GAURAV SHARAN SRIVASTAVA

+1 (510)-423-1790 | gaurav_sharan@berkeley.edu | www.linkedin.com/in/gauravsharans

EDUCATION

University of California, Berkeley Aug 2023 – Aug 2025
Master of Information Management and Systems, 3.7 GPA
Relevant Coursework: Deep Neural Networks, Generative AI, Machine Learning, Computer Vision, Bio-sensory Computing

New York University Jul 2022 – Oct 2022
Bridge to MS Computer Science, Distinction
Relevant Coursework: Discrete Math, Programming in C++, Data Structures, Operating Systems, Computer Networks

Kamla Nehru Institute of Technology, Sultanpur, India Aug 2013 – May 2017
Bachelor of Technology, Electrical Engineering, First Class

SKILLS

Languages	Python 3, R, SQL, Bash, MATLAB, C++, HTML, LaTeX
Frameworks & Tools	**ML/Deep Learning:** PyTorch, TensorFlow, NLTK, Ollama, OpenCV, Torchvision, Albumentations **MLOps & Platforms:** MLflow, Airflow, Spark, Hadoop, Flask, Docker, AWS, Azure, Git, Unity

EXPERIENCE

GetReal Labs, San Mateo, USA May 2024 – Oct 2024
Machine Learning/Computer Vision Engineering Intern

- Curated a high-fidelity deepfake corpus with a balanced domain/pose distribution using active-learning uncertainty scores on DeepFace to minimize label skew, and created a ~2M+ image dataset of 6 ethnic and 2 gender categories with > 98% confidence
- Fine-tuned a SWIN V2 backbone using a multi-GPU AWS instance (4×A10G) with PyTorch DDP, leveraging Sharpness-Aware Minimization (SAM) and cosine-annealed one-cycle LR to reach ~95% in-domain ROC-AUC while cutting training time by ~80%
- Replaced AWS Rekognition with an in-house extractor/classifier using RetinaFace with an IoU threshold of 0.95 that reduced per-frame cost > 10x and deployed the real-time detection solution delivering < 10 ms latency at 480p - 1080p 30 FPS streaming

Haas School of Business, University of California - Berkeley, USA Mar 2024 – Aug 2024
Graduate Student Researcher

- Fine-tuned BERT-base on ~2k CEO press releases, tripling the effective training set via nlpaug with contextual and character-level augmentations; achieved 89% F1-score (↑ 13 pp over baseline) accelerating analyst triage time 4x
- Built a high-precision entity-resolution pipeline for 100k company names across SEC 8-K filings using TF-IDF cosine + n-gram Jaccard similarity techniques, delivering 99% accuracy on the test set, and cutting manual deduplication hours by 90%

Tata Consultancy Services (TCS) - Research and Innovation Lab, New Delhi, India Jan 2018 – Jul 2023
Tech Lead, Machine Learning Engineer Oct 2019 – Jul 2023

Drug Development Insights: Intuitive Clinical Trial Monitoring platform

- Forecasted clinical-trial workload for site-managers across ~35k studies for J&J by building a Bi-LSTM model on multivariate site-activity logs with an MAE of 1.8 min (±2%), enabling automated CRA staffing and reducing overtime costs by > 20%
- Developed a Risk Analysis & Categorization Tool (RACT) in R to calculate dynamic site-level risk scores using Bayesian attrition priors and control-limit theory, improving audit readiness and reducing the impact of attrition and reporting biases by 70%
- Built an Airflow-orchestrated ETL mesh (SQL stored procs, window functions, partitioned indexes) automating end-to-end platform workflows including model re-training; reduced ad-hoc query latency, and ingestion time from 8 h to 55 min (–87%)

Signal Detection: Pharmacovigilance signal detection tool for effective drug safety

- Automated symptom-and-diagnosis extraction from 1.2M+ medical reports by training a Bi-LSTM + CRF sequence-tagger, and achieved an F1-score of 0.91 (↑ 23 pp over rule-based NER), reducing medical-reviewer annotation time by 70%
- Advanced a Drug Safety Tool with a regularized logistic-regression model (features: empirical-Bayes shrinkage, MedDRA hierarchy one-hots), that outperformed FDA's Gamma Poisson Shrinker (GPS), with an AUC-ROC score of 0.83 (+0.16 over GPS)

