SAURABH PUJAR

+1-201-469-1351

Scarsdale, NY saurabh.s.pujar@gmail.com Google Scholar Profile LinkedIn Profile

EDUCATION & CERTIFICATIONS

New York University, Courant Institute of Mathematical Sciences ***Sep 2014 – May 2016***

Master of Science in Computer Science GPA: **3.53**

Relevant Coursework: *Web Search Engines, Machine Learning & Computational Statistics, Foundations of Machine Learning, Real-time & Big Data Analytics, Statistical Natural Language Processing, Analysis of Algorithms, Operating Systems, Programming Languages.*

University of Mumbai ***Aug 2006 – June 2010***

Bachelor of Engineering (Information Technology) GPA: **3.45**

Relevant Coursework: *Automata Theory, Software Engineering, Computer Organization and Architecture, Advanced Database Systems, Data Warehousing and Mining, Project Management, Computer Simulation and Modeling, Data Structures and Algorithms, Management Information Systems, Robotics.*

Coursera: ML In Production ***Jan 2023 – Feb 2023***

Relevant Certifications: *Introduction to ML in Production*

Coursera: TensorFlow in Practice ***Jun 2019 – Aug 2019***

Relevant Certifications: *Introduction to TensorFlow for Artificial Intelligence, Machine Learning and Deep Learning; Convolutional Neural Networks in TensorFlow; Natural Language Processing in TensorFlow; Sequences, Time Series and Prediction.*

Coursera: Cloud Computing ***Jan 2018 – Mar 2018***

Relevant Certifications: *Cloud Computing Concepts, Cloud Systems and Infrastructure, Big Data Applications in Cloud, Cloud Networking.*

Coursera: Deep Learning ***Sep 2017 – Jan 2018***

Relevant Certifications: *Neural Networks and Deep Learning, Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization, Structuring Machine Learning Projects, Convolutional Neural Networks.*

AI WORK EXPERIENCE 7 Years

Senior Research Software Engineer, IBM Research, T. J. Watson Research Center, New York ***June 2016 onwards***

Artificial Intelligence (AI) for Natural Language to Code Generation - Product:

- Developed an English language to source code (Ansible YAML) generation system using AI (DL/ML).
- Developed a demo for this work as part of Project Wisdom, which was showcased at AnsibleFest 2022.
- Technical lead and ML Engineer for the development of the Wisdom model and the development, improvement and release of the Watson Code Assistant for Ansible model.
- Work was published as an invited paper at DAC 2023. A longer version of the paper is available on arXiv.
- Wisdom model was further improved and released as part of Ansible Lightspeed and Watson Code Assistant for Ansible technical preview in June 2023.
- The technical preview was very well received with high acceptance rate by thousands of users. We released an analysis of the user data and feedback on arxiv, which will be published in *ASE Industry Showcase 2024.*
- ***Domain:*** *Big Code, Deep Learning, Generative Models, AI for Code, Software Engineering*
- ***Skills:*** *Python, PyTorch, Multi-GPU, Transformers, Hugging- face, Hyperparameter Tuning, Neural Networks, Finetuning, Pre-training*

AI for Code LLMs - Research:

- In collaboration with Columbia University, we developed a new metric to measure LLM model consistency called IdentityChain. This work was published in ICLR 2024.
- In collaboration with UIUC, we analyzed transferability of learning between programming languages. Manuscript has been submitted to a conference for review and publication in 2024.
- In collaboration with Boston University, we analyzed LLM capability to reason about security vulnerabilities. This work was published in IEEE S&P 2024.
- ***Domain:*** *AI, NLP, Software Engineering, AI for Code* ***Skills:*** *Python, PyTorch, Transformers, HuggingFace*

BERT Model for Code - Research

- Helped the team develop a BERT model for C source-code called C-BERT. Work was published on arXiv and has many citations.
- Trained C-BERT to make it the top Defect Detection model on the CodeXGLUE leaderboard upon release.
- Trained the C-BERT model to perform well on multiple tasks in different programming languages. This work was published as part of Project CodeNet in proceedings of NEURIPS 2021
- Collaborated with ARiSE Lab of Columbia University to help develop a BERT model for code using structural and functional features. This work, known as DISCO was published in proceedings of ACL 2022.
- DISCO work was extended to develop CONCORD, which is clone aware contrastive learning for source code. The work was published in ISSTA 2023 and received the **Distinguished Paper Award**.
- ***Domain:*** *AI, NLP, Software Engineering, AI for Code* ***Skills:*** *Python, PyTorch, Abstract Syntax Tree, BERT, Transformers*

Varangian Augmented Static Analyzer - Prototype:

- Developed a prototype for Augmented Static Analysis using Machine Learning.
- Worked as Tech lead and ML engineer for the prototype.
- Protype was a Git-Bot which automatically scanned project code with Infer static analyzer and used ML to prioritize Infer output.
- I was responsible for ML pipeline implementation, designing the system, deployment and user feedback evaluation.
- Our work was published in the proceedings of the conference MSR 2022.
- ***Domain:*** *Big Code, Deep Learning, Machine Learning, Cloud Security* ***Skills:*** *GitHub, User feedback evaluation, UX Design*

Artificial Intelligence (AI) for Vulnerability Analysis (VA) - Research:

- Tech-lead and ML Engineer for project that used AI (DL/ML) to detect vulnerabilities in cloud source code.
- Was completely responsible for end-to-end ML pipeline implementation.
- Helped the team develop the D2A dataset for training the model.
- Work was published in proceedings of ICSE 2021.
- Led the development of a leaderboard that showcases performance of different models on our dataset.
- Our work on D2A Leaderboard and using BERT model for this problem was published in **Empirical Software Engineering** journal.
- ***Domain:*** *Big Code, Deep Learning, AI, AI for Code* ***Skills:*** *Python, PyTorch, Multi-GPU, Transformers, Hugging face, Hyperparameter Tuning*

