SHOIBOLINA KAUSHIK

GA, USA | +1 (510) 675-7443 | shoibolina.kaushik@gmail.com
https://www.linkedin.com/in/shoibolina-kaushik| https://github.com/shoibolina | https://orcid.org/0000-0003-0885-3836

SKILLS

- Programming languages- Python, SQL, R,C, C++, Java, Go, HTML, CSS, JavaScript
- Tools/Framework- Tensorflow, Transformers, Scikit-learn, NLTK, Spacy, BERT, d3.js, LLM, RAG, React.js, REST APIs, Flask, Django, Git, PostgreSQL, SQLite, GCP, AWS (EC2, S3), JIRA, WordPress

RESEARCH EXPERIENCE

Emory Center for Digital Scholarship (ECDS), Emory University, USA **Jan 2025 – May 2025**
Digital Scholarship Assistant

- Led the technical development of a 4-person, cross-functional team for an image segmentation pipeline to automate road extraction from 150+ years of historical maps of Atlanta, reducing manual tracing efforts using ArcGIS by 95%
- Fine-tuned NVIDIA SegFormer-B0 on "DHK 200 Turkey" map tiles (15-epoch AdamW + Dice-loss; validation Dice > 0.85) and warm-started that checkpoint on 1974 Atlanta maps for 200 more epochs, raising best Dice from 0.88 to 0.91
- Automated large-scale inference by stitching tile-level predictions into a dimension-matched GeoTIFF and hardened GPU workflows with proactive memory clears plus standardized mask polarity, eliminating out of memory (OOM) errors

Department of Biomedical Informatics, Emory School of Medicine, Emory University, USA **Jan 2024 – Jul 2024**
Graduate Research Assistant- Received Best Poster Award at Bio-STAR AI Symposium 2024

- Collaborated with Division of Physical Therapy in a 4-person team to develop a cost-effective, marker-less classification model for gait features to enable healthcare professionals to monitor mobility changes and detect early signs of gait abnormalities
- Generated biomarker keypoint datasets from over 300 mobile phone videos (~250 GB) using Python, and implemented ETL processes to prepare data for machine learning, achieving 91% classification accuracy on the best model
- Stored and managed the generated keypoint data on the department's high-performance computing cluster, enabling scalable access and streamlined processing for subsequent analysis workflows

PROJECTS

Global Burden of Diseases dashboard using D3.js **May 2025**

- Developed an interactive choropleth dashboard with D3.js to visualize global DALY rates by disease and year, leveraging IHME's GBD dataset for 200+ countries
- Consolidated 10+ CSV files into a unified dataset using Python for seamless integration with geospatial data layers
- Elevated user engagement by designing responsive UI components with dynamic tooltips, gradient legends, and filter controls, enabling real-time exploratory analysis across 30+ disease categories (git demo)

Upnext- Events and Venues management App **Feb 2025 – Apr 2025**

- Spearheaded a 5-person team to design a scalable Django REST backend + React frontend with PostgreSQL, implementing modular APIs for user authentication, event creation, venue listings, availability calendars, bookings, and comments
- Enforced secure JWT workflows and granular role-based access (organizer, venue owner, attendee), developed automated tests, contributed to CI/CD pipelines while driving collaborative development using GitHub (pull requests, code reviews) and Atlassian Jira Kanban boards for task division and progress tracking
- Integrated and optimized advanced platform features including real-time WebSocket messaging, QR-based ticketing, and live event dashboards for ticket scanning and attendance tracking, enhancing interactivity and operational control(git repo)

Biomedical Question Answering System **Oct 2024 – Dec 2024**

- Engineered a RAG style biomedical question-answering system leveraging PubMedBERT, fine-tuned on the BC5CDR dataset, achieving an F1 Score of 62% and an Exact Match score of 48% in extracting chemical-disease relationships
- Designed a TF-IDF-based retrieval mechanism to rank and retrieve relevant contexts from over 15,000 indexed biomedical passages, ensuring accurate and efficient information extraction for domain-specific queries
- Deployed the QA system using Gradio Spaces, enabling real-time interaction and query resolution with an average response time of under 2 seconds, facilitating accessible exploration for researchers and practitioners (demo link)

Synthetic iEEG Data Production for Epilepsy Analysis **Oct 2023 – Dec 2023**

- Engineered a Denoising Diffusion Probabilistic Model (DDPM) in a two-person team, generating 16-channel, 10-second synthetic iEEG samples at 400 Hz, achieving stable convergence and realistic signal replication
- Optimized data preprocessing pipelines for 3,224 interictal and 235 preictal EEG samples, enhancing classifier performance, with AdaBoost achieving an AUC of 0.73 and Naïve Bayes a macro F1-score of 0.54 when trained on real data
- Validated synthetic data fidelity through 5-fold cross-validation, hyperparameter tuning, and spectrogram analysis (0-40 Hz), ensuring signal characteristics aligned with real iEEG data while identifying gaps in high-frequency feature replication (git repo)

EDUCATION

Laney Graduate School, Emory University, GA, USA **May 2025**
Master of Science in Computer Science **(GPA 3.72/4.0)**

Manipal Institute of Technology, Manipal Academy of Higher Education (MAHE), India **May 2023**
Bachelor of Technology in Computer Science and Engineering *(minor spec in Big Data Analytics*) **(CGPA 9.04/10)**