. By adding more than one counterfactual point, the algorithm can help find a better general understanding of the feature area. In their approach, the authors use *MagneticSampling* to achieve this goal. This set of points is used in Phase 3 to find the decision

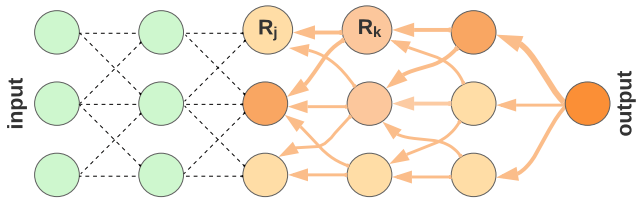


FIGURE 3. A visual depiction of Layer-wise Relevance Propagation. Relevance scores (R_j, R_k) are calculated backwards from the output for each layer (j and k represent neurons). Scores from each previous layer are used to score the next set of neurons with the final outcome being the importance of each input [110].

boundary. Phase 4 takes this approximated decision boundary and trains a *surrogate explainer model* for samples on both sides of the decision boundary. Phase 5 is the culmination of all of the previous work resulting in explanations, which include a feature importance explanation, a relative difference explanation, and a surrogate visualization module. The surrogate visualization can be done in a variety of ways; however, the authors choose to use a white box decision tree. The fidelity of their explainer is tested against LIME. Additionally, their method tests 2 varieties of explainers: a decision tree explainer and a linear explainer similar to LIME. The authors method performs better than LIME when *MagneticSampling* is used in Phase 2, but performs worse than LIME when random sampling is used. The tree performs better than the linear method and the authors believe that it is superior based on its performance and inbuilt explainability.

3) DECOMPOSITION BASED APPROACHES

Decomposition based approaches decompose the output of a model to create a relevance score. Layer-wise Relevance Propagation (LRP) is a technique where the scoring mechanism propagates backwards from the output node, highlighting activated neurons that impact predictions. According to the authors in [109], these approaches can either decompose the output or decompose the gradient of the model.

An explainable DNN using LRP has been proposed in [104]. The goal of this system is to give a confidence score of a prediction, give a textual explanation of a prediction, and the reasons why the prediction was chosen. Online, a user can see that an anomaly has been found and why it is considered an anomaly, while offline an expert can evaluate the explanations. The authors argue that the explanation for detected anomalies is provided to reduce the 'opaqueness' of DNN model and enhance 'human trust' in the algorithm. For their experiment, the authors create a partial implementation of their framework consisting of a Feed Forward DNN with explanations created by LRP. NSL-KDD is used for their experimentation. The tests are run using 4 different DNN configuration: two with three hidden layers and two with four hidden layers. Additionally, the dataset was separated into a 'simple' dataset (a smaller number of features) and a 'complete' dataset (all the features). The authors were able to achieve up to 97% accuracy from each of the implementations. The model performed better with the complete dataset

rather than with the simple dataset. The authors argue that the explainability of the simple dataset is worse than the complete dataset. This is because LRP chose a feature that would be difficult for a domain expert to verify. For example, binary features like 'flag' are more difficult to explain than continuous features. The most important features for the complete dataset contained continuous values that could more easily be determined to be anomalous (src byte count and destination host count). Although the authors do not create a complete implementation with full textual explanations, their methodology could prove useful to improving the trust of regular users.

To address the issue of decomposability of DL models, the authors in [111] propose an IDS system, based on CNNs, called GRACE (GRad-CAM-enhAnced Convolution neural nEtwork). They generate visual explanations for CNN decisions by utilizing the Gradient-weighted Class Activation Mapping (Grad-CAM) [96].

The authors use three different datasets including KDD-CUP-99, NSL-KDDCUP99, and UNSW-NB15 to train their model. The textual dataset is transformed using image encoding which converts the training sample from the 1D feature vector form $\mathbf{X}^{1D}$ with size $1 \times M$ to the 2D image form $\mathbf{X}^{2D}$ with size $m \times m$ (with $M \leq m_2$) and fed to the CNN model. The final convolution layer of 2D CNN is used to create heatmaps of class activations on input images, i.e., 2D grids of scores. Each pixel in the grid represents traffic characteristics, e.g., Destination port (X1) or idle max(X77). The scoring mechanism demonstrates how important each pixel (feature) is to a specific output class. This understanding of the most important features aids in the feature engineering process, resulting in a CNN model with higher accuracy.

To evaluate the performance of the model three evaluation metrics are used such as F1-Score (F1), Accuracy (A), and Computational complexity (T) (time spent to train the model). Of these, F1 and A metrics are used to compare against state-of-the-art such as CNN, GAN, LSTM, RNN, Triplet, DNN, MLP, and Autoencoder. Experimental results suggest GRACE generally outperforms its competitors. However, there are a few exceptions where the proposed model suffers slightly. For instance, when using the NSL-KDD dataset, the Triplet methods obtain 86.6% and 87.0% of A and F1, respectively, compared against 85.7% and 86.8% of the proposed model. The authors argue that this explanation approach can aid in the development of a more robust intrusion detection model.

Kauffmann et al. [112] propose another decomposition strategy aimed at verifying that a 'Clever Hans' strategy has not been adopted by the ML model. LRP is leveraged as an explainer module to aid in discovering this phenomenon. Three separate models are trained: a kernel density estimator, an autoencoder, and a deep one-class model. Image based anomaly detection datasets MNIST-C and MVTec are used for this experiment. A 'Clever Hans' score is adopted that is simply the difference between the detection accuracy and explanation accuracy. Detection accuracy is the ROC score, and explanation accuracy is the cosine similarity between the ground-truth and the pixel-wise explanation. It renders a score between 1 and -1 where 1 expresses a 'Clever Hans' phenomenon. Results from their testing show that, based on their scoring system, all of the models show some form of 'Clever Hans' logic. To address this problem, the authors propose a method of bagging anomaly detectors. This solution does not remove the phenomenon, but it does help to reduce it.

The previous authors also explore Deep Taylor Decomposition (DTD) for model explainability [105]. DTD is a technique that decomposes each neuron in each layer to determine feature relevance. A 'neuralized', one-class SVM is proposed that can be explained using DTD. The 'neuralized' form is a mapping of distance between the original sample and the SVM created support vectors as the first layer. The second layer is a soft min-pooling layer that calculates the 'outlierness' of samples. Samples can then be explained using DTD by decomposing each of the neurons in the prediction. In their experiment, they use image based datasets for finding anomalies. DTD is used to highlight anomalous pixels in each image.

4) HYBRID APPROACHES

A hybrid black box predictor, white box explainer has been created by Szczepanski et al. [106]. Their framework is built with principles from the ''XAI Desiderata'': Fidelity, Understandability, Sufficiency, Low Construction Overhead, and Efficiency [113]. The authors aim to contribute a system that is reliable, easy to understand, flexible, and meets all previous criteria without losing accuracy. With these goals in mind, a framework that uses local explanations is created. Their framework includes an ANN that predicts samples and a white box explainer that takes the output of the ANN and the original sample as input. The explainer is model agnostic and replaceable with any other explanation algorithm. The authors' explainer uses a clustering algorithm that uses a heuristic called Mean Distance to Average Vector. Clustering is done based on all of the attributes except the label. n centroids are computed for all features, then a model is trained for each centroid cluster created. Another distance based algorithm is used to find a centroid cluster that is both close to the predicted sample and gives the same prediction as the ANN. The selected cluster is then used as a visualized explanation for a prediction. The authors note that it is possible that the explainer may not return a valid tree and that the model should be trained on a feature rich, diverse dataset. The authors experiment using the CICIDS2017 dataset. The ANN is able to achieve an accuracy of 98%, and the explainer is able to achieve an accuracy of 99% with 200 clusters. The authors have created a system where there are effectively two predictors that are used to confirm and explain the other's prediction.

Pang et al. [107] create a framework based on Few-Shot Anomaly Detection (FSAD). The authors claim that their framework is interpretable and explainable through a prob-

ability based scoring method and an image demonstrating anomalous areas found in samples. One of the problems faced in IDS/Anomaly Detection is that models are generally trained on unsupervised, normal data. This makes it difficult for models to discern from normal and anomalous data. The authors aim using FSAD to improve detection rates. However, FSAD has difficulties learning a generalized representation of anomalies from a few samples and it is challenging to learn a robust representation of data with respect to anomolous data. To resolve this, the framework needs to be able to learn about anomalous samples but not learn that all anomalies are the same as the training samples. The authors achieve this by using a prior driven anomaly score and end-to-end optimization of anomaly scores with deviation learning based on the prior probability. The architecture of DevNet is composed of an Anomaly Scoring Network and a Reference Score Generator that outputs into a Multiple-Instance-Learning-based (MIL) deviation loss Score Learner. The Anomaly Scoring Network is a function ϕ that creates a scalar anomaly score for pieces of an input. In this case, the pieces of an input are parts of an image. The Reference Score Generator creates a reference score μ_r, which is a mean score of randomly selected non-anomalous samples. The reference score is derived from a prior F. The function $\phi(X)$, μ_r, and the standard deviation of μ_r are provided as input into the MIL Deviation Loss Learner whereby the goal is to optimize anomaly scores so that anomalies deviate significantly from normal samples.

The framework is tested on a variety of image datasets for identifying defects, planetary bodies, and medical anomalies. DevNet is tested against five other models and performs better on 7 out of 9 datasets. DevNet is able to achieve an AUC score between 80% to 98% amongst all of the datasets. As for explainability, the authors demonstrate that the algorithm can display the anomalous region on an image. DevNet generates both a black-white image of the location of the defect and an overlaid image showing where the defect lies on the original image.

B. WHITE BOX X-IDS MODELS

Models that can provide an explanation to expert users without utilizing additional models are referred to as *interpretable* or *white box models* [88]. A white box model's internal logic and programming steps are completely transparent, resulting in an interpretable decision process [114]. However, when the model is to be explained to non-expert users, it may demand post-hoc explainability, such as visualizations [64]. This interpretability, on the other hand, usually comes at a price in terms of performance [115].

A myriad of white box approaches are available for intrusion detection. Our survey will focus on the approaches most commonly used in the literature, as per our overview presented in Section III and the taxonomy showcased in Figure 2. Table 2 summarizes state-of-the-art research, challenges, and contributions with respect to white box approaches for intrusion detection systems.

1) REGRESSION

Linear Regression (LR) is a supervised ML technique that establishes a relationship between a dependent variable and independent variables by computing a *best fit* line. The linearity of the learned relationship puts LR under the umbrella of interpretable models.

Various regression-based IDS models exist in the literature. Subba et al. [122] deployed anomaly-based intrusion detection systems using two different statistical methods: Linear Discriminant Analysis (LDA) and LoR. While LR models are desirable for intrusion detection purposes, their performance is susceptible to outliers [123]. To mitigate the impact of outliers, the same authors proposed a robust regression method for anomaly detection [124]. The proposed method uses heteroscedasticity and a huber loss function instead of homoscedasticity and sum of squared errors.

While the existing approaches render promising outcomes, none of them were designed with *explainability* in mind. To overcome the issue of *explainability* in the area of *hardware performance counter* (HPC) – based intrusion detection, Kuruvila et al. [116] propose an explainable HPC-based Double Regression (HPCDR) ML framework. The study examines two distinct types of attacks: microarchitectural and malware. For the first type of attack, tests are conducted on five distinct datasets: Rowhammer, Flush+Flush, Spectre, Meltdown, and ZombieLoad. For the second attack, two distinct datasets are considered: Bashlite and PNScan. To minimize computational overhead, the proposed study employs Ridge Regression (RR) rather than Shapely values to generate interpretable results. First, the three ML models (RF, DT, and NN) are chosen to evaluate the classification accuracy. Second, the output from these models is perturbed and passed to the first RR model where HPCs are employed as features and weight coefficients are received. These furnished coefficients are run on the second RR model, which identifies the most malicious sample. The authors argue that by utilizing double regression techniques, their proposed method provides transparency, which enables users to locate malicious instructions within the program.

2) DECISION TREE AND RULE BASED

A Decision Tree (DT) is a tree structure with decision support system elements based on graph theory. In contrast to LR method, it works even when the relationship between input and output is nonlinear. In their simplest form, DT possesses three properties that make them interpretable [29]: simulatability, decomposability, and algorithmic transparency.

A simple rule is typically represented as a logical implication of IF-THEN statements by combining relational statements to form their knowledge [125]. These rules can be extracted from DT. Rule-based models are considered transparent because they generate rules to explain their predictions.

Mahbooba et al. [117] approach the task of developing an interpretable model to identify malicious nodes for IDS using a DT on the KDD dataset. They chose the Iterative

TABLE 2. A summary of the existing literature on white-box approaches to intrusion detection systems, with an emphasis on their scope, contribution, and limitations.

Paper Title	Focus/Objective	Contribution	Limitation
Regression based IDS			
Explainable Machine Learning for Intrusion Detection via Hardware Performance Counters [116]	To develop an explainable X-IDS technique based on the double RR technique and utilizing HPC as a feature.	• Proposes an explainable HPC-based Double Regression (HPCDR) framework for intrusion detection with human-interpretable results. • HPCDR is evaluated against real-world malware to determine whether it provides transparent hardware-assisted malware detection and to detect microarchitectural attacks with an indication of the malicious origin.	• DL models were not chosen to evaluate the optimal ML model. • Only Four HPC features were chosen for experimentation. • Other microarchitectural attacks (e.g. Prime+Probe) and malware (e.g. Rootkits) are not considered in the study.
Decision Tree and Rule based IDS			
XAI to Enhance Trust Management in Intrusion Detection Systems Using Decision Tree Model [117]	Focused on the interpretability in a widely used benchmark dataset KDD datasets.	• Addressed XAI concept to enhance trust management that human expert can understand. • Analyzed the importance of feature based on the entropy measure for intrusion detection. • Interpreted the rules extracted from the DT approach for intrusion classification.	• Information gain in decision trees is biased in favor of those attributes with more levels. • This behavior might impact prediction performance.
A Hybrid Approach for an Interpretable and Explainable Intrusion Detection System [118]	To design interpretable and explainable hybrid intrusion detection system to achieve better and more long-lasting security.	• Providing an IDS that stands out for its ML support on populating the knowledge base. • Focus on interpretability and explainability, since it justifies the suggested rules, and the diagnosis performed to each asset.	• DT only considered as ML model for system design. • Knowledge base is small.
Statistical and Probabilistic Models			
A New Method for Flow-Based Network Intrusion Detection Using the Inverse Potts Model [119]	To develop a new method for flow-based network intrusion detection using inverse statistical method.	• Implementation of a naturally interpretable flow classifier based on the inverse Potts model to be employed in NIDS. • Performance comparison with other ML based models using three datasets.	• Only binary classification is considered in the approach. • Applicability of the proposed methods in real world data.
Clustering based IDS			
Explainable unsupervised machine learning for cyber-physical systems [120]	Propose a novel Explainable Unsupervised Machine Learning (XUnML) approach using the Self Organizing Map (SOM) algorithm.	• Brief overview of Supervised Machine Learning (SML), Unsupervised Machine Learning (UnML), and XAI. • Exploring initial desiderata towards Explainable UnML (XUnML), defining XUnML terminology based on the terminology used for XAI, and exploring the necessity of XUnML for CPSs.	• Only clustering method is used.
ANNaBell Island: A 3D Color Hexagonal SOM for Visual Intrusion Detection [121]	Provide explanation to the outputs of SOM models using color scheme and island landscape analogy for different network traffics.	• Benign and malicious traffic is separated by color coding and zoning in the island. • Color and zone categorization of network traffic provides the explanation of the output.	• It is not clear if the temporal map maintain same basic landscape or change over time. • The proposed map seems to be specific to the tested network only.

