Sumon Biswas

Postdoctoral Researcher at Carnegie Mellon University
Institute for Software Research (ISR), School of Computer Science

TCS Hall, 4665 Forbes Avenue, Pittsburgh, PA 15213
+1 (515) 708-6166 • sumonb@cs.cmu.edu • sumonbis.github.io
sumonbis • sumonb • sumonbiswas

Computer science researcher with a goal to innovate the fundamentals and improve state-of-the-arts in the area of Software Engineering (SE), Programming Language (PL), and Artificial Intelligence (AI). Examining the algorithmic fairness and safety of machine learning software, and building SE/PL techniques for such high-assurance software systems.

Research Interest

Software Engineering	SE for AI, Reliability, Empirical Software Engineering, Big Code Mining
Programming Languages	Program Analysis, Verification, Program Synthesis, Program Evolution
Artificial Intelligence	Fairness of ML Software, Causal Analysis, DNN Verification

Academic Qualifications

Ph.D. in Computer Science — **Ames, IA**
Iowa State University — *2016–2022*
Advisor: Dr. Hridesh Rajan
Thesis: Reasoning and Verifying Fairness Properties of Machine Learning Pipeline

MS in Computer Science — **Ames, IA**
Iowa State University — *2016–2020*
Advisor: Dr. Hridesh Rajan
Thesis: Understanding Unfairness and its Mitigation in Open-Source Machine Learning Models

B.Sc. in Information Technology — **Dhaka, Bangladesh**
Jahangirnagar University — *2011–2015*
Advisor: Dr. Shamim Al Mamun
Thesis: Cloud Based Healthcare Application Architecture and Electronic Medical Record Mining

Employment

Carnegie Mellon University — **Pittsburgh, PA**
Postdoctoral Researcher at Institute for Software Research (ISR) — *May 2022 – Cont.*
Advisor: Dr. Eunsuk Kang

Iowa State University — **Ames, IA**
Research Assistant, Laboratory of Software Design — *May 2018 – May 2022*
Projects: Fairness in ML Pipeline, Dependable Data-Driven Discovery, Big-Code Mining & Analysis, Bug Detection & Repair

Iowa State University — **Ames, IA**
Teaching Assistant, Department of Computer Science — *August 2016 – May 2018*
Courses taught: · COMS 309 - Software Development Practices · COMS 327 - Advanced Programming Techniques in C & C++

Bangladesh University of Business and Technology — **Dhaka, Bangladesh**
Lecturer, Department of Computer Science & Engineering — *January 2016 – July 2016*

- **Courses instructed:** Computer and Programming Concepts · Structured Programming Language · Pattern Recognition

DataSoft Systems Bangladesh Limited — **Dhaka, Bangladesh**
Software Engineering Intern — *August 2013 - December 2013*

- **Responsibilities:** Develop Android Apps · Analyze mobile app market · Get training on coding conventions and SE practices i.e., agile, scrum, pair programming, etc. · Technologies include Java EE, Android SDK, Web APIs, etc.

Publications

Conference Paper

[1] David OBrien, **Sumon Biswas**, Sayem Mohammad Imtiaz, Rabe Abdalkareem, Emad Shihab and Hridesh Rajan. 23 Shades of Technical Debt: An Empirical Study on Machine Learning Software. In *Proceedings of the 30th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, **ESEC/FSE 2022**, page 1–13, 2022.

[2] **Sumon Biswas**, Mohammad Wardat and Hridesh Rajan. The Art and Practice of Data Science Pipelines: A Comprehensive Study of Data Science Pipelines In Theory, In-The-Small, and In-The-Large. *To appear in The 44th International Conference on Software Engineering*, **ICSE 2022**, page 1-13, 2022. [DOI] [Artifact] [Presentation]

[3] **Sumon Biswas** and Hridesh Rajan. Fair preprocessing: Towards understanding compositional fairness of data transformers in machine learning pipeline. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, **ESEC/FSE 2021**, page 981–993, 2021. [DOI] [Artifact] [Presentation]

[4] **Sumon Biswas** and Hridesh Rajan. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, **ESEC/FSE 2020**, page 642–653, 2020. [DOI] [Artifact] [Presentation]

[5] **Sumon Biswas**, Md Johirul Islam, Yijia Huang, and Hridesh Rajan. Boa meets python: A boa dataset of data science software in python language. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories*, **MSR 2019**, page 577–581, 2019. [DOI] [Slides]

