Anish Patnaik

San Diego, CA | (858) 220-6581 | apatnaik0@gmail.com | LinkedIn | GitHub

EDUCATION

University of California – San Diego, San Diego, CA **Sep 2025 – Dec 2026**
Master of Science – Data Science
Relevant Coursework: Web Mining and Recommender Systems, Numerical Linear Algebra, Probability and Statistics

Manipal Institute of Technology, Manipal, India **Jul 2018 – May 2022**
Bachelor of Engineering - Electronics and Communication Engineering, GPA: 9.39/10
Minor Specialization in Data Science
Relevant Coursework: Data Structures & Algorithms, Object-Oriented Programming, Data Science, Machine Learning

SKILLS

- **Programming Languages:** Python, C++
- **Software Engineering / ML:** Data Structures & Algorithms, OOP, API Development, Parallel Processing, Model Deployment, CI/CD, System Design, SQL, Machine Learning, Feature Engineering, Deep Learning, NLP, RAG, Generative AI & LLMs
- **Libraries & Frameworks:** NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn, PyTorch, TensorFlow, NLTK, Transformers
- **Cloud & Development Tools:** Microsoft Azure (Synapse Analytics, ML Studio, Function Apps, Networking, Blob Storage, SQL Database – via T-SQL & Serverless SQL Pools), GitHub, Docker, Containerization, Kubernetes, Rally, Power BI

PROFESSIONAL EXPERIENCE

Optum – UnitedHealth Group, Hyderabad, India **Jul 2022 – Aug 2025**
Associate Data Scientist

- Built and deployed an **end-to-end pipeline** to **identify & extract** relevant text from appeal PDFs via a **DiT** model, and classified appeal intent via the **OpenAI LLM API** with **97.5%** accuracy, reducing processing time by **15%** and saving **~$1.4M** annually.
- Designed and productionized a **real-time** statistical model with **event-driven architecture**, to assess risk in access approvals for **100k+** unique accesses, used **serverless Azure function APIs** to calculate and report risk scores with **<50ms latency**.
- Prototyped an access **recommendation model** with **96.1%** AOC using **random forest**, **engineering features** from peer access overlap and similarity scores to recommend application access for employees.
- Engineered and **scaled** an employee similarity score **microservice** using **ordinal encoding** and **Jaccard similarity** for **400k+** employees. Exposed the functionality via **internal APIs** to be used by multiple teams **across the organization**.
- Implemented **Exploratory Data Analysis** workflows on hiring data across **29** business lines to **influence business strategy**, reducing future hire attrition rate from **18%** to **11%**, estimated to save **~$0.5** million annually.

ConceptWaves Software Solutions, Hyderabad, India **Jan 2022 – May 2022**
Technical Intern

- Designed a **Random Forest** model to estimate university MBA admission chanced based on a student's academic credentials with an accuracy of **92%** on a dataset of **10k+** students.
- Improved model performance by **15%** through the **engineered features** (GPA trends, extracurricular involvement scores, and standardized test performance) & packaged the workflow in **reproducible Jupyter notebooks** for future integration.

ESDS Software Solutions, Pune, India **May 2020 – Jun 2020**
Machine Learning Research Intern

- Built a **POC** Web Application Firewall Algorithm using **XGBoost classifier** to detect malicious web requests with **90.8%** accuracy by processing raw traffic logs.
- Performed **feature importance analysis** to identify **key indicators** of malicious activity in web requests and logs, providing **actionable insights** for their internet security team to strengthen firewall policies.

ACADEMIC AND RESEARCH PROJECTS

Neonatal seizure detection using EEG Signals with Deep learning methods **Jan 2024 – Nov 2024**

- Designed a **deep learning** system to detect and predict neonatal seizures from **12,000+** EEG patient samples, achieving **97.49%** accuracy through a **CNN–attention ensemble** architecture, with potential applications in **real-time NICU monitoring**.
- Conducted extensive experimentation and **benchmarking** against **baseline models** (1D & 2D CNNs, ResNet50, and Attention Layers), demonstrating significant **performance improvements** and validating the effectiveness of the **ensemble** approach.
- Engineered processing workflows (**Individual Component Analysis (ICA), chunking, windowing, balancing, shuffling and robust normalization)** to optimize noisy EEG data for **deep learning** model consumption.

RAG Enhanced Authorship Classification for Essays **Jul 2025 – Aug 2025**

- Developed an **NLP** solution for **authorship classification** using **487K+** essays - to distinguish between AI-generated and human-written content using **Logistic Regression** and **Random Forest Classifier**
- Integrated a **Retrieval-Augmented Generation (RAG)** component to provide **natural language explanations** for predictions, enhancing interpretability and trust in results
- Engineered and optimized **TF-IDF text features** alongside **structured metadata features**, achieving **~99%** accuracy consistently across **5-fold cross-validation**, ensuring robustness and **minimizing overfitting**