Dichotomiser 3 (ID3) algorithm to ensure interpretability because it mimics a human-based decision strategy. The authors demonstrate that the algorithm can rank the relevance of features, provide explainable rules, and reach a level of accuracy comparable to state-of-the-art. Another explainable decision tree model is proposed in [125] and [126], with the latter being an extension of work in [127].

Sinclair et al. [128] extract rules using a DT and a Genetic Algorithm (GA) for improving the performance of the IDS model. The authors in [129] and [130] focus on optimizing

the IDS model by extracting rules using a GA. To add transparency to the decision process, Dias et al. [118] proposed an interpretable and explainable hybrid intrusion detection system. The proposed system integrates expert-written rules and dynamic knowledge generated by a DT algorithm. The authors suggest that the model can achieve explainability through the justifications of each diagnosis. Justification of certain predictions is provided in a tree-like format in the form of a suggested rule that provides a more intuitive and straightforward understanding of the diagnosis.

Snort is the world's most widely used open-source rule-based intrusion prevention system (IPS) [131]. It employs a set of rules that help define malicious network activity. These rules are then used to identify packets and generate alerts for users [131], [132].

3) STATISTICAL AND PROBABILISTIC METHODS

In statistics, the mean, standard deviation, and any other type of correlation are referred to as moments [133]. Statistical and probabilistic methods use this information to determine whether the given event is anomalous or not. The moment is predicted anomalous if they are either above or below a predefined interval. This approach is further divided into the univariate, multivariate, time series, parametric, non-parametric, operational and Markov models [133], [134], [135], [136].

Various IDS based on statistical and probabilistic models have been proposed. IDS based on the mean and standard deviation is explained in [137], while a study relating to multivariate modeling is proposed in [138]. Gyanchandani et al. [139] proposed an IDS based on the Markov process.

A different approach to intrinsically explainable statistical methods for network intrusion detection is proposed by Pontes et al. [119]. They introduce a novel Energy-based Flow Classifier (EFC) that utilizes inverse Potts models to infer anomaly scores based on labeled benign examples. This method is capable of accurately performing binary flow classification on DDoS attacks. They perform experiments on three different datasets: CIDDS-001, CIC-IDS2017, and CICDDoS19. Results indicate that the proposed model is more adaptable to different data distributions than classical ML-based classifiers. Additionally, they argue that their model is naturally interpretable and that individual parameter values can be analyzed in detail.

4) CLUSTERING

Clustering is the most widely used strategy for unsupervised ML. It classifies samples according to a similarity criterion. Clustering algorithms that can be explained have several advantages. The primary benefit of explainable clustering is that it summarizes the input behavior patterns within clusters, enabling users to comprehend the clusters' underlying commonalities [120]. As stated in Section III-A there are various clustering algorithms available. However, in the context of X-IDS, we will only focus on Self-Organizing Maps (SOMs).

SOMs are an unsupervised clustering technique within the artificial neural networks umbrella. It has two layers: an input layer that accepts high dimensional space and an output layer that generates a non-linear mapping of high-dimensional space into reduced dimensions. It is trained to produce a low dimensional representation of a large training dataset while preserving important topological and metric relationships of the input data [140].

An anomaly detection system using SOM techniques based on offline audit trail data is proposed in [141]. The major shortcoming of the proposed system is it does not allow for real-time detection. On the other hand, the authors in [142] propose Hierarchical SOMs (HSOM) for host-based intrusion detection on computer networks that are capable of operating on real-time data without requiring extensive offline training or expert knowledge. Another model based on HSOM is proposed in [143]. Wickramasinghe et al. [120] developed a novel model-specific explainable technique for the SOM algorithm that generates both local and global explanations for Cyber-Physical Systems (CPS) security. They used the SOMs training approach (winner-take-all algorithm) together with visual data mining capabilities (Histograms, t-SNE, Heat Maps, and U-Matrix) of SOMs to make the algorithm explainable.

A 3D color hexagonal SOM for visual intrusion detection called ANNaBell Island is proposed in [121] which is an extension of 1D ANNaBell reported in [144] and [145]. To make the SOM process and its output explainable to the users, the authors designed a hexagonal SOM in a meta-hexagonal layout, referred to as an island, that graphically displayed features of network traffic. The output of the SOM model was used to create a color-separated 3D landscaped island that represents various types of network traffic, distinguishing between malicious and normal behavior.

After surveying the current black box and white box approaches to X-IDS, we propose in the next section, a generic explainable architecture with a user-centric approach for designing X-IDS that can accommodate a wide variety of scenarios and applications without adhering to a specification or technological solution.

V. DESIGNING AN EXPLAINABLE IDS (X-IDS)

The purpose of an IDS is to continuously monitor a network for malicious activity or security violations known as incidents of intrusion. If found, intrusions are reported to the cybersecurity professional responsible for monitoring such systems. A significant problem with AI based IDS is their high false positive and false negative rates. Recently, many IDS based on ML/DL techniques have been proposed to address this issue, such as DNN [33], [146], RNN [147], [148], and CNN [149], [150]. These techniques yield unprecedented detection accuracy. However, the effective use of these approaches require using high-quality data, as well as a considerable amount of computing resources [151]. Additionally, this modeling approach has

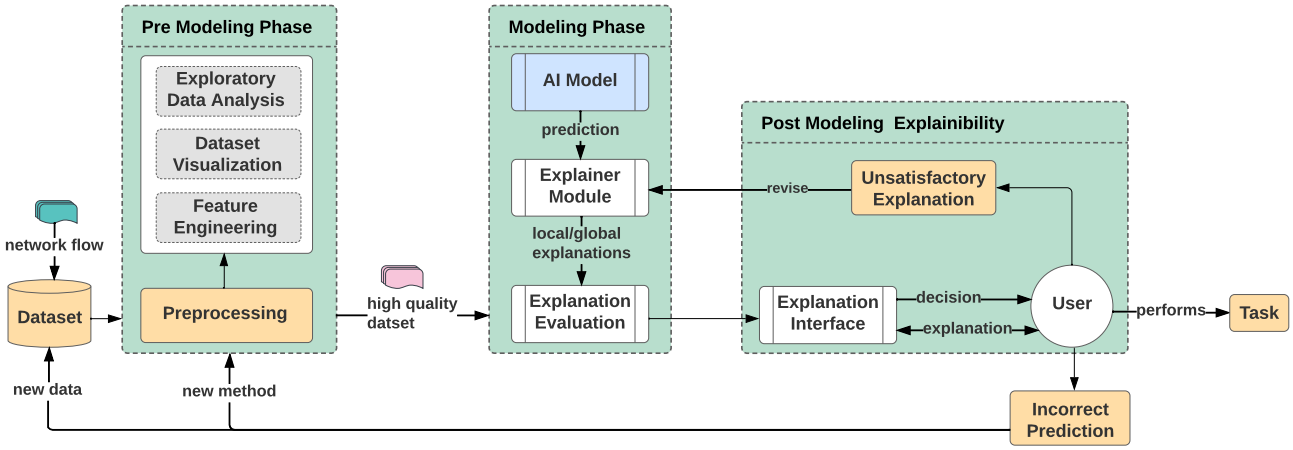


FIGURE 4. Recommended architecture for the design of an X-IDS based on DARPA [24]. The layered architecture is divided into three phases: pre-modeling, modeling, and post-modeling explainability. Each phase contributes to the development of an explanation for various stakeholders, thereby assisting in decision-making.

typically suffered from model bias, a lack of decision process transparency, and a lack of user trust.

The IDS systems based on ML/DL techniques are designed to generate event logs in the form of 'benign' or 'malicious' classification reports, that can be further analyzed by CSoC analysts. However, they do not showcase the connection between the inputs and output (i.e., they fail to indicate the reasoning behind the decision). To be more precise, a cybersecurity specialist serves as a user who reviews IDS results, but is not a component of the intrusion detection process [152]. In turn, this creates a larger problem for CSoC experts, as they are unable to optimize their decisions based on the model's decision process.

To address this semantic gap, one promising technique is to design X-IDS with a human-in-the-loop approach. Typically, methods that are retraceable, explainable, and supported by visualizations amplify cybersecurity analysts' understanding in managing cybersecurity incidents in both proactive and reactive manners.

In the following sub-sections, we explain the recommended architecture, as depicted in Figure 4, that could be used as guidance to design an Explainable Intrusion Detection Systems (X-IDS). The X-IDS architecture proposed in this paper is based on the DARPA recommended architecture for the design of XAI systems [24]. The layered architecture consists of three phases: *pre-modeling phase*, *modeling phase*, and *post-modeling explainability phase*. In each phase, different modules work in tandem to provide CSoC analysts with more accurate and explainable output. We believe that this architecture is sufficiently generic to accommodate a variety of scenarios and applications without adhering to a particular specification or technological solution.

A. PRE-MODELING PHASE

The first phase is a pre-modeling phase. The input of this module is raw network flow (dataset) and the output is a high-quality dataset. In the following subsections, we will first describe different benchmark datasets available for Intrusion Detection. We then present common data preprocessing techniques used in the literature.

1) DATASETS

While access to representative, labelled datasets for cybersecurity related AI tasks remains a challenge, a variety of publicly accessible datasets can be used to train and benchmark X-IDS. These are unprocessed network flows extracted from packet captures. To address privacy concerns, many of these datasets are generated in an emulated environment. NSL-KDD [153], based on the KDD-CUP-99 [154], is a dataset frequently present in the literature. Although old, its use allows comparisons with previous works. NSL-KDD is relatively small compared to other datasets in the field. A more modern dataset is CICIDS2017 [155], which contains more up-to-date attacks and network flows. In addition, it includes 3 million samples, which allows scalability testing. Another noteworthy dataset is UGR [156], a multi-terabyte dataset collected over the course of 5 months. This dataset is built to test IDS for long-term trends. The authors state that their dataset captures potential trends in daytime, nighttime, weekday, and weekend traffic.

These publicly available datasets, though good for benchmarking, are not suitable for deployable systems. We recommend, CSoC users deploying X-IDS systems evaluate these systems on organizational representative datasets.

2) EXPLORATORY DATA ANALYSIS (EDA) AND DATA VISUALIZATION

Data preprocessing is essential for increasing the likelihood of ML models producing accurate predictions. Using Exploratory Data Analysis (EDA), one can gain a general understanding of a dataset's key features and characteristics. To comprehend the features, visualization techniques such as heat maps, network diagrams, bar charts, and correlation matrices may be used. Once a comprehension of feature

space has been attained, the data is forwarded to the feature engineering model for further processing.

3) FEATURE ENGINEERING

The general trend in preprocessing IDS datasets is to normalize the numerical features and to One-Hot Encode (OHE) the categorical features. After the datasets are encoded, their feature space can be quite large which makes them computationally expensive. Two approaches to reducing dimensions are widely discussed in the literature: Feature Selection and Feature Extraction.

Feature selection techniques are used to reduce the feature space by selecting a subset of features without transforming them. There are three types of feature selection techniques popular in the IDS domain: filters, wrappers, and the embedded/hybrid method [157]. Apart from these, libraries such as Scikit-Learn [158] have also been used in published works for feature selection.

Another technique used in feature engineering is feature extraction (also known as dimensionality reduction). Feature extraction reduces the size of the feature space by transforming the original features while retaining most of their defining attributes. The most commonly used feature extraction technique in the literature is the Principal Component Analysis (PCA) [159]. PCA is an unsupervised method that does not require class knowledge to identify features. It also facilitates the identification of correlations and relationships between the features of a dataset.

B. MODELING PHASE

The second phase is the modeling phase. The input of this phase is the high-quality dataset generated in the pre-processing phase and the output is the explanations generated by the explainer module. First, the high-quality dataset is fed into the ML/DL model of choice. Second, the predictions generated by the model in use are passed through an explainer module. Third, these explanations are evaluated by an evaluation module. This process enables users to understand the reason behind certain predictions, which in turn, helps the CSoC analysts in their decision-making process.

1) AI MODEL

In Section IV, we discussed two different approaches which are currently being employed by different authors to create X-IDS: the black box and the white box. AI modules in these approaches generate predictions. However, there is a trade-off between the accuracy and the interpretability with these approaches. The white box approaches are popular for their interpretability, while the black box approaches are known for their prediction accuracy. In context of IDS, high prediction accuracy is required to prevent attacks. Moreover, black box models can capture significant non-linearity and complex interactions between data that white box models are not able to capture. For example, Recurrent Neural Networks (RNN) can capture temporal dependencies between samples. On the other hand, models like Support Vector Machine (SVM) and Deep Neural Network (DNN) can create their own representation of data. which might be helpful to discover unknown attacks. For this reason, we believe that future X-IDS should be built using black box models.

In our literature review, we found that authors use a variety of black box algorithms in their work, such as SVM, CNN, RF, and MLP, which prove to be quite effective. Another popular algorithm of choice in the intrusion detection domain is a variant of the RNN, referred to as Long Short-Term Memory (LSTM). Recently, Generative Adversarial Networks (GAN) have also become relatively popular. Consequently, there are a multitude of black box algorithms from which to select. Explainer modules then approximate the prediction generated by AI module employing a white box or black box algorithms.

2) EXPLAINER MODULE AND EVALUATION

The prediction generated by the model of choice in the AI module is then fed to the explainer module. The common explainers used from previous works include LIME, SHAP, and LRP. These out-of-the-box modules allow for quick testing on different algorithms and datasets. However, there are some problems with solely using these approaches in future X-IDS works. To begin, methods such as SHAP do not run in real-time. Therefore, it may be time-consuming to attempt to use SHAP on a simple Multi-Layer Perceptron classifier with a large feature space dataset. In X-IDS, both predictions and explanations must be made as quickly as possible. Secondly, these approaches are not always designed with X-IDS stakeholders in mind.

