Enabling Data Science Education in STEM Disciplines through Supervised Undergraduate Research Experiences

Yaser Banadaki
Department of Computer Science, Southern University, Baton Rouge, LA 70813, USA

Abstract

Data Science plays a vital role in sciences and engineering disciplines to discover meaningful information and predict the outcome of real-world problems. Despite the significance of this field and high demand, knowledge of how to effectively provide data science research experience to STEM students is scarce. This paper focuses on the role of data science and analytics education to improve the students' computing and analytical skills across a range of domain-specific problems. The paper studies four examples of data-intensive STEM projects for supervised undergraduate research experiences (SURE) in Mechanical Engineering, Biomedical science, Quantum Physics, and Cybersecurity. The developed projects include the applications of data science for improving additive manufacturing, automating microscopy images analysis, identifying the quantum optical modes, and detecting network intrusion. The paper aims to provide some guidelines to effectively educate the next generation of STEM undergraduate and graduate students and prepare STEM professionals with interdisciplinary knowledge, skills, and competencies in data science. The paper includes a summary of activities and outcomes from our research and education in the field of data science and machine learning. We will evaluate the student learning outcomes in solving big data interdisciplinary projects to confront the new challenges in a computationally-driven world.

Keywords: Data science, supervised undergraduate research experiences, computing, and analytical skills.

1. Introduction

All STEM students must have access to high-quality education to ensure our nation's economic growth [1]. STEM occupations account for more than 50% of the employment in major industries, which outpaces the production of STEM degrees in America [2]. It is now abundantly clear that the economic health of the nation depends on a robust infrastructure based on STEM foundations. Yet, we have a vast talent pool comprising underrepresented student groups (minority, female, and low-income) that, unfortunately, remain under-utilized because of the lack of opportunity to have access to high-quality education. The McKinsey Report [3] estimates a shortage of almost 200,000 people with deep analytical skills and 1.5 million managers and analysts to analyze big data and make decisions based on their findings in the United States. Similarly, it has been predicted [4] that in many states, mathematical and computing occupations will display the largest ten-year growth of any sector by 2022.

Furthermore, the National Center for Women & Information Technology [5] projected that up to 77% of future job openings could be filled by people with computing degrees. Despite the job opportunities, states with large minority populations like Louisiana had only 365 (18% female) and 455 (24% female) computer science (CS) graduates in 2015 and 2017, respectively [6]. Although the trend is positive, these numbers are far from the current 2,333 computing job openings in this state [6]. African Americans fill 5.5% of STEM jobs at levels disproportionate to their overall representation (11%) in the workforce [7]. On top of the small number of CS majors, most science and engineering graduates, especially among underrepresented minorities, do not acquire the essential skills and dispositions needed to succeed in a computationally-driven world.

Data Science (DS) makeup not only the largest occupation groups in STEM [2] but also enable underserved students to develop and apply the knowledge and skills in STEM disciplines. DS is often taught in a decontextualized manner in which these topics are presented in a CS class that is separated from their application in other content areas. This approach is inefficient in that students do not see the connection and usefulness of this field across their many applications in other disciplines. As such, the CS students do not perceive the significance of data science as creative and applicable topics that can help solve problems in various contexts. Judd et al. [8] showed that integrating advanced computing subjects in non-CS curricula promotes students' interest and inspires them to use the power of computing to solve real-world domain-specific problems. Similarly, the non-CS students do not have the opportunity to engage in science-based research for their capstone projects in data science. The number of courses offered in the areas of computational science in STEM disciplines is minimal, and mostly there is no data science courses offered in non-CS STEM programs. Also, enrollment in standard CS course sequences does not always serve these diverse STEM student populations well, so a few non-CS students enroll in advanced ML courses. The average number of students enrolling in technical CS courses from other STEM majors is limited, and most of the mandatory courses in STEM majors have no modules for data science and analysis techniques.

Acquiring computing and data science skills is essential for innovation and competitiveness that enables many underserved students to navigate successful STEM career pathways. Recent studies [9,

10] found that students' exposure to advanced CS techniques, such as machine learning (ML) as new research tools for a variety of majors, can potentially engage them in computing and place them on a trajectory for the future CS-related STEM education and employment. To maximize the impact of computing, AI, ML, and data science in other disciplines, the universities need not only carefully plan how to align CS courses for different STEM fields but also encourage computing-enhanced experiences for STEM capstone projects. The Next Generation Science Standards (NGSS) explicitly calls out CS topics in their practices as a tool to effectively engage students in CS experience [11]. Integrating ML and STEM not only advances the students' knowledge and skills in CS but also promotes deeper learning of STEM concepts. Applying DS techniques in STEM capstone projects improves how and what students learn in the classroom by supporting four fundamental characteristics of learning: active engagement, participation in groups, frequent interaction and feedback, and connections to real-world contexts.

