Nasim Mahmud Nayan

smnoyan670@gmail.com +8801742957256

Portfolio LinkedIn Google Scholar

Summary

ML Engineer with 3+ years building production AI systems that drive measurable impact. Deployed a wellness platform serving 100+ daily users and a RAG-powered sales agent handling 500+ inquiries/day. Specialized in cost-efficient GenAI solutions (LLMs, RAG, AI agents), achieving 94% cost reduction vs. competitors. Published researcher with 10+ papers and 100+ citations.

Skills

AI/ML	GenAI (LLMs, RAG, AI Agents), Computer Vision, NLP, RL, Predictive Modeling, Fairness & Explainable AI
Frameworks & Tools	PyTorch, TensorFlow, Scikit-learn, Hugging Face, LangChain, LlamaIndex, OpenCV, XGBoost
MLOps & Production	Docker, FastAPI/Flask, MLflow, CI/CD (GitHub Actions), Vector Databases (Pinecone, ChromaDB), PostgreSQL, MongoDB, Linux
Languages	Python (Expert), C++, C

Work Experience

Programming Hero (Remote) Dec 2023 – Sept 2025
ML Engineer

- Zenyora AI Wellness Platform (Microsoft Store)
  - Owned full ML lifecycle for production wellness platform serving 100+ daily users—from training to deployment to monitoring.
  - Built real-time posture detection using MediaPipe achieving 95% accuracy and reducing user eye strain by 40% (validated via user surveys, n=150).
  - Developed context-aware productivity coaching analyzing user behavior with sliding window analysis, processing 30 FPS video with sub-100ms latency.
    Stack: PyTorch, OpenCV, MediaPipe, GPT-4, FastAPI, Docker, PostgreSQL
- Automated Real Estate Sales Agent
  - Architected voice-based RAG agent with Pinecone vector database, handling data engineering for 2,000+ property listings with multi-modal interaction supporting voice (Whisper STT, ElevenLabs TTS).
  - Implemented hybrid retrieval (vector similarity + metadata filtering) achieving 95% query relevance with sub-2-second end-to-end voice response time.
  - Deployed conversational AI processing 500+ daily inquiries, increasing lead engagement 3x while reducing response time from 2 hours to real time.
    Stack: LangChain, RAG, Pinecone, GPT-4, Whisper, ElevenLabs, Twilio, FastAPI, Docker
- Student Performance Prediction System
  - Built predictive ML system forecasting student assignment scores with 85% accuracy, serving 10,000+ students and contributing to 15% average grade improvement.
  - Engineered features from behavioral data (submission patterns, time-on-task, prior performance) using XGBoost with Optuna hyperparameter tuning.
  - Created automated retraining workflow with MLflow tracking, cutting deployment time by 75%.
    Stack: XGBoost, Scikit-learn, FastAPI, PostgreSQL, MongoDB, MLflow

Primacy Infotech Ltd July 2023 – Nov 2023
AI Engineer

- Led AI tour planner for 6 UNESCO heritage sites, reducing itinerary creation time from 2 hours to 5 minutes through automated route optimization.
- Built complete data workflow: collected and processed 5,000+ visitor surveys, extracted insights via clustering analysis, and integrated results into personalized route generation algorithm.
- Deployed FastAPI + Streamlit system achieving 99.2% uptime, increasing tourist satisfaction by 35% (measured via post-visit surveys).
- Collaborated with CTO to define project milestones and deliver weekly technical presentations to stakeholders.
  Stack: GPT-3.5, FastAPI, PostgreSQL, Docker, Streamlit, Pandas

Rising Research Lab Aug 2022 – May 2023
Research Assistant

- Developed multi-disease prediction system (diabetes, Parkinson's, maternal health) achieving 92% average accuracy, serving as foundation for 4+ published papers.
- Engineered IoT air quality monitoring network across 5 locations with complete workflow from sensor data capture to real-time inference and visualization.
- Implemented AutoML solution using Optuna and Scikit-learn reducing model development time by 60% while maintaining performance across disease prediction tasks.
  Stack: PyTorch, TensorFlow, Scikit-learn, IoT sensors, Flask, PostgreSQL, Optuna

Entrepreneurship & Product Development

OgroPath (ogropath.com) Apr 2025 – Present
Founder & Lead AI Engineer

- Founded and launched AI-powered medical entrance exam preparation SaaS platform serving students across Bangladesh, democratizing access to medical education for underserved rural communities.
- Engineered end-to-end ML system featuring NLP-based question generation from PDF documents, integrated with 21,000+ question bank, adaptive study plans, and real-time national ranking system.
- Developed AI diagnostic engine analyzing student performance data to identify weaknesses and generate personalized topic-based assessments, validated through user feedback as providing "24/7 expert assistant" guidance.
- Architected full production infrastructure handling user authentication, payment processing, analytics dashboard, and multi-user concurrent access with sub-second response times.
  Stack: LLMs, NLP, FastAPI, PostgreSQL, React, payment integration, cloud deployment, Docker

Key Projects

- RAG-Based Website Generator
  - Architected cost-efficient AI website generator using RAG to assemble pre-built React components vs. generating code from scratch, achieving 92% token reduction and 94% cost savings ($0.03 vs. $0.50–$2.00 per site) compared to v0 and Lovable.
  - Implemented hybrid retrieval with PostgreSQL + pgvector for semantic search across 21 indexed components, using tree-sitter for AST parsing to extract reusable patterns and metadata.
  - Built FastAPI backend with vector embeddings + metadata filtering and Next.js 14 frontend, delivering production-ready TypeScript code in 30–40 seconds.
    Stack: Python, FastAPI, PostgreSQL, pgvector, Next.js, TypeScript, GPT-4o, Sentence Transformers, tree-sitter | GitHub
- Bengali Literature RAG System
  - Engineered RAG pipeline combining LaBSE embeddings with BM25 keyword search via Reciprocal Rank Fusion, achieving 82% MRR and 95% Hit@5 accuracy with 32ms response time on a 503-question dataset.
  - Implemented intelligent GPT-4o Mini enhancement with selective triggering based on retrieval confidence, reducing LLM API costs by 60% while maintaining answer quality.
  - Optimized FastAPI backend for sub-second response times through caching, lazy loading, and Docker containerization.
    Stack: Python, FastAPI, LaBSE, BM25, GPT-4o Mini, Docker, Pydantic | GitHub

Education

University of Information Technology and Sciences Jan 2019 – Jun 2023
B.Sc. in Computer Science and Engineering CGPA: 3.62/4.00
Relevant Coursework: Machine Learning, Computer Vision, Data Structures, Algorithms, Databases, Statistics

Selected Publications (10+ papers, 100+ citations)

- Nayan, N. M., et al. "An interpretable and balanced machine learning framework for Parkinson's disease prediction using feature engineering and explainable AI." PLOS ONE, 2025.
- Nayan, N. M., et al. "SMOTE Oversampling and Near Miss Undersampling Based Diabetes Diagnosis from Imbalanced Dataset with XAI." IEEE ISCC 2023. (Scopus)
- Full list: Google Scholar

Awards & Recognition

- Employee of the Month | Primacy Infotech Ltd | Aug 2023
- 2nd Place | Inter-university Programming Contest | UITS 2022
- 1st Place | PowerPoint Presentation Competition | UITS 2020
- Fastest Problem Solver | Victory Day Programming Contest | UITS 2021