Rissi Kumar Prabhakaran

rissikumarp@gmail.com | (303)-591-9117 | Boulder, CO (Open To Relocate)
https://github.com/RISSIKUMARP | https://linkedin.com/in/rissi-kumar-prabhakaran |
https://rissikumarp.github.io/rissikumar_portfolio

Education

University of Colorado Boulder — Boulder, CO
Master of Science in Data Science; GPA: 3.7/4.0 — *Aug 2024 - May 2026*

Anna University — Chennai, India
Bachelor of Technology in Artificial Intelligence and Machine Learning; GPA: 3.4/4.0 — *June 2020 - May 2024*

Experience

Kashmir World Foundation [Link] — May 2025 - Present
AI Data Scientist Intern — *Great Falls, Virginia*

- Deployed AI pipelines using **BirdNET** and **YamNET** for bioacoustic data preprocessing, processing **100,000+ audio samples** through segmentation and species analysis with spectrogram integration, improving model precision by **11%**.
- Developed an Autonomous Vehicle Assistant, a multi-agent system for drone operations utilizing **Agno framework** for agent orchestration, real-time navigation, telemetry monitoring, and multimodal data analysis across **3+ coordinated agent workflows**, generating structured status reports for mission decision support.
- Orchestrated end-to-end drone telemetry pipelines connecting **ArduPilot SITL** to LLM-based agents via **DroneKit**, **MAVLink protocol**, and TCP relay systems.
- Architected multi-agent architectures with **OpenRouter API**, reducing inference cost by **21%**, implementing context engineering and designing **RAG knowledge systems** with vector DB for agent context retrieval.

Vlog Innovations [Link] — January 2023 - June 2023
AI Engineer Intern — *Chennai, India*

- Engineered a document ingestion and vector indexing pipeline using **FAISS** and **sentence-transformer embedding models**, processing and embedding internal training documents through batch transformation scripts for an internal knowledge search prototype.
- Evaluated semantic search quality across multiple retrieval configurations, benchmarking **precision@k and recall metrics** using **Python (Pandas, NumPy)** across training materials to optimize retrieval relevance for downstream **GPT-3.5** query answering and delivering optimization recommendations.
- Designed and executed content moderation classification experiments comparing **DistilBERT** against **Logistic Regression** and **SVM** models, building pipelines and orchestrating experiment metadata with **MLflow**, tracking precision-recall tradeoffs and hyperparameter sensitivity across **50+ runs** to identify optimal tagging thresholds.
- Conducted comparative analysis of content moderation approaches using **scikit-learn** and **Matplotlib**, identifying architecture-hyperparameter configurations that minimized misclassification on underrepresented content categories.
- Configured **Weights & Biases** as centralized experiment tracking infrastructure, systematically comparing model performance metrics, accuracy degradation patterns, and latency benchmarks across model versions, generating performance benchmarking reports using **Excel** and **Python** that surfaced degradation trends.

Pantech E-Learning [Link] — June 2022 - November 2022
AI Engineer Intern — *Chennai, India*

- Built a content recommendation system using **RoBERTa** with **HuggingFace Transformers** and **PyTorch** for semantic course matching, analyzing semantic similarity patterns across **15,000+ learning resources** and computing similarity scores with **MySQL** and **Python**.
- Evaluated course-matching accuracy through **precision@k metrics**, conducted error analysis on low-confidence similarity scores to improve recommendation relevance, and optimized inference with **ONNX Runtime** for real-time serving through **FastAPI** endpoints.
- Produced a user engagement scoring pipeline using **XGBoost** and **LightGBM** with **scikit-learn** preprocessing pipelines, analyzing engagement data (session frequency, completion rates, interaction patterns) and performing **feature importance analysis with SHAP values**.
- Optimized decision thresholds through **Optuna** hyperparameter tuning to balance precision-recall for early intervention targeting, predicting learner churn risk and deploying the complete end-to-end pipeline as a REST API via **FastAPI** with **Docker** containerization, delivering risk-tier reports in **Power BI** that enabled the student success team to prioritize at-risk learners.
- Implemented a learner anomaly detection system using **Isolation Forest** and **PyTorch autoencoders** on platform usage logs, performing statistical profiling of engagement patterns across user segments, flagging unusual behavioral outliers through threshold optimization and distribution-based scoring, and scheduling batch inference jobs through cron-triggered **FastAPI** endpoints.

Technical Skills

Languages: Python, R, SQL, C, C++, JavaScript, HTML, CSS, Bash, MATLAB, Scala, Julia, PySpark, Java

ML/DL & NLP: TensorFlow, PyTorch, Keras, scikit-learn, XGBoost, LightGBM, Hugging Face, Transformers, LLMs, BERT, GPT, RAG, OpenCV, NLTK, SpaCy

Agentic AI: GPT-4, Claude, Gemini, Fine-tuning, Prompt Engineering, Multi-Agent Systems, LangChain, LlamaIndex, Agno AI, RAG Systems, Context Engineering

Data Science & Statistics: Pandas, NumPy, SciPy, Matplotlib, Seaborn, A/B Testing, Hypothesis Testing, Time Series Analysis, Feature Engineering, KPI Development, Regression, Predictive Modeling, Cohort Analysis, Funnel Analysis

MLOps & Big Data: Docker, Kubernetes, MLflow, Apache Spark, Apache Kafka, Hadoop, PySpark, Databricks, Apache Airflow, Prefect, dbt, HDFS, Hive, CI/CD

Cloud & Databases: AWS (SageMaker, S3, EC2, Glue, Lambda, Redshift, EMR), GCP (Vertex AI, BigQuery, Dataflow), Azure (ML, Data Factory, Synapse), PostgreSQL, MySQL, MongoDB, Snowflake, Redshift, BigQuery, Redis, Pinecone, ChromaDB, FAISS

