Yang Shi

 +1 706 207 8251 •   yshi26@ncsu.edu •   yshi.info

Education

North Carolina State University **Raleigh, NC**
Ph. D., Computer Science (GPA 4.0), Graduate Merit Award *August, 2019 – Current*

University of Georgia **Athens, GA**
M. S., Computer Science, Graduate Assistantship *August, 2015 – December, 2017*

Central South University **Changsha, China**
B. Eng., Automation, Graduate Excellence Award *September, 2011 – June, 2015*

Research Funding and Awards

- **Best Paper Award, EDM 2021** *(July 2021)*
- **Summer Graduate Merit Award** ($2,950) *(May 2021)*
- **Amazon Cloud Research Credits** ($6,000) *(Spring 2020)*
- **Graduate School Travel Award** ($600) *(Summer 2018)*

Publications

(*Italics are supervised Master students*)

- [13]. **Yang Shi**, Min Chi, Tiffany Barnes and Thomas Price: Knowing both when and where: Temporal-ASTNN for Early Prediction of Student Success in Novice Programming Tasks. To appear in The 15th International Conference on Educational Data Mining (EDM 2022), 2022. (Acceptance Rate 28.6%, 26/91 Full Papers)
- [12]. James Skripchuk, **Yang Shi** and Thomas Price: Identifying Common Errors in Open-ended Machine Learning Projects. The 53rd ACM Technical Symposium on Computing Science Education (SIGCSE 2022), 2022. (Oral Presentation)
- [11]. *Poorvaja Penmetsa*, **Yang Shi** and Thomas Price, "Identifying Struggling Students in Novice Programming Course with Knowledge Tracing." Work-In-Progress Track, CSEDM Workshop @ EDM'21 (Oral Presentation)
- [10]. **Yang Shi***, Ye Mao*, Tiffany Barnes, Min Chi and Thomas Price: More With Less: Exploring How to Use Deep Learning Effectively through Semi-supervised Learning for Automatic Bug Detection in Student Code. The 14th International Conference on Educational Data Mining (EDM 2021), 2021. (Oral Presentation, Combined Acceptance Rate 27.2%, 44/162 Short Papers)
- [9]. Ye Mao, **Yang Shi**, Samiha Marwan, Thomas Price, Tiffany Barnes and Min Chi: Knowing both when and where: Temporal-ASTNN for Early Prediction of Student Success in Novice Programming Tasks. The 14th International Conference on Educational Data Mining (EDM 2021), 2021. (Oral Presentation, Acceptance Rate 22%, 22/100 Full Papers)
- [8]. Samiha Marwan, **Yang Shi**, Ian Menezes, Min Chi, Tiffany Barnes and Thomas Price: Just a Few Expert Constraints Can Help: Humanizing Data-Driven Subgoal Detection for Novice Programming. The 14th International Conference on Educational Data Mining (EDM 2021), 2021. (Oral Presentation, Acceptance Rate 22%, 22/100 Full Papers, **Best Paper Award**)
- [7]. **Yang Shi**, *Krupal Shah*, Wengran Wang, Samiha Marwan, Poorvaja Penmetsa, Thomas Price: Toward Semi-Automatic Misconception Discovery Using Code Embeddings. The 11th International Conference on Learning Analytics & Knowledge (LAK 21), 2021. (Oral Presentation, Acceptance Rate 29.3%, 29/99 Short Papers)
- [6]. Wengran Wang, Yudong Rao, **Yang Shi**, Alexandra Milliken, Chris Martens, Tiffany Barnes, Thomas W Price: Comparing Feature Engineering Approaches to Predict Complex Program-

ming Behaviors: Comparing Feature Engineering Approaches to Predict Complex Programming Behaviors. CSEDM Workshop @ EDM'20, 2020.

- o [5]. **Yang Shi**, Fangyu Li, Wenzhan Song, Xiang-Yang Li, Jin Ye, Energy Audition based Cyber-Physical Attack Detection System in IoT. ACM SigMobile China, 2019.
- o [4]. Fangyu Li, **Yang Shi**, Shinde Aditya, Jin Ye, WenZhan Song, Enhanced Cyber-physical Security in Internet of Things through Energy Auditing. IEEE Internet of Things Journal.
- o [3]. Fangyu Li, Shinde Aditya, **Yang Shi**, WenZhan Song, Xiang-Yang Li, System Statistics Learning based IoT Security: Feasibility and Suitability. IEEE Internet of Things Journal.
- o [2]. **Yang Shi**, Fangyu Li, Tianming Liu, Fred R. Beyette, WenZhan Song, Dynamic Time-frequency Feature Extraction for Brain Activity Recognition. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).
- o [1]. Minhui Zou, **Yang Shi**, Chengliang Wang, Fangyu Li, WenZhan Song, Yu Wang, PoTrojan: powerful neural-level trojan designs in deep learning models. (*Available on arXiv*)

