# Rahul Yadav

+1(716)238-0835 | rahulyad@buffalo.edu | linkedin.com/in/rahul-yadav-704 | Github | Website

## EDUCATION

**University at Buffalo, Buffalo** — Aug. 2022 – Dec. 2023
*Master of Science in Computer Science and Engineering* — New York, US
Coursework: NLP, Computational Linguistics, Pattern Recognition, Machine Learning, Algorithms Design — GPA: 3.82 / 4

**Indian Institute of Technology (Banaras Hindu University), Varanasi** — Jul. 2014 – May 2018
*Bachelor of Technology in Electronics Engineering* — Uttar Pradesh, India

## EXPERIENCE

**Hilabs** — Feb. 2024 – May 2024
*Senior Data Scientist | Applied Data Science Team* — Washington, D.C.

**GenAI Driven Contract Analyzer: Streamlining Claim Processing**

- Led development of scalable end-to-end system for processing contract documents of insurance providers, emphasizing entity extraction to facilitate pricing configuration in claim processing applications
- Engineered custom document processing pipeline with LayoutLM extracting entities with images, tables, and text elements
- Developed Langchain service for indexing extracted elements and managed metadata within OpenSearch to optimize retrieval
- Implemented a robust infrastructure leveraging self-hosted fine-tuned Mistral AI LLM and asynchronous processing via SQS to efficiently handle requests, while prioritizing the utmost security and privacy of sensitive legal contracts
- Successful POCs followed by integration for 3 clients to validate solution effectiveness and potential for widespread adoption

**Apexanalytix** — May. 2023 – Feb. 2024
*Data Scientist | ML Engineer Intern | Analytics Team* — Remote, New York

**Generative Knowledge Specific Chatbot**

- Developed advanced Retrieval Augmented Generation(RAG) Chatbot with LLM for intelligent knowledge access across teams
- Engineered efficient retrieval pipeline with Parent Child document indexing using LangChain and Chroma VectorDB
- Designed agile RAG pipeline, integrating MMR scoring for chunk retrieval and Azure OpenAI LLM for response generation
- Implemented user feedback collection to monitor chatbot performance and gather data for iterative refinement and fine-tuning
- Achieved 89% approval in human evaluations and integrated technology into 12 internal and 18 external client applications

**Oracle** — Sept. 2020 – Aug. 2022
*Senior Application Engineer(ML) | Oracle Financial Machine Learning Team* — Bengaluru, India

**Preemptive Anomaly Prediction in Corporate Billing**

- Implemented in-memory multivariate anomaly prediction system for corporate billing, addressing monthly billing challenges
- Leveraged Oracle in-database ML for Semi-Supervised classification with both local and global model explainability
- Optimized service with indexing and parallelism, processing 1.2M bills and 5M segments in 20 mins with 92% precision
- Integrated services with Oracle Revenue Management and Billing (ORMB) product, **USPTO patent granted** [US17/710745]

**Wipro Limited** — June 2018 – Sept 2020
*Project Engineer(AI) | AI Research Team* — Bengaluru, India

**Chatbot Services for Employee Helpline Portal's Ticketing System**

- Developed Employee Helpline Portal Chatbot with effective retrieval of historical ticket resolutions for enhanced user support
- Implemented query intent classification and BERT-powered semantic search to deliver precise responses
- Successfully integrated the chatbot into the portal and achieved the target human agent intervention reduction of 70%
- Achieved 79% accuracy score in evaluation and decreased wait times from 18 to 4 minutes, improving overall user experience

## PATENTS and PUBLICATIONS

**Paper Published:** Virtual Conversation with Real-Time Prediction of Body Moments/Gestures — **ICMLIP 2019**
*Gopichand Agnihotram, Rajesh Kumar, Pandurang Naik, **Rahul Yadav*** — springerLink

**Patent Granted:** Method And System For Multimodal Analysis Based Emotion Recognition — **US16/795840** [Link]
**Patent Granted:** Technology System For Assisting Financial Institutions In Debt Collection — **US17/659017** [Link]

## SKILLS

**Languages** : Python, Java, C++, C, PLSQL, SQL, MongoDB

**Technologies** : Retrieval Augmented Generation(RAG), Large Language Models(LLMs), Natural Language Processing(NLP), Machine Learning, Deep Learning, Web Services, Data Structures, Algorithms, Prompt Engineering, Indexing, Quantization

**Cloud** : Amazon Web Services (AWS), Azure, Google Cloud Platform (GCP), Oracle, Atlas, OpenAI

**Frameworks and Libraries** : Pytorch, Tensorflow, MLFlow, Kafka, SentenceTransformers(embeddings, re-rankers), Databricks, Snowflake, PySpark, Langchain, LlamaIndex, VectorDBs(ChromaDB, Faiss, elasticsearch), MLFlow, Git, Jenkins, Big Data, Hadoop, Flask, Docker, Kubernetes, Pandas, Streamlit, Flask-RESTful, FastAPI, YOLO, Django, XGBoost, GAN, ActiveMQ, Springboot