Sanitation Inspection: AI-powered automated sanitation inspection app for state public schools

- Co-developed a hierarchical sanitation-inspection model trained on 120k images using a dual-head MobileNet-V2 SSD, classifying hygiene compliance for washroom components across 45k+ schools, achieving ~90% accuracy and ~92% recall on unclean cases
- Built the end-to-end ML pipeline with geo-tagged capture, real-time retake prompts, and image validation processing ~0.5M images/day, powering dashboards and alerts, leading to a ~28% drop in sanitation-linked girl's absenteeism as per gov. reports

Data Engineer Jan 2018 – Sep 2019

- Designed hybrid data model and DB components improving data quality by 95% and reducing non-compliance by 83%
- Built API interfaces using Jupyter Gateway to serve ML models, enabling integration of 4 standalone analytics solutions
- Conceptualized a Linear Programming algorithm using Knapsack method to help site managers decide clinical trial visits

PROJECTS

Cross-Generator Deepfake Detection & PEFT Generalization

- Benchmarked seven PEFT methods on a CLIP-based detector across 10 k deepfakes from five generators; while LoRA hit the best in-domain ROC-AUC (99.7% with 0.26% parameters), the BayesTune adapter delivered the strongest zero-shot generalization, averaging +4 to +8 pp higher accuracy on unseen generators than other selective PEFTs while tuning just 0.04% of weights
- Designed a LoRA rehearsal curriculum that maintained +5.4 pp higher accuracy on prior domains and +6.1 pp on unseen domains compared with full fine-tuning, mapping how training on one generator helps or hurts other generators

ADHD EEG Experimentation

- Designed a 30-session, single-subject protocol with an OpenBCI Ganglion (4-channel EEG), capturing baseline, post-medication, and post-chanting states; logged 15 k+ 90s epochs to track theta, alpha, beta power and the theta/beta ratio (TBR)
- Medication cut the frontal TBR by 38% (5.3 to 3.3, $p < 0.01$) and boosted low-beta power of 22%; chanting raised alpha power by 31% and sustained attentional improvements (−27% Conners score) after two weeks. Random-Forest classifier distinguished the three cognitive states with an 87% cross-validation accuracy, pointing to viable real-time neurofeedback markers

RAG Benchmarks on Ollama

- Built a reproducible RAG test-bed that runs small (Mistral-7B), medium (Mixtral 8×7B), and large (Llama 3-70B) GGUF LLM models through Ollama, backed by a FAISS HNSW vector store (8k docs, MiniLM-all-v2 embeddings); GPT-4 scoring shows the large model boosts factual F1 by 12 pp but triples latency, while the medium model hits a sweet spot (89% F1, 2.1s response)
- Optimized local inference with 4-bit QLoRA quantization + passage RRF fusion, cutting end-to-end latency ≈ 40 % and memory footprint ≈ 55 % versus the naive baseline, enabling laptop-grade RAG demos without cloud GPUs

Building Damage Assessment from Satellite Imagery

- Implemented an attention-enhanced Siamese CNN that ingests paired pre-/post-disaster tiles from Google Earth & Maxar; added CBAM attention, multi-scale ResNeXt-50 encoders, and heavy geo-augmentations (CutMix-geo, random shadows)
- Reached F1 of 0.69 on the xView2 leaderboard (↑ 18 pp vs. vanilla U-Net) and delivers ≈ 55 ms inference per 1024 × 1024 tile on a single TeslaT4 GPU, enabling near-real-time damage heat-maps for first responders

PUBLICATIONS

- Srivastava et al., 2025, Beyond the Benchmark: Generalization Limits of Deepfake Detectors (Manuscript in preparation)
- Evolution of Multimodal AI (3-part series), Medium, 2025. https://medium.com/@gaurav.sharan1003
- Chen et al., 2024, Fairness in Post-Disaster Response: A Case Study of Aid Distribution During the 2018 Sulawesi Earthquake and Tsunami. (Manuscript in preparation)
- Indani et al., 2020, Preparing for Intelligent Infrastructure for Mass Immunization Related Supply Chain and Monitoring Process, Journal of Hospital Pharmacy. DOI JOHP-ISSN:2348-7704
- Adopting a Technology-Driven Approach to Implementing EDC for Medical Devices, 2020, Applied Clinical Trials