An Analysis of Prerequisites for Artificial Intelligence/Machine Learning-Assisted Malware Analysis Learning Modules

Portia Pusey
Porita Pusey, LLC
edrportia@gmail.com
0000-0002-8672-9083

Maanak Gupta
Dept. of Computer Science
Tennessee Technological University
TN, USA
mgupta@tntech.edu
0000-0001-9189-2478

Sudip Mittal
Dept. of Computer Science & Engineering
Mississippi State University
MS, USA
mittal@cse.msstate.edu
0000-0001-9151-8347

Mahmoud Abdelsalam
Dept. of Computer Science
North Carolina Agricultural and Technical State University
NC, USA
mabdelsalam1@ncat.edu
0000-0001-5627-5239

***Abstract*—This paper presents the findings of action research conducted to evaluate new modules created to teach learners how to apply machine learning (ML) and artificial intelligence (AI) techniques to malware data sets. The trend in the data suggest that learners with cybersecurity competencies may be better prepared to complete the AI/ML modules' exercises than learners with AI/ML competencies. We describe the challenge of identifying prerequisites that could be used to determine learner readiness, report our findings, and conclude with the implications for instructional design and teaching practice.**

Keywords—cybersecurity, machine learning, artificial intelligence, education, teaching

I. REVIEW OF LITERATURE

It would seem logical that learners with prerequisite competencies for a particular course, or learning activity, would perform better than learners without. Anecdotally, many educators could recount an unfortunate story of a learner who did not have the prerequisite competencies to meet the learning objectives of their course. In fact, research suggests that prerequisite competencies for an instructional activity correlate to positive student outcomes [1] [2]. However, the body of literature on learning science has no guidance for educator-authors of newly created modules, for newly created courses, in a new domain that combines two relatively new fields. This paper presents the findings of action research conducted to evaluate new modules created to teach learners how to apply machine learning (ML) and artificial intelligence (AI) techniques to malware data sets. We describe the challenge of identifying prerequisites that could be used to determine learner readiness, report our findings, and conclude with the implications for instructional design and teaching practice.

II. EASE OF USE

A. Increase Demand for AI/ML Skills

Threat actors are using AI/ML to improve their effectiveness and efficiency. The result is a greater need for automation and adaptation in risk management and other cybersecurity fields; a quick job search reveals that AI/ML is now one of the most sought-after skills in the security industry [3]. Capgemini Research Institute published findings from a survey of 850 senior executives from seven industries and ten countries, including the United States [4]. Of their respondents, 42% currently use, or plan to use AI assisted cybersecurity products, 28% of the respondents said that they use AI embedded security products, while 30% use proprietary AI algorithms [4]. The [4] report recommended that organizations prepare their cybersecurity analysts to be "AI-read" based on their finding that 63% of the respondents report planning to use AI-related technologies by the following year [4].

B. Increase in AI/ML Courses in Higher Education

In response to this surge in demand, institutions of higher education have increased their machine learning offerings. [5] describes changes to IA course offerings in higher education from 2018 to 2021. In their survey of 207,000 programs from 3,700 universities in over 120 countries, they found a 102.9% increase in the number of AI courses at the undergraduate level and a 41.7% increase at the graduate level [5]. And while cybersecurity researchers have actively developed novel AI and ML solutions for, cyber threat intelligence, malware analysis, malware classification datasets and instructional materials for AI/ML cybersecurity classes are slow to reach the cybersecurity education classrooms. [6] describe the datasets as old, "noisy, incomplete, insignificant, imbalanced, or may contain inconsistency instances related to a particular security incident [pp.16]."

C. AI/ML Modules and Prerequisites with Malware Datasets

The authors of this paper have designed, taught, and/or studied the implementation of six modules at the intersection of AI/ML and malware [7]. The six modules include (1) Cyber Threat Intelligence (CTI) and malware attack stages, (2) Malware knowledge representation and CTI sharing, (3) Malware data collection and feature identification, (4) AI assisted malware detection, (5) Malware classification and attribution, (6) Advanced malware research topic and case

studies. Each module consists of multiple lectures, background/technical readings/videos, and lab sessions (each with an appropriate data set). Each lab is self-contained so that it can be offered independently from the other five modules. We recommend two prerequisite courses or the equivalent competencies. The two courses are Introduction to Machine Learning/Artificial Intelligence and Introduction to Cybersecurity. Since the modules can be used at the undergraduate or graduate level, work experience and certifications also would provide the foundational AI/ML and cybersecurity competencies needed to successfully complete the hands-on assignments and performance assessments. Specifically, learners should have a fundamental understanding of cybersecurity foundations, cybersecurity principles, and IT systems components as specified in the NSA/DHS CAE-CDE designation requirements [8] and the content covered in most introductory ML/AI textbooks.