At present, there are no set standard metrics to evaluate explanations. Several authors have attempted to evaluate explanations in various ways. In Section II-C we described different ways to evaluate the explanations. Metrics such as application grounded evaluation, human-grounded evaluations, and function-grounded evaluation proposed by Doshi et al. [81] can be used as a baseline to evaluate the explanation generated by X-IDS. A noteworthy method to evaluate the effectiveness of explanations is proposed by authors in [24]. Figure 6 illustrates their approach.

C. POST MODELING EXPLAINABILITY PHASE

The third and final phase is the post-modeling explainability phase. This phase has two major components: the explanation interface and users. The recommendation, decision, or action generated by the AI module, explained by an explainer module, and evaluated by an explanation evaluation module is rendered in a graphical user interface (explanation interface). The users, on the other hand, use this interface to make an informed decision.

1) EXPLANATION INTERFACE

The custom visual dashboards are created to help the user to understand the X-IDS. An excellent approach to building such an explanation interface is found in the work by [102] and [98]. The engineers who design X-IDS can use this approach as guidance to create their explanation interface.

Furthermore, this paper also recommends that future X-IDS developers make custom explainers built for specific stakeholders to help improve explainability. Open-source toolkits and libraries are also available that create a visual dashboard and explain the prediction. One such library is Shapash [160]. It is an overlay package to other intelligibility libraries, such as Shap and Lime, that are dedicated to the interpretability of models. Another example of the library for quickly building interactive dashboards for analyzing and explaining the predictions and workings of sci-kit-learn machine learning models is explainerdashboard [161].

2) USERS

For this paper, the stakeholders will consist of developers, defense practitioners, and investors. Section VI-B discusses the need for defining the identity of the stakeholders of an X-IDS system. The developers are tasked with creating, modifying, and maintaining the X-IDS. The defense practitioners guard the assets of the investors. Lastly, the investors make budgeting decisions for the benefit of the X-IDS system and other assets. These three audiences have distinct tasks and explainability requirements that must be addressed differently by the X-IDS. An explanation interface designed from the user's perspective can bridge this gap.

If an explanation is unclear or unhelpful, the stakeholders will need a way to voice that opinion. For example, a set of explanations is too complicated for a group of investors. Investors may ask for additional explanations that simplify or even link to a web page that can teach them more on a subject. In such a situation, the developers can revise the explainer module to fit the users request or needs. This could also include making a new explainer module or updating to a new state-of-the-art module like those in AIX360 [80]. For the same reasons, incorrect predictions and explanations need to be corrected and updated. The developers or defense practitioners will then need to introduce new data to the model. Moreover, a different method of data preprocessing may be required to augment the efficacy of the model.

To make the recommended X-IDS architecture as shown in Figure 4 a reality, researchers need to study different aspects of the three phases. To this end, in the next section we discuss various challenges inherent in designing the proposed X-IDS architecture and make research recommendations aimed at effectively mitigating these challenges for future researchers interested in developing X-IDS.

VI. RESEARCH CHALLENGES AND RECOMMENDATIONS

The sub-domain of explainable AI based Intrusion Detection Systems is still in its infancy. Researchers working on X-IDS must be made aware of the issues that hinder its development. The issues that we described in Section II such as finding the right notion of explainability, generating explanations from a stakeholder's perspective, and lack of formal standard metrics to evaluate explanations are prevalent in the X-IDS domain as well. Existing X-IDS research is primarily focused on the goal of making algorithms explainable. Explanations are not being designed around stakeholders, and researchers need to quantify useful evaluation metrics. Apart from these challenges, issues pertaining to IDS may also pose a problem for X-IDS. There are many promising avenues of exploration, in this section we detail some existing research challenges and give our recommendations.

A. DEFINING EXPLAINABILITY FOR INTRUSION DETECTION

The first problem faced by researchers designing X-IDS is the lack of consensus on the definition of explainability in IDS. The research community needs to agree on a common definition of explainability for IDS. To find common ground, we can leverage the foundational XAI definition proposed by DARPA [31]. However, an X-IDS definition needs more security domain-specific elements. The inclusion of the CIA principles may be a good start for cementing a definition that combines aspects of cybersecurity and XAI.

Questions relevant to the X-IDS that researchers need to answer include: ''What is explainability when used for intrusion detection?'', ''How do we effectively create explanations for IDS?'', and ''Who are we creating explanations for?''. Other questions such as ''How can Confidentiality, Integrity, and Availability benefit from explanations?'' and ''How do we categorize X-IDS algorithms?'' should be reassessed by X-IDS researchers as well. Current work is extremely narrow in its scope and limits its objective to explaining each sample in an IDS dataset. These works also do not consider the type of audience when building X-IDS.

B. DEFINING TASKS AND STAKEHOLDERS

The second challenge is to define the task and the stakeholders of the X-IDS ecosystem. After formalizing the definition of 'explainability' for X-IDS, we need to create explanations tailored to the stakeholders. Figure 5 demonstrates a simple user and explanation taxonomy. We consider three major stakeholders based on their roles in this taxonomy including *IDS developers*, *security analyst*, and *investors*. Each of the stakeholder categories necessitates a different degree of explanation and visualization. Developers and CSoC members are more familiar with the field and may want more complex explanations. Investors and managers, on the other hand, may be more satisfied with summarized visualizations. Each user group performs varying tasks based on the explanations. Programmers will work to debug and increase the efficacy of the AI model. CSoC members will be tasked with protecting investor assets. Indirectly, investors will need to make hiring or budgeting decisions. Take for example a corporate, network security system. The set of stakeholders consists of IDS developers, security analysts, upper-level managers, and team managers. In such a scenario, IDS developers will be creating and/or updating the corporation's IDS. These developers will want the IDS to return which features from local and global explanations are making the most impact. Additionally, as attacks change over time, these

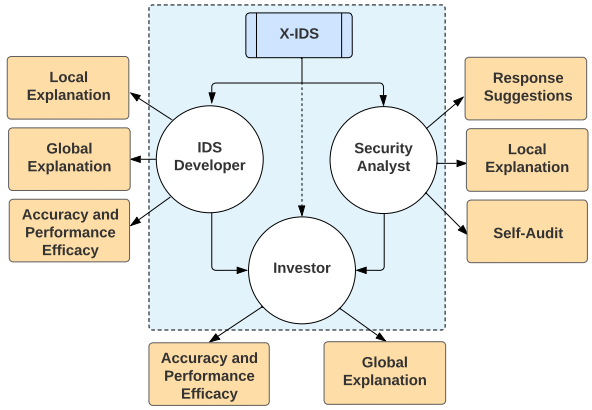


FIGURE 5. A simple taxonomy illustrating the importance of tailoring explanations to specific stakeholders based on their roles in CSoCs.

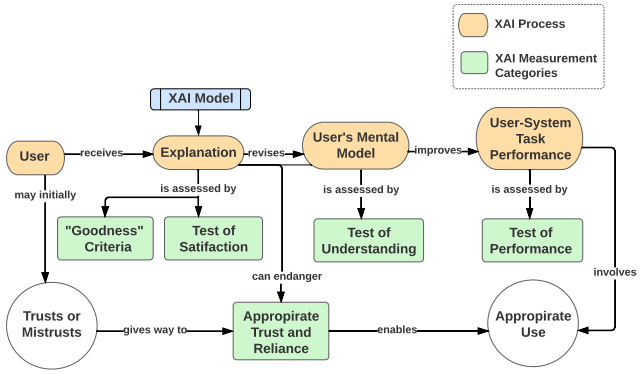


FIGURE 6. Different categories for assessing the effectiveness of explanations in the IHCM psychological model with detailed explanation process [24].

individuals will want explanations potentially keeping track of these changes. New attacks can be known to exist if the performance begins to degrade, so accuracy and performance metrics will be integral to maintaining the IDS. Security analyst or server admins would benefit in a similar way to IDS developers. Having potential leads to base their actions on could lower system down time. The different managerial levels could be using certain other metrics or explanations to aid them in leadership decisions. Hiring more staff, adding more funding, or deciding to pivot to a new system would be some actions managers could take. The needs of each group are different and future research is needed to determine the best types of explanations that will benefit each group the most.

C. EVALUATION METRICS

The third challenge in designing X-IDS is evaluating the explanation generated by the 'explainer module'. Finding the best explanation for each stakeholder category requires customized evaluation metrics. Currently, there is no consensus on metrics for explanations. In Section II-C we described a body of literature proposing various evaluation metrics that could be used towards evaluating explanations. In particular, we recommended evaluation metrics proposed by authors in [81] to evaluate explanations for X-IDS in Section V-B. Another notable work that could serve as a baseline for evaluating explanations is the psychological model of explanation created by the Florida Institute for Human and Machine Cognition (IHMC) [24]. The proposed model is illustrated in Figure 6. The user receives an explanation from the XAI model. This explanation can be tested for ''goodness'' and the satisfaction of the user/stakeholder. The user then revises their mental model of the XAI system. Their understanding of the system can be tested. Tasks are performed based on the explanation. The IHMC model merges the purpose of the XAI model, with the task and mindset of the user.A noteworthy method to evaluate the effectiveness of explanations is proposed by authors in [24]. Figure 6 illustrates their approach.

D. ADVERSERIAL AI

Adversarial AI refers to the use of artificial intelligence for malicious purposes, including attacks on other artificial intelligence systems to evade detection [162] or to poison data [163]. Malicious actors can potentially attack the classifiers that are used to generate predictions and cause misclassification. In context of X-IDS, the explanations generated by the explainer module may become a new point of attack for malicious actors. Attackers may add, delete, or modify explanations to evade detection [164]. Attackers may also attack training datasets to alter the explainer's behavior. The methods and effects of these attacks will need to be explored. Defense techniques must be created to correct attacked explanations. Studies to defend IDS against adversarial attack include [165], [166], and [167], etc. Study-specific to adversarial approach for X-IDS is discussed in [6].

E. MISLEADING/INCORRECT EXPLANATIONS

An explanation does not have to be attacked to be misleading. The explanation itself may be misleading, or the user may interpret the explanation incorrectly. This may lead to circumstances where the model is correct and the user is the problem. The explainer will need to be modified to prevent user error in such situations.

Explanations that are misclassified either by an attack or due to the poor quality of data can have a significant negative impact on CSoCs. CSoCs security analysts should always critically analyze the reasoning behind the prediction. Moreover, methods for auditing previously incorrect explanations should be created. Ideally, the X-IDS should be able to audit itself and generate explanations for the audit.

F. SCALABILITY AND PERFORMANCE

Performance is of utmost importance for an IDS. CSoCs can incur losses for lost time. Explanations should not needlessly slow down an IDS. So how do we optimize an X-IDS? One approach is that the explainer could generate explanations for every sample it sees, or it could strategically choose which samples to explain. A comprehensive analysis of the CPU, RAM, and disk usage should be run on current and future explainers.

VII. CONCLUSION

The exponential growth of cyber networks and the myriad applications that run on them have made CSoC, Cyber-physical systems, and critical infrastructure vulnerable to cyber-attacks. Securing these domains and their resources through the use of defense tools such as IDS, is critical to combating and resolving this issue [10], [11]. Recent AI-based IDS research has demonstrated unprecedented prediction accuracy, which is helping to lead to its widespread adoption across the industry. CSoC analysts largely rely on the results of these models to make their decision. However, in most cases, decision-making is impaired simply because these opaque models fail to justify their predicted outcomes. A solution to this problem is to embrace the concept of 'explainability' in these models. This, in turn, may facilitate quick interpretation of prediction, making it more feasible for CSoC analysts to accelerate response times.

A systematic review of current state-of-the-art research on 'XAI' or 'explainability' highlighted some key challenges in this domain, such as the lack of consensus surrounding the definition of 'explainability', the need to formalize explainability from the user's perspective, and the lack of metrics to evaluate explanations. We propose a taxonomy to address this problem with a focus on its relevance and applicability to the domain of intrusion detection.

In this paper, we present in detail two distinct approaches found in the body of literature which address the concern of 'explainability' in the IDS domain, including the white box approach and the black box approach. The white box approach makes the model in use inherently interpretable, whereas the black box approach requires post-hoc explanation techniques to make the predictions more interpretable (e.g., LIME [71], SHAP [32]).While the former approach may provide a more detailed explanation to assist CSoC members in decision-making, its prediction performance is in general outperformed by the latter. Nevertheless, the field of IDS requires a high degree of precision to prevent attacks and avoid false positives. Bearing this in mind, a black box approach is recommended when developing an X-IDS solution.

In addition, we also propose a three-layered architecture for the design of an X-IDS based on the DARPA recommended architecture [24] for the design of XAI systems. This architecture is sufficiently generic to support a wide variety of scenarios and applications without being bound by a particular specification or technological solution. Finally, we provide research recommendations to researchers that are interested in developing X-IDS.

ACKNOWLEDGMENT

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Army ERDC or the U.S. DoD.