[6] **Sumon Biswas**, M. S. Kaiser, and S. A. Mamun. Applying ant colony optimization in software testing to generate prioritized optimal path and test data. In *2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pages 1–6, 2015. [DOI]

[7] Manan Binth Taj Noor and **Sumon Biswas**. A secure data security infrastructure for small organization in cloud computing. In *2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pages 1–6, 2015. [DOI]

[8] **Sumon Biswas**, Anisuzzaman, Tanjina Akhter, M. S. Kaiser, and S. A. Mamun. Cloud based healthcare application architecture and electronic medical record mining: An integrated approach to improve healthcare system. In *2014 17th International Conference on Computer and Information Technology (ICCIT)*, pages 286–291, 2014. [DOI]

Under Review

[1] **Sumon Biswas** and Hridesh Rajan. Farify: Fairness verification of deep neural network, *Submitted*, pages 1–14, 2022.

[2] Rangeet Pan, **Sumon Biswas**, Vu Le, Sumit Gulwani and Hridesh Rajan. Can Program Synthesis Suggest Fixes for Deep Learning Crash Bugs? *Submitted*, pages 1-11, 2022.

[3] Sayem Mohammad Imtiaz, **Sumon Biswas**, Shibbir Ahmed and Hridesh Rajan. How Do Deep Learning Models Evolve? A Comprehensive Study of the Evolution of Deep Learning Model Code in Open Source Repositories, *Submitted*, pages 1-12, 2022.

[4] Rangeet Pan, **Sumon Biswas**, Mohna Chakraborty, Breno Dantas Cruz and Hridesh Rajan. An Empirical Study on Bugs Found while Reusing Pre-trained Models in Natural Language Processing, *Submitted*, pages 1-12, 2022.

Honors & Awards

- **Research Excellence Award**: Awarded by Iowa State University recognizing outstanding research *2022*
- **CAPS Award**: Awarded by ACM SIGSOFT to attend in-person ICSE conference at Pittsburgh, USA *2022*
- **Publication Award**: Awarded by Computer Science Department at Iowa State University for publication in the top-tier venues in consecutive years *2022*
- **CAPS Award**: Awarded by ACM SIGSOFT to attend ESEC/FSE conference *2021*
- **Publication Award**: Awarded by Computer Science Department at Iowa State University for publication in the top-tier venues *2021*
- **Panelist @ ESEC/FSE'20**: Selected as a panelist in the session on Fairness at ESEC/FSE 2020 *2020*
- **PLMW Scholar @ PLDI'19**: Awarded by Programming Language Mentoring Workshop (PLMW) at PLDI *2019*
- **Professional Advancement Grants @ ISU**: Awarded by Graduate and Professional Student Senate at Iowa State University *2019*

- **ACM Travel Grant**: Awarded to attend IMA Workshop organized by Institute for Mathematics & its Applications (IMA), University of Minnesota *2019*
- **NST Fellowship**: Awarded by Ministry of Science & Technology for Research Excellence, Bangladesh *2015*
- **University Merit Scholarship**: Top 5% in Undergraduate Level, Jahangirnagar University, Dhaka, Bangladesh *2014*

Services

Organizing Committee

- **WiDS Ambassador**: Served as the Women in Data Science (WiDS) Ambassador to organize and promote the event at ISU as part of the annual WiDS Worldwide Conference organized by Stanford University. *2022*
- **SPLASH'21**: Served as the Accessibility Chair of the ACM SIGPLAN conference on Systems, Programming, Languages, and Applications (SPLASH 2021) with OOPSLA and REBASE *2021*
- **Web Chair**: Served as the web chair of the Midwest Big Data Summer School, organized by Iowa State University, May 17-20, 2021 *2021*
- **SPLASH'20**: Served as the Accessibility Chair of the ACM SIGPLAN conference on Systems, Programming, Languages, and Applications (SPLASH 2020) with OOPSLA, ECOOP and REBASE *2020*

Reviewer & Program Committee (PC) Member

- **ASE'22**: Serving in the Late Breaking Results track of ASE 2022 *2022*
- **ASE'22**: Serving in the Student Research Competition track of ASE 2022 *2022*
- **ESEC/FSE'22**: Serving in the Program Committee of Doctoral Symposium track of ESEC/FSE 2022 *2022*
- **Journal Reviewer**: Serving as the reviewer for IEEE Transactions on Software Engineering (TSE) *2021*
- **MSR'21**: Shadow PC member of the International Conference on Mining Software Repositories (MSR) *2021*
- **OOPSLA'21**: PC Member of Artifact Evaluation Committee (AEC) at OOPSLA *2021*

