Deepak Kumar Kushwaha

LEAD MACHINE LEARNING ENGINEER (MLE) - DELOITTE AI COE

9582854595, 9650540890 | deepakkushwaha818@gmail.com | https://www.linkedin.com/in/deepak-kushwaha818 | Delhi - NCR

SUMMARY

- **8.8 years** of experience as a **Lead & Senior Machine Learning Engineer** specializing in designing, developing, and deploying scalable, production-grade AI/ML solutions using robust MLOps best practices.
- Expert in **NLP**, **Generative AI, Large Language Models (LLMs), Vector Databases, Retrieval-Augmented Generation (RAG), Text-to-SQL**, and Industry data integration with LLMs.
- Extensive experience **fine-tuning, optimizing, and deploying LLMs** using state-of-the-art tools and frameworks including **AWS Bedrock, HuggingFace, LangChain, OpenAI APIs, TensorFlow, PyTorch, and NVIDIA RAPIDS.**
- Expertise in **Graph Neural Networks (GNNs), Reinforcement Learning (RL), ensemble methods (XGBoost, Random Forests), and anomaly detection techniques**, optimized for cybersecurity applications, threat detection, and behavioral analytics.
- Skilled in designing **scalable, distributed, cloud-based AI infrastructure (AWS, GCP)** leveraging GPU acceleration for training and real-time inference.
- **Demonstrated leadership in managing cross-functional AI teams**, driving technical strategy, and successfully delivering high-impact AI solutions for global industry leaders including **Airbus, Amazon, NVIDIA, AIG, NYL, and CrowdStrike.**
- **Published and co-authored 5 peer-reviewed research papers** in Cybersecurity and Reinforcement Learning, with active collaboration with academic institutions to pioneer novel AI-driven solutions.

PROFESSIONAL EXPERIENCE

Lead MLE (AI Centre of Excellence)

Deloitte (Offices of US in India)

05/2021 - Present | Gurugram

- Built and deployed **4 LLM, RAG, VectorDB, GenAI-powered** solutions focusing on **threat alert summarizations, vulnerability detection, document compliance management, and intelligent chatbot systems**
- Used Tools & **Service like AWS Bedrock and OpenSearch, Neo4J, GraphRAG to** deploy LLM solutions at scale
- Developing **real-time, high-impact Cyber GenAI solutions with third party tools like NVIDIA Morpheus, NIMs, NeMO** enhancing decision-making
- Designed & Built **4 Novel Cyber Threat detection Machine Learning solutions (LM,ZeroDay and Ransomware detection)** using Unsupervised **Parametric learning and Deep Graph** approaches
- Optimized ServiceNow ticket assignment processes using advanced **optimization algorithms**
- **Managed and mentored teams**, overseeing project planning, task allocation, and successful execution within deadlines
- Presented **Cyber AI solutions to US-DoD and at NVIDIA GTC**

Senior Data Scientist

Sopra Steria India

04/2019 - 04/2021 | Noida

- Improved pipeline orchestration efficiency by 25% by **benchmarking 5 state-of-the-art AutoML tools** for structured and unstructured data
- Increased efficiency by {15%} by leading research in NLP and pipeline orchestration technologies for MLOps tools
- Interfaced with AIRBUS (client), optimizing product delivery timelines through client-facing engagements

WHERE DO I SPEND MY TIME

Strategic Project Planning, Architectural designing & Identifying Novel techniques

Data Exploration, ML Algorithmic Design, Pipeline Development & Deployment

Team Leading, Mentoring, Stakeholder Engagement & Presentation

TECHNICAL SKILLS

Generative AI Frameworks

LLM, RAG, Graph RAG, LighRAG, Text-to-SQL, Agentic AI, Prompt Engineering

Generative AI Tools

AWS Bedrock, Langchain, OpenAI APIs, NVIDIA NIMS, NeMO

Databases [SQL, NoSQL and Vector]

MongoDB, AWS Athena, GCP BigQuery, PostgresSQL, Qdrant, Milvus

NLP (Natural Language Processing)

