

Amaan Vora

amaanvora@gmail.com | 7326403024 | linkedin.com/in/amaan-vora | github.com/deadven7 | amaan-vora.com

EXPERIENCE

Data Scientist – Virtua Health – Marlton, NJ | Jan 2026 – present

- **Clinical prediction models:** Built multimodal ablation risk models using EHR data, notes, and **473K imaging studies**, evaluating many ML and deep learning approaches. Achieved **0.81 ROC, 0.60 PR-AUC** across **100K+ patient encounters**.
- **Model optimization & evaluation:** Developed ensemble modeling framework using LightGBM, XGBoost, and CatBoost with feature engineering over **4,192 variables**. Improved robustness through temporal validation and SHAP interpretability.
- **Medical imaging modeling:** Applied deep learning (VideoMAE, EchoNet) to multi-frame DICOM data, integrating imaging-derived features with structured clinical data to improve prediction performance.
- **Clinical NLP pipeline:** Built GPT-based system to process **50K+ clinical notes**, using structured prompting, semantic chunking, and parallel processing to convert unstructured text into model-ready features.
- **Enterprise AI assistants:** Developed GPT-4o-based HR and Access Center chatbots for over **12K+ documents**, enabling retrieval-augmented responses with source-linked citations and improved knowledge access across the organization.
- **Clinical documentation automation:** Rebuilt ICD documentation pipeline using Spark and Azure OpenAI across **50K+ records**, reducing processing time from **4 days to 20 hours** and reducing misclassification from **700+ to 57 cases**.
- **Clinical decision support:** Built and deployed SIRS/sepsis monitoring pipeline across **3 hospital locations**, processing up to **750K records per run** and generating **800+ clinician-reviewed alerts bi-monthly**.
- **Data engineering:** Built scalable Spark and SQL pipelines on Epic Clarity data to consolidate longitudinal patient records, enabling downstream ML and reporting workflows.
- **Technical ownership:** Sole developer across all AI initiatives; regularly presented model results and system updates to physicians and executive leadership.
- **Tools:** *Python, PyTorch, scikit-learn, LightGBM, XGBoost, CatBoost, Random Forest, VideoMAE, EchoNet, LangChain, GPT-4o, Azure OpenAI, Azure AI Foundry, Azure AI Search, Microsoft Fabric, Spark, SQL Server, Epic Clarity, Azure Blob Storage, OneLake, Docker, Kubernetes, FastAPI, MLflow, Git, GitHub Actions*

Machine Learning Research Engineer – Rutgers NJ Medical School – Newark, NJ | Oct 2024 – Dec 2025

- **Molecular property prediction:** Built modular DMPNN model pipeline with hyperparameters, regularization and domain features achieving 14% improvement in ROC-AUC and 0.82 PR-AUC against XGBoost and SVM baseline for supervised learning tasks. Reduced validation time for 4000+ molecules from hours to minutes, directly informed drug discovery.
- **Statistical validation framework:** Designed stratified k-fold cross-validation and quantifiably compared fine-tuning optimization with Ray Tune, Optuna and HyperOpt testing 200+ hyperparameter configurations.
- **Domain-adapted LLM pipeline:** Integrated GPT, Claude API for literature analysis, accelerating candidate identification 40% through knowledge extraction and semantic search across 10,000+ papers with sub-second query latency using FAISS.
- **Fine-tuning for scientific text:** Built BERT and SciBERT fine-tuning pipelines for document classification enabling semantic search with 91% accuracy on domain-specific benchmarks.
- **Tools:** *PyTorch, scikit-learn, Ray Tune, PostgreSQL, Docker, AWS*

Lead Developer – The Daily Targum – New Brunswick, NJ | Aug 2023 – May 2024

- **Engagement prediction models:** Trained LSTM + Attention model on 2 years of behavioral data forecasting article performance. Editorial used score for homepage and topic-based prioritization, lifted engagement 20% and ad revenue 15%.
- **BERT-based recommendation engine:** Built semantic search and recommender, processing 50,000+ documents with inference using FAISS, ChromaDB and sentence embeddings. Increased discoverability by 35% and session duration by 40%.
- **Fine-tuned classification models:** Trained RoBERTa for content tagging and multi-label categorization, reducing manual editorial processing by 50%. Implemented continuous evaluation monitoring classification drift and precision degradation.
- **Stakeholder reporting:** Presented analytics through Plotly and Gradio dashboards to 12-person editorial board, translating statistical findings into 6-month content strategy and budget allocation decisions for partner CMS provider.
- **Tools:** *Python, PyTorch, HuggingFace, FAISS, ChromaDB, AWS Lambda, PostgreSQL*

