

Sumanth D

+1 (469) 850-1729 | desireddysumanth77@gmail.com | www.linkedin.com/in/sumanth-d-773562361

PROFILE SUMMARY

Results-driven software engineer with a Master's in Computer Science from the University of New Haven, specializing in full-stack and Generative AI development. Proficient in Python, Flask, Spring Boot, FastAPI, JavaScript, Vue.js, Node.js, and LangChain. Experienced in designing multi-agent AI systems using Crew.ai, implementing RAG pipelines with LangChain, Pinecone, and Hugging Face embeddings, and integrating LLMs such as GPT-4, LLaMA, and Mistral. Skilled in fine-tuning models using LoRA/QLoRA. Strong background in deploying scalable applications on AWS (Lambda, ECS), containerization with Docker, and managing data with MySQL and MongoDB. Adept in Agile workflows, DevOps practices, CI/CD pipelines, and building robust, real-time, and user-centric AI applications in cross-functional environments.

CORE SKILLS

Languages: Python, Core Java, JavaScript, SQL, HTML, CSS

Frameworks & Libraries: Spring Boot, Flask, FastAPI, Node.js, Vue.js, LangChain, LangGraph, Crew.ai

AI/ML: GPT-4, GPT-3.5, LLaMA, Mistral, Pinecone, FAISS, Hugging Face, RAG, Scikit-learn, NLTK, OpenAI Whisper

Cloud & DevOps: AWS (Lambda, ECS, EC2, S3), Docker, Prometheus, Grafana, Netlify

Databases: MySQL, MongoDB

Tools: Git, Postman, XAMPP, VSCode, VMware

Other Skills: Agile/Scrum, Critical Thinking, Team Collaboration, Problem Solving.

PROFESSIONAL EXPERIENCE

Software Engineer - Generative AI

KKRGenAI Innovations | Sep 2024 - Present
Alpharetta, GA.

KKRGenAI Innovations is a healthtech startup focused on building domain-specific Generative AI solutions to automate clinical workflows and enhance patient care.

- Collaborated with cross-functional teams to design and develop domain-specific Generative AI-based virtual health assistants using GPT-4, LLaMA, and Mistral, supporting clinical workflows such as appointment booking, patient support, and triage automation.
- Developed intelligent symptom triage workflows by orchestrating AI agents using Crew.ai and LangGraph, integrating LangChain and Pinecone for efficient retrieval and reasoning over clinical data.
- Integrated speech recognition capabilities using OpenAI Whisper, allowing patients to interact via voice input. This improved accessibility and user experience by converting spoken symptoms and queries into accurate text for AI processing.
- Participated in fine-tuning large language models using LoRA and QLoRA, ensuring improved contextual accuracy and domain relevance.
- Used Hugging Face embedding models (e.g., sentence-transformers) alongside OpenAI embeddings to generate semantic representations of clinical documents for vector search in the RAG pipeline.
- Part of backend team in which we built secure Flask-based APIs, containerized with Docker, and deployed on AWS Lambda and ECS for cloud-native scalability and fault tolerance.
- Engaged in iterative testing and refinement cycles based on collaborative feedback from product managers, QA engineers, and healthcare domain experts.

AI/ML - Generative AI

RiteCare | July 2023 - Aug 2024
Irvine, CA.

RiteCare is an AI-driven healthcare research company developing intelligent virtual assistants and NLP solutions for improving healthcare access and efficiency.

- Worked as part of a research team to build a healthcare-focused chatbot powered by GPT-3.5, leveraging LangChain and Pinecone for semantic search and dynamic reasoning.
- Utilized FAISS for local vector database testing and semantic retrieval during chatbot prototype development.
- Developed backend APIs using FastAPI to handle real-time model interactions.
- Took part in prompt engineering sessions and experimental tuning efforts, aligning model behavior with medical data requirements.
- Evaluated chatbot performance across domain-specific use cases and documented LLM behavior.
- Took part in prompt engineering sessions and experimental tuning efforts, aligning model behavior with medical data requirements

Software Development Intern

Intellectuals AI Solutions | Aug 2021 - July 2022
Hyderabad, India.

Intellectuals AI Solutions is a technology company specializing in developing AI-driven and web-based software solutions, with a focus on backend systems, API development, and scalable application architecture for enterprise clients.

