Mahmoud Jahanshahi

https://mahmoudjahanshahi.com

Profile

Experienced Research/Data Scientist with a demonstrated track record in Empirical Software Engineering research. Adept at applying Machine Learning (ML), Deep Learning, and Natural Language Processing (NLP) to complex, real-world problems. Strong foundation in Statistics and Operations Research (OR/Optimization) to uncover insights and guide data-driven decisions. Recognized for key competencies including:

- Excel at rapidly learning and applying new tools, concepts, and techniques.
- Tackle complex, intellectually demanding problems with analytical depth and minimal guidance.
- Thrive in cross-functional teams and clearly communicate complex ideas across technical and non-technical audiences.
- Take initiative in designing and leading original research projects, from problem formulation to publication.

Education

Ph.D. | Computer Science | May 2021 - May 2025

University of Tennessee, Knoxville, USA

- Dissertation: Copy-Based Reuse and its Implications in Open Source Software Supply Chains.
- Advisor: Dr. Audris Mockus
- GPA: 4.00/4

Master of Science | Industrial Engineering | Sep 2011 - Sep 2013

Sharif University of Technology, Tehran, Iran

- Thesis: The Influence of Information Presentation and Risk Attitude on Asset Allocation in Financial Markets.
- Advisor: Dr. S. T. Akhavan Niaki

Bachelor of Science | Industrial Engineering | Jan 2007 - Jul 2011

Mazandaran Institute of Technology, Babol, Iran

- Ranked 6th nationwide in the Industrial Engineering Graduate Admissions Exam.

Awards

- LLM4Code Best Paper Award for “Cracks in The Stack: Hidden Vulnerabilities and Licensing Risks in LLM Pre-Training Datasets” at the International Workshop on Large Language Models for Code (LLM4Code). 2025.
- ACM SIGSOFT Distinguished Paper Award for “Understanding the Response to Open-Source Dependency Abandonment in the npm Ecosystem” at the International Conference on Software Engineering (ICSE). 2025.

Certificates

Deep Learning Specialization by DeepLearning.AI on Coursera | **Dec 2024**

1. Structuring Machine Learning Projects, 2. Neural Networks and Deep Learning, 3. Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization, 4. Convolutional Neural Networks, 5. Sequence Models

Skills

- **Programming Languages |** Bash (Scripting), R, Python, C
- **AI/ML |** TensorFlow, Keras, PyTorch, Scikit-learn, NumPy, Pandas, NLTK
- **Generative AI |** Large Language Models (LLMs), Transformer Architectures, Text Generation, GPT
- **Data Analysis |** SQL, MongoDB, Tableau, Power BI, MATLAB, Matplotlib, Seaborn, ggplot2
- **Big Data |** Hadoop, Apache Spark
- **DevOps & Productivity |** AWS, Azure, GCP, Git, Docker, Jira, MS Project, COMFAR
- **Languages |** English (Fluent), German (Working Knowledge), Persian (Native)

Professional Experience

Graduate Research Assistant | University of Tennessee Knoxville | May 2021 - May 2025

- Built large-scale OSS analysis pipelines (Bash, Python, R) using ML and NLP (e.g., Winnowing) to detect 24B+ copy instances across 1B copied files from 130M+ projects. Enabled the first ecosystem-scale measurement of copy-based reuse, impacting 80% of OSS, with implications for security, licensing, and maintenance.

- Developed an automated LLM dataset curation pipeline that flagged 19,944 vulnerable code blobs linked to 6,947 distinct CVEs in "The Stack". Found 17% of files with newer, fixed versions (patched or bug-free), improving LLM pre-training data safety, compliance, and maintainability.
- Contributed core functionality to the World of Code infrastructure, enabling blob-level provenance tracking and metadata analysis across 16B files and 108M de-forked OSS projects. Produced key research outputs including a copy-based reuse dataset, project-to-license map, and an interactive author/project collaboration network tool.

Senior Data Scientist | Mobile Communications Company of Iran | May 2019 - Apr 2020

- At the ERP department, data warehousing & business intelligence office, bridged the gap between business units and data warehouse teams by acting as a liaison, translating business needs into technical requirements for the design, testing, and development of data warehouse solutions.
- Developed and maintained reports, dashboards, and analyses using Oracle Business Intelligence Enterprise (OBIEE) and SQL, leveraging ETL pipelines to extract, transform and analyze enterprise-scale datasets for decision support.

Strategic Investments Lead | Mobile Communications Company of Iran | Feb 2018 - May 2019

- At the mergers and acquisitions department, feasibility studies office, led a team of five professionals in managing complex investment projects, including business plan development, feasibility analysis, valuation of acquisition targets and due diligence.
- Collaborated with C-level management and the board of directors to negotiate contract terms and cooperation models with potential counterparts, such as strategic partners or business sellers.

International Investment Analyst | Mobile Communications Company of Iran | Feb 2016 - Feb 2018

- At the mergers and acquisitions department, international investments office, developed financial models to project earnings and assess the viability of acquisition opportunities in international markets.
- Screened markets for potential investment targets, interpreting financial statements and other data to evaluate risk and profitability.

Publications

- **Jahanshahi M.**, Mockus A. "Cracks in The Stack: Hidden Vulnerabilities and Licensing Risks in LLM Pre-Training Datasets". *Accepted in the Second International Workshop on Large Language Models for Code (LLM4Code).* 2025. **Won the *LLM4Code Best Paper Award*.**
- **Jahanshahi, M.**, Reid, D., & Mockus, A. "Beyond Dependencies: The Role of Copy-Based Reuse in Open Source Software Development". *Accepted in ACM Transactions on Software Engineering and Methodology (TOSEM).* 2025.
- **Jahanshahi, M.**, Reid, D., McDaniel, A., & Mockus, A. "OSS License Identification at Scale: A Comprehensive Dataset Using World of Code". *Accepted in 2025 IEEE/ACM 22st International Conference on Mining Software Repositories (MSR).* IEEE, 2025.
- Miller, C., **Jahanshahi, M.**, Mockus, A., Vasilescu, B., & Kästner, C. "Understanding the Response to Open-Source Dependency Abandonment in the npm Ecosystem". *Accepted in the 47th International Conference on Software Engineering (ICSE).* 2025. **Won the *ACM SIGSOFT Distinguished Paper Award*.**
- Thakur, A., Milewicz, R., **Jahanshahi, M.**, Paganini, L., Vasilescu, B., & Mockus, A. "Scientific Open-Source Software Is Less Likely To Become Abandoned Than One Might Think! Lessons from Curating a Catalog of Maintained Scientific Software". *Accepted in The ACM International Conference on the Foundations of Software Engineering (FSE).* 2025.
- **Jahanshahi, M.** & Mockus, A. "Dataset: Copy-based Reuse in Open Source Software". *2024 IEEE/ACM 21st International Conference on Mining Software Repositories (MSR)* (pp. 42-47). IEEE, 2024.
- Reid, D., **Jahanshahi, M.**, & Mockus, A. "The extent of orphan vulnerabilities from code reuse in open source software". *Proceedings of the 44th International Conference on Software Engineering (ICSE)* (pp. 2104-2115). 2022. **Nominated for the *ACM SIGSOFT Distinguished Paper Award*.**
- Lyulina, E., & **Jahanshahi, M.** "Building the collaboration graph of open-source software ecosystem". *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)* (pp. 618-620). IEEE, 2021.