REFERENCES

[1] D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim, and K. J. Kim, "A survey of deep learning-based network anomaly detection," *Cluster Comput.*, vol. 22, pp. 949–961, Jan. 2019.
[2] S. Sridhar, A. Hahn, and M. Govindarasu, "Cyber–physical system security for the electric power grid," *Proc. IEEE*, vol. 100, no. 1, pp. 210–224, Jan. 2011.
[3] R. Rajkumar, I. Lee, L. Sha, and J. Stankovic, "Cyber-physical systems: The next computing revolution," in *Proc. 47th Design Autom. Conf. (DAC)*, 2010, pp. 731–736.
[4] A. Cardenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, and S. Sastry, "Challenges for securing cyber physical systems," in *Proc. Workshop Future Directions Cyber-Phys. Syst. Secur.*, vol. 5, 2009.
[5] M. Ahmadian and D. C. Marinescu, "Information leakage in cloud data warehouses," *IEEE Trans. Sustain. Comput.*, vol. 5, no. 2, pp. 192–203, Apr. 2018.
[6] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable AI in intrusion detection systems," in *Proc. 44th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct. 2018, pp. 3237–3243.
[7] V. Cardellini, E. Casalicchio, S. Iannucci, M. Lucantonio, S. Mittal, D. Panigrahi, and A. Silvi, "An intrusion response system utilizing deep Q-networks and system partitions," 2022, *arXiv:2202.08182*.
[8] K. Sane, K. P. Joshi, and S. Mittal, "Semantically rich framework to automate cyber insurance services," *IEEE Trans. Services Comput.*, early access, Sep. 16, 2021, doi: 10.1109/TSC.2021.3113272.
[9] A. McDole, M. Gupta, M. Abdelsalam, S. Mittal, and M. Alazab, "Deep learning techniques for behavioural malware analysis in cloud iaas," in *Malware Analysis Using Artificial Intelligence and Deep Learning*. Springer, 2021.
[10] S. Wali and I. Khan, "Explainable AI and random forest based reliable intrusion detection system," Tech. Rep., 2021.
[11] M. Wang, K. Zheng, Y. Yang, and X. Wang, "An explainable machine learning framework for intrusion detection systems," *IEEE Access*, vol. 8, pp. 73127–73141, 2020.
[12] *Cyber Security Operations Center (CSOC)*, Raytheon, 2017.
[13] J. P. Anderson, "Computer security threat monitoring and surveillance," Tech. Rep., 1980.
[14] D. E. Denning, "An intrusion-detection model," *IEEE Trans. Softw. Eng.*, vol. SE-2, no. 2, pp. 222–232, Feb. 1987.
[15] R. G. Bace and P. Mell, "Intrusion detection systems," Tech. Rep., 2001.
[16] S. X. Wu and W. Banzhaf, "The use of computational intelligence in intrusion detection systems: A review," *Appl. Soft Comput.*, vol. 10, no. 1, pp. 1–35, Jan. 2010.
[17] B. Mukherjee, L. T. Heberlein, and K. N. Levitt, "Network intrusion detection," *IEEE Netw.*, vol. 8, no. 3, pp. 26–41, May 1994.
[18] W. Lee, S. J. Stolfo, and K. W. Mok, "A data mining framework for building intrusion detection models," in *Proc. IEEE Symp. Secur. Privacy*, May 1999, pp. 120–132.
[19] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2015.
[20] M. Belouch, S. El Hadaj, and M. Idhammad, "Performance evaluation of intrusion detection based on machine learning using apache spark," *Proc. Comput. Sci.*, vol. 127, pp. 1–6, Jan. 2018.
[21] E. Aminanto and K. Kim, "Deep learning in intrusion detection system: An overview," in *Proc. Int. Res. Conf. Eng. Technol. (IRCET)*, 2016.
[22] K. Kim and M. E. Aminanto, "Deep learning in intrusion detection perspective: Overview and further challenges," in *Proc. Int. Workshop Big Data Inf. Secur. (IWBIS)*, Sep. 2017, pp. 5–10.
[23] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable AI: A brief survey on history, research areas, approaches and challenges," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.* Springer, 2019, pp. 563–574.
[24] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, Jun. 2019.
[25] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," 2017, *arXiv:1708.08296*.
[26] M. Alaa, "Artificial intelligence: Explainability, ethical issues and bias," *Ann. Robot. Autom.*, vol. 5, no. 1, pp. 34–37, Aug. 2021.
[27] R. A. Berk and J. Bleich, "Statistical procedures for forecasting criminal behavior: A comparative assessment," *Criminol. Pub. Pol'y*, vol. 12, p. 513, Jun. 2013.

[28] Q. P. Nguyen, K. W. Lim, D. M. Divakaran, K. H. Low, and M. C. Chan, "GEE: A gradient-based explainable variational autoencoder for network anomaly detection," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Jun. 2019, pp. 91–99.

[29] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.

[30] M. V. Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in *Proc. Nat. Conf. Artif. Intell.* Menlo Park, CA, USA: MIT Press, 2004, pp. 900–907.

[31] *Broad Agency Announcement Explainable Artificial Intelligence (XAI)*, DARPA-BAA-16–53, DARPA, 2016, pp. 7–8.

[32] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[33] K. Amarasinghe and M. Manic, "Improving user trust on deep neural networks based intrusion detection systems," in *Proc. 44th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct. 2018, pp. 3262–3268.

[34] J. Haspiel, N. Du, J. Meyerson, L. P. Robert, D. Tilbury, X. J. Yang, and A. K. Pradhan, "Explanations and expectations: Trust building in automated vehicles," in *Proc. Companion ACM/IEEE Int. Conf. Hum.-Robot Interact.*, Mar. 2018, pp. 119–120.

[35] M. P. S. Lorente, E. M. Lopez, L. A. Florez, A. L. Espino, J. A. I. Martínez, and A. S. de Miguel, "Explaining deep learning-based driver models," *Appl. Sci.*, vol. 11, no. 8, p. 3321, Apr. 2021.

[36] Y. Li, H. Wang, L. M. Dang, T. N. Nguyen, D. Han, A. Lee, I. Jang, and H. Moon, "A deep learning-based hybrid framework for object detection and recognition in autonomous driving," *IEEE Access*, vol. 8, pp. 194228–194239, 2020.

[37] J. Martinez-Cebrian, M.-A. Fernandez-Torres, and F. Diaz-De-Maria, "Interpretable global-local dynamics for the prediction of eye fixations in autonomous driving scenarios," *IEEE Access*, vol. 8, pp. 217068–217085, 2020.

[38] T. Ponn, T. Kröger, and F. Diermeyer, "Identification and explanation of challenging conditions for camera-based object detection of automated vehicles," *Sensors*, vol. 20, no. 13, p. 3699, Jul. 2020.

[39] A. Deeks, "The judicial demand for explainable artificial intelligence," *Columbia Law Rev.*, vol. 119, no. 7, pp. 1829–1850, 2019.

[40] O. Loyola-González, "Understanding the criminal behavior in Mexico City through an explainable artificial intelligence model," in *Proc. Mexican Int. Conf. Artif. Intell.* Springer, 2019, pp. 136–149.

[41] Q. Zhong, X. Fan, X. Luo, and F. Toni, "An explainable multi-attribute decision model based on argumentation," *Expert Syst. Appl.*, vol. 117, pp. 42–61, Mar. 2019.

[42] C. S. Vlek, H. Prakken, S. Renooij, and B. Verheij, "A method for explaining Bayesian networks for legal evidence with scenarios," *Artif. Intell. Law*, vol. 24, no. 3, pp. 285–324, Sep. 2016.

[43] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" 2017, *arXiv:1712.09923*.

[44] K. Gade, S. C. Geyik, K. Kenthapadi, V. Mithal, and A. Taly, "Explainable AI in industry," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 3203–3204.

[45] S. Itani, F. Lecron, and P. Fortemps, "A one-class classification decision tree based on kernel density estimation," *Appl. Soft Comput.*, vol. 91, Jun. 2020, Art. no. 106250.

[46] L. Lindsay, S. Coleman, D. Kerr, B. Taylor, and A. Moorhead, "Explainable artificial intelligence for falls prediction," in *Proc. Int. Conf. Adv. Comput. Data Sci.* Springer, 2020, pp. 76–84.

[47] E. Pintelas, M. Liaskos, I. E. Livieris, S. Kotsiantis, and P. Pintelas, "Explainable machine learning framework for image classification problems: Case study on glioma cancer prediction," *J. Imag.*, vol. 6, no. 6, p. 37, May 2020.

[48] E. Prifti, Y. Chevaleyre, B. Hanczar, E. Belda, A. Danchin, K. Clément, and J.-D. Zucker, "Interpretable and accurate prediction models for metagenomics data," *GigaScience*, vol. 9, no. 3, Mar. 2020.

[49] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K. W. Low, S. F. Newman, J. Kim, and S. I. Lee, "Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery," *BioRxiv*, Jan. 2017, Art. no. 206540.

[50] L.-C. Huang, W. Yeung, Y. Wang, H. Cheng, A. Venkat, S. Li, P. Ma, K. Rasheed, and N. Kannan, "Quantitative structure–mutation–activity relationship tests (QSMART) model for protein kinase inhibitor response prediction," *BMC Bioinf.*, vol. 21, no. 1, pp. 1–22, Dec. 2020.

[51] A. Anguita-Ruiz, A. Segura-Delgado, R. Alcalá, C. M. Aguilera, and J. Alcalá-Fdez, "EXplainable artificial intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research," *PLOS Comput. Biol.*, vol. 16, no. 4, Apr. 2020, Art. no. e1007792.

[52] S. M. Muddamsetty, M. N. Jahromi, and T. B. Moeslund, "Expert level evaluations for explainable ai (XAI) methods in the medical domain," in *Proc. Int. Conf. Pattern Recognit.* Springer, 2021, pp. 35–46.

[53] M. Graziani, V. Andrearczyk, S. Marchand-Maillet, and H. Müller, "Concept attribution: Explaining CNN decisions to physicians," *Comput. Biol. Med.*, vol. 123, Aug. 2020, Art. no. 103865.

[54] I. Rio-Torto, K. Fernandes, and L. F. Teixeira, "Understanding the decisions of CNNs: An in-model approach," *Pattern Recognit. Lett.*, vol. 133, pp. 373–380, May 2020.

[55] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[56] Y. E. Chun, S. B. Kim, J. Y. Lee, and J. H. Woo, "Study on credit rating model using explainable AI," *J. Korean Data Inf. Sci. Soc.*, vol. 32, no. 2, pp. 283–295, Mar. 2021.

[57] M. Han and J. Kim, "Joint banknote recognition and counterfeit detection using explainable artificial intelligence," *Sensors*, vol. 19, no. 16, p. 3607, Aug. 2019.

[58] E. Amparore, A. Perotti, and P. Bajardi, "To trust or not to trust an explanation: Using LEAF to evaluate local linear XAI methods," *PeerJ Comput. Sci.*, vol. 7, p. e479, Apr. 2021.

[59] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, "Evaluating XAI: A comparison of rule-based and example-based explanations," *Artif. Intell.*, vol. 291, Feb. 2021, Art. no. 103404.

[60] K. Sokol and P. Flach, "Explainability fact sheets: A framework for systematic assessment of explainable approaches," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 56–67.

[61] T. Rutkowski, K. Łapa, and R. Nielek, "On explainable fuzzy recommenders and their performance evaluation," *Int. J. Appl. Math. Comput. Sci.*, vol. 29, no. 3, pp. 595–610, Sep. 2019.

[62] X. Wang, D. Wang, and C. Xu, "Explainable reasoning over knowledge graphs for recommendation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Aug. 2019, pp. 5329–5336.

[63] G. Zhao, H. Fu, R. Song, T. Sakai, Z. Chen, X. Xie, and X. Qian, "Personalized reason generation for explainable song recommendation," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 4, pp. 1–21, Jul. 2019.

[64] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[65] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018.

[66] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.

[67] K. Gade, S. Geyik, K. Kenthapadi, V. Mithal, and A. Taly, "Explainable AI in industry: Practical challenges and lessons learned," in *Proc. Companion Proc. Web Conf.*, Apr. 2020, pp. 303–304.

[68] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Inf. Fusion*, vol. 76, pp. 89–106, Dec. 2021.

[69] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," 2020, *arXiv:2006.11371*.

[70] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020.

[71] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[72] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–9.

[73] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *J. Amer. Stat. Assoc.*, vol. 113, no. 523, pp. 1094–1111, Jul. 2018.

[74] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[75] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," in *Proc. Int. Conf. Artif. Neural Netw.* Springer, 2016, pp. 63–71.

[76] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, and F. Viegas, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2668–2677.

[77] C. Yang, A. Rangarajan, and S. Ranka, "Global model interpretation via recursive partitioning," in *Proc. IEEE 20th Int. Conf. High Perform. Comput. Communications; IEEE 16th Int. Conf. Smart City; IEEE 4th Int. Conf. Data Sci. Syst. (HPCC/SmartCity/DSS)*, Jun. 2018, pp. 1563–1570.

[78] M. A. Valenzuela-Escárcega, A. Nagesh, and M. Surdeanu, "Lightly-supervised representation learning with global interpretability," 2018, *arXiv:1805.11545*.

[79] A. Rosenfeld and A. Richardson, "Explainability in human–agent systems," *Auto. Agents Multi-Agent Syst.*, vol. 33, no. 6, pp. 673–705, Nov. 2019.

[80] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, "One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques," 2019, *arXiv:1909.03012*.

[81] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[82] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2018, pp. 80–89.

[83] D. A. Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–10.

[84] R. Luss, P.-Y. Chen, A. Dhurandhar, P. Sattigeri, K. Shanmugam, and C.-C. Tu, "Generating contrastive explanations with monotonic attribute functions," Tech. Rep., 2019.

[85] I. Butun, S. D. Morgera, and R. Sankar, "A survey of intrusion detection systems in wireless sensor networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 266–282, 1st Quart., 2014.

[86] A. Sharma and S. K. Sahay, "Evolution and detection of polymorphic and metamorphic malwares: A survey," 2014, *arXiv:1406.7061*.

[87] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009.

[88] O. Loyola-Gonzalez, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view," *IEEE Access*, vol. 7, pp. 154096–154113, 2019.

[89] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[90] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *J. Roy. Stat. Soc., B, Stat. Methodol.*, vol. 82, no. 4, pp. 1059–1086, Sep. 2020.

[91] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *J. Comput. Graph. Statist.*, vol. 24, no. 1, pp. 44–65, 2015.

[92] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *Ann. Appl. Statist.*, vol. 2, no. 3, pp. 916–954, Sep. 2008.

[93] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," 2018, *arXiv:1806.07421*.

[94] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," 2016, *arXiv:1605.01713*.

[95] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.

[96] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[97] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, May 2017.

[98] S. R. Islam, W. Eberle, S. K. Ghafoor, A. Siraj, and M. Rogers, "Domain knowledge aided explainable artificial intelligence for intrusion detection and response," 2019, *arXiv:1911.09853*.

[99] M. Sarhan, S. Layeghy, and M. Portmann, "An explainable machine learning-based network intrusion detection system for enabling generalisability in securing IoT networks," 2021, *arXiv:2104.07183*.

[100] L. Antwarg, R. M. Miller, B. Shapira, and L. Rokach, "Explaining anomalies detected by autoencoders using SHAP," 2019, *arXiv:1903.02407*.

[101] I. A. Khan, N. Moustafa, D. Pi, K. M. Sallam, A. Y. Zomaya, and B. Li, "A new explainable deep learning framework for cyber threat discovery in industrial IoT networks," *IEEE Internet Things J.*, vol. 9, no. 13, pp. 11604–11613, Jul. 2022.

[102] C. Wu, A. Qian, X. Dong, and Y. Zhang, "Feature-oriented design of visual analytics system for interpretable deep learning based intrusion detection," in *Proc. Int. Symp. Theor. Aspects Softw. Eng. (TASE)*, Dec. 2020, pp. 73–80.

[103] N. Burkart, M. Franz, and M. F. Huber, "Explanation framework for intrusion detection," in *Machine Learning for Cyber Physical Systems*. Berlin, Germany: Springer, 2021, pp. 83–91.

[104] K. Amarasinghe, K. Kenney, and M. Manic, "Toward explainable deep neural network based anomaly detection," in *Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI)*, Jul. 2018, pp. 311–317.

[105] J. Kauffmann, K.-R. Müller, and G. Montavon, "Towards explaining anomalies: A deep Taylor decomposition of one-class models," 2018, *arXiv:1805.06230*.

[106] M. Szczepanski, M. Choras, M. Pawlicki, and R. Kozik, "Achieving explainability of intrusion detection system by hybrid oracle-explainer approach," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[107] G. Pang, C. Ding, C. Shen, and A. van den Hengel, "Explainable deep few-shot anomaly detection with deviation networks," 2021, *arXiv:2108.00462*.