The goal of this paper is to encourage the implementation of interdisciplinary DS-intensive projects for STEM students. The rest of the article is organized as follows. Section 2 presents the guideline to implement an interdisciplinary DS-intensive STEM supervised undergraduate research experience (SURE). This section will also include the summary of activities and outcomes from our research and education in the fields of data science and machine learning. Section 3 provides examples of four interdisciplinary projects to highlight the critical role of ML literacy across all non-CS STEM disciplines and encourage faculty and students in the adoption of ML techniques for real-world problem-solving. The four ML-intensive STEM projects include the applications of ML for improving additive manufacturing, automating microscopy images analysis, identifying the quantum optical modes, and detecting IoT network intrusion. The last section draws summarizing conclusions and future work.

2. Data Science in STEM SURE

Data Science is increasingly central to innovation across a wide range of disciplinary domains. Thus, it is crucial to encourage STEM students to explore a connection between their domain-specific capstone projects and the interdisciplinary potentials of the data science field. To support effective DS-enhanced domain-specific capstone projects, supervised undergraduate research experiences (SURE) can be designed. The SURE provides the opportunity for STEM students to have both the CS and non-CS faculty advisors. The interdisciplinary research subjects bridge CS and STEM domains to increase the research experience and participation in computational and data science and transfer the contents of computer science, computational science, and data analytics to STEM students. The SURE research can be part of their senior capstone project or thesis required for graduation supervised by both CS and non-CS faculty. To implement effective SURE research, the CS mentors need first to learn the students' computing skills to develop an effective interdisciplinary DS research project and modify/adjust the computational tasks accordingly. Mentoring relationships need to consider students' personal goals, needs, previous educational experiences, and learning styles to effectively promote STEM students' interests in ML research and education. The SURE's interdisciplinary projects will establish foundations for a strong, trusting, and supportive relationship between students and mentors that flows both directions instead of only from mentors down to students. As such, the faculty mentors could effectively stress the importance of CS skills for developing their theoretical and experimental concepts in their disciplines. The common goal of the CS faculty mentor is to support computing literacy, computational thinking, data analysis, and programming skills to advance the adoption of DS techniques for solving and modeling real-world problems in STEM disciplines. As students become more proficient with the fundamentals of ML applications and data science, the faculty mentors need to increase attention to their progress both as researchers by acting as a consultant and as professionals by suggesting lines of inquiry and options for solving domain-specific problems.

The ML-intensive SURE project harnesses the role of future ML developments in STEM disciplines. It empowers undergraduate students to confront the challenges in computational and data-enabled sciences to analyze and make decisions based on their findings. The SURE projects focusing on DS application could provide a forum through which undergraduate students acquire the confidence and scientific preparation required to enroll in a graduate program and pursue a professional career related to data science. SURE's DS-intensive interdisciplinary projects could improve a shortage of workforce with deep analytical skills in the United States. Even though the students could learn how to use computational tools for their domain-specific applications through attending CS-related classes, many technical details remain a black box for many students. For instance, choosing the optimal hyperparameters for neural networks is critical to have high performance and precision. However, it requires research experience to apply extensive trial and error to find the optimal values for the ML model. Without having enough skills, it becomes a huge problem for the students starting in the field. Enhancing the interdisciplinary DS-based projects, such as four examples of SURE projects designed in this paper, increase the number of qualified graduates with deep analytical skills for the nation's workforce needs.

In addition to individual mentors-mentees relationships, students will attend ML-related

