Shudhanshu Ranjan

+1 551 344 2982 | s.ranjan0304@gmail.com | linkedin.com/in/shudhanshu-ranjan | github.com/git4sudo | sudoai.org

EDUCATION

Stevens Institute of Technology, Hoboken, NJ
Master of Science in Computer Science | GPA: 3.63/4.0 **Sep 2022 - May 2024**
Coursework: Algorithms, NLP, CV, Applied Statistics with App/ in Finance, Reinforcement Learning and Sequential Decision Making

Stevens Institute of Technology, Hoboken, NJ
Graduate Certificate in Machine Learning | GPA: 3.75/4.0 **Nov 2023 - May 2024**
Coursework: Artificial Intelligence, Machine Learning, Statistical Machine Learning, Deep Learning

Presidency University, Bangalore, India
Bachelor of Technology in Computer Science and Engineering | CGPA: 8.41/10 **Aug 2018 - May 2022**
Coursework: Data Structures, Algorithms, DBMS, Data Visualization, Image Processing, Neural Networks, Graph Theory

SKILL

Technical: Python - (Pandas, Numpy, Scikit-learn, MatPlotLib, SciPy, Statsmodels, ARIMA, Seaborn, NLTK, CV2, XGBoost, LightGBM), Gen AI - (GAN, VAE, GPT, BERT, T5, LLaMA, RAG, ViT), C++, Tensorflow, PyTorch, Keras, MatLab, SQL, GCP - (Vertex AI, BigQuery ML, Bigtable, LookML, AutoML), Git, Jupyter Notebook, MLflow, TensorRT, Weights & Biases, Hadoop, Kubernetes, PySpark, Docker.

Certification: TensorFlow Developer Certificate (by Tensorflow); 60+ GCP skill badges of 200+ hours worth of training from **cloudskillsboost** (ML Infrastructures, Serverless Cloud Run Development, Cloud Dataflow, Pub/Sub, BigQuery, Cloud Architecture).

WORK EXPERIENCE

Software Engineer, Neurability Foundation, Remote, US **Dec 2024 – Present**
- Architected the end-to-end backend for an AI-driven productivity tool using **FastAPI**, **PostgreSQL**, **Pinecone**, with **Claude Sonnet & GPT-4o mini API**, deploying via **Cloud Run** with integrated observability for **latency, memory, and error rate alerts**.
- Refactored LLM inference pipeline from 3rd-party APIs to **Vertex AI**, deploying two fine-tuned models (**LLaMA 1B & 7B instruct**), reducing system latency by **~70%** from ~650ms to <200ms while improving inference reliability and scalability.
- Trained and deployed custom LLMs to power **task decomposition and prioritization** for ADHD users, generating **3-level nested task trees**, dynamic system prompts, context-aware reminders, and a **batch email summarization agent**.

Machine Learning Engineer, Health Innovators, Remote, US **Dec 2024 – Present**
- Built and deployed an end-to-end **multimodal medical assistant** using HealthGPT models, enabling X-ray, CT, MRI comprehension, translation, reconstruction, and super-resolution; supported 12+ medical tasks and processed **100+ high-res images/day**.
- Fine-tuned and deployed an LLaMA-2-7B chatbot for patient-doctor conversations using **PEFT (LoRA)** with **4-bit quantization**; integrated **Firestore** for memory with **conversational logging**; explored **FAISS-based RAG** for long-term medical context retention.
- Developed a **hybrid NER pipeline** combining **Stanza**, **BioBERT**, and **Gemini Pro** to extract and normalize medical entities from clinical notes; mapped over **50+ disease mentions** to **SNOMED codes** with **>92% accuracy** using embedding-based similarity and ontology linking via **Qwen2-1.5B (Mixture of Experts) model**.

Graduate Research Assistant, Stevens Institute of Technology, Hoboken, NJ **Jan 2024 – Present**
- Extracted 10k tweets using the **Twitter v2 REST API** to develop a **misinformation classifier**, enhancing **AI content moderation**.
- Conducted literature review on **crowdsourced fact-checking models** (**Matrix Factorization**, **Difference in Differences**, and **Regression Discontinuity Design**), noting **Community Notes' improvement** from **4%** to **12.5%**, increasing **annotation relevance**.
- Assisted PhD students and postdocs in co-authoring a **research paper** on **parameter-efficient fine-tuning** & **prompt engineering** by analyzing **10+** NLP **datasets** with various **LLMs** to benchmark performance on **time complexity**, optimizing **model efficiency**.
- Researched **style transfer** in NLP using **GPT-2** and **T5 LLMs** from the **Hugging Face API**, achieving a **2.4x** improvement in **BLEU score** on the **GYAFC** and **XFORMAL** datasets with over **110k sentence pairs**, utilizing **PyTorch** and **CUDA** for efficient computation.

Deep Learning Research Intern, Prayogpeti, Bangalore, IN **Sep 2021 – Jun 2022**
- Constructed a **comparative study** between 3 inductive **Graph Neural Networks**, namely TexTING, In-GCN, and In-GAT, to determine the best-performing model for **text classification**. Published in **IEEE**: **DOI**: 10.1109/ASSIC55218.2022.10088315.
- Modified **Inductive GAT models**, finding higher **entropy** in smaller datasets like IMDB (~50k datapoints) and lower initial entropy in larger datasets like DBPedia (~630k datapoints), which led to an increase in model **accuracy** to **98.10%** on 14 different classes.

Machine Learning Engineer Intern, UAV Team, iNeuron Intelligence PVT LTD, Bangalore, IN **May 2021 – Aug 2021**
- Designed & implemented **UAV simulations** in ROS and Gazebo, and performed **real-time object detection** on those simulations.
- Optimized by converting a **YOLOv5** model to a **MobileNetV3** architecture, retrained on **TPU**, resulting in a **5.5x increase** in **inference speed** on **CPU** and improved frames per second (**FPS**) performance.

Machine Learning Engineer Intern, SAT IMG Team, iNeuron Intelligence PVT LTD, Bangalore, IN **Nov 2020 – May 2021**
- Adapted an innovative satellite **image masking** method and presented a **U-Net** with **Inception CNN model** on a 21.69 GB dataset. Devised a novel image subtraction **algorithm** and a **prototype** to outline the differences between multiple reference images.
- Retrieved **MultiPolygons detailed masks** for both 3-band and 16-band format images, used said MultiPolygons to detect 10 different object classes in the image, and evaluated the performance by **Jaccard similarity** on the region of interest.

PROJECT

RL-Based Stock Trading System Utilizing Sentiments Analysis [LINK] **Feb 2024 – Present**
- Developed a comprehensive automated **trading** system by training 5 distinct **RL agents** (A2C, DDPG, PPO, SAC, TD3) using the **FinRL** and **OpenAI Gym**, leveraging **20 years** of data and trading 10 company **stocks** across 5 sectors for 4 years.
- Incorporated emotion and sentiment data from tweets, applied **SMOTE**, and used **BiLSTM** to **impute** missing values with **75% accuracy**, yielding **1.53x avg returns** over counterpart agents and improving system performance in **real-time** market conditions.
- Benchmarked against **DOW, S&P 500, NASDAQ, and MVO** across all the stock options, the best model (**PPO**) demonstrated superior performance in volatile markets, achieving a **2.98x return** with a **1.33 Sortino Ratio** for a fixed 3.5% yearly risk-free rate.