D. Prerequisites

[1] describes two kinds of prerequisites based on function. The two types of prerequisites are prerequisites for sequential courses or nonsequential courses. For example, [1] identifies math and English as foundational and critical for preparing learners to succeed in all their higher education courses whether a student matriculates as a math or English major. Sequential prerequisites are usually related to a learner's major and are dependent on each other (e.g., Cybersecurity 100 and Cybersecurity 200). This paper addresses sequential prerequisites with a twist – the prerequisites come from two different fields: cybersecurity and machine learning.

Schools and educators walk a tightrope when deciding whether a learner has adequate competency to successfully complete the learning outcomes for a course. Part of the problem is that a student's grade is the only quantitative way that schools, educators, and researchers can determine successful completion of a course. The problem with this method is that there is wide variance among educators, course materials (books, labs, content), assessments and schools; not to mention, student characteristics such as motivation, study skills, and attendance [1] [9]. Furthermore, in studies of prerequisites, researchers have noted that individual differences identified in prior studies, such as gender or socioeconomic status, can be explained by scores on standardized tests (ACT or SAT) or domain-specific assessments (concept inventories or advanced placement tests).

We enter the discussion about prerequisites by adding additional challenges faced by decision makers. Most of the studies of prerequisites were conducted in well-defined domains (e.g., physics [10], biology [11], accounting); Cybersecurity and IA/ML are ill-defined domains. The fields of cybersecurity, AI, and ML have matured over the past 70 years; but tools and techniques are constantly changing [12] [13]. Each field comes with its own vocabulary and methods none which overlap. Cybersecurity risks produce data related to threats, weaknesses, and impacts [14] [15]; AI/ML uses cybersecurity data to create models which can then make predictions or decisions "without being explicitly programmed to do so [16]."

The flexible implementation of the AI/ML for Malware modules creates another set of challenges for deciding how much prerequisite competency is required for the learners to meet the learning objectives of the module. To date, the modules have been used as workshops at several conferences, and in courses. A single module was used for each of the conference workshops; as few as two and as many as six modules were used in courses with undergraduate and graduate students; some learning together in the same course. We collected data to try to answer the following research question: What prerequisite competencies are required to successfully meet the learning objectives for the course?

III. METHOD

The first implementation and evaluation of the modules were at several conference workshops. Modules 4 and 5 were presented at several workshops between Fall of 2021 and 2022. We administered surveys to attendees after each workshop to determine (1) how confident attendees were that they could (a.) detect, (b.) collect, and (c.) identify malware. In addition to asking attendees for suggestions to improve the modules, we also asked what hindered the attendee's confidence in detecting, collecting and identifying malware.

Anecdotally, when looking at the qualitative responses of individuals who reported low confidence in using the conference workshop methods, prerequisite knowledge appears to be a factor. One respondent with the lowest rating of their confidence wrote the following response, "The major factor that hindered my understanding is my lack of knowledge about machine learning." Another respondent with low confidence suggested that their background knowledge might have inhibited their confidence, "I'm a complete newcomer to A.I. and Malware Analysis. I also doubt my ability to work with scripts."

We implemented changes that addressed the attendee feedback during the next workshop we conducted. We created a resource document that was distributed prior to the workshop. The attendees who used the resources said that it supported their competency development. We also implemented several interventions from learning science to address learner confidence. We

- asked learners what they already know about a topic before teaching. This enables instructors to address competency gaps while they are teaching the modules.
- conducted checks for understanding and recorded learner responses. We used this information to refine the resource document provided to learners.

All of this feedback was incorporated into the implementation of the modules in classrooms in spring 2022. Spring 2022 was the first opportunity we had had to evaluate the modules across different courses with different instructors and different schools.

- Instructor 1 had 40 students in his course at a public land-grant research university. The course title is Artificial Intelligence for Cybersecurity and includes all six modules.
- Instructor 2 included modules 3, 4, 5, & 6 in a course called AI Assisted Malware Analysis at historically black land-grant research university. This course has three prerequisites: (1) Graduate senior status; (2) Basic knowledge of cybersecurity and AI/ML concepts; (3) Ability to use/learn the following technologies: Python, ML libraries (e.g., Pytorch, Tensorflow, Scikit-learn, Keras, etc.)
- Instructor 3 included modules 4, 5 in his course for graduate students on the doctoral degree pathway at a public research university.

