Vedant Sahai

📞 814-769-0201　✉ vedantsahai18@gmail.com　in linkedin.com/vedantsahai18　○ github.com/vedantsahai18

Experience

Julep AI Inc.　　　　　　　　　　　　　　　　　　　　　　　　　　　Oct 2024 – Present
Machine Learning Engineer　　　　　　　　　　　　　　　　　　　　　　New York City, NY
- Enhanced Julep AI's platform infrastructure ○ usability by creating 15+ cookbooks thereby streamlining workflows.
- Improved platform performance by 15 % through rigorous experimentation with algorithms and optimization techniques.
- Reduced processing time by 20% by developing and deploying routes for seamless AI Agent interactions with external integrations.

JP Morgan Chase & Co.　　　　　　　　　　　　　　　　　　　　　　Jun 2023 – Aug 2023
AIML Summer Associate　　　　　　　　　　　　　　　　　　　　　　　Wilmington, DE
- Increased the Transaction Detection Rate gain by 100-150 basis points by implementing the Online ML XGBoost algorithm for the Transaction Risk model (TRS).
- Attained 90% accuracy by programming a TabNet-based Deep Neural Network as a challenger for the TRS XGBoost model.
- Ensured 95% code coverage by implementing a PyTest-based testing framework for the TRS Feature Engineering codebase.

Plexflo　　　　　　　　　　　　　　　　　　　　　　　　　　　　　Oct 2021 – Jul 2022
ML Engineer　　　　　　　　　　　　　　　　　　　　　　　　　　　　Mumbai, India
- Developed Evidence, a Meter Data Management & Analytics (MDMS) software with a latency of less than 90ms, by leveraging ITRON, Sensus Xylem and Siemens data streams, powered by AWS, Apache Flink, and a custom ML model.
- Achieved an F1 score of 85% for Plexflo AI, an open-source Python library by leveraging a Variational Autoencoder for Non-Intrusive Load Monitoring.
- Scaled a FastAPI + AWS Timestream backend to support up to 20,000 IoT devices for MDMS over Grafana.
- Enhanced rooftop solar assessment accuracy by 30%, using Mask R-CNN and Geo-Spatial Image Processing techniques.

Sync Energy AI　　　　　　　　　　　　　　　　　　　　　　　　　Jul 2020 – Sept 2021
ML Research Intern　　　　　　　　　　　　　　　　　　　　　　　　Mumbai, India
- Optimized the power outage extraction, resulting in a 40% faster response, by deploying an AWS Lambda-Python-REST API.
- Improved accuracy to 83% in estimating the locations of utility poles from Google Street View images by employing Mask R-CNN and Image Processing.
- Boosted research capabilities by generating Neo4J-based knowledge graphs from research papers on wildfires and their effects.

Technical Skills

Languages: Python, C, JavaScript, Java, HTML5/CSS3, LaTeX
ML Frameworks: PyTorch, Keras, TensorFlow, RASA, Scikit, XGBoost, HugginFace, NLTK, Spacy, Pandas, Langchain, Autogen
Databases: MySQL, PostgreSQL, MongoDB, Neo4J
Cloud: Amazon Web Services [Ec2, S3, Lambda, API Gateway, Sagemaker, IAM]
Technologies: Django, React.JS, Flask, FastAPI, PyTest, Elasticsearch, Git, Docker, PySpark, Grafana, Sphinx, Temporal

Education

Pennsylvania State University　　　　　　　　　　　　　　　　　　　May 2024
Master of Science in Computer Science & Engineering (GPA: 3.75 / 4.00)　　　University Park, PA
- **Teaching Assistant:** CMPSC 132: Programming and Computation II: Data Structures
- **Relevant Coursework:** Computer Vision, Operating Systems, Data Structures & Algorithms, NLP, Intro to Deep Learning, Machine Learning and Algorithmic AI, Vision & Language, Computer Security, Topics in Computer Architecture

University of Mumbai　　　　　　　　　　　　　　　　　　　　　　May 2021
Bachelor of Engineering in Computer Engineering (CGPA: 9.58 / 10.00)　　　Mumbai, India

Projects

Datacertus ○　　　　　　　　　　　　　　　　　　　　　　　　　　May 2022
Reacts. JS, AWS APIs, Python, Docker, AWS ELB, AWS Lambda, DynamoDB, Scikit, RASA, PyTorch, Keras, Scikit-Learn
- Trained a BERT summarization model to summarize updates on disaster information with 75% ROGUE-L.
- Reduced update frequency by over 50% by developing a pipeline using Lambda, Selenium, and the BERT model to track natural calamities.
- Made data processing 10% faster by integrating data pipelines using Lambda, S3, and API Gateway-based trigger events.

Conversational AI for Secure Healthcare Assistance ○　　　　　　　　　　May 2021
Docker, Node.JS, MongoDB, RASA, Blockchain, Python, Express.JS, jQuery, AJAX, Vault, Nginx
- Handled over 10,000 EHR records concurrently by architecting a Docker-NodeJS-MongoDB-Vault-based software.
- Attained the intent prediction confidence up to 95% by developing and integrating a RASA-powered therapy chatbot.
- Integrated the BigchainDB + IPFS blockchain database with the RASA for behavior analysis and secure medical record storage.

Publication

Vedant S., Jason D., Mayank S., Mahendra M., Dhananjay K. (2021) Leveraging Deep Learning and IoT for Monitoring COVID19 Safety Guidelines Within College Campus. In: Garg D., Wong K., Sarangapani J., Gupta S.K. (eds) Advanced Computing. IACC 2020. Communications in Computer and Information Science, vol 1367. Springer, Singapore. https:doi.org10.1007978-981-16-0401-0_3