Aravind Balachandar

New York City | (716) 400-7802 | balacha2@buffalo.edu | https://www.linkedin.com/in/aravind4 | GitHub | Website

WORK EXPERIENCE

Zummit Africa *Wilmington, Delaware*
Machine Learning Engineer Intern *May 2024 – Aug 2024*

- Achieved a 40% increase in user interaction rates by implementing **distributed** and **scalable Retrieval-Augmented Generation** (**RAG**) system with **Pinecone** and **Elasticsearch**, driving a 15% revenue growth.
- Engineered and deployed a secure chatbot using **fine-tuned LLMs,** including **LLAMA 2** and **GenAI,** and applied advanced **NLU**, **NLP,** and **transformer** techniques, achieving a 90% accuracy in generating contextually relevant user responses.
- Designed **GAN-based** data augmentation pipeline, boosting model robustness and accuracy by 12% on imbalanced datasets.
- Customized a **diffusion model** for image synthesis, optimizing generation time by 30% to streamline visual content creation.

Quickplay Media | *Client - AMD, Rogers Sports & Media* *Chennai, India*
Software Engineer | Python & Golang Developer *Sep 2022 – Aug 2023*

- Developed **multi-camera stream switching** for live events, optimizing load balancing to reduce latency by 30% and enhance seamless viewing experience with **adaptive bitrate switching**.
- Spearheaded the design and architecture of a highly scalable **REST API microservice** for an OTT platform, reducing program guide data retrieval time and enhancing user-experience for **10M+ users**.
- Implemented real-time **data streaming pipeline** using **Apache Kafka**, **Spark Streaming**, and **AWS Kinesis,** improving data ingestion rates by 70% and enabling near-instantaneous processing of live OTT feeds.
- Enhanced **NoSQL DB** performance by 64% through **query optimization** and function refactoring, reducing application load time.

Accenture | *Client - British Telecom* *Chennai, India*
Application Development Associate | Python & Golang Developer *Nov 2020 – Sep 2022*

- Collaborated with **cross-functional teams** to design and build a **content aggregator** using **Kafka, Java**, and **Spring Boot**, yielding 71% reduction in code churn and improving **system scalability**.
- Optimized telecom data processing efficiency using **Apache Spark's** distributed **parallel processing**, improving processing speed by 45% for **real-time analytics**, facilitating quicker **decision-making**.
- Architected a high-availability **custom caching system** for **OAuth tokens** using **Vault, Redis** and **Nginx**, leading to 40% decrease in data access time and a 50% boost in system responsiveness under heavy load.
- Streamlined JotForm processing using **AWS Lambda, S3**, and **API Gateway**, curtailing costs by 30% through elimination of an **EC2 instance** reliance and effectively communicated cost-saving measures to stakeholder.

TECHNICAL SKILLS

Languages: Python, Go, Java, R, SQL, C (Data Structures and Algorithm), C++, JavaScript, React, Nodejs, Django, TypeScript.
Cloud & Databases: AWS (EC2, Lambda, SageMaker, RDS, EKS, IAM), HashiCorp, Azure, Postgres, MySQL, Redis, MongoDB.
OS & DevOps: Linux, Gitlab, Docker, Jira, Kubernetes, Jenkins, Kafka, Ansible, Redshift, Flask, RabbitMQ, Terraform, CI/CD.
Machine Learning tools: LLM, PyTorch, Spark, Hadoop, NumPy, NLP, Scikit-learn, Matplotlib, OpenCV, Tableau.

EDUCATION

University at Buffalo, The State University of New York *Buffalo, New York*
Master of Science in Data Science | GPA: 3.92/4.0 *Aug 2023 – Dec 2024*

Anna University *Chennai, India*
Bachelor of Engineering in Computer Science | GPA: 3.8/4.0 *Aug 2017 – Apr 2021*

PROJECTS

Traffic flow optimization using multi-agent RL | ***Tech stack: Reinforcement learning, DQN, A2C***

- Reduced vehicle wait time by 30%, increased traffic flow efficiency by 25% using **CUDA**-accelerated **Deep Q Network (DQN)** and **A2C algorithms** on **GPU**, and achieved 95% simulation accuracy with **SUMO** and **OpenStreetMap**.

Netflix Movie and Show Recommender | ***Tech stack: Python, TF-IDF, Streamlit***

- Led the development of **high-performance**, **multi-threaded** content-based recommendation engine using **TF-IDF** and **cosine similarity**, generating an 8000-feature similarity matrix for personalized recommendations of top 25 contents.

Multimodal Emotion Recognition on Facial Expression and EEG | ***Tech stack: Computer Vision, Deep Learning***

- Built a real-time emotion recognition system achieving 30 FPS with **OpenCV's Haar cascades**. Applied **transfer learning** with **ResNet15V2** to achieve 94% emotion classification accuracy after optimizing model architecture and hyperparameters.

PUBLICATIONS

Multimodal Emotion Recognition Based on Speech, Facial Expression and EEG.