Ajay Addada

+1 (917)-517-3757 - ajayvarma.addada2001@gmail.com - LinkedIn - GitHub

EDUCATION

New Jersey Institute of Technology **Newark, New Jersey, USA**
Master of Science in Computer Science(GPA: 3.8) *2024 - 2026*

TECHNICAL SKILLS

Programming: Python, R, SQL, Java, C++
ML & AI: PyTorch, TensorFlow, Scikit-learn, HuggingFace, Transformers, YOLOv8, LLMs, RAG, Deep Learning
Data Engineering: ETL/ELT, dbt, PySpark, Dataiku, MySQL, Oracle, Snowflake
MLOps / Infrastructure: Docker, Kubernetes, Git, CI/CD, Streamlit, Gradio, APIs
Cloud: AWS(EC2, S3, Lambda, Redshift), Azure Synapse
Tools: Power BI, Tableau, Alteryx, Jupyter, Git

WORK EXPERIENCE

Student Office Assistant
New Jersey Institute of Technology/ part-time *Mar 2025 - Present*

- Streamlined administrative reporting by managing and updating 1,000+ academic and operational records using Excel and Access, improving data accuracy and retrieval efficiency.
- Automated spreadsheet workflows to reduce manual errors and repetitive tasks, increasing reporting accuracy by 15% and saving significant staff time each week.
- Built data-driven tracking systems for event planning, enhancing attendance consistency and reducing resource shortages by 15–20% through forecasting and allocation.

ML Engineer
GeekBull Consulting, Hyderabad, India *2023 - 2024*

- Designed and deployed end-to-end ML pipelines for data ingestion, feature engineering, model training, and evaluation, reducing model development time by 30% and improving experiment reproducibility across teams.
- Built, fine-tuned, and optimized ML/LLM models(NLP, CV, supervised/unsupervised) and deployed them as scalable APIs using FastAPI, Docker, and CI/CD workflows—cutting inference latency by 20–35% in production-like environments.
- Engineered high-performance retrieval and inference systems using vector databases, optimized embeddings, and ETL pipelines, increasing query accuracy by 25% and supporting low-latency real-time ML applications.
- Improved model accuracy and reliability through extensive hyperparameter tuning, A/B testing, and prompt engineering(for LLMs), leading to 15–20% uplift in model performance and more stable production metrics.

PROJECTS

- **Retrieval-Augmented Generation(RAG) Chatbot** - Python, Streamlit, Ollama, FAISS Vector DB

  -Built a RAG-based chatbot using Python, LangChain, FAISS, and Ollama, achieving a 75–80% user engagement rate and reducing document search time by 50–60%.
  -Improved retrieval relevance and model reliability by refining chunking strategy, embedding generation, and vector search, ensuring accurate answers even for complex queries.
  -Real-world Use Case: Policy lookup, internal knowledge search, academic document Q&A.

- **Real-Time Object Detection with YOLOv8** - Python, OpenCV, Ultralytics YOLOv8

  -Developed a YOLOv8 detection pipeline achieving 85–88% mAP50 and 20–25 FPS on an M2 Mac using Metal acceleration, optimized preprocessing, and targeted augmentation.
  -Refined model robustness by 12–15% and applied the system to security surveillance, traffic monitoring, and warehouse automation to reduce manual monitoring effort by 35%.
  -Real-world Use Case: Security surveillance, traffic monitoring, warehouse automation, weed detection in fields.

- **Deepfake Detection System** - PyTorch, Transformers, Gradio UI

  -Engineered a CNN + ViT-based deepfake classifier trained on FaceForensics++, achieving 90%+ accuracy(92% precision, 88% recall, 90% F1-score) and deployed via a Gradio web interface.
  -Reduced manual media-verification time by 40% and increased authenticity-check reliability by leveraging transformer-based artifact detection for high-risk content.
  -Real-world Use Case: Media authenticity checks for news, cybersecurity, and digital forensics.

CERTIFICATIONS & ACTIVITIES

- AWS Machine Learning Specialty
- **Volunteer** at New Jersey Institute of Technology for 2024 NJ Science Olympiad - Jan 12th 2024
- Continuous learning in AI-driven analytics and cloud-based data pipelines