BI & Visualization: Power BI, Tableau, Excel (Pivot Tables, VBA, Power Query), Looker, Google Analytics, Plotly, Dash

Data Modeling & APIs: Star Schema, Snowflake Schema, Dimensional Modeling, Data Normalization, RESTful APIs, GraphQL, Flask, FastAPI, Microservices

Projects

Synthetic Data Generation Platform With Hallucination Detection [Link]

- Structured a multi-agent validation system with an orchestrator agent coordinating **4 specialized agents** (rule-based, statistical, LLM semantic, RAG similarity) featuring inter-agent communication, autonomous decision-making, and self-healing pipeline capabilities including automatic **CTGAN re-generation** on quality threshold failures.
- Designed and built an end-to-end data generation pipeline ingesting **284K credit card transactions**, training a **CTGAN** model, and producing **100K synthetic records** preserving complex feature distributions including a heavily imbalanced fraud class **(0.17%)** and **28 PCA-transformed features**.
- Constructed hierarchical hallucination detection across two levels, catching **CTGAN output hallucinations** and detecting **LLM reasoning hallucinations** through a meta-validation layer that validates **GPT-4o-mini's** own validation logic.
- Validated synthetic data quality through **statistical testing** (Kolmogorov-Smirnov tests, correlation matrix comparison, distribution analysis) achieving **95%+ fidelity** against real data baselines established through comprehensive **EDA**.
- Integrated **RAG-based similarity validation** using **Pinecone vector DB** and **OpenAI text-embedding-3-small**, enabling nearest-neighbor anomaly detection with automated escalation to the LLM semantic agent for deeper review.
- Implemented privacy-preserving mechanisms including **differential privacy** and **k-anonymity (k=100)** ensuring zero PII leakage while maintaining **95%+ statistical fidelity** against source dataset.
- Built RESTful API layer using **FastAPI** to expose generation and validation endpoints, with structured logging producing quality reports (PDF), hallucination logs (JSON), and synthetic datasets (CSV) as pipeline artifacts, containerized with **Docker** and **Docker Compose** for consistent deployment.

Retrieval Based Question Answering Model for Medical Drugs [Link]

- Architected a **RAG pipeline** for medical drug QA, engineering an end-to-end data pipeline that web-scraped, cleaned, and structured drug information for **2,755 medications** from MayoClinic.com, handling nested HTML deduplication and hierarchical JSON-to-CSV flattening across **45,697 document chunks**.
- Built a **FAISS-indexed vector store** with **MiniLM-V6 embeddings** for efficient approximate nearest neighbor search, enabling sub-second retrieval across **45K+ document chunk corpus**, achieving **87.50% retrieval accuracy**.
- Embedded a two-stage retrieval system with weighted intent-aware query vectorization and **Sentence-BioBERT reranking**, improving **F1@3 to 60.96%**.
- Benchmarked **MiniLM-V6, MiniLM-V3, BGE-Small and BM25** with a custom 40-query evaluation set, and integrated **Llama-4** via **Groq API** for grounded answer generation that eliminates LLM hallucination.
- Designed a queryable search and retrieval system enabling users to find relevant drug information across the structured database, achieving an **87.5% top-3 retrieval success rate**.

Smart Fields - Enhancing Agriculture with Machine Leanring [Link] *PyTorch, ResNet, Random Forest, Flask*

- Tackled crop yield prediction challenges by building **LSTM networks** with **weather data integration** (temperature, rainfall, humidity), achieving **90% accuracy** via **multi-variable correlation analysis** on **15+ environmental parameters**.
- Engineered **CNN-based disease detection** system using **ResNet-9** architecture across **38 disease categories** from **10,000+ leaf images**, achieving **94% F1-score** with **k-fold cross-validation**.
- Created **Random Forest** crop recommendation engine processing **7 soil parameters** with **real-time weather API** integration for data-driven farming decisions.

Task Insights and Workload Overview Analysis for Denver International Airport [Link]

- Built an integrated **Power BI dashboard** connecting **ServiceNow** enhancement requests with **Azure DevOps** tasks for Denver International Airport, providing real-time visibility into **53 active requests** across **27 customer stakeholders**.
- Designed automated priority-based timeline calculations, reducing manual tracking time by **15 minutes per analysis**, and implemented **NLP keyword extraction** for request summarization with calculated fields for cross-team dependency tracking, workload distribution, and monthly trend analysis.

Fire in Focus: A Deep Learning Approach to Analyzing Wildfires [Link]

- Assessed **25 years of wildfire patterns** across the United States using **NASA satellite imagery**, collecting and cleaning **1 million temporal records** across **10 Southern California counties** and performing EDA using **Python**.
- Built predictive classification models combining satellite hotspot detection with **40+ weather variables** to forecast wildfire probability, applying **Random Forest** and **XGBoost**, achieving **83.5% accuracy** and **0.87 AUC-ROC** with optimized KNN.

Spatiotemporal Deep Learning for High-Resolution Flood Mapping [Link]

- Built a data pipeline to extract, transform, and analyze **NASA satellite** and geospatial flood records for Houston's urban zones, processing raw **GeoTIFF rasters** into structured tabular datasets with spatial coordinates, timestamps, and flood-extent labels.
- Conducted comparative statistical analysis across **4 modeling approaches**, evaluating performance through precision, recall, and overlap metrics **(IoU)**, identifying that post-processing spatial filters improved boundary accuracy by **12%**.
- Tuned model hyperparameters using **Bayesian optimization (Optuna)**, reducing validation loss by **9.6%** over baseline.

Publications

Smart Fields: Enhancing Agriculture with Machine Learning [Link] – IEEE AIMLA 2024