Arshitha Basavaraj

Email: arshithab@proton.me — **LinkedIn:** linkedin.com/in/arshitha — **GitHub:** github.com/Arshitha

Skills

- **Programming Languages:** Python, Bash, SQL, C/C++, Java
- **Cloud & Infrastructure:** High Performance Computing (HPC), AWS, Apptainer, Docker
- **Tools & Technologies:** Git, GitHub, DataLad, HuggingFace, PyTorch
- **Generative AI techniques:** Prompt engineering, Retrieval-Augmented Generation (RAG)

Experience

International Institute of Information Technology, Bangalore (IIIT-B) **Jan 2024 – Present**
Data Engineer/Scientist *Bangalore, IND*

- Leading data science efforts for the Indian Council of Medical Research's (ICMR) PARAM project—a multi-site, multi-modal, longitudinal developmental cohort study on resilience and mental health.
- Designed and implemented a pipeline to ingest, transform, and store both tabular and imaging data on a centralized server for downstream analysis.
- Building a reporting pipeline to automate questionnaire scoring, data de-identification, site-wise participant reports, and QC insights.
- Collaborating closely with clinicians and non-technical stakeholders to align technical solutions with research objectives.

National Institute of Mental Health, National Institutes of Health (NIMH/NIH) **Sep 2019 – Sep 2024**
Data Engineer *Bethesda, USA*

- Transformed over 15,000 participants' datasets to the widely adopted Brain Imaging Data Structure (BIDS) specification, from study-specific formats.
- Implemented data warehousing pipelines for multi-modal (imaging, tabular and genomic) health data for over 15 studies within NIMH.
- Collaborated with cross-functional teams and clinicians to streamline data cleaning, processing and curation.
- Developed git/GitHub training curriculum and conducted workshops for senior researchers to improve adoption of FAIR data principles by research groups within the NIMH.

BIDS contributions

- Active open-source contributor to the BIDS standard. Two key contributions: 1. Extension proposal for tabular data curation guidelines, 2. An automated PDF document generator of the specification.

Anatomical Scans Defacer

- Automated the process of de-identifying structural MRI scans by removing facial features.
- Tested existing defacing programs on two neuroimaging datasets containing over **2000** scans.
- Integrated visual inspection and rating tool with an existing de-identification program to flag and correct failures efficiently.
- *Significance:* De-identification of MRI scans is a crucial and high-effort final step before datasets can be shared openly. Automating the process of defacing scans, visual QC, and failure correction has reduced the timeline for data sharing by weeks.

dataxu (acquired by Roku) **May 2018 – Aug 2018**
Engineering Intern *Boston, USA*

- Ported the legacy CRON-based data transfer program to an event-triggered, cleaner, and more efficient process.
- Boosted data transfer efficiency by **95%**, while also improving scalability.
- Developed a proof-of-concept solution to integrate Sphinx and automate documentation generation for all the Python scripts within a repo with minimal developer input.

Indian Statistical Institute **May 2015 – Aug 2015**

Research Assistant *Bangalore, IND*

- Recognized and classified six facial expressions from frontal face images using 2-layer feed-forward neural network.
- Improved classification efficiency by **2%** overall with respect to existing literature.
- Discovered a simple, novel method using a combination of Mathematical Morphological and Image Processing techniques for feature extraction.

Education

Boston University **Jul 2017 – May 2019**

Master of Science in Electrical and Computer Engineering *Boston, USA*

National Institute of Technology, Karnataka (NITK) **Jul 2012 – May 2016**

Bachelor of Technology in Electrical and Electronics Engineering *Surathkal, IND*

Publications

- Taylor Paul A.,..., **Basavaraj Arshitha**, et al.,Editorial: Demonstrating quality control (QC) procedures in fMRI, Frontiers in Neuroscience, Sec. Brain Imaging Methods, Volume 17, 31 May 2023, doi: 10.3389/fnins.2023.1205928
- Allison C. Nugent,..., **Arshitha Basavaraj**, et al., (2022). The NIMH intramural healthy volunteer dataset: A comprehensive MEG, MRI, and behavioral resource. Scientific Data, 9(1). doi: 10.1038/s41597-022-01623-9
- A. Apte, **A. Basavaraj**, et al., Efficient Facial Expression Recognition and classification system based on morphological processing of frontal face images, 2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS), 2015, pp. 366-371, doi:10.1109/ICIINFS.2015.7399039.
- More on **Google Scholar**

Projects

Plain-language Summarization of Radiology Reports using LLMs **Apr 2025 – May 2025**

- Developed a unified framework for the BioLaySumm 2025 shared task to translate complex biomedical articles and radiology reports into layperson-friendly summaries.
- Engineered solutions using advanced prompting techniques (zero-shot, few-shot, role-based) with SOTA Large Language Models, including Llama-3.3-70B-Instruct and GPT-4.1.
- Implemented a RAG pipeline to enrich summaries by integrating external biomedical knowledge from the Unified Medical Language System (UMLS).
- Achieved competitive results as part of the 5cNLP team, securing **2nd place** in Radiology Report Translation (Subtask 2.1) and **3rd place** in Summarization with External Knowledge (Subtask 1.2).

Language Usage Checker **Feb 2019 – May 2019**

- Developed a multi-threaded web crawler using BFS to collect text data from websites, optimizing crawling speed with 8 parallel threads and URL deduplication using a database-backed set.
- Implemented multilingual tokenization and parts-of-speech (POS) tagging using Stanford CoreNLP, supporting English, Chinese, Arabic, and German with custom tokenizers.
- Designed a NoSQL document and graph-based database, which stored words as vertices and bigrams as edges, with frequency-based metadata. This achieved **99.9%** compression efficiency for crawled data storage.
- Built a statistical usage checker that analyzes new text input by comparing bigram frequencies from crawled data.
- **GitHub:** Language Correction

Toxic Comments Classification **Oct 2018 – Dec 2018**

- Developed classical ML and deep learning models to classify toxic comments in the Jigsaw Toxic Comments dataset, a dataset of comments from Wikipedia's talk page edits.
- Evaluated model performance differences with word embedding techniques such as count vectorizers, tf-idf and word2vec.
- **GitHub:** Jigsaw Toxic Comment Classification