

Professional Summary

- AI/ML Engineer with 2.5+ years of experience in designing, developing, and deploying scalable machine learning solutions across NLP, computer vision, forecasting, and generative applications, delivering measurable impact in production environments.
- Expert in creating full-stack ML pipelines using Python, Scikit-learn, Pandas, MLflow, and XGBoost; experienced in data cleaning, feature selection, model tuning, and containerized deployments with Docker, FastAPI, and AWS SageMaker.
- Skilled in building and fine-tuning deep learning models using PyTorch, TensorFlow, and Keras for OCR, image recognition, time series analysis, and large-scale classification; optimized for inference speed and GPU memory efficiency.
- Proficient in applying transformer architectures such as GPT-4, BERT, LLaMA, and T5 for NLP tasks like semantic search, document summarization, text classification, and Retrieval-Augmented Generation (RAG) pipelines using LangChain and Hugging Face.
- Strong knowledge of Generative AI tools including OpenAI API, Stable Diffusion, DALL-E, and prompt engineering techniques; coupled with advanced understanding of MLOps, version control (DVC), experiment tracking, SQL, and visualization using Power BI and Tableau.

Education

Master of Science in Computer Science (GPA: 3.42/4) |Purdue University ,Hammond, Indiana

Skills

Languages & IDEs: Python, R, Java(core), SQL, Jupyter Notebook, Google Colab, VS Code, SSMS

DPackages and Frameworks: NumPy, Pandas, Matplotlib, Scikit-learn, Seaborn, TensorFlow, Keras, NLTK, XGBoost, PyTorch, OpenCV, CUDA

Machine Learning: Linear, Logistic Regression, Decision Trees, Random Forests, A/B Testing, Naive Bayes

Deep Learning: CNN, RNN, LSTM, NLP, OCR, Large Language Model (LLM), LangChain, CrewAI, LangGraph, Hugging Face Transformers (BERT, GPT-4, GPT-3, LLaMA, T5)

Generative AI : Retrieval-Augmented Generation (RAG), Prompt Engineering, Fine-Tuning, Semantic Search, Stable Diffusion, DALL-E

Cloud/Visualizations: AWS(EC2,SQS,SNS,Code Deploy,CloudWatch,API Gateway), GCP(Vertex AI,Google Cloud Storage), Tableau, Power BI

MLOps & Tools: Docker, FastAPI, MLflow, DVC, CI/CD

Statistical Techniques: Hypothesis Testing, Data Visualization, Data Modelling, A/B testing, Model Evaluation

Database and Tools: SQL Server, MySQL, PostgreSQL, Redis

Certifications: Oracle Cloud Infrastructure 2025 Certified Generative AI Professional (Oracle), DevOps for Data Scientists, Data Analysis with Python,Python (TCS YouthEmployment Program), Cloud Computing (Cognizant Digital Nurture)

Work Experience

CyberInfo Tech, Hagerstown, Maryland

Aug 2025 – Present

Software Developer Intern (AI/ML)

- Developed and deployed **LLM-powered intelligent applications** using OpenAI GPT-4 and LangChain, integrating advanced prompt engineering techniques for high-quality, context-aware AI responses.
- Built and orchestrated **multi-agent collaboration systems** with CrewAI and LangGraph, enabling intelligent coordination between autonomous agents for data analysis, summarization, and strategic decision support.
- Implemented retrieval-augmented generation (RAG) pipelines to enhance contextual accuracy and reduce hallucinations in AI outputs.
- Designed RESTful APIs using FastAPI for seamless AI service integration within enterprise workflows.
- Collaborated with cross-functional teams to optimize model performance, scalability, and latency, improving response efficiency by 20%.

Wipro Limited, India

May 2022– Sep 2023

Project Engineer(AI/ML Engineer)

- Successfully deployed multiple machine learning models on AWS EC2 by building automated retraining workflows, setting up monitoring tools, and integrating CI/CD pipelines. This led to a 30% improvement in deployment stability across production environments.
- Refined a BERT-based sentiment analysis model on over 20,000 customer feedback entries. The project boosted classification accuracy by 22% and equipped the UX and product teams with key emotional insights for better decision-making.
- Embedded OpenAI GPT-3 with prompt engineering and few-shot learning to generate real-time summaries of lengthy customer logs. This initiative reduced average response time by 41% and significantly enhanced customer support performance.
- Engineered a scalable document classification pipeline combining OCR technologies with TensorFlow CNNs. This system processed more than 9,000 scanned documents monthly and reduced compliance-related turnaround time by 50%.
- Created custom NLP pipelines using Hugging Face Transformers and NLTK for entity recognition and keyword extraction, enabling faster analysis of qualitative survey data and helping marketing teams identify emerging consumer trends 18% quicker.
- Designed and maintained interactive Tableau dashboards to present NLP outputs such as sentiment drift, topic clusters, and model confidence levels. These dashboards provided business stakeholders with actionable insights at a glance.

Wipro Limited, India

Mar 2022 – May 2022

Intern(AI/ML Engineer)

- Assisted in developing and deploying deep learning models using CNNs and attention-based layers to extract structured data from handwritten medical forms, helping automate 7,000+ patient records and reduce manual entry by 65%.
- Contributed to building LSTM and OpenCV-based surveillance models for ICU and elderly care units, supporting early detection of fall risks and abnormal movements, which reduced manual video review time by over 50%.
- Helped streamline the model training workflow by automating hyperparameter tuning in TensorFlow, improving training efficiency and model iteration speed by 40%.
- Created visual analytics dashboards in Power BI to monitor model accuracy and prediction trends, improving performance visibility for the clinical informatics team.

Projects & Trainings

- Pre-Train a GPT from Scratch to Build VetGPT:** Pre-trained a GPT model from scratch using PyTorch and CUDA on 4 GPUs, achieving 82% accuracy on veterinary diagnostic data, and curated a dataset of 50,000+ records using custom web crawlers for training.
- TCS Youth Employment Program (YEP) – Software Development Training (2021):** Completed Python programming and software development training, covering coding fundamentals, OOP concepts, and small-scale web application development.
- Jio Robotics Training – Robotics & ML Integration (2025):** Worked on robotics applications with AI/ML integration for real-time automation, implementing sensor data processing and model-based control systems to improve robotic accuracy and efficiency