Param Jain

Boston, MA | +1-302-409-9135 | jain.param@northeastern.edu | linkedin.com/in/paramsachinjain

PROFESSIONAL SUMMARY

AI enthusiast with over 10 projects specializing in AI and machine learning software applications. A collaborative problem solver driven by a genuine curiosity for innovation and a vision for meaningful impact in the digital space. Seeking an internship opportunity to continue building intelligent solutions and contribute to forward-thinking teams

EDUCATION

M.S. in Computer Science | Northeastern University, Boston, MA — Expected May 2027
B.Tech in Computer Science and Business Systems | Medi-Caps University, Indore, India | GPA: 8.2/10 — Aug 2021 – May 2025

PROFESSIONAL EXPERIENCE

AI/ML ENGINEER INTERN, DevQAExpert Solutions Pvt Ltd — Jan 2025 – Apr 2025

- Optimized Python inference pipelines for production-grade systems, improving latency by 30% and scalability by 25% on distributed deployments
- Built 5 prototypes integrating NLP and computer vision using PyTorch and HuggingFace; collaborated with cross-functional teams to ship 3 to production
- Applied model optimization techniques including quantization and pruning, taking ownership of edge deployment solutions on resource-constrained devices

AI/ML SOFTWARE ENGINEER INTERN, Genesis Technology — Jul 2024 – Aug 2024

- Deployed scalable image classification API with production-grade code serving 5K+ users, optimizing model size by 40% while maintaining 94% accuracy
- Profiled and optimized Transformer attention mechanisms in CUDA, reducing inference overhead by 15% for real-world system deployment
- Built cross-platform model serving infrastructure using TFLite and ONNX across AWS cloud platforms adopted as standard pipeline for 2 production projects

PROJECTS

AI-POWERED DOCUMENT ANALYZER — Nov 2025 – Dec 2025

- Developed GenAI application integrating LLMs (OpenAI API, Gemini, Mistral) for intelligent document analysis across 10+ unstructured formats using Flask backend
- Implemented RAG system with ChromaDB vector database and embeddings, enabling multi-turn conversational AI agents for contextual Q&A and information retrieval
- Built production-grade Python backend with Streamlit interface, automated data pipelines, and batch processing for scalable workflow automation

ELARA – NLP SYSTEM FOR EDGE DEPLOYMENT — May 2025 – Sep 2025

- Built NLP question-answering system fine-tuning DistilBERT with HuggingFace Transformers, achieving 91% F1 on SQuAD 2.0 with 35% faster inference
- Compressed model from 2GB to 500MB (75%) via knowledge distillation, ensuring scalable and maintainable deployment on edge devices
- Created benchmarking pipeline evaluating inference engines across 8 configurations, iterating based on feedback to guide technical decisions

DIABETIC RETINOPATHY DETECTION – MEDICAL IMAGE CLASSIFICATION — Nov 2024 – Dec 2024

- Trained EfficientNet-B4 on 92K medical images using PyTorch and Scikit-learn, achieving 89.7% accuracy through iterative model tuning and clean code practices
- Applied quantization-aware training to compress model by 60%, delivering production-ready ML solution with maintainable abstractions
- Scaled training across 4-GPU distributed system using CUDA and parallel data pipelines, reducing training time by 61% (18h to 7h)

TECHNICAL SKILLS

Languages: Python, JavaScript, TypeScript, C/C++, SQL, CUDA, Bash, Java
Machine Learning: PyTorch, TensorFlow, Scikit-learn, HuggingFace Transformers, LLM Fine-tuning, Model Optimization
GenAI & Agents: LLMs, OpenAI API, RAG Systems, Embeddings, AI Agents, Multi-turn Conversationality, LangChain
Backend & APIs: Flask, FastAPI, REST APIs, PostgreSQL, Redis, Distributed Systems, Data Pipelines
Frontend: React, Streamlit, Data Visualization
Infrastructure & Cloud: Git, Docker, Linux, AWS (EC2, S3), Airflow, CI/CD, Version Control

PUBLICATIONS

"Application of AI in Interior Design," Global Intl. Journal of Eng. and Tech., Jun 2025. GAN-based generative design evaluated on 300K+ samples with 24% accuracy improvement over baseline

"Enhancing HCI through Advanced NLP," Intl. Journal of Scientific Research in Eng. and Mgmt., Jun 2024. Trans- former attention visualization methods improving interpretability by 18% on user study metrics