technical meetings and conferences. The program exposes students to additional areas of study and provides a guide for the development of their communication and presentation skills. Students have the opportunity to write ML-related domain-specific papers and present them to a large audience at conferences. It boosts undergraduate students' confidence and creates a sense of ownership. The program makes students interested in pursuing a graduate degree in their discipline with intensive computational and data sciences. It ignites a long-term interest in students to pursue the ML-knowledgeable STEM workforce. For STEM students, CS conferences will improve their understanding of the significance of employing computational and AI techniques to solve real big data problems. The students are encouraged and supported to participate in regional and national conferences, seminars, and workshops, where they present posters or papers and expose to the research in AI, ML, computational, and data science. The students have the opportunity to network with scholars who are leaders in the field, which positively impacts their career development. The CS conferences also update undergraduate students on the latest employment trends and internship opportunities in AI, ML, computational, and data science. Also, attending CS conferences provides a good knowledge of professional standards, ethical issues, job environment responsibility, balancing career and personal life. The STEM SURE's goal is to provide a forum through which undergraduate students acquire the confidence and scientific preparation required to enroll in a graduate program and pursue a professional career related to AI, ML, computing, and data sciences. STEM SURE guides students toward becoming independent creators of knowledge or users of ML concepts to apply them to their disciplines, prepare for the career paths of their choice, and be ready to move on to the next phase of professional life.

The SURE component of our computing and data education (CoDE) project increases the interest, engagement, and participation of African American students in interdisciplinary fields. 63% of the audience and 43% of the student presenters in the 1st SU-CoDE symposium were female. The students also presented six presentations in the proceedings of the 2020 undergraduate research conference, Louisiana council on excellence in undergraduate research (LaCOUR). Three of the extended abstracts are published in the LaCOUR journal. Seven students also presented in the 95th annual meeting of the Louisiana Academy of Sciences (LAS). All the students affirmed the research program's strong impacts on their ability to identify appropriate graduate programs, communicate ideas, develop effective presentations and research manuscripts, and speak in front of a group. Similarly, our findings show the program strongly impacted all the students' desire to work in a STEM field and use computing and analytical skills in their careers. 86% of students mentioned that the SURE program increased their knowledge of interdisciplinary research, computational research, and data science. The students learned the essential data science concepts and skills needed to succeed in their successful careers. The impact of interdisciplinary ML-based SURE projects on economic development is through producing qualified STEM graduates with the ability to collaboratively apply ML skills in their domain-specific problems and across a range of contexts and challenging issues in STEM. Improving the quality and number of underrepresented graduates will help fill the needs of more qualified STEM professionals in the nation.

3. Examples of Interdisciplinary Machine Learning Projects for STEM SURE

3.1. CS + Mechanical Engineering SURE Project: Data Science for additive manufacturing

Additive manufacturing (AM) [12] or 3D printing is a technique of blending or depositing materials layer upon layer in precise geometric shapes. As AM is rapidly gaining viability in mass production, it is essential for students entering the workforce to understand how to use this technology. The parts built using the state-of-the-art powder-bed AM, however, have remarkable, unpredictable mechanical properties. The attempts to improve the geometry and mechanical properties of final products in the 3D printing process were usually limited to the development of a typical control system using feedback from sensor measurement. Current AM systems only have limited sensing capability, and most of them are inaccessible to the users. Future AM needs to be a smart system that can perform self-monitoring, self-calibrating, and real-time quality self-controlling. In this SURE research project, students learned how to employ the ML approach through a Deep Convolutional Neural Network (DCNN) [13] to detect the defects in printing the layers automatically. The students understood how to use a transfer learning approach based on Google's open-source Inception-v3 model in the Tensorflow framework. The SURE project trained a new generation of engineers who bring ML-related skills to the AM workforce and any manufacturing business. The students who learn ML and apply it to AM can bring creativity to design and production.

3.2. CS + Biology SURE Project: Data Science for microscopy images analysis

The general cell quantification tools are manual or semi-automated techniques that are time-intensive, cumbersome, and prone to human errors. Available analysis software is based on the assessment of fixed immunolabelled tissue samples, making it impossible to follow the dynamic development of neurite outgrowth. This SURE project develops an accurate and fully automated technique for the quantitative analysis of microscopy images. ML approach through

a DCNN model has recently shown remarkable success in image-based data analysis resulting in a tremendous improvement in automated detection of complex morphologies [14]. In this SURE project, the students learn how to employ an ML approach to generate an accessible bio-imaging analysis tool for biologists to detect and quantitatively analyze cells in high-content microscopy images accurately.

3.3. CS + Physics SURE Project: Data Science for automated quantum optical mode recognition

Computational analysis and data science are required knowledge for students to design, model, and create optical sensing security products, light detection products, and products that use lasers. The CCD or CMOS cameras are usually used to capture the mode information such as centroid and radius of a beam profile, but identifying the modes with human eyes is challenging, especially in higher modes. In this SURE project, students learn how to apply ML algorithms to reconstruct and detect the quantum properties of optical systems [15]. Students learn how the marriage of the ML and quantum physics may give birth to a new research frontier that could automate the laboratory procedure and facilitate the experimental techniques. The project motivates students to use ML to discover new fundamental concepts in physics and effectively enhance basic conceptual understanding of physics. This strategy paves the way to transform the educational aspects of our current laboratory experiments into new domains.

