



Mehedi Hasan

AI/ML Engineer

☎ (+880) 1521323549

✉ abir.aust.102@gmail.com

🌐 mhabir.dev

🌐 linkedin.com/in/mehedihasan102

🐙 github.com/Mehedi-Hasan-Abir

📍 Dhaka, Bangladesh

PROFESSIONAL SUMMARY

Senior AI/ML Engineer with 5+ years of experience designing and deploying production-grade AI systems from business requirements through full production deployment. Proven track record **translating client and stakeholder requirements into scalable AI architectures**; building and integrating AI models, data pipelines, APIs, and distributed services; and delivering reliable solutions across constrained and cloud-scale environments. Experienced across **LLM adaptation, agentic systems, Retrieval-Augmented Generation (RAG)**, data engineering, and production AI infrastructure. Published IEEE researcher in **Natural Language Processing (NLP)** and Bengali language AI.

PROFESSIONAL EXPERIENCE

Senior Software Engineer (AI/ML), Technonext

Jun 2025 – Present

- ▷ **CRM Platform – Agentic AI Customer Care Chatbot:** Designed a multi-step agentic customer-care system with LangGraph and hybrid RAG, integrating OpenAI models through MCP-based tooling to expose domain-specific services and data sources through a standardized interface. Implemented metadata-driven tool and resource filtering to tailor available capabilities across client platforms and user contexts while maintaining a common agent architecture, with OpenAI/Gemini APIs and local vLLM model fallbacks for flexible deployment. [LangGraph](#) [Hybrid RAG](#) [MCP](#) [OpenAI API](#) [Gemini API](#) [vLLM](#)
- ▷ **Candidate Speech Assessment Service:** Built a multi-tenant voice assessment pipeline evaluating reading accuracy, prosody, fluency, breath management, and task-specific tone using API-based LLM scoring with Pydantic-validated scoring schemas. Integrated S3, PostgreSQL, Redis, RabbitMQ, and Celery Beat with end-to-end task/job tracking and operational visibility for scheduled asynchronous processing. [Google Gemini 3.1 Pro](#) [Pydantic](#) [S3](#) [PostgreSQL](#) [Redis](#) [RabbitMQ](#) [Celery Beat](#)
- ▷ **Food Delivery application – Cuisine Prediction Service:** Designed and deployed a gRPC-based microservice predicting cuisines from user intent, retrieving relevant dishes/restaurants using Sentence Transformers, FAISS vector search, PostgreSQL, and Redis-powered personalization/trending signals. Optimized inference with ONNX, achieving **100 RPS on 2 vCPU**. [gRPC](#) [Sentence Transformers](#) [FAISS](#) [PostgreSQL](#) [Redis](#) [ONNX](#)
- ▷ **E-commerce – Visual Product Search:** Built end-to-end image search on constrained 2-core infrastructure using CLIP embeddings, FAISS retrieval, Elasticsearch ranking, and catalog indexing. Developed complete data ingestion, training, validation, and inference pipelines with Kafka batch processing. Built Apache Spark ETL pipelines to migrate and transform production data into curated datasets for model fine-tuning and downstream model ingestion. Optimized models via ONNX and OpenVINO for efficient production deployment. [CLIP](#) [FAISS](#) [Elasticsearch](#) [Kafka](#) [Apache Spark](#) [ONNX](#) [OpenVINO](#)
- ▷ **Passport MRZ Scanner:** Built an end-to-end ML pipeline for MRZ extraction with real-time ONNX-based inference, using Redis Streams, a custom asynchronous scheduler (Python/Go) with micro-batching, and a CPU-isolated multiprocessing worker pool for low-latency production document processing. [ONNX](#) [Redis Streams](#) [Python](#) [Golang](#) [Multiprocessing](#)
- ▷ **Infrastructure & DevOps:** Deployed all services using Docker containerization with Docker Compose and Kubernetes orchestration, integrated into CI/CD pipelines under version control, and monitored via Prometheus and Grafana for production observability, reliability, and automated scaling. [Docker](#) [Docker Compose](#) [Kubernetes](#) [CI/CD](#) [Prometheus](#) [Grafana](#) [Git](#)