Chatbot, NER, BERT, Transformers, LLMs

Accelerated and Distributed AI/ML

Rapids cuDF, cuML, NVIDIA TensorRT, Dask-cuDF, Dask, Apache Spark

Graphs

GNN, Neo4J, PyTorch Geometric, NetworkX

Machine Learning Frameworks

TensorFlow, PyTorch, XGBoost, Keras, MLFlow, Pandas, Scikit-Learn

Cloud Platforms & Service

AWS, Azure, GCP, DataBricks, Lambda, Sagemaker, Glue, Athena, Redshift, Bedrock

Model Deployment and Serving

TF Serving, NVIDIA Triton Inference Server, AWS Infrentia, Flask API

Containerization and Orchestration

Docker, GitHub Actions, Kubernetes, Kubeflow, NVIDIA Morpheus, Airflow

PROFESSIONAL EXPERIENCE

Member and Senior Member of technical staff

NEC Technologies India

10/2016 - 03/2019 Noida

- Led a team of 5 in AI solution design and deployment, driving innovation
- Converted client requirements into tailored AI solutions using Python and Deep Learning, NLP and Reinforcement Learning
- Applied **NLP, Word embeddings and Similarity analysis** to solve NLP problems

PROJECTS

LLM Powered Automated Vulnerability prioritization and remediation + Chatbot

Deloitte

Building a **RAG powered solution** to **prioritize vulnerabilities, suggest remediations, and assist the Vulnerability Manager** through an **interactive chatbot** deployed at the client site

- Reduced Patch resolution time by {20%} using AWS Bedrock, OpenSearch, Lex, and Lambda integration
- Used correlation and meta-data to improve retrieval accuracy

LLM enabled Graph RAG pipeline for Client Policy document assessment & recommendation

Deloitte

Utilizing a LightRAG pipeline with LLMs (Claude 3.5 & Titan text embedding), Milvus, Neo4J and PostgresSQL to assess client policy documents, identifying current state, gaps, and recommendations by comparing them against established security frameworks like NIST

- Processed complex documents across formats like PDF, XLSX, and DOCS
- Improved global context retrieval accuracy by {25%} utilizing LightRAG

Threat Report Generation and Summarization using NVIDIA LLMs and RAG pipelines

Deloitte

Led the team to build and deploy real time threat alert prioritization, summarization and enrichment to enhance SoC Analysts experience and expedited threat hunting

- Integrated AI models, reducing data processing time by {5x}, using NVIDIA Morpheus, Llama LLMs, and Reranker models
- Deployed the model on 4 clients in production and caught 5 detections

Real Time Lateral Movement Detection in Enterprise Networks

Deloitte

Built and deployed end-to-end solution for AI enabled Lateral Movement threat detection capturing users, assets and accounts behaviours to identify malicious behaviours using graphs

- Reduced incident triage time by {70%} for {5} clients, improving threat detection speed significantly
- State-of-the-art User and Entity Behaviour modelling to identify indicators of compromise in real time

Zero Day Threat Detection in Enterprise Networks

Deloitte

Built and deployed an AI model, implementing state-of-the-art encoder decoder architectures coupled with graph features Zero Day Threat Detection

- Increased data throughput by {20%} by optimizing log processing with NVIDIA Morpheus
- Deployed successfully on 3 clients including IOC(International Olympics)

TECHNICAL SKILLS

Model Monitoring Tool

MLFlow, Robust Intelligence

CERTIFICATION

NVIDIA DLI Certificate – Building AI-Based Cybersecurity Pipelines

NVIDIA Deep Learning Institute

NVIDIA DLI Certificate - Fundamentals of accelerated computing with CUDA Python

NVIDIA Deep Learning Institute

NVIDIA DLI Certificate – Fundamentals of Accelerated Data Science

NVIDIA Deep Learning Institute

PUBLICATIONS

Discovering exfilteration paths using reinforcement learning with attack graphs

IEEE Conference on Dependable and Secure Computing (DSC)

01/2022 https://arxiv.org/abs/2201.12416

Exposing surveillance detection routes via reinforcement learning, attack graphs, and cyber terrain