With IRB approval, data were collected by surveys, instructor notes, and learner records. There were 29 completed surveys with 24 learners providing their grades for the module evaluation.

IV. RESULTS

Participation was optional however we had 29 responses to the survey. There were 16 male, 12 female, and one participant who preferred not to say their gender. There was an almost even amount of Asian and white students. There were 11 Asian students and 12 white students. Two students reported being black or African American and four students preferred not to say. No students reported being Hispanic/Latinx/Spanish and three declined to say.

A. Experience in AI/ML, Cyber, & Computer Science

We asked respondents to let us know how much experience they have had with computer science, machine learning and cybersecurity. Most of the respondents have had some work experience in computer science and/or taken a course in computer science. Eleven of the respondents are in a CS doctoral program, 9 are in a master's program and 4 are in a bachelor's degree program. 11 of the respondents also have another 1-3+ years of CS work experience.

When we asked a similar question about machine learning, there was less experience among the respondents. Twenty-two of the respondents have taken an introduction to machine learning/ artificial intelligence course. However, only 1 respondent is currently seeking a doctoral degree in ML/AI and 2 are in a master's degree program. There are a total of 7 respondents who say they have work experience that includes ML/AI.

And lastly, we asked the respondents about their exposure to cybersecurity courses or work experiences. Nineteen respondents said they have taken an introductory course. Yet 10 said they had no cybersecurity work experiences and 8 said they had no course experience. Five respondents are in a master's degree program and 6 are in a doctoral degree program.

In general, the demographic survey indicates that about a third of respondents are taking advanced computer science courses and have computer science work experience. Up to 7 respondents have some ML/AI work experience and 8 respondents have cybersecurity work experience.

B. Module Feedback on Prerequisite Knowledge

We also asked students to respond to a survey about prerequisites. We should have had over 180 responses based on the demographic survey. Fifty-six students responded to the feedback survey. Eleven students responded about module one, 7 for module two, 8 for module three, 18 for module four, seven for module 5 and 6. Part of this distribution is related to the number of modules taught and the number of students in each course.

One hundred percent of the learners said that they had the prerequisite competencies needed to complete the laboratory exercise. There is no qualitative data that clarifies these answers. A hypothesis that may explain this unanimous response is that the survey was implemented after the learners successfully completed the assessment. Thus, they associated successful completion with pre-requisite knowledge.

C. Likelihood of Successfully Repeating Performance on the Exercise

A good measure of what students learned is to ask them how confident they are that they could complete the lab exercise associated with the module again. Three respondents said that they were not confident that they could complete the lab exercises again. This is not statistically significant from the number of learners who said that they could complete the exercise with or without the resources provided. Eight respondents suggested that they used additional resources to complete the exercises.

V. DISCUSSION

Given the background data we have collected for the implementation of modules 4 and 5, and the feedback we incorporated into these modules, we did statistical analysis on the survey and grade results for these two modules only. There was almost no variation in the grade data. For each module there were two learners who earned half credit and one that earned a near perfect score (87.67% and 97%). Therefore, there were no statistically significant differences between instructors and student prerequisite knowledge or learning experiences when comparing learners to learners using instructor, experience, confidence, grade, or resources as the grouping variable. However, there were slight trends in the distribution of the data.

- Less experienced learners used extra resources.
- More students who took Introduction to Machine Learning needed resources than students who took Introduction to Cybersecurity courses.
- But when looking at both ML and Cyber intro classes, there were 9 students who took both intro courses and did not use resources. There were 6 ML intro course takers who used resources. Three of those had also taken intro to cyber courses.

- The five of the six students who used extra resources, had perfect scores on modules. One of the students who used extra resources earned 50%

The trends suggest that learners with cybersecurity competencies may be better prepared to complete the modules' exercises than learners with AI/ML competencies; and having competencies in both fields seems to reduce the need for extra resources. Since we did not ask about using resources from past classes, students might not have considered them resources that they had to find themselves. So, this is a limitation of the action research which conclusions and inferences only apply to the modules and instructors in this one semester.

A result that is not statistically significantly different does not mean that observations of good instructional practices do not apply. Because despite the diversity of the students, almost all of them earned a perfect score. So, one could ask the question, **how was student performance achieved with the many differences in prerequisite competencies?**

VI. RECOMMENDATIONS

The three classes examined for this work were heterogeneous in terms of the degrees, background knowledge and experience of the students. It is likely that many cybersecurity courses are equally diverse. Several research-based learning strategies were used in the implementation of the modules that can be applied in all classrooms to mitigate the difference in prerequisite competencies.