Employment

- o **Research Intern**
  Microsoft Research, Redmond, WA. *March 2022 –*
- o **Summer Instructor**
  North Carolina State University, Raleigh, NC. *Summer 2020, Summer 2021*
  I teach CSC116, Introduction to Java course over the summer semesters. These classes are functioned as the first class of CS major, minor and certificate students. The class size has been about 60 students, while the summer semesters are three months. The course content covers basics about Java programming, including basic operations, logic, loops, objects and file management.
- o **Research Scientist Intern**
  Stratifyd Inc., Charlotte, NC. *May 2019 – August 2019*
  This internship focuses on design and development of the algorithms for natural language processing. The main research project is stock earning call transcript sentiment analysis. In this project, I collect transcript data and over 10 different financial and sentiment dictionaries, and leverage spaCy tool for text analysis. The sentiment classification accuracy is 80%, and is 5% higher than state of art statistic models in a pilot dataset, and further improved to 84% with Snorkel framework. I also implement Thai language tokenization and sentiment analysis in the platform, and integrate attentive bidirectional LSTM model in the product line.

Selected Research Projects

- o **Semi-Supervised Learning (SSL) for Student Bug Detection**
  The lack of data has been an issue for deep learning in automated tasks to help student programming. In this project, I use semi-supervised scheme to train Code2Vec and ASTNN models to compare their performance with classic machine learning models in automated bug detection tasks. I use student programming data from Virginia Tech, and use the deep models to detect three types of common bugs. This project shows that the two deep learning models perform better in detecting bugs even with small amount of data, with the help of SSL. This work has been published in EDM 2021.
- o **Discovery of Student Bugs with Attention Weights**
  This project has focused on using programming data to assist student learning process by giving automated feedback. I use data from a programming class in NCSU, and apply Code2Vec model to automatically grade the student submissions, while using the attention layer representations to explain student common mistakes. One paper has been published with this project in LAK 2021.
- o **Trojan Attacks in Deep Learning Neurons**
  Received Media Coverage from The Register.
  This project is built from a novel idea that we can manipulate the output to the target given access to the model. I mainly participate in idea generation, feasibility discussion, and the experiments. Multiple experiments are built in Tensorflow framework and the models are examined with the accuracy of controlling

output in different layers. In experiment, we achieve around 99% successful rate in controlling the Alexnet, VGGnet, etc. models.

o **Energy Audited IoT Security**
This project is aimed at proposing a novel framework, including a series of algorithm design in using energy meter data for security applications in IoT systems. Different problem modeling and machine learning algorithms are designed and tuned for the characters of the energy data, including shallow models such as HMM, SVM, and deep models. So far, we have submitted three papers using statistical analysis, HMM method, and deep learning methods, and the experiments are all written in Python. The deep learning framework is Keras.

o **Dynamic Time-frequency Feature Extraction for Brain Activity Recognition**
In this project, a framework is proposed to classify the brain intention of left or right more accurately. The time series data is analyzed in time-frequency (TF) domain. This framework applies the idea that in TF domain, time information can be used for aligning the features in order to capture the event more efficiently. The experiments are performed with real data collected with EEG headset, and processed in MATLAB. The algorithm achieves an accuracy of 91%.

Academic Services

Seminar Talk:..
o CEI Brown Bag: Department of Computer Science, North Carolina State University (April 2021)

Organizer/Co-Organizer/Chair..
o Educational Data Mining in Computer Science Education (CSEDM) Workshop (2020 - 2022)

Reviewer:..
o Educational Data Mining (EDM) (2022)
o ACM annual conference on Innovation and Technology in Computer Science Education (ITiCSE) (2022)
o ACM CHI 2022 Conference on Human Factors in Computing Systems (2022)
o Handbook of Artificial Intelligence in Education (2021)
o IEEE Transactions on Network and Service Management (2020, 2021)
o IEEE Transactions on Signal and Information Processing over Networks (2020)
o IEEE Access Journal (2019)
o IEEE SENSORS Conference (2019)
o IEEE Internet of Things Journal (IoTJ) (2018 - 2022)

External reviewer:...
o JEDM, SPLICE workshop, ACM ICER, EDM, AIED (2020 - 2022)
o IEEE INFOCOM, IEEE FWC (2018 - 2019)

Teaching Experience

As Instructors..
North Carolina State University
o AI Academy Data Mining   *Fall 2021*
o CSC 116 Intro to Programming – Java   *Summer 2020, Summer 2021*

As Teaching Assistants..
North Carolina State University
o CSC 116 Intro to Programming – Java   *Summer 2020, Summer 2021*
o CSC 422/522 Automated Learning and Data Analysis   *Spring 2020*

o CSC 510 Software Engineering *Fall 2019*
University of Georgia
o CSCI 3360 Data Science I *Spring 2018, Spring 2017*
o CSCI/PHIL 4550/6550 Introduction to Artificial Intelligence *Fall 2016*