Sanat B Singh

AI Engineer | India | +91 8114342153 | sanat.b.singh99@gmail.com | linkedin.com/in/sanatbsingh | github.com/sanatsingh | sanatsingh.github.io

SUMMARY

AI Engineer and Tech Lead co-leading product development at ETS GCC (the org behind TOEFL and GRE), with a focus on agentic AI, interactive assessment simulation, and generative AI systems. Experienced in LLM fine-tuning and pre-training on HPC clusters (Nvidia DGX A100/H100), edge AI deployment on Jetson AGX Orin, and building multi-agent systems at scale. Previously at Deloitte USI as a Data Scientist and Machine Learning Engineer across NLP, Computer Vision, and Edge AI. Personal research interest in controllable world models and agentic system design.

EXPERIENCE

AI Engineer (AI Consultant) | **ETS GCC / Firstsource Solutions (FSL)** | *Sept 2024 - Present*

- Working on Product Research & Development at ETS GCC Hyderabad - set up in collaboration with Firstsource Solutions (FSL).
- Building an end-to-end platform and framework/library for interactive assessment - spanning simulation development, skill practice, and deployment at scale.
- Working with SLMs and PEFT techniques to build efficient scoring and feedback models.
- Designed and implemented a multi-agent automated prompt generation and evaluation system.
- Developed and deployed a Text-to-Image and Image+Text-to-Image generation pipeline.
- Implemented benchmarking and cost-saving measures for model selection and deployments, resulting in ~10x cost reductions across use cases.

Data Scientist / MLE (Consultant) | **Deloitte USI** | *June 2023 - Aug 2024*

- Spearheaded end-to-end development while leading a small team for GenerativeAI (GenAI) and Computer Vision (CV) projects, successfully handling problem sets including Text Classification and Summarization, Q&A, Object Detection & Classification, and Image Generation.
- Fine-tuned and evaluated multiple open-source Language Models (LLMs) such as LlaMA 2/3.1 and Mixtral 8x7B using PEFT techniques like LoRA and QLoRA. Experienced with Pre-training techniques like Domain Adaptive Pre-Training (DAPT) using Nvidia Nemo, HuggingFace, and DeepSpeed on HPCs by Nvidia like DGX A100 and H100.
- Developed and deployed an API endpoint for the ControlNet pipeline on Nvidia Jetson AGX Orin for Image Re-Styling and Generation.
- Using Nvidia Deepstream and TernsorRT built a real-time traffic congestion analyzer using Yolov8 and deployed on Nvidia Jetson AGX Orin.
- Developed a Finance Analyst Assistant using DAPT and LoRA which solved complex financial queries by planning and executing steps to achieve high-quality insights and results, model training and deployments were done using Huggingface and Nvidia TensorRT-LLM on Nvidia DGX H100 8-node HPC cluster.
- Built Knowledge Retrieval Q&A chatbots (RAG), Brochure Generation Systems, and similar applications, taking ownership of the development, containerization (using Docker), and deployment of solutions across cloud services like AWS, Azure, and HPCs and edge devices by Nvidia - Nvidia DGX A100/H100, Jetson AGX Orin and Jetson Orin.
- Utilized frameworks and tools such as PyTorch, Keras, HuggingFace, Nvidia NeMo, TensorRT, DeepStream, detectron2, FastAPI, Docker, and Kubernetes to streamline development processes and enhance project outcomes.
- Collaborated closely with cross-functional teams to ensure seamless integration of AI solutions into existing systems and workflows.

Data Scientist / MLE (Analyst) | **Deloitte USI** | *July 2021 - May 2023*

- Level: Analyst, Data Scientist / ML Engineer (Role in Projects) working on CV, NLP & EdgeAI.
- Successfully worked on multi-modal problems, such as Visual Question Answering (VQA) - benchmarking and fine-tuning the UNITER model to achieve state-of-the-art performance.
- Applied NLP techniques to solve problems like Intent Slot Classification and developed Conversational Chatbots utilizing Nvidia Nemo for model development and Nvidia Riva for deployment.
- Developed computer vision solutions like object detection and tracking for a retail shelf management system using PyTorch and Keras, integrating with SAP for seamless data management. MaskRCNN and Alexnet models were fine-tuned on Nvidia T4 GPUs and converted to ONNX format for deployment on edge devices.
- Ensured clear and concise communication of project objectives, methodologies, and results to stakeholders, avoiding technical jargon and industry-specific terms.
- Maintained consistent formatting and tense usage throughout project descriptions, enhancing readability and professionalism.