1. Assess prerequisite knowledge
2. Have students work in teams
3. Provide supplementary resources
4. Use Classroom Assessment Techniques [17]

A. Assess Prerequisite Knowledge

Starting to teach without understanding what students know is like driving in a new city without a map. You may know where you need to end up, but getting there will be frustrating – for you and your students. Prerequisite knowledge is assessed by asking students what they know about instructional content before beginning instruction and learning activities; and there are many methods of doing this. Understanding a student's prior knowledge enables educators to address gaps in knowledge and misconceptions while teaching rather than while grading related assignments. It also helps learners recall what they have learned as preparation to learn the new content.

There are several ways to identify learner background knowledge. The first is to ask the students. Start the class session by asking the students directly what they know about the topic or whether they have experience with the learning activity. Another method is to assign a homework or in class activity that will assess students' prior knowledge. Or, run a speed test. Offer 10 true false questions that students must answer in a short period of time to identify common errors. Alternatively, instructors could provide a graphic of a process that students describe. Or provide teams of students multiple graphics that they must put in order and explain their decisions.

B. Teamwork

Teamwork is often despised by learners; however, it is the nature of their future professional work. For educators, teamwork supports peer learning. It is not unusual for each learner to hold a piece of the puzzle that completes a learning assignment This enables each learner to take the lead on tasks that contribute to solving laboratory exercises. Faculty who taught these modules found that teamwork helped even out the gaps in prerequisite knowledge.

Teams should be created by an instructor so that the gap in competencies between team members is not too large. Vygotsky's Zone of Proximal Development [18] requires learners to be close in competencies for their interaction to result in learning. If the gap is too large, advanced learners can become bored or end up doing all the work; novice learners get frustrated and may give up.

While teamwork comes naturally in an established professional setting, there are strategies that can make teamwork less painful in a classroom setting. Teams should set guidelines for their accountability and performance. Educators can make suggestions, however, by setting their own rules, teams build a habit for communicating their expectations. Educators should also provide authentic work roles for each member of the team. This better defines the contribution of each team member. Educators should also take the time to lead teams in reflecting on their learning experience and put what they have learned into action. However, teams should decide for themselves what are their strengths and what they can do to improve their collective performance.

C. Provide Supplementary Resources

When preparing to teach a class session educators have much to consider: the interests of their students, their learning styles, their background knowledge, and experiences, in addition to time limitations and course content requirements. It is impossible to personalize instruction for each learner, so supplemental resources help each individual student successfully meet the learning objectives for learning activities. Build a resource list by asking students what supplementary materials/resources they used to complete each learning activity. This saves students the time and frustration of searching for their own resources which may or may not help them solve the laboratory exercise.

Best practices for supplementary resources include providing a list for each learning activity (e.g., laboratory exercise, quiz, test) not for the course as a whole. Keep the list in an online document that can be updated and annotated by students. This keeps the list up-to-date and links functional. Provide for different learning styles. Include resources that can be read (e.g., books, articles, web pages), watched (e.g., videos, conference presentations), or listened to (e.g., podcasts, audiobooks). Integrate videos or interviews with professionals who can explain "why" learners are

building specific competences and give them tips to improve their performance.

This work introduced supplementary resources as a way of addressing the challenges workshop participants expressed because of a lack of prerequisite competencies. Workshop participants who received and used the supplemental information said that they were helpful.

D. Use Classroom Assessment Techniques

Educators often forget that they can improve their instruction and student performance by asking students directly. Classroom Assessment Techniques (CAT) [17] are a teacher improvement and classroom research technique created by Angelo and Cross [dates]. CAT "involves students and teachers in the continuous monitoring of students' learning. It provides faculty with feedback about their effectiveness as teachers, and it gives students a measure of their progress as learners. [17]."

CATs are successful because students respond anonymously. CATs should be conducted as a formative assessment. In other words, students should see their feedback in action while they are still taking the course. Asking for feedback after the class is over will not benefit the current students or enable educators to test out the new practice on the learners who made the recommendation. Acting on learners' feedback immediately reminds the learners that their time is well spent responding to the CAT prompts.