[108] Q.-V. Dang, "Understanding the decision of machine learning based intrusion detection systems," in *Proc. Int. Conf. Future Data Secur. Eng.* Springer, 2020, pp. 379–396.

[109] L. Arras, G. Montavon, K.-R. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," 2017, *arXiv:1706.07206*.

[110] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and A. K.-R. Müller, "Layer-wise relevance propagation: An overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. 2019, pp. 193–209.

[111] F. P. Caforio, G. Andresini, G. Vessio, A. Appice, and A. D. Malerba, "Leveraging grad-CAM to improve the accuracy of network intrusion detection systems," in *Discovery Science*. 2021.

[112] J. Kauffmann, L. Ruff, G. Montavon, and K.-R. Müller, "The clever Hans effect in anomaly detection," 2020, *arXiv:2006.10609*.

[113] L. K. Hansen and L. Rieger, "Interpretability in intelligent systems—A new concept?" in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 41–49.

[114] E. Pintelas, I. E. Livieris, and P. Pintelas, "A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability," *Algorithms*, vol. 13, no. 1, p. 17, Jan. 2020.

[115] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable artificial intelligence approaches: A survey," 2021, *arXiv:2101.09429*.

[116] A. P. Kuruvila, X. Meng, S. Kundu, G. Pandey, and K. Basu, "Explainable machine learning for intrusion detection via hardware performance counters," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, early access, Feb. 7, 2022, doi: 10.1109/TCAD.2022.3149745.

[117] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model," *Complexity*, vol. 2021, pp. 1–11, Jan. 2021.

[118] T. Dias, N. Oliveira, N. Sousa, I. Praça, and O. Sousa, "A hybrid approach for an interpretable and explainable intrusion detection system," 2021, *arXiv:2111.10280*.

[119] C. F. T. Pontes, M. M. C. de Souza, J. J. C. Gondim, M. Bishop, and M. A. Marotta, "A new method for flow-based network intrusion detection using the inverse Potts model," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 2, pp. 1125–1136, Jun. 2021.

[120] C. S. Wickramasinghe, K. Amarasinghe, D. L. Marino, C. Rieger, and M. Manic, "Explainable unsupervised machine learning for cyber-physical systems," *IEEE Access*, vol. 9, pp. 131824–131843, 2021.
[121] C. Langin, M. Wainer, and S. Rahimi, "ANNaBell Island: A 3D color hexagonal SOM for visual intrusion detection," *Int. J. Comput. Sci. Inf. Secur.*, vol. 9, no. 1, pp. 1–7, 2011.
[122] B. Subba, S. Biswas, and S. Karmakar, "Intrusion detection systems using linear discriminant analysis and logistic regression," in *Proc. Annu. IEEE India Conf. (INDICON)*, Dec. 2015, pp. 1–6.
[123] Z. Wang, J. Yang, and F. Li, "A new anomaly detection method based on IGTE and IGFE," in *Proc. Int. Conf. Secur. Privacy Commun. Netw.* Springer, 2014, pp. 93–109.
[124] Z. Wang, J. Yang, Z. ShiZe, and C. Li, "Robust regression for anomaly detection," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.
[125] O. Loyola-Gonzalez, A. E. Gutierrez-Rodriguez, M. A. Medina-Perez, R. Monroy, J. F. Martinez-Trinidad, J. A. Carrasco-Ochoa, and M. Garcia-Borroto, "An explainable artificial intelligence model for clustering numerical databases," *IEEE Access*, vol. 8, pp. 52370–52384, 2020.
[126] N. Frost, M. Moshkovitz, and C. Rashtchian, "ExKMC: Expanding explainable K-means clustering," 2020, *arXiv:2006.02399*.
[127] S. Dasgupta, N. Frost, M. Moshkovitz, and C. Rashtchian, "Explainable K-means and K-medians clustering," in *Proc. 37th Int. Conf. Mach. Learn.*, Vienna, Austria, 2020, pp. 12–18.
[128] C. Sinclair, L. Pierce, and S. Matzner, "An application of machine learning to network intrusion detection," in *Proc. 15th Annu. Comput. Secur. Appl. Conf. (ACSAC)*, 1999, pp. 371–377.
[129] A. Ojugo, A. Eboka, O. Okonta, R. Yoro, and F. Aghware, "Genetic algorithm rule-based intrusion detection system (GAIDS)," *J. Emerg. Trends Comput. Inf. Sci.*, vol. 3, no. 8, pp. 1182–1194, 2012.
[130] K. Chadha and S. Jain, "Hybrid genetic fuzzy rule based inference engine to detect intrusion in networks," in *Intelligent Distributed Computing*. Springer, 2015, pp. 185–198.
[131] M. Roesch, "Snort: Lightweight intrusion detection for networks," in *Proc. LISA*, vol. 99, 1999, pp. 229–238.
[132] B. Caswell, J. Beale, and A. Baker, "Snort intrusion detection prevention toolkit," Syngress, Tech. Rep., 2007.
[133] A. Qayyum, M. H. Islam, and M. Jamil, "Taxonomy of statistical based anomaly detection techniques for intrusion detection," in *Proc. IEEE Symp. Emerg. Technol.*, Sep. 2005, pp. 270–276.
[134] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, pp. 1–22, Dec. 2019.
[135] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: Methods, systems and tools," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 303–336, 1st Quart., 2013.
[136] A. T. Tran, "Network anomaly detection," in *Future Internet (FI) and Innovative Internet Technologies and Mobile Communication (IITM) Focal Topic: Advanced Persistent Threats*. 2017.
[137] A. B. Ashfaq, M. Javed, S. A. Khayam, and H. Radha, "An information-theoretic combining method for multi-classifier anomaly detection systems," in *Proc. IEEE Int. Conf. Commun.*, May 2010, pp. 1–5.
[138] W. Sha, Y. Zhu, T. Huang, M. Qiu, Y. Zhu, and Q. Zhang, "A multi-order Markov chain based scheme for anomaly detection," in *Proc. IEEE 37th Annu. Comput. Softw. Appl. Conf. Workshops*, Jul. 2013, pp. 83–88.
[139] N. Ye, "A Markov chain model of temporal behavior for anomaly detection," in *Proc. IEEE Syst., Man, Cybern. Inf. Assurance Secur. Workshop*, vol. 166, Jun. 2000, p. 169.
[140] T. Kohonen, E. Oja, O. Simula, A. Visa, and J. Kangas, "Engineering applications of the self-organizing map," *Proc. IEEE*, vol. 84, no. 10, pp. 1358–1384, Oct. 1996.
[141] A. J. Hoglund, K. Hatonen, and A. S. Sorvari, "A computer host-based user anomaly detection system using the self-organizing map," in *Proc. IEEE-INNS-ENNS Int. Joint Conf. Neural Netw. Neural Comput., New Challenges Perspect. New Millennium*, 2000, pp. 411–416.
[142] P. Lichodzijewski, A. N. Zincir-Heywood, and M. I. Heywood, "Host-based intrusion detection using self-organizing maps," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2002, pp. 1714–1719.
[143] H. G. Kayacik, A. N. Zincir-Heywood, and M. I. Heywood, "A hierarchical SOM-based intrusion detection system," *Eng. Appl. Artif. Intell.*, vol. 20, no. 4, pp. 439–451, 2007.
[144] C. Langin, H. Zhou, and S. Rahimi, "A model to use denied internet traffic to indirectly discover internal network security problems," in *Proc. IEEE Int. Perform., Comput. Commun. Conf.*, Dec. 2008, pp. 486–490.
[145] C. Langin, H. Zhou, S. Rahimi, B. Gupta, M. Zargham, and M. R. Sayeh, "A self-organizing map and its modeling for discovering malignant network traffic," in *Proc. IEEE Symp. Comput. Intell. Cyber Secur.*, Mar. 2009, pp. 122–129.
[146] J. Kim, N. Shin, S. Y. Jo, and S. H. Kim, "Method of intrusion detection using deep neural network," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2017, pp. 313–316.
[147] M. Sölch, "Detecting anomalies in robot time series data using stochastic recurrent networks," Tech. Rep., 2015.
[148] C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," *IEEE Access*, vol. 5, pp. 21954–21961, 2017.
[149] M. Azizjon, A. Jumabek, and W. Kim, "1D CNN based network intrusion detection with normalization on imbalanced data," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Feb. 2020, pp. 218–224.
[150] R. Vinayakumar, K. P. Soman, and P. Poornachandran, "Applying convolutional neural network for network intrusion detection," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2017, pp. 1222–1228.
[151] A. Holzinger, "From machine learning to explainable AI," in *Proc. World Symp. Digit. Intell. Syst. Mach. (DISA)*, Aug. 2018, pp. 55–66.
[152] H. Liu, C. Zhong, A. Alnusair, and S. R. Islam, "FAIXID: A framework for enhancing AI explainability of intrusion detection results using data cleaning techniques," *J. Netw. Syst. Manage.*, vol. 29, no. 4, pp. 1–30, Oct. 2021.
[153] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE Symp. Comput. Intell. Secur. Defense Appl.*, Jul. 2009, pp. 1–6.
[154] (1999). *KDD Cup 1999 Data the UCI KDD Archive*. Accessed: Apr. 9, 2022. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[155] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. 4th Int. Conf. Inf. Syst. Secur. Privacy*, vol. 1, Jan. 2018, pp. 108–116.
[156] G. Maciá-Fernández, J. Camacho, R. Magán-Carrión, P. García-Teodoro, and R. Therón, "UGR-16: A new dataset for the evaluation of cyclostationarity-based network IDSs," *Comput. Secur.*, vol. 73, pp. 411–424, Mar. 2018.
[157] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *Proc. Sci. Inf. Conf.*, Aug. 2014, pp. 372–378.
[158] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
[159] F. R. S. K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Philosoph. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.
[160] *Shapash Welcome to Shapash's Documentation*. Accessed: Jul. 23, 2022. [Online]. Available: https://shapash.readthedocs.io/en/latest/overview.html
[161] O. Dijk. (2019). *ExplainerDashboard Starting the Default Dashboard*. Accessed: Jul. 22, 2022. [Online]. Available: https://explainerdashboard.readthedocs.io/en/latest/dashboards.html
[162] H. S. Anderson, J. Woodbridge, and B. Filar, "DeepDGA: Adversarially-tuned domain generation and detection," in *Proc. ACM Workshop Artif. Intell. Secur.*, 2016, pp. 13–21.
[163] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," 2017, *arXiv:1712.05526*.
[164] N. Rastogi, S. Rampazzi, M. Clifford, M. Heller, M. Bishop, and K. Levitt, "Explaining RADAR features for detecting spoofing attacks in connected autonomous vehicles," 2022, *arXiv:2203.00150*.
[165] D. Han, Z. Wang, Y. Zhong, W. Chen, J. Yang, S. Lu, X. Shi, and X. Yin, "Evaluating and improving adversarial robustness of machine learning-based network intrusion detectors," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2632–2647, Aug. 2021.
[166] M. Pawlicki, M. Choraś, and R. Kozik, "Defending network intrusion detection systems against adversarial evasion attacks," *Future Gener. Comput. Syst.*, vol. 110, pp. 148–154, Sep. 2020.
[167] A. Hartl, M. Bachl, J. Fabini, and T. Zseby, "Explainability and adversarial robustness for RNNs," in *Proc. IEEE 6th Int. Conf. Big Data Comput. Service Appl. (BigDataService)*, Aug. 2020, pp. 148–156.

SUBASH NEUPANE received the bachelor's degree in computer engineering from Kathmandu University, Nepal, in 2011, the M.S. degree in information technology (professional computing) from the Swinburne University of Technology, Melbourne, Australia, in 2016, and the M.S. degree in information systems and security management from Tuskegee University, AL, USA, in 2020. He is currently pursuing the Ph.D. degree in systems and security with Mississippi State University. He was a Telecom Network Engineer with Tandem Corporation, Melbourne, before moving to the USA. His current research interests include systems and security, machine learning, and blockchain.

JESSE ABLES (Graduate Student Member, IEEE) received the B.S. degree in software engineering, in 2016, and the M.S. degree in cyber security, in 2019. He is currently pursuing the Ph.D. degree in autonomous cyber security with Mississippi State University (MSU). He worked as a Teacher in computer science field at MSU and as an English Teacher in South Korea. He is currently working as a Research Assistant with the Computer Science and Engineering Department, MSU. His research interests include autonomous security, the Internet of Things, and anomaly detection.

WILLIAM ANDERSON received the B.S. degree in computer science from Mississippi State University, in 2019, where he is currently pursuing the M.S. degree in artificial intelligence. In 2020, he worked as a Research Assistant at NSPARC, developing natural language processing applications for workforce development. He is also working as a Research Assistant with the Computer Science and Engineering Department, Mississippi State University. His research interests include natural language processing, anomaly detection in the fields of cybersecurity and healthcare, and explainable AI.

SUDIP MITTAL (Member, IEEE) received the Ph.D. degree in computer science from the University of Maryland Baltimore County, in 2019. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Mississippi State University. His primary research interests include cybersecurity and artificial intelligence. His goal is to develop the next generation of cyber defense systems that help protect various organizations and people. At Mississippi State University, he leads the Secure and Trustworthy Cyberspace (SECRETS) Laboratory and has published over 60 journals and conference papers in leading cybersecurity and AI venues. He has received funding from the NSF, USAF, USACE, and various other the Department of Defense Programs. He also serves as a program committee member or the program chair of leading AI and cybersecurity conferences and workshops. His work has been cited in the LA Times, Business Insider, WIRED, and Cyberwire. He is a member of the ACM.

SHAHRAM RAHIMI (Member, IEEE) is currently a Professor and the Head of the Department of Computer Science and Engineering, Mississippi State University. Prior to that, he led the Department of Computer Science, Southern Illinois University, for five years. He is also a Recognized Leader in the area of artificial and computational intelligence with over 220 peer-reviewed publications and a few patents or pending patents in this area. He has served as the Editor-in-Chief for two leading *Computational Intelligence* journals and sits on the editorial board of several others. He is a member of IEEE New Standards Committee in Computational Intelligence and provides advice to staff and administration at federal government on predictive analytics for foreign policy. He was a recipient of 2016 Illinois Rising Star Award from ISBA, selected among 100's of other highly qualified candidates. His intelligent algorithm for patient flow optimization and hospital staffing is currently used in over 1000 emergency departments across the nation and was named top ten AI technology for healthcare, in 2018, by *HealthTech Magazine*. He has secured over $20M of federal and industry funding as a PI or a co-PI in the last 20 years. He has also organized 15 conferences and workshops in the areas of computational intelligence and multi-agent systems over the past two decades.

