

Balraj Ashwath

Machine Learning Engineer | [in](#) LinkedIn | [🔗](#) Google Scholar

✉ balrajashwath98@gmail.com

☎ +91-9964952869

EXPERIENCE

- ServiceNow** Bengaluru, IN
 - Senior ML Engineer** Sep 2023 – Aug 2026
 - MCP Server | Core Platform Frameworks Engineering (AI Control Tower)** Jun 2025 – Aug 2026
 - Built ServiceNow's MCP Server from the ground up – a protocol gateway exposing enterprise AI capabilities (Now Assist Skills, REST APIs, Workflows) as MCP tools for clients such as Claude and AI Agent Studio. Delivered features end-to-end across the FastAPI application, Kubernetes workload & ServiceNow Store application.
 - Reliability & Observability:** Founded the service's end-to-end telemetry, alerting, and on-call infrastructure (Prometheus, Grafana, Splunk). Shaped the MCP-S release and deployment process, and built core reliability primitives. Stood up isolated production, sub-production infra with capacity planning for customer MCP-S traffic.
 - Core Feature Engineering:** Engineered and shipped key MCP-S subsystems – Redis caching, async rate limiting, request throttling, circuit breakers, per-user session management, JWT authentication, and service guardrails. Drove monthly releases, defect triage and resolution, and shipped performance improvements.
 - Platform Extensibility & Security:** Architected MCP configurations to survive Update Set migrations and instance clones. Extended the tool-creation framework to expose REST APIs as MCP tools. Built a file-attachment MCP tool enabling document uploads to instance records. Spearheaded security hardening – resolved high-severity vulnerabilities, minimized container image findings, and shipped FIPS-compliant builds for regulated markets.
 - Customer Adoption & Documentation:** Authored **20+** Knowledge Base articles and client integration guides, driving MCP-S adoption across customers and internal BUs through faster issue resolution and setup. Kickstarted internal docs for release, on-call & alerting workflows. Resolved **25%** of customer-reported cases for MCP-S team.
 - Core AI (prev. CoreLLM) / Advanced Technology Group** Sep 2023 – Jun 2025
 - Drove work across the full LLM development lifecycle for ServiceNow's foundation models (NowLLM family) and fine-tuned open-source LLMs – from synthetic data generation through SFT and DPO / RLHF alignment to large-scale evaluation across model checkpoints, AI skills, and languages.
 - Alignment Data Generation:** Designed **15+** categories of complex reasoning and instruction-following datasets for SFT and DPO / RLHF training, generating them at scale on SyGra, a custom LangGraph-based pipeline co-developed alongside this work. Published preference-pair datasets that fed directly into model training.
 - Model Evaluation Infrastructure:** Built end-to-end auto-evaluation pipelines combining embedding-based representative sampling (cutting evaluation cost at scale without sacrificing statistical confidence) with an automated LLM-as-judge framework that pinpointed failure modes per metric (Faithfulness, Completeness, Relevance, Coherence) to drive targeted training iterations. Onboarded **12+** product AI skills to the evaluation platform, expanding coverage across summarization, generation, search, and workflow domains.
 - Now Assist Summarization:** Built and optimized preprocessing pipelines that improved Now Assist summarization quality and reduced hallucinations across HR, ITSM, and CSM business units in production. Conducted systematic evaluation studies on real-world records to track quality improvements and regressions.
- Qualcomm** Bengaluru, IN
 - Senior ML Engineer** *(Org: Corporate R&D Qranium AI SW)* Apr 2022 – Jun 2023
 - Optimized ML model accuracy for Qualcomm's AI Cloud division as part of the ML Accuracy Team, targeting deployment on the Cloud AI 100 Inference Accelerator (AIC 100). Measured, analyzed, and optimized INT8 quantization accuracy across a range of CV and NLP models – profiling and tuning quantization parameters to consistently exceed **99%** of FP32 accuracy baselines while maximizing the performance-accuracy trade-off.
 - AI Model Efficiency Toolkit (AIMET):** Developed new learnable calibration schemas that improved accuracy in Quantization-Aware Training (QAT). Researched and benchmarked Post-Training Quantization (PTQ) techniques – Auto Mixed Precision, CLE, BC, and Adaptive Rounding – identifying optimal configs per model. Evaluated AIMET / AIMET Pro across CV & NLP architectures. Supported enabling more QAT models inside AIMET zoo.
 - Led the migration of mixed-precision model configs from the AIC 100 stack to QNN (Qualcomm's cross-device ML API) as part of QNN-AIC Convergence. Owned end-to-end development, ran model evaluations, and triaged accuracy regressions surfaced in Daily Accuracy & Performance Digests.