- Built RESTful APIs using Spring Boot (Java) and Flask (Python) for core application modules.
- Collaborated with the database team to design schemas and write optimized queries for MySQL and MongoDB, enhancing data consistency and retrieval speed.
- Participated in Agile ceremonies including **daily stand-ups**, **sprint planning**, and **retrospectives**, ensuring alignment across development and QA teams.
- Engaged in peer code reviews and collaborative debugging sessions to maintain code quality and resolve production issues.
- Coordinated with the frontend team to ensure seamless API integration, improving application responsiveness and usability.

ACADEMIC PROJECTS

Web-Database Application Development

(2024)

University of New Haven

- Developed a full-stack web application to manage and interact with dynamic data.
 - Built the frontend using Vue.js for a responsive and interactive user interface.
 - Implemented backend services with Node.js and Express to handle data requests efficiently.
 - Created and exposed REST APIs for seamless data interaction.
 - Utilized MongoDB for data storage and management.
 - Deployed the application on Netlify for easy access and scalability.
 - Showcased expertise in full-stack web development, including frontend, backend, and database integration.
- Github link: https://github.com/sumanthreddy8910/sumanth_vue.git
https://github.com/sumanthreddy8910/sumanth_node.git

Facebook Spam Detection Using Data Mining Techniques

(2024)

University of New Haven

- Designed and implemented a machine learning system to detect and categorize spam on Facebook.
- Utilized natural language processing (NLP) to analyze user-generated content such as comments and messages.
- Trained and evaluated models using a Facebook public dataset to identify spam patterns and behaviors.
- Applied data preprocessing techniques to handle missing values and optimize dataset quality.
- Tested multiple classifiers, achieving high accuracy in detecting spam across various models.
- Visualized results using boxplots, heatmaps, and ROC curves to evaluate model performance.
- Conducted comprehensive evaluation to compare classifier performance using AUC and other metrics.

Github link: https://github.com/sumanthreddy8910/Final-Report-Data_Mining.git

Full-Stack Web Application

(2024)

University of New Haven

- Developed a full-stack web application enabling users to interact with and manage data dynamically.
- Designed a responsive and intuitive user interface for seamless user interactions.
- Built backend services to handle data management and integrated REST APIs for smooth operations.
- Implemented efficient data storage and retrieval using MongoDB.
- Deployed the application on Netlify, ensuring scalability and easy access.

Live Demo Link: <https://sumanth-reddy.netlify.app/>

Steganography – A technique to hide data

(2023)

University of New Haven

- Explored and implemented steganography techniques for secure data communication.
- Embedded sensitive data within media formats like images and audio files to maintain confidentiality.
- Analyzed real-world applications, advantages, and challenges of steganography in fields like cybersecurity and military communication.
- Experimented with detection methods to evaluate the robustness of data hiding techniques.
- Demonstrated the project using BMP and WAV media formats for testing data embedding.

Github link: <https://github.com/sumanthreddy8910/Steganography---Research-Paper.git>

Dentist-Polyclinic-Management-System-DBMS

(2022)

University of New Haven

- Developed a comprehensive dentist polyclinic management system to automate patient registration, appointment scheduling, and billing.
- Implemented features like patient history tracking, automated billing with discounts, and detailed reporting to improve operational efficiency.
- Designed a MySQL database with schema design, ER modeling, and SQL queries to manage and manipulate patient data.
- Streamlined workflow, reducing administrative overhead and minimizing errors in data handling.

Github link: <https://github.com/sumanthreddy8910/Dentist-Polyclinic-Management-System-DBMS-.git>

SRM University

- Developed a machine learning-based system to detect fake news using multiple textual features.
- Implemented an ensemble model incorporating Naive Bayes, Logistic Regression, and K-Nearest Neighbors for classification.
- Achieved 94.5% accuracy using cross-validation on real-world datasets.
- Used Python libraries such as Scikit-learn, Pandas, and NLTK for data processing and model development.
- **Published in:** International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 10, Issue III, March 2022.
- **Certifications:** Received a certification for successful publication and contribution to the research community.
Github link: <https://github.com/sumanthreddy8910/Learning-Based-Fake-News-Detector-Using-Multiple-Features-Using-Machine-Learning.git>

EDUCATION**University of New Haven**

Master of Science in Computer Science
West Haven, CT | 2022 - 2024
GPA: 3.71/4.0

SRM University

Bachelor of Technology in Computer Science and Engineering
Chennai, India | 2018 - 2022
CGPA: 9.45/10.0

CERTIFICATIONS**Oracle Cloud Infrastructure 2025 Certified Generative AI Professional**

Oracle — Credential ID: 103014699OCI25GAIOCP
Issued: October 2025