This work included *CAT 44: Group Instructional Feedback Technique GIFT*. This CAT is designed to answer the following three questions: What do students think is helping them learn? What is hindering the students' learning? What specific suggestions do the students have for improving learning? A google form was created and CAT 44 was administered after several of the modules. The feedback we received was used to improve the modules. The data we collected provided evidence that suggested that the learners needed more resources to compensate for gaps in their prerequisite competencies.

VII. CONCLUSION

While research has been conducted to examine the connection between learner outcomes and prerequisite information the results of these studies are conflicting. Some research and anecdotal evidence suggest that prerequisites improve learner performance, while other research suggests the differences can be explained by looking at additional variables. Other research suggests that prerequisites do not make a difference in learner outcomes.

Despite vast differences in preparedness, this action research showed no statistically significant differences between learners with different amounts of prerequisite competencies and experiences. This same group of learners achieved almost unanimously perfect scores on the assessments for the modules examined. In this work we provide recommendations for evidence-based instructional interventions that we applied in our modules to address gaps in prerequisites. Future studies will investigate how instructional interventions can address the diverse needs of heterogeneous classes in cybersecurity. Given the multiple fields involved in these modules, future research could inform instructional practice for cybersecurity, machine learning, and potentially computer science educators.

ACKNOWLEDGEMENT

This work was supported by National Science Foundation awards 2025682, 2150297, 2133190.

REFERENCES

[1] F. Abou-Sayf, "Prerequisite courses: statistical considerations," *Community College Review*, pp. 34-38, 1999.

[2] F. Abou-Sayf, "Does the elimination of prerequisites impact course success?," *Community College Review*, vol. 36, no. 1, pp. 47-62, 2008.

[3] B. Roussey, "The 8 most in-demand cybersecurity skills for 2019," TechGenix, 14 December 2018. [Online]. Available: https://techgenix.com/in-demand-cybersecurity-skills/.

[4] Capgemini Research Institute, "Reinventing cybersecurity with artificial intelligence: the new frontier in digital security," 2019. [Online]. Available: https://www.capgemini.com/wp-content/uploads/2019/07/AI-in-Cybersecurity_Report_20190711_V06.pdf.

[5] J. Clark and R. Perrault, "Artificial Intelligence Index Report 2022," 2022. [Online]. Available: https://aiindex.stanford.edu/report/.

[6] I. Sarker, A. Kayes, S. Badsha, H. Alqahtani, P. Watters and N. Ng, "Cybersecurity data science: an overview from machine learning perspective," *Journal of Big Data*, vol. 2020, no. 41, pp. 7-41, 2020.

[7] M. Gupta, S. Mittal and M. Abdelsalem, "Collaborative research: SaTC: EDU: artificial intellegence assisted malware analysis," [Online]. Available: https://sites.google.com/view/nsfsatcaima/home.

[8] Center of Academic Excellence in Cyber Defense Education , "(CAE-CDE) Designation Requirements," [Online]. Available: http://www.iad.gov/NIETP/documents/Requirements/CAE-CD_2019_Knowledge_Units.pdf

[9] C. Kauffman and D. Gilman, "Are prerequisite courses necessary for success in advanced courses," ERIC ED475157, ED475157, 2002.

[10] E. Burkholder, G. Murillo-Gonzalez and C. Wieman, "Importance of math prerequisites for performance in introductory physics," *Education Research*, vol. 17, no. 1, 2021.

[11] F. K. Abou-Sayf, "Challenges inherent in the assessment of the impact of prerequisite courses," *Journal of applied research in the Community College*, vol. 17, no. 1, pp. 9-15, 2009.

[12] C. Smith, B. McGuire, T. Huang and G. Yang, "The history of artificial intelligence," 2006. [Online]. Available: https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf.

[13] A. Fradkov, "Early history of machine learning," *IFAC-Papers On Line*, vol. 53, no. 2, pp. 1385-1390, 2020.

[14] A. Jun and Z. Pala, "Information sharing in cybersecurity: a review," *Decision Analysis*, vol. 16, no. 3, pp. 172-196, 2019.

[15] E. Fisher, "Cybersecurity issues and challenges: in brief," Congressional Research Service, Washington, DC., 2014.

[16] Wikipedia, "Machine Learning," n.d.. [Online]. Available: https://en.wikipedia.org/wiki/Machine_learning#cite_note-2.

[17] T. Cross and K. Angelo, Classroom assessment techniques, Jossey Bass Wiley, 2012.

[18] R. Walker, "Zone of Proximal Development," in *International Encyclopedia of Education (Third Edition)*, Science Direct, 2010.