21st IEEE International Conference on Machine Learning and Applications (ICMLA)

12/2022

https://ieeexplore.ieee.org/document/10069285

Lateral Movement Detection Using User Behavioral Analysis

arXiv preprint arXiv:2208.13524

08/2022

Zero Day Threat Detection Using Graph and Flow Based Security Telemetry

International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2022

https://ieeexplore.ieee.org/iel7/10037246/10037222/10037596.pdf

Cross-Platform Lateral Movement Detection via Unsupervised Graph Machine Learning

International Symposium on Digital Forensics and Security (ISDFS 2025)

PROJECTS

LLM Chatbot to Enhance Client Onboarding workshop experience

Deloitte

Built a chatbot using Microsoft Bot Framework, RASA and LLMs to enable real time responses for the workshop on aligning clients on proposed security roles and data access

- Enhanced end user's experience and reduce the workshop time by 50%

RL for Pentesting

Deloitte

Exposing Surveillance Detection Routes via Reinforcement Learning, Attack Graphs, and Cyber Terrain

- Generated {20%} increase in academic citations by publishing research materials
- Published 3 papers as eminence for the research work

AI enabled Requirement Doc generation

Deloitte

AI assisted Auto-generation of requirements and test cases to accelerate app integration/onboarding

- Reduced client onboarding time by {80%}, saving approximately {2-5 days} per client, resulting in increased client satisfaction.

Custom NER for Invoices

Sopra Steria

- Enhanced invoice processing efficiency by {30%} through custom CNN-CRF entity extraction techniques
- Improved entity prediction accuracy by {25%} using CNN-CRF model on labeled invoice datasets

Image Denoising using Auto-Encoders for Airbus

Sopra Steria

- Increase sharpness of image, Remove unwanted lines and Remove blurriness in invoice images by 70% using Auto-encoders

Product Delivery Date prediction for Airbus

Sopra Steria

- Created a Deep learning-based solution for prediction of spare parts delivery dates with 95% Accuracy
- Enhanced operational efficiency, integrating prediction engine API (FastAPI) into Airbus UI, reducing processing time by {30%}

AutoML Tools Benchmarking

Sopra Steria

Benchmarked multiple AutoML tools available in the market (offline tools as well as AutoML as Service) over different domains of ML problems (Supervised learning, Unsupervised learning, Natural Language processing)

- Generated Benchmarking Report for tools like H2O.ai, TPOT, Dataiku, AutoKeras, AWS Autopilot, Cloud ML and Auto AI

Automated System Integrations Tool

NEC Technology

Automated Software identification process with provided data (Functional and Non-Functional Requirements) using ML & DL.

- Enhanced AI accuracy by {15%} using user feedback with reinforcement learning.
- Enhanced user requirement accuracy by {25%} using NLP and Entity Extraction Automation.

KEY ACHIEVEMENTS

- **Deployed Lateral movement threat detection capability to IOC (International Olympics Council) used in 2024 Olympic Games**
- **Work on Cyber AI was presented in NVIDIA GTC 2024 and 2025**
- **Presented cybersecurity solutions to U.S. Department of Defence (D.o. D) and secured 2nd position in the list to get contract**
- **International exposure during my on site work in JAPAN**
- **Worked with NVIDIA to collaborate on using and improving NVIDIA Morpheus orchestration tool**

EDUCATION

MBA in Business Analytics

IMT Ghaziabad

2021 - 2023

B.Tech. Information Technology

H.B.T.U Kanpur

2012 - 2016

Senior Secondary Examination

Kendriya Vidhyalaya

2010 - 2011

AWARDS

- **Pinnacle award at Sopra Steria India (2019)**
- **SPOT Award and Best Debut Award at NEC Technologies (November 2017)**

TRAININGS / COURSES

Generative AI with Large Language Models

deeplearning.ai with AWS

Implementing Azure Cognitive Services with Microsoft Azure Bot Framework

Udemy with Packt

Learn SQL for Data Analysis with Google Big Query

Udemy online course