IOANA BANICESCU (Life Senior Member, IEEE) received the Diploma degree in engineering (electronics and telecommunications) from the Polytechnic University of Bucharest and the M.S. and Ph.D. degrees in computer science from New York University—Polytechnic Institute. Between 2009 and 2017, she was the Director of the Center for Cloud and Autonomic Computing, Mississippi State University (MSU), and also the Co-Director of the National Science Foundation Center for Cloud and Autonomic Computing. She is currently a Professor with the Department of Computer Science and Engineering, MSU. Her research interests include parallel algorithms, scientific computing, scheduling theory, load balancing algorithms, performance modeling, analysis and prediction, autonomic computing, performance optimization for problems in computational science, and graph analytics. She has given many invited talks at universities, government laboratories, and at various national and international forums in the USA and overseas. She was a recipient of a number of awards for research and scholarship from the National Science Foundation (NSF). She served and continues to serve on numerous research review panels for advanced research grants in the USA and Europe, on steering and program committees of a number of international ACM and IEEE conferences, symposia, and workshops, and on the Executive Board and the Advisory Board of the IEEE Technical Committee on Parallel Processing (TCPP). She was an Associate Editor of the *Cluster Computing* journal and the *International Journal on Computational Science and Engineering*. Over the years, she was recognized with many distinctions for her scholarly contributions.

MARIA SEALE received the B.S. degree in computer science from the University of Southern Mississippi and the M.S. and Ph.D. degrees in computer science from Tulane University. She is currently a Computer Scientist at the Information Technology Laboratory, U.S. Army Engineer Research and Development Center (ERDC). She has over 20 years of experience in research, development, and teaching in computer science. She has held positions at the Institute for Naval Oceanography, the U.S. Naval Research Laboratory, and various private companies, and a tenured associate professorship at the University of Southern Mississippi. Her experience has included work with ocean modeling, underwater seismic data collection and processing, geographical information systems design, natural language processing, and machine learning. At ERDC, she has been involved with research in making scalable machine learning algorithms available on high performance computing platforms and expanding the laboratory's capabilities to manage and analyze very large data sets.

Selected Publication 4

EDITOR: Amit Sheth, amit@sc.edu

DEPARTMENT: KNOWLEDGE GRAPH

Knowledge-Enhanced Neurosymbolic Artificial Intelligence for Cybersecurity and Privacy

Aritran Piplai, *University of Texas at El Paso, El Paso, TX, 79968, USA*

Anantaa Kotal, Seyedreza Mohseni, and Manas Gaur, *Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Catonsville, MD, 21250, USA*

Sudip Mittal, *Department of Computer Science, Mississippi State University, Starkville, MS, 39762, USA*

Anupam Joshi, *Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Catonsville, MD, 21250, USA*

Neurosymbolic artificial intelligence (AI) is an emerging and quickly advancing field that combines the subsymbolic strengths of (deep) neural networks and the explicit, symbolic knowledge contained in knowledge graphs (KGs) to enhance explainability and safety in AI systems. This approach addresses a key criticism of current generation systems, namely, their inability to generate human-understandable explanations for their outcomes and ensure safe behaviors, especially in scenarios with unknown unknowns (e.g., cybersecurity, privacy). The integration of neural networks, which excel at exploring complex data spaces, and symbolic KGs, which represent domain knowledge, allows AI systems to reason, learn, and generalize in a manner understandable to experts. This article describes how applications in cybersecurity and privacy, two of the most demanding domains in terms of the need for AI to be explainable while being highly accurate in complex environments, can benefit from neurosymbolic AI.

Neurosymbolic artificial intelligence (*AI*) refers to the integration of neural network-based methods with symbolic knowledge-based approaches. It combines the strengths of both approaches to leverage the representational power of neural networks and the logical reasoning capabilities of symbolic approaches. Neural networks are excellent at large-scale data processing and the extraction of intricate patterns and characteristics from unprocessed input. However, they often struggle to provide explicit explanations for their decisions.[1] This is one of the reasons why, after a promising decade of work from the mid-1980s, neural networks never made it beyond academic and industrial research labs. On the other hand, symbolic knowledge-based approaches, such as rule-based systems or expert systems, utilize explicit knowledge representations and logical reasoning mechanisms. They can capture domain-specific knowledge and provide transparent explanations for their conclusions.[1,12] However, these approaches may struggle with handling uncertain or incomplete information and have limited capacity to learn from large-scale data.[1]

A combination of these two paradigms, called *neurosymbolic AI*, has begun to see popularity among the AI community over the last five years. The idea for this combination is not new, with the term *neural symbolic* being used at least as far back as the early 2000s.[2] In the 1990s, for instance, there were several efforts to marry connectionist approaches with fuzzy rules.[11] Indeed, it can (and has) been argued that the kernel of this idea can be found in the McColloch and Pitts' paper "A Logical Calculus of the Ideas Immanent in

1089-7801 © 2023 IEEE
Digital Object Identifier 10.1109/MIC.2023.3299435
Date of current version 14 September 2023.

Nervous Activity." There are multiple reasons for this renewed popularity. We describe these reasons in the context of cybersecurity.

First, a combination of symbolic reasoning and data-driven methods can be used to extract the sequence of steps or events that triggered the conclusion that the model reached. This is a great motivation for neurosymbolic approaches to be used in cybersecurity and privacy, especially in solving problems such as threat detection and analysis, which require not just patterns to be detected, but for such detected patterns from disparate systems across time to be put into a common context.[5] Neurosymbolic approaches can do this while preserving privacy (e.g., incorporating privacy policies, regulations, and compliance). For example, a neurosymbolic model can reason about sensitive network flow data usage by the neural network detector based on explicit privacy policies, and ensure compliance by incorporating privacy-preserving techniques such as differential privacy or secure multiparty computation.

Second is keeping AI systems *secure and safe.* The rise in data-driven models for automating vulnerability assessment of systems limits safety as the system learns only the vulnerabilities on which it is trained. In a neurosymbolic approach, the AI-based software systems would be trained with experts acting as adversaries and the AI model would learn to infer policies/rules dynamically.[1] Further, with knowledge in security specification documents explicitly captured using symbolic approaches and used as constraints on behavior, the AI system would be more robust and safe. This is of immediate interest to regulators and legislators in many countries as advanced AI systems have a high probability for generating risky/harmful information, without the control of human knowledge/expertise.

Another reason why a combination of rules and data-driven methods may be useful is the lack of high-quality data to make reliable inferences. This problem can be found in domains where the data for experiments are either hard to obtain or difficult to share because they may be sensitive. However, alternate sources such as text descriptors of the sensitive data may be available. Rules can be derived from these alternate sources that are shareable. If the available data alone are insufficient to infer reliable conclusions, these rules can be used to augment the conclusions derived from the data.[12] They can also be provided as an input to the data-driven model during the learning process.

Some domains may also be very dynamic, and so the data may be representative for only a brief period of time. The conclusions derived from the data may also be valid for a brief period of time. This is a problem in domains such as cybersecurity and fraud detection. The patterns that we derive from our existing dataset may be useful for some cyberattacks that are happening right now, but they may not be useful in the future. Combining deep network-based detectors with explicit rules that capture data drift or the temporal limits on the usefulness of a model can help in such situations.

NEUROSYMBOLIC AI IN CYBERSECURITY

In the context of cybersecurity, neurosymbolic AI can be applied to enhance various aspects of security systems, such as intrusion detection, malware analysis, vulnerability assessment, and threat intelligence. Very broadly, it can assist in creating the next-generation Security Operations Center (SoC), which combines AI approaches with a human, either in or on the loop.

Let us consider a scenario of security analysts who work in an SoC and play a major role in ensuring the security of an organization. The amount of background knowledge they have about evolving and new attacks makes a significant difference in their ability to detect attacks from the output of deep neural networks or machine learning (ML)-based systems that today analyze the sensed data stream. We can assist an analyst by capturing information available in open source threat intelligence sources, such as text descriptions of cyberattacks or threat feeds, and store them in a structured fashion in a cybersecurity knowledge graph (CKG). We describe two methods in which the structured cyber information present in CKGs can be used for downstream tasks, with a focus on explainability (reasoning and inference). The first method involves creating sophisticated rules based on real data, and an existing knowledge engine (*rule-based framework*). The second method involves using existing rules in downstream data-driven AI models and creating new policies for cybersecurity (*knowledge-guided models*).

In the *rule-based framework*, the ultimate goal is to create the strongest and closest rules for target machines to protect them from any type of threats and adversary behaviors. The rules can be simple to complex and will be consumed by any system or subsystem that needs protection. In the *knowledge-guided models*, we aim to tackle novel cyberthreats or mutated versions of older cyberthreats that do not exist in existing datasets for data-driven experimentation. Exploratory modeling techniques, such as reinforcement learning (RL), are needed to discover new adversaries that can further lead to new defenses. In our experiments, we see that CKGs can guide these exploratory learning strategies to be faster, more effective, and explainable.

Modeling Cyber Events

A plethora of information is available in unstructured text for cybersecurity. This information can come from various sources such as social media posts, user-written blogs, or published reports from large organizations. Through our research, we were able to extract this unstructured information and convert it into structured knowledge.[3] To achieve this, we utilize semantic triples, which consist of a subject-predicate-object relationship. In other words, when encountering two entities in a text, we aimed to deduce the relationship between them. We use Bidirectional Encoder Representations from Transformers (BERT) embeddings, as well as neural models, to generate these semantic triples. By analyzing 474 technical reports and numerous smaller technical posts, we construct a KG using these semantic triples.

In addition, software companies release information concerning cybersecurity threats to help consumers identify vulnerabilities in their products. These data can be accessed through *Trusted Automated eXchange of Intelligence Information* servers and integrated into a centralized KG.

We model our CKG ontology based on the concepts used in the *Structured Threat-Intelligence Exchange*, an industry standard for exchanging cyberthreat information. We further enrich this ontology with system-attribute concepts that describe the malware's behavior. The *rule-based framework* and the *knowledge-guided models* for cybersecurity can use this ontology for generating new rules and policies and employ it in downstream models for explainable intrusion and malware detection.[7]

Reasoning and Inference Examples

The semantic triples extracted from open source text are asserted with a CKG. This CKG can be leveraged, not just by human security analysts but also by other data-driven models. We can see some examples of how security analysts can use this CKG to uncover important insights about malware using the following SPARQL queries:

```
SELECT ? x where {
  ? x a FusedCKG:Malware;
    FusedCKG:uses
      FusedCKG:588f41bbc[ ... ].}
```

This query asks what malware(s) match a particular hash value.

```
SELECT DISTINCT ?x ?y ?z WHERE {
  ?x a FusedKG:Malware.
    FusedKG:hasHash [b9d ..... ]
    FusedKG:uses ?y.
  ?y a FusedKG:Attack-Pattern.
  ?z FusedKG:parameterchange
    FusedKG:increases_meanchange.}
```

The query asks which malware has a particular hash and the associated attack pattern for that malware. It also asks which system parameters show an increase when the malware is active. If a framework employs a KG where such dynamic information, in the form of observations, are recorded, then an AI model can leverage such structured knowledge in creating rules when defending against attacks.

Rule-Based Framework

The rule-based framework is one such architecture that employs existing AI models and transforms them into rule generators. AI models focus on the knowledge of events, their interlinking with other events or malware in the CKG, and a method for generating hypotheses or rules for inference (see Figure 1). The three broad sections of this framework are as follows:

1) *Event parsing engine*: The network block grants access to uncontrolled and unregulated networks. The packets are classified and distinguished to organize and handle data according to their specific types. Administrators, typically network admins, establish administrative policies that contribute to the control and dependability of network traffic. These policies assist in the creation of effective rules for the system.
2) *Symbolic engine*: The KG constructor serves as a graph builder, while the CKG stores the contextual knowledge used to generate hypotheses. The CKG constructor ensures the provision of valid and accurate context knowledge. The observation constructor produces clear and consistent observations derived from network packets and admin policies for integration into the KG. Also, the knowledge test and verification process will be applied to test and verify the alignment (e.g., simply using similarity measures) of both knowledge datasets and observations prior to their incorporation into the CKG.
3) *Reasoning engine*: In a neurosymbolic framework, the knowledge extractor component retrieves hypotheses from a KG and combines them with observations to form entailments. These entailments, organized with start and separator tokens, are then processed by a reasoning engine composed of transformer-based AI models [e.g., Generative Pre-Trained Transformer (GPT)]. This engine generates rankings and selects the most suitable hypothesis.[4] The neurosymbolic rule-based framework collects and categorizes network packets for use in symbolic engines, generating hypotheses

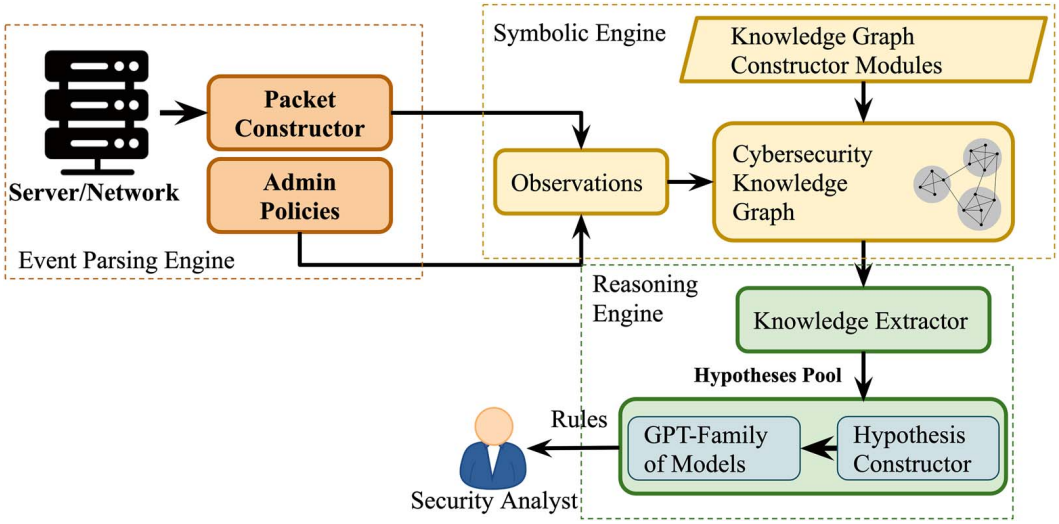


FIGURE 1. Neurosymbolic rule-based framework for dynamic rule inference and generation in cybersecurity. admin: administrative; GPT: Generative Pre-Trained Transformer.

through observation sequences and correlation scores. These hypotheses are combined with the GPT-Family models to create cybersecurity rules, providing improved accuracy, dependability, and explainable behavior in AI malware detection, surpassing traditional rule-based systems like SNORT.