- Microland Limited - Automation Team** Bengaluru, IN
 • **Senior ML Engineer & Data Scientist** [Assoc Tech Arch - SW Dev] Jul 2019 – Apr 2022
 - **Metric Time Series Analysis & Modeling:** Owned the design and build of **Forecasting & Anomaly Detection** systems for edge-device metric data end-to-end, from data pipeline to deployment. Ensembled *ARIMA*, *Prophet*, *NeuralProphet* into a high-accuracy forecasting system, and engineered an AutoEncoder-based anomaly detector reaching **81%** reconstruction F1 – giving Ops teams reliable, low-noise visibility into edge environments.
 - **Alert Clustering:** Built an unsupervised NLP system that organized a completely unlabeled alert corpus into coherent topics, benchmarking K-Means, LDA (TFIDF), doc2vec, and BERTopic to find the strongest representation. Shipped a tagging pipeline that let Ops Engineers label a cluster once and have every incoming alert auto-routed to its nearest match, with cluster-correlation and new-alert-type detection pushing accuracy past **90%** – removing the need for hand-labeled training data entirely.
- Vision and AI Lab, Indian Institute of Science** Bengaluru, IN
 • **Research Intern** (Mentors: Dr. R. Venkatesh Babu & K. Ram Prabhakar) May 2019 – Jul 2019
 - **HDR Image Deghosting:** Designed a novel CNN architecture suppressing ghosting artifacts from camera and object motion in high-resolution HDR fusion. Beat SOTA on the Kalantari SIGGRAPH 2017 benchmark by +1.2 dB PSNR while running **60%** faster – making high-quality deghosting practical at megapixel scale. [ECCV 2020]
- Computer Vision & Deep Learning Lab, NIT-K** Surathkal, IN
 • **Undergraduate Researcher** (Mentors: Dr. Shyam Lal & Karuna K. Eerapu) Jul 2018 – Apr 2019
 - **Road Extraction from Satellite Imagery:** Developed novel segmentation architecture to identify road pixels in aerial imagery based on concepts from U-Net, dense & residual networks. The proposed model improved SOTA on Massachusetts Roads dataset by **2%** with an *order-of-magnitude fewer* parameters. [IEEE Access Journal Paper]
- RBL Bank** Mumbai, IN
 • **Summer Intern (Machine Learning)** May 2018 – Jul 2018
 - **e-KYC Automation:** Built lightweight CNN-based OCR models to extract KYC fields from Aadhaar, PAN cards, replacing manual data entry at onboarding – **83%** F1 at **sub-30ms** inference on low-compute edge devices.

EDUCATION

- National Institute of Technology Karnataka** Surathkal, IN
 • *B.Tech in Electronics & Communications Engineering* | GPA: 7/10 | Senior Thesis: 10/10 Aug 2015 – Apr 2019
- Narayana Pre-University College** Bengaluru, IN
 • *Karnataka State Pre-University* | Percentage: 90% | JEE Mains: 99th %ile Jul 2013 – May 2015

PUBLICATIONS

- **Dense Refinement Residual Network for Road Extraction from Aerial Imagery Data**, Karuna K. Eerapu*, **Balraj Ashwath***, Shyam Lal, Fabio D'Acqua & A.V. Narasimhadhan, *IEEE Access Journal*, 2019. [* - equal contbⁿ]
- **Towards Practical and Efficient High-Resolution HDR Deghosting with CNN**, K. Ram Prabhakar, Susmit Agrawal, Durgesh Singh, **Balraj Ashwath**, & R. Venkatesh Babu, *European Conference on Computer Vision*, 2020.

SKILLS

Languages: Python, JavaScript, SQL, Java, C++

AI / ML: PyTorch, HuggingFace, Transformers; LLM Fine-tuning & Alignment (SFT, DPO, RLHF), Synthetic DataGen, LLM Evaluation; Retrieval-Augmented Generation (RAG), Prompt Engineering, Agentic AI & Tool Use, LangGraph; Inference, Serving & Quantization; Computer Vision (CNNs, Vision-Language Models, Multimodality); MLOps

Infrastructure & Tools: Docker, Kubernetes, Helm, SGLang, vLLM, Model Context Protocol (MCP); Prometheus, Grafana, OpenTelemetry, Splunk; FastAPI, Pydantic, Redis, Git / GitHub Actions, Postman

HONORS & AWARDS

- Shortlisted in Microsoft's Incubathon 2021, Silver Medalist in Kaggle's NFL Impact Prediction 2021 - Top 10%
- Intel Edge AI Scholarship (2019), HackWithInfy Finalist (2018), Merit-cum-Means Scholarship, FFE Scholarship