Research Intern | **University of Houston Downtown** | *2020 - 2021*

- Worked as a research intern under Dr. Hong Lin on EEG signal based sleep stage classification at the Department of Computer Science and Engineering Technology, University of Houston Downtown, USA.

ML Engineer & Co founder | **Zyik.ML** | *2019 - 2021*

- Founded Zyik.ML along with my friend & colleague Aayush Kumar to provide affordable AI-based healthcare services.
- Took decisions as one of the founding partners, and successfully led the team, which resulted in us getting selected for the AWS Activate program. Registered as a startup under the Ministry of Micro, Small and Medium Enterprises - Government of India.
- Developed Computer Vision-based CADx Systems and Proof of Concept projects/prototypes.

Machine Learning Instructor (Core Team) | **Konnexions - KIIT** | *2018 - 2021*

- Taught and mentored around 120 undergrad students in a semester long Machine Learning course.
- Designed course curriculum, prepared presentation, conducted hands-on sessions and workshops.

PROJECTS

Data Independent Analysis with Peptide Sequences using Deep Learning | **Independent Research**

- Conducted independent research in collaboration with Dr. Hong Lin and a team of researchers, utilizing deep learning techniques to predict iRT and Ion Mobility based on peptide sequences.
- Developed and implemented novel deep learning models using Nvidia RTX 3070, achieving significant improvements in the accuracy and efficiency of data analysis.

CADx System for Chest X-Ray Diagnosis | **Academic Project**

- Designed and developed a web application for interpretation/diagnosis of chest X-rays. Transfer Learning was applied, a pretrained DenseNet121 was used. Major issue tackled in the project was of Data Imbalance present in the dataset which solved by using Weighted Cross-Entropy Loss, weights were calculated based on class frequencies to ensure equal contribution to the loss by each class.

ACL Tear Detection | **Zyik.ML**

- A CADx system to detect ACL (Anti cruciate ligament) tear in MRI scans. A CNN classifier is built using Alexnet on the MRNet dataset released by Stanford ML group. Data augmentation was applied while training to deal with less number of data samples. The AUC achieved was 0.858 on the train set and 0.876 on the validation set.

Computer Aided Diagnostic System for Malaria Detection | **Zyik.ML**

- A lightweight Computer-Aided Diagnostic System with the aim of easing the weary task of detection of malaria-infected cells by examination of blood smears under a microscope using deep learning. A custom lightweight ConvNet is implemented with less than 8 million parameters that come close to Densenet121 in terms of parameters but shows 10x faster inference time with far fewer usage of resources on CPU deployment thus eligible for deployment on edge devices. Validation accuracy achieved was 95.6% (Updated Mish Version attained 97.8%).

PUBLICATIONS

- MOSQUITO-NET: A deep learning based CADx system for malaria diagnosis - Wiley Expert Systems (SCI) - DOI: 10.1111/exsy.12695. Also presented as a poster at the ICML 2020 Machine Learning for Global Health Workshop.
- Using Multi-Student, Generative AI Teaching Simulations as Practice Spaces for Facilitating Science Discussions - stand-alone paper, 2026 NARST Annual International Conference, Seattle, WA.
- Zyik.ML - A CADx System - Published as a book chapter in proceedings of Project Innovation Contest 2021.

CERTIFICATIONS

Computer Vision for Industrial Inspection (Nvidia DLI); Model Parallelism: Building and Deploying Large Neural Networks (Nvidia DLI); TensorFlow Developer Certificate (Google); Computer Vision Nanodegree (Udacity); Deep Learning Specialization (Coursera); Machine Learning (Coursera).

ACHIEVEMENTS

- Ranked #11 (Top 20) at VibeCon India 2026, placing in the top 20 out of 25,000+ competing teams, organized by Emergent - with Soumik Rakshit and Atanu Sarkar.
- Recognized by ETS GCC leadership for proactively delivering AI product innovations, including a multi-agent prompt generation system and generative AI pipelines, with measurable impact across multiple use cases.
- Received multiple Applause Awards at Deloitte USI for accelerating delivery of complex LLM and computer vision models under tight timelines across client engagements.
- Selected for AWS Activate program with Zyik.ML, recognized by Amazon for building viable AI-powered healthcare prototypes and CADx systems.
- Presented poster at ICML 2020 Machine Learning for Global Health Workshop for the MOSQUITO-NET deep learning CADx system (published in Wiley Expert Systems, SCI).

EDUCATION

Bachelor of Technology (Computer Science & Engineering) | **Kalinga Institute of Industrial Technology, Bhubaneswar** | *2017 - 2021*

- Graduated with CGPA 8.99