Research Experience

Research Projects

- **Fairness in ML Pipeline**: I studied algorithmic fairness of machine learning models in three main projects: 1) Fairness as a software engineering property, 2) Reasoning about compositional fairness, and 3) Fairness verification. First, I conducted an empirical study to understand different SE aspects of fairness e.g., metrics, trade-offs, mitigation techniques, and their impacts in real-world ML models collected from Kaggle (ESEC/FSE'20). Second, I proposed causal reasoning in ML pipeline to measure component-level fairness, which helped to identify unfair stages in the pipeline and instrument them to mitigate bias (ESEC/FSE'21). Third, I proposed individual fairness verification for ReLU based DNNs using sound neural pruning and heuristics (submitted). Furthermore, I worked on employing AutoML techniques such as Baysian multi-objective optimization in Auto-SkLearn to improve fairness and accuracy together (submitted). Currently, I am extending the compositional reasoning of fairness in ensemble learning methods such as functional or dependent ensemble.
- **Dependable Data Driven Discovery (D^4)**: D^4 is an interdisciplinary data science hub at Iowa State led by my Ph.D. advisor and funded by NSF-TRIPODS grant. I contributed to the (D^4) project from its inception, to grant writing, to conduct research on dependability of data science (DS) software. My recent work on *Art and practice of DS pipeline* analyzed DS pipelines in theory (literature), in-the-large (mature DS projects in GitHub), and in-the-small (standalone notebooks in Kaggle). I identified the typical stages of the pipelines, how they are connected, and their differences (to be appeared in ICSE'22).
- **Big-Code Mining**: I built infrastructure within Boa framework to mine millions of ML programs and Jupyter Notebooks. A curated dataset containing the project metadata and all the versions of source files (commits) was published in MSR'19. I worked and collaborated on two different directions using the mined data: 1) Understanding evolution of deep learning models i.e., analyzing several changes made such as tuning, choice of API, refactoring, check-pointing, versioning, etc., 2) Studying ML technical debts such as ML knowledge debt, model interpretability debt, dependency debt, etc.
- **Bug Detection and Repair in DS software**: We leveraged program synthesis to fix deep learning functional bugs reported by underlying library, e.g., Keras, TensorFlow. Our key insights are: 1) fixes are repetitive, 2) similar error messages lead to similar fixes. We designed a DSL using PROSE framework, and proposed an online synthesis algorithm that learns and update rules from one or more example fixes. In another project, we conducted an empirical study on the new kinds of bugs found in pre-trained NLP models such as BERT, CTRL, GPT-2. We mined 9,420 GitHub issues from 10 most popular models, created a benchmark of 865 bugs, and built a taxonomy of bug types, their root causes, and impacts using open-coding scheme.

Research Mentoring

May 2020 – May 2022 **David OBrien, Undergraduate student, University of Northern Iowa**
Mentored him as an undergraduate REU student on mining software repositories using Boa, and analyzing machine learning pipelines. His research progress as an undergrad helped him to get accepted at Iowa State University as a graduate student. I mentored him further as a Ph.D. student on *analyzing technical debts in machine learning programs*, which was accepted in ESEC/FSE 2022.

January 2021 – May 2021 **Mohna Chakraborty, Ph.D. student, Iowa State University**
Mentored her on software engineering research, and analyzing bugs from StackOverflow posts. She published an student research competition short-paper under the mentorship in ESEC/FSE 2021 on the *bugs found in pretrained NLP models*. [DOI]

May 2020 – May 2022 **Usman Goher, Ph.D. student, Iowa State University**
Mentored him on empirical software engineering research, collecting benchmark, designing experiments, and answering research questions related to *SE for AI*.

August 2019 – May 2020 **Senior Design Project, Electrical and Computer Engineering, Iowa State University**
Mentored a five-member senior design team for their two semester long project – "*Analysis of GitLab projects using Boa*". Team members are: Diego Realpe (team leader), Adrian Hamill, Benjamin Carland, Megan Miller and Yi-Hsien Tan. The goal of the project was to mine source code and metadata from GitLab using the Boa infrastructure. [Project website]

Research Grants

2019 – 2022 **NSF TRIPODS:** Assisted to write grant proposal and contributed to the goal of the project – HDR TRIPODS: D4 (Dependable Data-Driven Discovery) Institute.

