

Shantanu Sharma

Delhi, India | [LinkedIn](#) | [Github](#) | [Portfolio Website](#) | shantanu100.sharma@gmail.com | +91-9990193322

WORK EXPERIENCE

Becton Dickinson

Gurugram, India

Data Analyst, Strategy & Transformation Team

June 2023 – May 2024

- Developed time-series forecasting models (Neural Prophet, XGBoost) for predicting product inventory.
- Automated data processing pipeline to minimize time and labor up to 50%, while eliminating human error.
- Analyzed supply chain data to provide stakeholders with actionable insights and drive business growth.

EdgeNeural.ai

Pune, India

Machine Learning Intern

July 2022 – October 2022

- Trained object detection models and optimized them to gain up to 2x FPS boost on edge devices.
- Built a Tflite optimization tool for automating model optimization and integrated it with Docker.
- Benchmarked lightweight detection models (SSD7, FastestDet, NanoDet etc) & tested deployment on Android.

Carscan.ai

Pune, India

Machine Learning Intern

August 2021 – November 2021

- Trained neural network (EfficientNet) to identify physical damage in car images.
- Optimized models using model pruning and ONNX inference engine for fast inference.
- Leveraged reflection removal techniques and video stabilization using deep learning to improve data quality.

RELEVANT PROJECTS

Progressive Distillation of Diffusion Models (Image Generation) | [Github](#)

July 2024

- Built a diffusion model with DDIM sampler from scratch, trained it for 50 hrs & improved image generation quality.
- Distilled (student - teacher method) the model from 256 to 2 sampling steps resulting in faster inference.

Diffusion Models from scratch (Image Generation) | [Github](#)

June 2024

- Implemented the research paper “Denoising Diffusion Probabilistic Models” from scratch.
- Created a UNet with Self-attention, trained it for 40+ hrs on Stanford Cars dataset to generate car images.

Smol - GPT (LLM) | [Github](#)

May 2024

- Built a 22 million parameter LLM based on GPT2 architecture from scratch. Trained it on Wikitext3 dataset.
- Learned & implemented transformer architecture & causal self-attention from scratch in Pytorch.

Speaking - Doc: RAG Application | [Github](#)

April 2024

- Developed a full stack Retrieval Augmented Generation (RAG) application & deployed it using Streamlit.
- Enabled users to chat with any document (.pdf, .txt, .docx). Used Mistral 7B LLM and LlamaIndex.

Building Area Estimation Using UNets | [Research Paper](#)

August 2023

- Analyzed & benchmarked multiple deep learning models for segmenting buildings from satellite images.
- Led the entire project and published a research paper as the lead author.

SKILLS

Core: Machine Learning | Neural Networks | Computer Vision | Generative AI | LLMs | Data Structures & Algorithms | Statistics

Languages/Tools: Python, C, Pytorch, Tensorflow, OpenCV, Langchain, LlamaIndex, Github, Numpy, Pandas, Matplotlib, HuggingFace, Docker, Android Studio, Excel, PowerBI, HTML, ROS, SQL, FastAPI

EDUCATION

Bharati Vidyapeeth's College of Engineering

Delhi, India

B.Tech in Electronics and Communication (CGPA: 8.8)

2019 - 2023

- Chairperson of IEEE RAS (Robotics and Automation Society): Organized robotics and computer vision related workshops and competitions.
- Certificate of Completion - e-Yantra Robotics Competition (eYRC) IIT Bombay 2020-21: Simulated Autonomous Ground Vehicle (AGV), implemented 2D & 3D mapping & autonomous navigation using ROS.

CERTIFICATIONS

- **Generative AI with Large Language Models | Coursera:** Supervised Fine Tuning, LoRA, PEFT, RLHF.
- **Deep Learning Specialization | Coursera:** Neural Networks, Sequence Models, Convolutional Neural Networks.