KAUSHIK KUMAR

520-609-4957 Tucson, AZ, USA Portfolio kaushikkumar.208@gmail.com linkedin

EDUCATION

UNIVERSITY OF ARIZONA **Aug 2024 - Dec 2025**
Master of Science in Data Science, GPA: 4.0 Tucson, AZ

VELS INSTITUTE OF SCIENCE & TECHNOLOGY **Jun 2019 - May 2023**
Bachelor of Engineering Computer Science and Engineering, Chennai, TN

EXPERIENCE

GRADUATE RESEARCH ASSISTANT **Mar 2025 - Present**
UNIVERSITY OF ARIZONA (ACT LAB) & GRAD TUTOR (SALT) Tucson, AZ

- Led end-to-end development of ShieldNN Adaptive Margins, a novel RL system for autonomous vehicle safety combining dual-agent PPO architecture with Control Barrier Functions; implemented an adaptive margin agent with a 6-component reward design, variational autoencoder (VAE) state encoding, and ONNX integration for real-time safety filtering in the CARLA simulator.
- Built ML pipelines from MATLAB safety networks to RL training, with CUDA acceleration reaching 4× speedup (3 vs 20 days) validated via Monte Carlo rollouts and adversarial tests, sustaining 99% safety, 98% completion, and reproducible benchmarks.

SOFTWARE ENGINEER INTERN (GEN AI) **Jun 2025 - Aug 2025**
SELECTOR AI Santa Clara, CA

- Engineered ingestion pipelines (rsyslog, Promtail, Loki, Kafka, cloud networks) with real-time normalization and NER-based clustering, Pytest validations(50+ tests) ensuring reliable event detection across 1000K+ logs/day. Deployed containerized services to a private GCP registry and delivered to clients via Selector's S2AP platform and Grafana Dashboards, improving client MTTD.
- Migrated Logminer model from Random Forest to custom-built RAG -BM25S retrieval engine (Lucene/Numba) with optimized tokenization, IDF scaling, having 99.2% accuracy, inference throughput 6K logs/sec and reduced memory footprint of < 1GB.

DATA SCIENTIST **Jun 2023 - Jul 2024**
JOHNSON ELECTRIC Chennai, India

- Developed Kalman Filter–based predictive models with SQL-driven feature extraction to estimate mass outflow in Tesla AGP water pumps; deployed via Docker containers on VMs for integration with EOL testing systems, reducing test times by 72% (80W) and 50% (50W) with cost savings of $74K/month.
- Built Gradient Boosting + Random Forest models (93% accuracy, 83% faster cycles) and Power BI dashboards with anomaly detection (JSD/KL) & leakage modeling (SVR + skew-normal), reducing test time 160s→45s (0.93 corr., zero false positives).
- Architected enterprise AI initiatives including a RAG pipeline for PLM Teamcenter chatbot (Streamlit, Azure Blob Storage, Microsoft Graph API, Marqo AI, Solr, Azure OpenAI GPT) with PostgreSQL feedback logging, and pioneered synthetic defect generation using GANs, boosting anomaly detection accuracy on production lines; hosted solutions on Azure VMs/AVD.

RESEARCH DATA SCIENCE INTERN **Dec 2022 - Feb 2023**
NATIONAL UNIVERSITY OF SINGAPORE (Collaboration with HPE) Singapore

- Researched multilingual hate speech & bias detection in underrepresented languages using n-gram analysis, cross-lingual embeddings, and mBERT+GRU with attention; achieved $> 92\%$ F1, 15% fewer false positives, confirmed via ablation studies.
- Built an enterprise-ready moderation platform (Flask API, React dashboards, Chrome extension on Azure) processing 10K+ items/day; led 6 researcher interns and presented to 15+ HPE executives on responsible AI deployment.

PROJECTS

REAL-TIME VOICE CLONING SYSTEM (GitHub) **Mar 2025 - May 2025**

- Developed end-to-end voice cloning system using GE2E speaker encoder, FastSpeech2 synthesizer, and HiFi-GAN vocoder, processing 26K+ VCTK utterances with MFA alignment to achieve < 100ms inference latency
- Engineered ML pipeline combining contrastive learning, sequence-to-sequence modeling, and adversarial training, implementing PyTorch-based architecture with 256-dimensional speaker embeddings and 10x real-time throughput.

IMAGE DIFFUSION FOR SYNTHETIC DATA GENERATION (GitHub) **Jan 2025 - May 2025**

- Built a DDPM pipeline with baseline and attention-enhanced U-Net (ResNet blocks, linear attention, group norm, SiLU) on 10K+ Oxford Flowers; optimized with beta scheduling + sinusoidal embeddings to reduce reconstruction error 21% and latent size 78%, achieving FID 12.4 (–37%) and IS 4.8, deployed on Streamlit for sampling and +17% downstream defect classification.

TECHNICAL SKILLS

Programming & Development : Python, SQL, R, MATLAB, C++, Bash, JavaScript; PyTorch, TensorFlow, Scikit-learn, Hugging Face, Keras, spaCy, ONNX, FastAPI, Flask, H3 Streamlit; Serialization.
Data Science & ML: Ensemble Models, Kalman Filters, Deep Learning, mBERT, Transformers, RL (PPO, SAC, DQN), Diffusion Models, GANs, Time-Series Forecasting, Anomaly Detection, A/B Testing, RAG tools, LLMs, NLP, LangChain, LangGraph, Context Engineering; PCA, Autoencoders, , Agentic systems; Statistical Modeling.
Data Engineering & Cloud: MLOps, Snowflake, MySQL, PostgreSQL, NoSQL, dbt, Kafka, Redis, Promtail, Loki, rsyslog, Matplotlib, Seaborn, PyQt5, Flask, FastAPI, Airflow; AWS (Lambda, S3, API Gateway, EC2), Azure (VMs, Blob, ML, Synapse, OpenAI), GCP (Kubernetes, GKE, Registry), DigitalOcean; OAuth2, JWT, Docker, GitHub Actions, Kubernetes, CI/CD pipelines
Visualization & Research Tools: Power BI, Tableau, Excel (Power Query, Pivot Tables), Plotly, Matplotlib, Seaborn; CARLA Simulator, OpenAI Gym; ONNX Runtime, CUDA/GPU Optimization, Microservices, Parallel Pipelines, Job Scheduling