Rules For Knowledge-Guided RL in Cybersecurity

The semantic knowledge embodied in KGs has the potential to direct algorithms like RL through rewards. Although KGs encompass well-established facts regarding cyber incidents, they might not possess details about emerging threats in unpredictable environments. Because of the exploratory nature of RL algorithms, new rules can be inferred based on the static knowledge in KGs, systems' data, and rules extracted from the rule-based framework.[5] The reward-based learning in RL allows the intrinsic deep learning AI algorithm to be expressive over the rules and explainable in prediction.

In the scenario of malware detection, we show that knowledge-guided RL sees even faster convergence, better efficiency, and better response time.[5] The acquired knowledge is incorporated into the exploration phase of the RL algorithm, and higher rewards are assigned to states that align with our knowledge sources. Specifically when we integrate prior knowledge, we observe an 8% reduction in average episode time.

Additionally, we conduct experiments with offline RL algorithms to examine the influence of prior knowledge. We discover that incorporating prior knowledge leads to a 4% rise in detection in three out of four malware families.[6]

In another study by Piplai et al.,[7] we show that employing knowledge-guided RL with rules on extensive Packet Capture files, which ranged from 3 to 4 GB in size, leads to more accurate decision making in countering attackers. In a two-player zero-sum game setup, both the attacker and defender were simulated using a knowledge-guided RL algorithm. By incorporating informed actions based on this approach, we demonstrate a remarkable preservation of 78% network availability, in comparison to a mere 25% when knowledge guidance was not utilized.[7] In both of the studies, the knowledge-guided RL strategy would preserve trace to knowledge sources for providing explanations to analysts.

CYBERSECURITY AND PRIVACY WITH KNOWLEDGE-GUIDED AI AND A RULE-BASED FRAMEWORK

Ensuring privacy in AI models is vital, similar to the importance of cybersecurity. AI models that unintentionally reveal personally identifiable information (PII) during defense against attacks can leave systems vulnerable to future unpredictable attacks. The challenge lies in training AI models to handle various cyberattacks while optimizing response time and preserving privacy. Reliable datasets for privacy-preserving training are scarce as organizations are hesitant to share data that contain PII or sensitive insights.

To overcome this, generative modeling methods like generative adversarial networks (GANs) can produce surrogate datasets that protect privacy while remaining useful for learning tasks. We utilized conditional GANs (CGANs) along with the t-closeness principle to preserve privacy in tabular data containing continuous and discrete variables.[8] However, training standard CGANs on sensitive data has limitations as they struggle to model conditionally continuous variables and can only repeat the discrete variable values seen in the original dataset.

To address these limitations, we propose a dual approach using privacy-preserving deep learning models, combining generative modeling and symbolic KGs that express domain knowledge. Domain-specific KGs like the Unified Cyber Ontology can guide the generative model by providing standardized information.[9] By querying the KG, the CGAN can be trained using a mix of original dataset and discrete values, ensuring that the generated dataset contains observed values and other alternatives from the KG. This approach enhances privacy preservation in generated datasets for downstream ML tasks.

CONCLUSION AND FUTURE WORK

This article discussed two main approaches of neurosymbolic AI in cybersecurity and privacy: 1) the rule-based framework and 2) knowledge-guided AI using RL. Both approaches intrinsically involve partitioning the CKG for solving tasks in our concerned domains and ensuring that the AI system is explainable. Our perspective is based on the noticeable improvements over traditional AI systems. However, additional research endeavors are required to develop reasoning engines in cybersecurity and privacy for optimal and explainable decision making that focuses on the users, such as security analysts. The combination of rules and ML/RL models can further be improved by focusing on which information is more beneficial to the model. The application of transformers, in this case by selecting specific paths in the graph based on real data, is a promising idea. We can also improve this with the help of graph embeddings of state spaces appended to data representation during training time.

Neurosymbolic AI has the potential to address critical challenges in the future, especially in domains like privacy in health care, where there is a growing demand for explainability. The techniques we explored can be expanded to biomedicine, where medical treatments and procedures are confidential. When limited to specific domains, GANs can only replicate treatment procedures based on the information upon which they were trained, however, incorporating a biomedical KG can enhance the GAN model by providing a broader range of knowledge.[10] By delving into such methods of knowledge infusion, not only can we contribute to the cause we can effectively tackle the limitations associated with purely data-driven approaches.

REFERENCES

1. A. Sheth, K. Roy, and M. Gaur, "Neurosymbolic artificial intelligence (Why, What, and How)," *IEEE Intell. Syst.*, vol. 38, no. 3, pp. 56–62, May/Jun. 2023, doi: 10.1109/MIS.2023.3268724.
2. S. Bader and P. Hitzler, "Dimensions of neural-symbolic integration - A structured survey," 2005, *arXiv:cs/0511042v1*.
3. A. Piplai, S. Mittal, A. Joshi, T. Finin, J. Holt, and R. Zak, "Creating cybersecurity knowledge graphs from malware after action reports," *IEEE Access*, vol. 8, pp. 211,691–211,703, 2020, doi: 10.1109/ACCESS.2020.3039234.
4. D. Paul and A. Frank, "Social commonsense reasoning with multi-head knowledge attention," 2020, *arXiv:2010.05587v1*.
5. A. Piplai, P. Ranade, A. Kotal, S. Mittal, S. N. Narayanan, and A. Joshi, "Using knowledge graphs and reinforcement learning for malware analysis," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, 2020, pp. 2626–2633, doi: 10.1109/BigData50022.2020.9378491.
6. A. Piplai, A. Joshi, and T. Finin, "Offline RL+ CKG: A hybrid AI model for cybersecurity tasks," in *Proc. AAAI Make*, 2023.
7. A. Piplai, M. Anoruo, K. Fasaye, A. Joshi, T. Finin, and A. Ridley, "Knowledge guided two-player reinforcement learning for cyber attacks and defenses," in *Proc. Int. Conf. Mach. Learn. Appl.*, 2022, pp. 1342–1349, doi: 10.1109/ICMLA55696.2022.00213.
8. A. Kotal, A. Piplai, S. S. Laya Chukkapalli, and A. Joshi, "PriveTAB: Secure and privacy-preserving sharing of tabular data," in *Proc. ACM Int. Workshop Secur. Privacy Analytics*, 2022, pp. 35–45, doi: 10.1145/3510548.3519377.
9. Z. Syed, A. Padia, M. L. Mathews, T. Finin, and A. Joshi, "UCO: A unified cybersecurity ontology," in *Proc. AAAI Workshop Artif. Intell. Cyber Secur.*, 2016.
10. L. Elluri, A. Piplai, A. Kotal, A. Joshi, and K. P. Joshi, "A policy-driven approach to secure extraction of COVID-19 data from research papers," *Frontiers Big Data*, vol. 4, Aug. 2021, Art. no. 701966, doi: 10.3389/fdata.2021.701966.
11. A. Joshi, N. Ramakrishman, E. N. Houstis, and J. R. Rice, "On neurobiological, neuro-fuzzy, machine learning, and statistical pattern recognition techniques," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 18–31, Jan. 1997, doi: 10.1109/72.554188.

12. S. Kambhampati, "Polanyi's revenge and AI's new romance with tacit knowledge," *Commun. ACM*, vol. 64, no. 2, pp. 31–32, Feb. 2021, doi: 10.1145/3446369.

ARITRAN PIPLAI is an assistant professor at the University of Texas at El Paso, El Paso, TX, 79968, USA. Contact him at apiplai1@umbc.edu.

ANANTAA KOTAL is a Ph.D. candidate in the Department of Computer Science and Electrical Engineering, the University of Maryland, Baltimore County, MD, 21250, USA. Contact her at akkotal1@umbc.edu.

SEYEDREZA MOHSENI is a Ph.D. candidate in the Department of Computer Science and Electrical Engineering, the University of Maryland, Baltimore County, MD, 21250, USA. Contact him at mohseni1@umbc.edu.

MANAS GAUR is an assistant professor with the Department of Computer Science and Electrical Engineering, the University of Maryland, Baltimore County, MD, USA. Contact him at manas@umbc.edu.

SUDIP MITTAL is an assistant professor with the Department of Computer Science at Mississippi State University, Starkville, MS, 39762, USA. Contact him at mittal@cse.msstate.edu.

ANUPAM JOSHI is the Oros Family Professor with the Department of Computer Science and Electrical Engineering, the University of Maryland, Baltimore County (UMBC), MD, 21250, USA, and acting dean of the College of Engineering and Information Technology at UMBC. He is also the director of UMBC's Center for Cybersecurity. Contact him at joshi@umbc.edu.

Selected Publication 5

Risks to Zero Trust in a Federated Mission Partner Environment

Keith Strandell
Dr. Sudip Mittal

ABSTRACT

Recent cybersecurity events have prompted the federal government to begin investigating strategies to transition to Zero Trust Architectures (ZTA) for federal information systems. Within federated mission networks, ZTA provides means to minimize the potential for unauthorized release and disclosure of information outside bilateral and multilateral agreements. But when federating with mission partners, there are potential risks that may undermine the benefits of Zero Trust. This article explores risks associated with integrating multiple identity models and proposes two potential avenues to investigate mitigation of these risks.

INTRODUCTION & BACKGROUND

Within days following the cyberattack on the Colonial Pipeline, U.S. President Joseph R. Biden Jr., signed into effect Executive Order 14028: Improving the Nation's Cybersecurity.[1] Prompted by recent "sophisticated and malicious" cyberattacks, the order acts as a catalyst for federal agencies to take necessary and immediate steps to coordinate with industry on improving information sharing, adopting best practices, and migrating federal information systems from perimeter-based security to a Zero Trust Architecture (ZTA). The foundational elements of Zero Trust are micro-segmentation and a well-informed trust algorithm. When effectively implemented with data tagging, Zero Trust provides a strong compartmentalization model that lends itself to federated mission partner environments. However, in an environment where mission partners are responsible for bringing to the table their own identity models, consideration must be given to risks associated with federating multiple mission partners.

Keith Strandell is the Materiel Leader for the Royal Saudi Air Force AWACS Modernization Program and previously served as the Materiel Leader for life cycle management of enterprise applications within the Enterprise Services portfolio for the United States Air Force's Enterprise Information Technology (EIT) construct, including capabilities such as the Air Force's Office 365 instance in the Impact Level 5 environment. Keith started his career as a Communication and Information Officer in the Air Force, leading the Network Security and Information Assurance offices at Travis Air Force Base. From there, he transitioned to a Technical Program Management role leading various hardware and software acquisition/development efforts related to battle control, intelligence, and Enterprise Information Technology systems. Included in this time frame were four years managing Foreign Military Sales cases across the Pacific, European, and Central Commands. Keith is currently pursuing a Ph.D. in Computer Science & Engineering at Mississippi State University.

In this article, we investigate the risks associated with a multi-partner environment built on ZTA that federates with each mission partner's identity model. For purposes of isolating the impact of federated identities, the operating assumption is that the environment has fully implemented micro-segmentation and data tagging such that the primary risks are associated with the integration of multiple identity models. In addition to assessing the risks, we recommend two potential areas of investigation that may alleviate some of the risks associated with this architecture.

MISSION PARTNERS AND DATA PROTECTION

Combatant Commands (COCOMs) work with a variety of international mission partners, the most obvious being foreign militaries. However, there is a significant degree of cooperation that occurs with other agencies. In January 2010, U.S. Southern Command (USSOUTHCOM) responded to a request for earthquake relief support by Haiti. This Humanitarian Assistance and Disaster Response (HA/DR) operation required coordination with multiple international organizations, including foreign government agencies, nongovernment agencies, and foreign militaries. In order to share information effectively, data were kept unclassified to the maximum extent possible and public platforms were used for dissemination.[2] Another example of cooperation with international partners can be found in a recent partnering among U.S. Africa Command (USAFRICOM), the International Criminal Police Organization (INTERPOL), and local law enforcement from several West African nations. The operation targeted illegal fishing and "other maritime crimes" along the West African coast.[3] Not only do these mission sets require sharing of unclassified data, but they also demonstrate the potential for both persistent and transient user bases operating in the same environment.

Sudip Mittal is an Assistant Professor in the Department of Computer Science & Engineering at Mississippi State University. He graduated with a Ph.D. in Computer Science from the University of Maryland Baltimore County in 2019. His primary research interests are cybersecurity and artificial intelligence. Mittal's goal is to develop the next generation of cyber defense systems that help protect various organizations and people. At Mississippi State, he leads the Secure and Trustworthy Cyberspace (SECRETS) Lab and has published over 70 journal articles and conference papers in leading cybersecurity and AI venues. Mittal has received funding from the NSF, USAF, USACE, and other agencies within the U.S. Government. He also serves as a Program Committee member or Program Chair of leading AI and cybersecurity conferences and workshops. Mittal's work has been cited in the Los Angeles Times, Business Insider, WIRED, the Cyberwire, and other venues. He is a member of the ACM and IEEE.

Attempting to create a collaborative environment to facilitate data sharing that allows for multiple missions and user bases increases the need for effective controls to prevent the unauthorized release and disclosure of information such as Controlled Unclassified Information (CUI). For example, data controlled as Not Releasable to Foreign Nationals (NOFORN) are not releasable to foreign mission partners; however, these data may need to reside in this environment due to a need to release to non-foreign entities such as the Federal Emergency Management Agency. Similarly, data controlled as "CUI//REL TO USA, FVEY" are releasable to members of the Five Eyes alliance.[4] A comparable protection requirement exists for mission partner data. The Mission Partner Environment framework is designed to facilitate collaboration and sharing with "participants within a specific partnership or coalition."[5] The implication is a requirement to ensure data are shared only within designated groups. For example, assume there are existing agreements among the United States, country A, and country B, as depicted in Figure 1. In this image, the overlapping areas represent shared data based on these partnerships. Each country contributing data to the environment expects the information it uploads to the system to be protected accordingly. That is to say, data transferred to the United States as part of a bilateral agreement with country A must not be released to country B without the express consent of country A.

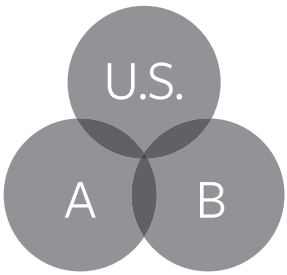


Figure 1: Multi-country data sharing partnerships.

ZERO TRUST AND FEDERATION

The operating assumption in ZTA is that the network is compromised and therefore steps must be taken to minimize the potential impact of unauthorized access. Through micro-segmentation and data tagging, a ZTA can provide a framework in which compartmentalization is baked into the security model. The result is smaller trust zones, which reduce the potential for lateral movement of an adversary exploiting a vulnerability (see Figure 2). However, to realize the benefits fully, the ZTA must also implement a trust algorithm that takes in relevant data feeds to provide continuous authentication and authorization decisions on access requests. A robust trust algorithm will have access to contextual information on the requesting entity and device, the target resource, resource access policies, and threat intelligence.[6] Access to these information feeds provides a more complete view of the request and associated risks. For example, consider the ability to access data related to the requesting entity's device configuration to compare those data with data on known configurations and thus to predict the level of vulnerability associated with the device.[7] Such an assessment increases the insight into the risk of a given request.