2015 – 2021 **Boa:** I contributed to the development of Python language support for the project grant – Boa: Enhancing Infrastructure for Studying Software and its Evolution at a Large Scale.

2015 – 2021 **Facebook Probability and Programming Award:** I contributed to writing the grant proposal and thereafter my dissertation was partially supported by the grant.

Research Highlights

2022 **Data Science Pipeline:** My research explained how data science pipelines are designed in theory and practice.

2021 **Fair ML:** My research identified prevalence of unfairness in machine learning models.

2020 **Dependability:** My association and contribution towards the NSF TRIPODS grant has been highlighted.

Teaching Experience

Teaching Assistant

COMS 327 - Advanced Programming Techniques in C & C++ Class size: ∼230
Undergraduate level, Iowa State University *Fall 2016, Spring 2017, Fall 2019*

Topics Differences between managed (Java) and unmanaged languages (C/C++) · Design and build large programs from specification · Memory management in C and C++ · Templates and standard library · Concurrent and network programming

Responsibilities Assist students with large programming projects and assignments · Debugging and pair programming · Prepare and grade tests · Teach usage of the tools and technologies

Technologies GDB, Valgrind, Ncurses, Build systems e.g., Make

COMS 309 - Software Development Practices Class size: ∼250
Undergraduate level, Iowa State University *Fall 2017, Spring 2018, Fall 2018*

Topics Develop complex software in a team: from idea to release · Software development criteria: client-server architecture, relational database, multi-user setting, concurrent features e.g., online chat · Using SE tools, IDE, source-control e.g., Git · SE lifecycle

Responsibilities Weekly lecture (1 hour) in a class section of ∼30 students on design pattern, version-control, server configuration, database design, etc. · Prepare and record screencasts · Supervise 4-member teams developing software projects throughout the semester

Screencasts: Create short videos for the tools and technologies used in the SE lifecycle · Example videos: develop first Android app, how to host project to Linux server, Using MySQL Workbench for to remote database server, etc.

Project supervision: Supervised 8 out of ∼55 teams in each semester · Team under my supervision won best project award in Spring 2017 and Fall 2018 · Weakly meeting: explain software requirements and deliverables, solve problems · Technologies: Android SDK, Spring Boot, .Net, RESTful API, Angular JS, MySQL, Java Socket, MVC, etc.

Project evaluation: Evaluate demonstrations · UI design · Code review: functionality, quality, bugs, etc. · Teamwork · Weakly project reporting

Supervised projects:
- **Fall'18:** Business QR, CookBuddy, ISU Service, Twenty One, Dog Matcher, Project X (best project award), Movie App, Campus Connections
- **Spring'18:** HabiTracker, Smart Art, Image Guesser, CyBike, Next Generation 911 (best project award), CyChat, Time Flies
- **Fall'17:** Battle of the Worlds, Pre Park, Run Samurai, CyDisc, Songusoid, Cute and Fluffy

Lecturer

Primary Instructor — Class size: ∼50
Bangladesh University of Business and Technology — *January 2016 - July 2016*

Responsibilities: Primary instructor · Design courses · Create tests · Conduct programming lab sessions

Courses instructed:
- **Computer and Programming Concepts:** Instruct two sections: computer science major and non computer science major
- **Structured Programming Language:** Instruct theory and lab sessions
- **Pattern Recognition:** Instructor for the seniors in computer science major

Selected Academic & Skill Based Projects

Uncertainty in DNN Hyperparameter Optimization — Python
Project for 'Advanced Topics in Programming Languages' Course | 3-Member Team | — *Aug. 2019 – Dec. 2019*
In this project, we provided aid to the deep learning programmers to quantify uncertainty in the model hyperparameters and help them to make informed decision while initializing hyperparameters. We have leveraged a first order type Uncertain<T> to represent the uncertainty in the random hyperparameters and choose the best value by performing statistical tests.

New Semantics using Lambda Calculus — Coq proof assistant
Project for Programming Languages Course | — *Mar. 2019 – May 2019*
I extended Lambda calculus using Coq to implement a core language that includes two Python features i.e., compound comparison statement and generator function. I described the syntax, operational semantics and type system of the features.