3.4. CS + Cybersecurity SURE Project: Data Science for IoT network intrusion detection

The internet is one of the most used inventions to date, and every day billions of people access the web to handle a wide variety of tasks such as email, banking, and storing data. Internet of Things (IoT) has grown up rapidly that making security and central privacy subjects for network design. In this SURE project, students learn how the application of data science ensures the protection and monitoring of information and data in a network by detecting anomalies and intrusions in networks in real-time. The project helps students to understand how the marriage of two CS concepts, ML and cybersecurity, can create a smarter network by analyzing the traffic with better accuracy, increasing the correct alerts of bad and good network activities.

4. Conclusion

Data science is an ideal research field to promote innovation across a wide range of disciplinary domains. The author examines the educational strategies for engaging STEM students in ML research through developing supervised undergraduate research experiences (SURE) program and demonstrates how numerous interdisciplinary capstone projects can be developed in disciplinary domains to highlight the significant role of data science in the undergraduate STEM disciplines. The paper provides the essential guidelines to effectively educate the next generation of STEM students who can efficiently solve big data problems in their disciplines and confront the new challenges in a computationally driven world.

Acknowledgment:

This work was funded by the NSF HBCU-UP project (NSF Award # 2011900): Targeted Infusion Project: Southern University – Computing and Data Education (SU-CoDE).

References

[1] Langdon D., McKittrick G., Beede D., Khan B., and Doms M. (2011), "STEM: Good Jobs Now and for the Future. ESA Issue Brief# 03-11," *US Department of Commerce*.

[2] Fayer S., Lacey A., and Watson A. (2017) "STEM occupations: Past, present, and future," *Spotlight on Statistics,* pp. 1-35.

[3] McKinsey G. (2011) "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Global Institute*, pp. 1-6.

[4] Jindal R. (2014) "Louisiana Workforce Commission Employment forecasts by occupation and industry workforce investment council,"*http://www.laworks.net/Downloads/PR/WIC/WICPresentation20140610.pdf*.

[5] Education C. (2011) "Future Jobs: A Look at National, State, and Congressional District Data," *National Center for Women & Informaton Technology*.

[6] Code.org (2019) "Support K-12 Computer Science Education in Louisiana," *Advocacy Coalition and CSTA State of Computer Science Education,* https://code.org/advocacy/state-facts/LA.pdf.

[7] Charleston L., Adserias R. P., Lang N. M., and Jackson J. F. (2014) "Intersectionality and STEM: The role of race and gender in the academic pursuits of African American women in STEM," *Journal of Progressive Policy & Practice,* 2(3), pp. 273-293.

[8] Judd B. C. and Graves C. A. (2012) "Cellular STEM: Promoting interest in science, technology, engineering, and math education using cellular messaging, cloud computing, and web-based social networks," *12th IEEE/ACM Intl Symp. on Cluster, Cloud and Grid Computing*, pp. 799-804.

[9] Wang F., Kinzie M. B., McGuire P., and Pan E. (2010) "Applying technology to inquiry-based learning in early childhood education," *Early Childhood Education Journal,* 37(5), pp. 381-389.

[10] Davis D. (2014) "10 Years of Advanced Placement Exam Data Show Significant Gains in Access and Success; Areas for Improvement," *The College Board Communications Office,* 11.

[11] Lee I., Martin F., Denner J., Coulter B., Allan W., and Erickson J., and Werner, L. (2011) "Computational thinking for youth in practice," *Acm Inroads,* 2(1), pp. 32-37.

[12] B. Mueller (2012) "Additive manufacturing technologies–Rapid prototyping to direct digital manufacturing," *Assembly Automation*, Springer.

[13] Simonyan K. and Zisserman A. (2014) "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*.

[14] Sadanandan S. K., Ranefall P., Le Guyader S., and Wählby C. (2017) "Automated training of deep convolutional neural networks for cell segmentation," *Scientific reports,* 7(1), pp. 1-7.

[15] Ndagano B., Mphuthi N., Milione G., and Forbes A. (2017) "Comparing mode-crosstalk and mode-dependent loss of laterally displaced orbital angular momentum and Hermite–Gaussian modes for free-space optical communication," *Optics letters,* 42(20), pp. 4175-4178.