AI/ML Graduate Research Assistant – Rutgers University – New Brunswick, NJ | Jan 2023 – July 2023

- **Reproducible ML pipelines:** Standardized training workflows (SVM, Random Forest, XGBoost) with stratified CV and parallel processing. Achieved 95% accuracy on adversarial datasets, reduced runtime 35%, became department standard.
- **Causal inference:** Applied propensity score matching and sensitivity analysis on NumPy data, reduced RMSE by 15% over regression by accounting for confounding variables. Assisted professor with Two Stage Curvature Identification research.
- **Tools:** *Python, scikit-learn, PyTorch, TensorFlow, XGBoost, LangChain, AWS*

SKILLS

Programming: Python, R, SQL, C++

Statistical Methods: Hypothesis testing, A/B testing, regression analysis, causal inference, survival analysis, experimental design

ML/Modeling: scikit-learn, XGBoost, LightGBM, PyTorch, Random Forest, SVM, ensemble methods, time series forecasting

Data Analysis: Pandas, NumPy, feature engineering, EDA, cohort analysis

Visualization: Tableau, Plotly, Matplotlib, Seaborn, ggplot2

Tools: Git, Docker, AWS (SageMaker, EC2, Lambda), PostgreSQL, FAISS

PROJECTS

S&P 500 Sector Analysis & Stock Forecasting

Analyzed 10+ years of market data across 11 sectors using statistical modeling and ARIMA/SARIMA forecasting. Applied Box-Ljung and Chi-Square tests for validation, identified investment opportunities with 15% annual returns. Built correlation analysis and volatility models in R with ggplot2 visualizations.

Healthcare Claims Analysis with LLM Reasoning

Designed agentic workflow for payer claims combining schema validation with LLM reasoning over policy documents using vector retrieval. System flags NCCI conflicts, bundling edits, and medical necessity gaps through rule checks before generating denial summaries. Outputs explainable adjudication reasoning maintaining compliance, proof-of-concept reducing manual review loops in claims processing.

PDF Intelligence System

Built local PDF parser extracting text, tables, figures, and metadata with section segmentation. Added spaCy NER for entity extraction, TF-IDF search, and summarization using BART and FLAN-T5 on CPU/CUDA. Created Gradio interface streaming progress and rendering per-section summaries, entities, and extracted assets, turning PDF papers and reports into searchable, structured output in minutes.

La Ghigliottina NLP Solver

Built solver for Italian word puzzle using 277K semantic relations from CNN/IMDb/Taboo/Wiki plus WordNet hypernyms and meronyms. Baseline: Word2Vec Skip-Gram with cosine similarity intersection ranker. Scaled to Seq2Seq LSTM (64 units, pretrained embeddings) and fine-tuned BART/T5 for masked clue generation. Pipeline embeds five clues, mines overlapping neighbors, validates candidates, showcasing multi-clue semantic reasoning.

Privacy-Preserving Face Recognition

Built facial recognition with differential privacy: DP-SGD on LFW dataset, ResNet50 frozen backbone with dense classifier. Applied CKKS homomorphic encryption via TenSEAL for encrypted inference. Achieved 86.2% accuracy vs 91.5% baseline, quantified utility-privacy tradeoff and HE overhead. Demonstrates practical privacy-by-design approach balancing recognition performance with data protection.

A Computational Linguistics Approach to Clustering Scientific Research ([link](#))

Clustered 250,000 research papers using SciBERT and BERTopic discovering 50+ coherent topic groups. Applied t-SNE for 2D visualization and validated clusters with domain experts. Published methodology at International Conference on IoT, Cloud and Data Science receiving Best Paper award for combining transformer embeddings with unsupervised learning.

EDUCATION

M.S. in Statistics (Data Science) – Rutgers University, NJ, USA

- Focus: Statistical Analysis, Statistical Inference, Machine Learning, Statistical Modeling, Deep Learning

B.S. in Computer Science – SRM Institute of Science & Technology, Chennai, India

- Focus: Algorithms, Data Structures, Machine Learning

CERTIFICATION

Google Data Analytics Professional Certificate | Deep Learning (NVIDIA DLI) | LLMops (DeepLearning.AI) | Machine Learning Fundamentals (University of Washington) | Epic Clarity/Caboodle/Cogito Certifications