Question Answering using AI - Research:

- Worked with Question Answering team to build a QA system for technical questions called TechQA. I wrote the training code and created the SQUAD baseline models by training with PyTorch transformers on GPUs.
- TechQA Dataset was published in the proceedings of ACL 2020.
- Worked with the QA team to develop a PyTorch transformer-based system for participating in the SuperGLUE challenge. I was the primary engineer for the MultiRC sub-challenge and was responsible for the design, training, evaluation and experiments.
- IBM was third in the ranking at the time of submission on the SuperGLUE leaderboard.
- ***Domain:*** *Question Answering, Deep Learning, NLP* ***Skills:*** *Python, PyTorch, GPUs, Transformers, BERT, Hugging face, Hyperparameter Tuning*

Electronic Medical Record Analysis - Product:

- Wrote a custom Random Forest 2 class classifier that classified features from Weka arff files, which *improved the precision of a Random Forest based classifier by 2%* and was twice as fast. Reengineered a J48 based classifier with custom Random Forest classifier while maintaining performance of downstream apps.
- Replaced an ML pipeline consisting of Random Forest, Conditional Random Fields (CRF), and Support Vector Machine (SVM) classifiers with a single CRF classifier. The new classifier was 7 times faster and increased F1 score by 20 points.
- Czar for 3 of the 5 sprints that took the application to production. Tasks involved integration run of the different ML pipelines, validating the accuracy of models with regression analysis, launching and load testing production services.
- Split the CRF classifier into a CRF + SVM classifier, based on changes in data attributes. Led to 2-point improvement of 10 of the 12 downstream classifier. Involved in continuous error analysis and improvement of these classifiers.
- ***Domain:*** *Medical, Machine Learning, NLP, Software Engineering, DevOps* ***Skills:*** *Java, Python, UIMA, Apache DUCC, Tokenization*

Data Science Intern, **CY Data Science, New York, New York** ***Jan 2016 – May 2016***

- Analyzed Twitter data to predict economic trends using python libraries like NLTK, Spacy and Gensim.
- Used emojis to label tweets and used this as training data. Trained a classifier to classify tweets with no emoji.
- Compared change in sentiment trend with the change in consumer confidence and actual sales of cars, employment figures.
- Obtained better correlation to employment trends (0.69), compared to survey-based techniques by US Bureau of Labor Statistics and The Conference board (0.46). The *software was successfully sold to a New York based financial data company*.
- ***Domain:*** *Nowcasting, Economics, Machine Learning, NLP.* ***Skills:*** *Python, SpaCy, NLTK, Gensim, scikit-learn, Word Vectors*

PATENTS

Contextual embeddings for improving static analyzer output

- Pujar, Saurabh, Luca Buratti, Alessandro Morari, Jim Alain Laredo, Mihaela Ancuta Bornea, Jeffrey Scott McCarley, and Yunhui Zheng. "Contextual embeddings for improving static analyzer output." U.S. Patent 11,765,193, issued September 19, 2023.

Building pre-trained contextual embeddings for programming languages using specialized vocabulary

- Pujar, Saurabh, Luca Buratti, Alessandro Morari, Jim Alain Laredo, Alfio Massimiliano Gliozzo, and Gaetano Rossiello. "Building pre-trained contextual embeddings for programming languages using specialized vocabulary." U.S. Patent 11,429,352, issued August 30, 2022.

System and method to share and utilize healthcare data

- Malvankar, Abhishek, Saurabh Pujar, Edward A. Epstein, Louis Degenaro, and Burn Lewis. "System and method to share and utilize healthcare data." U.S. Patent 11,250,937, issued February 15, 2022.

SOFTWARE ENGINEERING WORK EXPERIENCE 4 Years

Technology Summer Analyst, **Morgan Stanley, New York, New York** ***Jun 2015 – Aug 2015***

- Application I worked on was designed to capture, store, and analyze range of data about corporate actions.
- The web services in the application implemented a hybrid REST-SOAP API.
- Completed a proof of concept to turn the web services into a purely RESTful service.
- Developed a framework to implement all the CRUD operations. Only retrieve was initially supported.
- ***Domain:*** *Finance, Investment Banking, Web Application Development, Software Engineering* ***Skills:*** *Java, REST, SOAP, Spring, Ajax, Ext-JS*

Associate Technical Analyst, **Oracle Financial Services Software Ltd., Mumbai, India** ***Jun 2012 – May 2014***

- Worked on maintenance and enhancement of internet and mobile banking applications covering corporate and retail banking.
- Developed batch files to partially automate the testing of certain enhancements reducing testing time from 2 days to half a day.
- Won “We Applaud” award in Aug 2013 at Oracle for developing tools which cut down testing time by 75%.
- ***Domain:*** *Finance, Web App Dev, Mobile App Dev, Software Engineering* ***Skills:*** *Java, Servlets, Ajax, Oracle DB, Web security Automation Testing*

Systems Engineer, **Infosys Ltd., Pune, Maharashtra, India** **Jan 2011 – May 2012**

- Development of Asset and Vendor Management applications including Portfolio management, Invoice generation and delivery for Union Bank of Switzerland.
- ***Domain:*** *Finance, Banking, Asset Management, Web App Dev, Software Engineering* ***Skills:*** *Java, Struts, JSP, Oracle DB, MVC Architecture, SQL*