Varun Totakura

(848) 667-6729 — varun.totakura@gmail.com — varuntotakura.github.io — linkedin/in/varuntotakura

Professional Summary

Passionate Software/ML Engineer with over 4 years of professional experience. Excellent team player and individual contributor with strong interpersonal skills, committed to achieving project objectives efficiently.

Technical Skills

- Frameworks: PyTorch, TensorFlow, Scikit-Learn, Keras, LangChain, Hugging Face
- Languages: Python, JavaScript, R, SQL
- Modules: Dask, Numpy, Pandas, PySpark, FastAPI, Ollama, OpenAI, Gemini
- MLOps: Databricks, Jenkins, Docker, Airflow, Kafka, Vertex AI, AutoML, Azure AI
- Databases: PostgreSQL, MySQL, BigQuery, Teradata, MongoDB, Cassandra, VectorDB - Pinecone, Chroma
- Clouds: Google Cloud, Microsoft Azure, Amazon Web Services, ServiceNow
- Visualization and Reports: Matplotlib, Seaborn, Plotly, Power BI, Tableau
- Version Control: Git, GitHub, GitLab

Professional Experience

AI Engineer - Lead Assistant Manager, EXL Service Inc., New York, NY (Remote) *October 2024 – Present*

- Engineered 15 CI/CD ML pipelines deployment using Jenkins and Apache Airflow DAGs on GCP, launching DataProc clusters to process data and generate predictions, with results stored in BigQuery tables for downstream process.
- Optimized text processing code by parallelizing tasks with Dask and Multi-threading, reducing execution time from 4 hours to under 2 hours while maintaining accuracy.
- Productionized two Kubernetes-based applications, optimizing container orchestration for scalability and achieving good uptime across distributed healthcare systems.
- Automated deployment of local Python scripts to GCP-compatible Airflow workflows, slashing deployment time from over 24 hours to under 30 seconds, boosting team productivity by 50%.
- Refactored and optimized a data pipeline processing 1M+ records daily, enhancing code readability and accelerating execution from 9 hours to under 5 minutes through efficient resource utilization.
- Developed a Generative AI solution for fax data extraction using OCR, integrating a RAG system with OpenAI GPT-4o to achieve 97% accuracy in context-aware query responses.
- Engineered a Generative AI pipeline leveraging OpenAI Whisper to transcribe audio calls, enhancing a RAG-based model to deliver context-driven answers with 20% improved response relevance.
- Designed a GenAI-driven customer support tool powered by a RAG framework and internal knowledge bases, reducing query resolution time by 40% while ensuring precise, actionable solutions.

AI/ML Engineer - Graduate Research Assistant, Florida State University, Tallahassee, FL *August 2022 – August 2024*

- Built a Gen AI assistant with Advanced RAG to process insurance PDFs, extracting key details and enabling real-time user queries, cutting review time by 40%.
- Designed a groundbreaking hybrid AI architecture combining TCN, CNN, and GRU to analyze Alzheimer's Disease patterns, achieving a 35% performance boost through domain disparity reduction.
- Developed a state-of-the-art Active Learning Algorithm for Imperfect Oracles with LLMs, optimizing performance through parallel computing with Dask to achieve 4x speedup in model training.
- Pioneered a scalable deep learning pipeline for model correctness across prediction, processing over 300,000 records of 10 diverse datasets with high accuracy.
- Implemented an NLP solution using BERT and GAN networks for precise bug localization in code, analyzing 150,000+ social media posts with high precision.

AI Software Developer - Systems Engineer, Tata Consultancy Services, Hyderabad, India *July 2020 – September 2022*

- Developed a cutting-edge Customer Service Chatbot using Azure Bot Framework and Python, deployed on Microsoft Teams to interact with ServiceNow via REST API; reducing operational times by 40% and enhancing user satisfaction.
- Implemented a REST API - based service to integrate data from external systems into centralized SQL database to process as input to SQL-driven analytics dashboard for monitoring operations and issues in ServiceNow.
- Spearheaded the creation of AI-powered chatbots with Virtual Agent, integrating 9+ data sources to transform user engagement, achieving a 50% increase in interaction efficiency.
- Orchestrated seamless integrations between ServiceNow and 5 external platforms using REST and SOAP protocols in an Agile SDLC environment.

- Engineered scalable ServiceNow automations, leveraging intelligent process optimization to boost productivity by 45% across critical workflows.
- Led a high-stakes data migration project, leveraging AI-driven integrations to transfer 15 million records with unprecedented efficiency and reducing manual effort by 70%.
- Architected and maintained 15+ robust applications, incorporating advanced workflows, Service Portal, ACLs, and AI-enhanced scheduled jobs.
- Achieved 95% accuracy in cross-platform data synchronization, ensuring flawless information flow across multiple ServiceNow instances and external systems.
- Pioneered R&D initiatives by designing ML-driven automation solutions with AI Studio, amplifying task efficiency by an impressive 20x.

Research Publications

- Active Learning with Imperfect Oracles using Text Data. *TPAMI - Expected 2025*
- Analyzing the behavioural patterns in elderly people suffering with Alzheimer's Disease. *TACCESS - Expected 2025*
- MetaErr: Predicting Error Patterns in Deep Neural Networks using the Meta Model Data. *Expected 2025*
- Master's Thesis - A Study on Deep Learning Models for Real World Applications. *August 2024*
- Prediction of Transmittable Diseases Rate in a Location Using ARIMA and GARCH. *November 2021*
- Improved Safety of Self-Driving Car using Voice Recognition through CNN. *October 2021*
- Prediction of Stock Trend for Swing Trades using Long Short-Term Memory Neural Network Model. *March 2020*
- Prediction of Animal Vocal Emotions using Convolutional Neural Network. *February 2020*
- An Integrated Approach to Sentiment Analysis using Machine Learning Algorithms. *April 2020*
- Published another 5 research papers and survey papers. *2019 - 2020*

Awards and Achievements

- Three times Star of the Month, Tata Consultancy Services. *2021 - 2022*
- Five times On the Spot Awards, Tata Consultancy Services. *2021 - 2022*
- Learning Achievement Award, Tata Consultancy Services. *2021 - 2022*
- Ranked 2nd Topper in Academics, Guru Nanak Institutions. *2018 - 2019*
- Peer Reviewed three research papers - ORCID ID - https://orcid.org/0000-0002-5114-5205 *2019 - 2024*
- Google Scholar - https://scholar.google.com/citations?user=ZbVUfuYAAAAJ *2019 - 2024*

Certifications and Trainings

- Completed Deep Learning Specialization taught by Andrew Ng and other instructors. *2020*
- Completed altogether 10 Coursera courses on Machine Learning, Customer Analytics and more. *2020*
- Cyber Security Internship at Hyderabad City Police, trained on spam email detection, and more. *October 2019*
- Ethical Hacking, IOT, and Web Development training by Imparta. *June 2019*

Education

Florida State University, Tallahassee, Florida, United States
Master of Science in Computer Science (Thesis) *August 2024*

Jawaharlal Nehru Technological University – Guru Nanak Institutions, Telangana, India
Bachelor of Technology in Computer Science & Engineering *September 2020*