Near Duplicate Detection Using Simhash — Java
Course Project for Probabilistic Methods in Computer Engineering Course | — *Aug. 2017 – Dec. 2017*
I extended Lambda calculus using Coq to implement a core language that includes two Python features i.e., compound comparison statement and generator function. I described the syntax, operational semantics and type system of the features.

Performance Benchmarking for Link Prediction Algorithms in Social Networks — Java, MySQL, Neo4J
Research Project for Database Systems Course | 3-Member Team | — *Aug. 2017 – Dec. 2017*
From a given snapshot of a social network database, we predicted whether a person can be potentially connected to another person, by analyzing existing links. We used two datasets (Facebook dataset from Stanford Large Network Dataset Collection and bibliography dataset from DBLP) and import that into MySQL, and Neo4J (Graph based DB).

Tank Battle (Android Shooting Game) — Java, Android SDK, AndEngine
Semester Project for Game Development Course | — *Sep. 2014 - Dec. 2014*
Tank Battle is an Android game where you have to shoot at the enemy tanks at a high speed from different angles to destroy them. Enemy tanks comes from the opposite direction and try to hit the gamer's tank. If any enemy tank passes you are defeated and the game is over.

Talks

1. **ICSE'22**: Presented paper both virtually and in-person at 44th ACM/IEEE ICSE conference, 2021 — *Pittsburgh, PA*
2. **Invited Talk**: Presented my research in the CREATE SE4AI group participated by Concordia University, Polytechnique Montreal, Queen's University, and University of Alberta, February 2022 — *Virtual*

3. **NSF PI Meeting**: Presented my recent work on *Fairness Verification of DNN* in the TRIPODS monthly PI meeting chaired by Lenore J. Cowen from Tufts University, December 2021 — *Virtual*
4. **ESEC/FSE'21**: Paper presentation in the virtual conference of 29th ACM ESEC/FSE, 2021 — *Athens, Greece*
5. **ESEC/FSE'20**: Paper presentation in the virtual conference of 28th ACM ESEC/FSE, 2020 — *Sacramento, CA*
6. **TADS Presentation**: Presented my recent research on "*Fairness of Machine Learning Models*" in front of Theoretical and Applied Data Science (TADS) Group of D4 Institute, Iowa State University, 2020 — *Ames, IA*
7. **MSR'19 Presentation**: Data showcase and paper presentation in Mining Software Repository Conference (MSR), 2019 — *Montreal, Canada*
8. **Panelist @ ESEC/FSE'20**: Selected as a panelist in the session on Fairness at ESEC/FSE 2020 — *Sacramento, CA*

Professional Activities

- **PLDI'19 and PLDI'20**: Attended Programming Language Design and Implementation Conference PLDI 2019 at Phoenix, Arizona and PLDI 2020 (virtual) — *Phoenix, Arizona*
- **ICSE'19-22**: Attended International Conference on Software Engineering ICSE 2019 at Montreal, Canada, ICSE 2020-21 (virtual) and ICSE 2022 at Pittsburgh, PA — *Montreal, Canada*
- **Summer School**: Completed professional development program in Midwest Big Data Summer School — *Ames, Iowa*
- **IMA Workshop**: Attended 4-day workshop on "*Recent Themes in Resource Tradeoffs: Privacy, Fairness and Robustness*" organized by Institute for Mathematics & its Applications (IMA), University of Minnesota, 2019 — *Minneapolis, MN*
- **Turing Lecture**: Attended Turing Lecture at FCRC 2019 by Geoffrey Hinton and Yann LeCun on "*The Deep Learning Revolution*" — *Phoenix, AZ*
- **Finalist**: National Hackathon 2014 organized by ICT Ministry, Bangladesh — *Dhaka, Bangladesh*
- **Finalist**: Google Developer Group (GDG) DevFest Hackathon 2013 organized by GDG Bangladesh — *Dhaka, Bangladesh*

Affiliation

- **ACM**: Student member of Association for Computing Machinery (ACM) — *2019–22*
- **SIGSOFT & SIGPLAN**: Member of the ACM SIGSOFT and ACM SIGPLAN — *2020–22*
- **IEEE Student Branch**: Served as the vice chair of IEEE Student Branch, Jahangirnagar University — *2015–16*