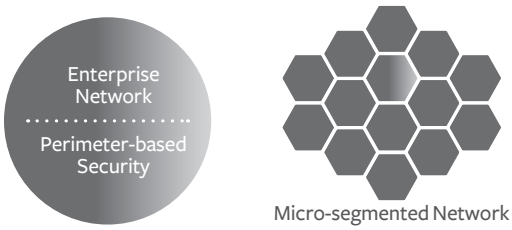


Figure 2: Lateral Movement in Perimeter Network vs. Zero Trust.

Federation in ZTA poses an interesting conundrum because, in an architecture that strives to remove trust, it introduces an inherent trust among the federated organizations. Here, the mission partners act as identity providers and are responsible for authenticating their users. Once authenticated, the identity is securely transferred to support the trust algorithm making the authorization decision. This model eliminates the need for the user to maintain security information related to a separate identity, which can reduce the risk of compromise associated with user behavior. However, it can undermine the rigor of the trust algorithm by preventing access to contextual information related to the requesting entity.

RISKS FEDERATING WITH MISSION PARTNER IDENTITY SOLUTIONS

Zero Trust touts a robust trust algorithm rooted in the ability to verify a user's identity. However, in a federated model, the algorithm is only as strong as the weakest identity solution. Figure 3 depicts a simplified model of a federated ZTA environment. The network supports countries A, B, and C. The entities in each country (e.g., military, law enforcement) have their own distinct identity models that are federated with the system. The Registered

User List provides a means for restricting users, standardizing attributes, and providing redirects to the appropriate identity system for authentication.

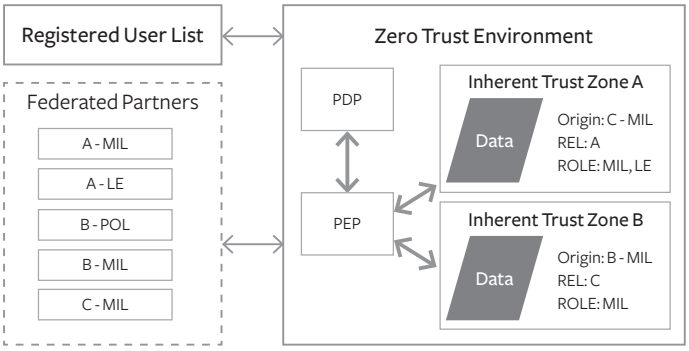


Figure 3: Notional Federated ZTA Environment

Ideally, each identity system would adhere to a minimum baseline that supports a context-rich authentication model. However, when balancing risk and mission requirements, mission may take priority and drive risky behavior or decisions. As such, there is the potential for integration with substandard models, which increases the risk of unauthorized access to the system. Assuming an ideal implementation of ZTA in the model above, access to a given Inherent Trust Zone will be restricted to an appropriate user base, and users granted access to a given zone cannot move laterally. Hence, a user who accesses Zone A above will not have access to Zone B without having gone through a separate access request. Therefore, in the model provided, the "C – MIL" identity system offers the widest potential reach for a threat, because it is the only one whose user base has authorized access to both Inherent Trust Zones.

Successful implementation of Zero Trust is predicated on a robust Trust Algorithm with access to contextual information around a given access request. For example, relevant information for an access request could include multiple authentication factors, device registration check, and device health status. The ability to access the contextual information around the requesting entity is a challenge in federated models.[8] This model assumes contextual checks occur within the authentication pipeline managed by the mission partner, and therefore the ZTA environment is effectively blind to the degree of rigor used to authenticate a user. If it is assumed the authentication model for the "C – MIL" identity system is strictly a username and password, it becomes a prime target for adversaries looking to access the system. Given the strength of ZTA in containerizing information, an adversary should only have access to those Inherent Trust Zones to which the compromised account has access. In this model, the obvious impact of a compromised "C – MIL" user account would be the unauthorized disclosure of data in Inherent Trust Zones B and A. However, there is also the potential to upload misinformation and malicious code that could compromise entities in countries A and B.

The effective establishment and enforcement of a common, robust authentication process increase the security of the system; however, they do not address vulnerabilities in the supply chain. On December 8, 2020, it was discovered that SolarWinds had been compromised. The hack, attributed to Russia, affected approximately 17,000 SolarWinds clients, including several federal agencies such as the Department of Homeland Security and the Department of Defense. The attackers targeted a third-party vendor, Orion, that had a long-standing relationship with SolarWinds. Because Orion had been infiltrated, when clients of SolarWinds updated their software they inadvertently loaded malware onto their devices and thus gave hackers access to their networks, which in many instances resulted in significant data breaches.[9] The attack is significant in that it focused on popular network infrastructure devices, which allowed for the vast attack surface. A comparable attack on the identity components used by "C – MIL" could provide an adversary with the ability to hijack existing credentials or bypass authentication processes. There also exists the potential for introducing fraudulent credentials, but, that can be mitigated with the effective implementation of a Registered User List.

This risk is amplified by strategic competitors' ability to leverage the Diplomatic, Informational, Military, and Economic (DIME) framework to deliberately position state-sponsored technology that provides them covert access. Strategic competitors such as the People's Republic of China (PRC) and Russia are actively exercising DIME strategies to advance their influence in regions around the world. General Stephen J. Townsend noted in his statement to the House Armed Services Committee that both countries have an "inside track" in central and southern Africa. He also stated that Russia is actively buying influence in the region and the PRC is investing billions in infrastructure and development in Africa.[10] Within USSOUTHCOM's area of responsibility (AOR), the PRC holds $165B in loans and is using COVID-19 as a pretext to indebt nations in the region further while enhancing its integration with their infrastructure and technology. For example, as part of their COVID-19 response, the PRC was offering to donate Huawei technology.[11] The significant investments these competitors are infusing into the region provide the pretext to gain access to senior government officials with the leverage to secure deals that further embed their technology or allow insight/access to processes like identity management. The infrastructure investments in these regions serve, at a minimum, as a method to increase reliance and influence. However, they also introduce the potential supply risk noted above. Specific to the PRC, there are concerns related to the Military-Civil Fusion Strategy and how involved vendors such as Huawei are with the People's Liberation Army and the extent of their collaborations.[12]

This concern is furthered by incidents which suggest not only security issues but the intentional inclusion of surveillance capabilities that lend themselves to espionage.[13] While the Huawei push is focused on 5G, the concern extends to any presumably state-sponsored technology that may serve as critical infrastructure for mission partner networks. Coupled

with the potential for growing an insider threat, there is the potential to undermine the processes of the Registered User List and reintroduce the risk of fraudulent accounts.

POTENTIAL ENHANCEMENTS

Reducing the risks associated with federating multiple identity providers in the model depicted requires introducing an additional layer into the authentication process that provides contextual data. Two promising designs for consideration are blockchain and Adaptive Neuro-Fuzzy Inference System (ANFIS). The former shows promise in reducing the likelihood of compromised credentials while the latter has the potential to identify and flag behaviors that deviate from the norm.

Blockchain

Blockchain first gained popularity as the digital ledger supporting bitcoin transactions; more recent implementations have shown its promise as a mechanism for augmenting or replacing existing authentication systems. The strength of blockchain lies in its immutable, secure nature, which comes from the combination of Merkle Tree hashing, encryption, distributed architecture, and consensus protocol.[14] Smart contract implementations of blockchain can support authentication models through its abilities both to store data and to automate processes. It has been proposed, for example, as an authentication model for a cloud-centric database that requires access from both internal and external users.[15] Blockchain has been shown to be capable of storing digital identities and data necessary to support authentication. It has also been shown to be capable of authenticating devices in an Internet of Things (IoT),[16] which may be leveraged to support an agent-based model that allows a user to register a limited number of devices. Within a federated network, blockchain has the potential to introduce a layer of managed context that decreases the likelihood of an account being compromised.

Adaptive Neuro-Fuzzy Inference System

Adaptive Neuro-Fuzzy Inference System is a machine learning framework that couples the learning capabilities of adaptive neural networks with the fuzzy inference system's ability to detect ambiguities in decision-making criteria. This combination makes it well-suited for applications such as nonlinear analysis, control systems, and expert systems. The ANFIS framework has been leveraged in areas such as an improving pattern password authentication performance for touchscreens,[17] anomaly classification to support intrusion detection in a vehicular ad hoc network,[18] and a continuous authentication system for mobile devices.[19] The latter utilizes ANFIS to learn passive and active patterns of use for a given mobile user in order to define a behavioral model. This allows the authentication system to monitor behaviors continuously and support implicit authentication while also flagging deviations.[20]

For the purposes of a federated ZTA environment, ANFIS has the potential to be leveraged in both a client-side and server-side model. A client-side model could potentially generate a confidence score specific to a user's behaviors that is passed as context for authentication. This could support identifying a compromised device. A server-side variant may support identification of anomalous behavior relative to an archetype based on attributes. For example, if a certain user base only logs in periodically to check email and a specific user's behavior is significantly more active, that anomalous behavior may represent an insider threat or compromised credentials.

CONCLUSION

The transition from perimeter-based cybersecurity to ZTAs should result in significant improvements in the overall security posture of enterprise networks. Specifically, it shows promise in the realm of multinational operations in which cooperation can often be born out of necessity and built on a tentative trust among mission partners. The inherent compartmentalization of a robust ZTA lends itself well to an environment rooted in mission partners' trust that their data are protected from unauthorized release and disclosure. Unfortunately, the benefits of Zero Trust can be undermined by the federating of multiple identity models, a risk made worse by actions of strategic competitors to employ the DIME framework to enhance their regional footprints, advance their influence, and deploy state-sponsored technologies. These activities increase the opportunities for social engineering, political influence, and clandestine cyber operations. Some of the risks can be mitigated by limiting federation to mission partners with known, trusted architectures and limited ties to strategic competitors while offering to host all other partners. This model, however, has the potential to be compromised when mission requirements outweigh the cybersecurity risks. To secure the environment's security posture further, additional measures should be investigated.

Two promising options for enhancing the authentication model that could be investigated as augmenting technologies are blockchain and Adaptive Neuro-Fuzzy Inference Systems. Blockchain gained popularity as the digital ledger supporting bitcoin transactions. However, recent efforts go well beyond that, using blockchain for authentication as part of a self-sovereign identity model. In 2019, a group of credit unions piloted the use of blockchain and noted the improvement in the authentication model could reduce a credit union's annual fraud expenses by $150K just by reducing the authentication risks tied to call centers.[21] ANFIS is a machine learning model that integrates adaptive neural networks with a fuzzy inference system. In a study on its potential use to support "continuous implicit authentication" on mobile devices, ANFIS was used to learn user behaviors for supporting implicit user authentication and identification of both informed and uninformed adversary attacks. While the model showed a 5% increase in user recognition, the improvement in informed adversary attacks was negligible and it underperformed on identifying uninformed adversary

attacks.[22] The ANFIS architecture does show promise for user authentication on mobile devices. However, if paired with an identity model, it may be used as part of an enterprise authentication solution that focuses on learning archetype behaviors to identify when a user's behavior deviates from the normal behaviors of users assigned to the same role, or from the user's own behavior pattern.

DISCLAIMER

The views expressed in this work are those of the authors and do not reflect the official policy or position of the United States Military Academy, the Department of the Army, the Department of the Air Force, or the Department of Defense.

NOTES

1. Joseph Biden, "Exec. Order No. 14028," 3 C.F.R. 86 (May 2021).
2. Gary Cecchine et al., *The U.S. Response to the 2010 Haiti Earthquake* (RAND Corporation, 2013).
3. U.S. Africa Command Public Affairs, "AFRICOM and law enforcement cooperation enhances maritime security in West Africa," April 2022.
4. *Controlled Unclassified Information Markings,* Office of the Under Secretary of Defense for Intelligence & Security and Director for Defense Intelligence (Counterintelligence, Law Enforcement & Security), September 2020.
5. J-6, Chairman of the Joint Chiefs of Staff, *Requirements Management Process for Mission Partner Environment*, Instruction CJCSI 6290.01 (September 2019).
6. Scott Rose et al., "NIST Special Publication 800-207," August 2020.
7. Paulo Shakarian, Jana Shakarian, and Kazuaki Kashihara, Systems and methods for vulnerability-based cyber threat risk analysis and transfer (US Patent 2022/0078203 A1, filed March 2021).
8. Koudai Hatakeyama, Daisuke Kotani, and Yasuo Okabe, "Zero Trust Federation: Sharing Context under User Control towards Zero Trust in Identity Federation," in *2021* IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops) (2021), 514-519.
9. Rahaf Alkhadra et al., "Solar Winds Hack: In-Depth Analysis and Countermeasures," in 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) (2021), 1–7.
10. Stephen Townsend, "Hearing to Receive Testimony on United States Central Command and United States Africa Command in Review of the Defense Reauthorization Request for Fiscal Year 2022 and the Future Years Defense Program," April 2021, www.aldersonreporting.com.
11. Craig Faller, "Statement of Admiral Craig S. Faller, Commander, United States Southern Command Before the 117th Congress Senate Armed Services Committee," 2021,
12. Elsa Kania and Lorand Laskai, "Myths and Realities of China's Military-Civil Fusion Strategy," Center for a New American Security, January 2021.
13. Elsa Kania, *Securing our 5G future: The Competitive Challenge and Considerations for U.S. Policy* (Center for a New American Security, November 2019).
14. Harsh Sheth and Janvi Dattani, "Overview of blockchain technology," *Asian Journal For Convergence In Technology* (AJCT), ISSN-2350-1146, 2019.
15. Gaurav Deep et al., "Authentication protocol for cloud databases using blockchain mechanism," Sensors 19, no. 20 (2019): 4444.
16. Mohamed Tahar Hammi et al., "Bubbles of Trust: A decentralized blockchain-based authentication system for IoT," *Computers & Security* 78 (2018): 126-142.
17. Orcan Alpar, "Intelligent biometric pattern password authentication systems for touchscreens," *Expert Systems with Applications* 42, no. 17 (2015), 6286-6294, https://www.sciencedirect.com/science/article/pii/S0957417415002948.
18. B. Karthiga et al., "Intelligent Intrusion Detection System for VANET Using Machine Learning and Deep Learning Approaches," *Wireless Communications and Mobile Computing* (2022).
19. Feng Yao et al., "Continuous implicit authentication for mobile devices based on adaptive neuro-fuzzy inference system," in 2017 International Conference on Cyber Security And Protection Of Digital Services (Cyber Security) (IEEE, 2017), 1-7.
20. Yao et al., 3-5.
21. Hyperledger Foundation, "Case Study: How CULedger protects credit unions against fraud with Hyperledger Indy," September 2020.
22. Yao et al., 6.