Senior AI Engineer, Next Solution Lab

Jun 2023 – Jun 2025

- ▷ Led **English DeepICR** (document understanding for contracts/invoices): built, trained, and tested **Text Detection**, **Layout Detection**, **Text Recognition**, and **Data Extraction**. Complex noise filtering (checkmarks/underlines), handwriting handling; document spell-correction via fuzzy + dictionary (Levenshtein) tuned to client lexicons. [Mask R-CNN](#) [OpenCV](#) [YOLOv7](#) [DBNet](#) [RCNN+Attention](#) [LCNN](#) [LayoutLM](#) [BROS](#)
- ▷ Optimized training and inference pipelines via model fine-tuning and parallel processing to improve throughput and latency. [PyTorch DDP](#) [CUDA/cuDNN](#) [joblib](#)
- ▷ Enabled **GPU multi-training** and **PDF batching** for large documents. [Keras](#) [PyTorch](#) [TensorFlow](#)
- ▷ Standardized evaluation with **automated reports** for PM/business sign-off. [XlsxWriter](#) [pandas](#)
- ▷ Packaged services with **Docker**, integrated **modular monolith** server-side APIs with client systems, and managed deployment/version control via **AWS CodeCommit**, **S3**, **EC2**, **ECR**, **Elasticsearch**, and **CloudWatch**. [Node.js](#) [React](#) [Material-UI](#) [AWS SDK](#)
- ▷ Monitored Japanese DeepICR platform; supported model updates and maintained routine regression tests.
- ▷ **LLM Adaptation & Fine-Tuning**: Implemented LoRA/PEFT-based LLM fine-tuning workflows and optimized inference acceleration pipelines for efficient model adaptation and deployment. [HuggingFace](#) [Transformers/TRL](#) [BitsAndBytes](#) [ONNX](#)
- ▷ **Japanese Legal RAG Search**: Built a RAG search system for Japanese legal documents with memory and inference caching, enabling low-latency conversational Q&A. [LangChain](#) [ChromaDB](#) [Elasticsearch](#) [FastAPI](#)

AI Engineer, Next Solution Lab

Jun 2022 – Jun 2023

- ▷ Built the **Japanese Text-recognition** training pipeline with augmentation/evaluation and ~5,000-character coverage; created an **automated data-engineering & reporting** flow; maintained for new-data sprints. [RCNN+CTC](#)
- ▷ **English DeepICR**: trained/tested models for new datasets every sprint; standardized onboarding & regression tests.
- ▷ **DocQA (pre-LLM)**: Q&A for contracts/invoices documents; [BERT/RoBERTa](#) [Flask](#) [Streamlit](#)
- ▷ **Multilingual OCR**: Document OCR for (Arabic, Vietnamese, Thai, Indonesian). [Google Vision API](#)
- ▷ **Deepfake Detection, STT/TTS & Speaker Verification**: Exploration of deepfake detection approaches; STT/TTS pipeline evaluations; speaker verification. [FastAPI](#) [NVIDIA NeMo](#)

Associate AI Engineer, Next Solution Lab

Jun 2021 – Jun 2022

- ▷ **Key-Value Extraction Module**: deterministic OpenCV rule engine extracting **128 fields** on **44** document clusters; authored **SRS**, exposed **API**, and defined/tested plans; owned **text-line & text-box update** integration for readability. [OpenCV](#) [REST API](#)
- ▷ **Smart Farming (POC)**: camera-only cattle monitoring under a no-RFID policy; delivered fast-deploy models and proposal. [Detectron2](#) [SORT](#)
- ▷ **ID/OCR**: driving licence extraction. [YOLOv5](#) [FastAPI](#) [Streamlit](#)

TECHNICAL SKILLS

Programming Languages	Python , JavaScript, Node.js, C#, Java
Frameworks & Libraries	PyTorch , TensorFlow , Keras, FastAPI , LangGraph , LangChain , HuggingFace Transformers , Streamlit, OpenCV, Spacy
LLM Tooling	vLLM , Model Context Protocol (MCP), Retrieval-Augmented Generation (RAG), Natural Language Processing (NLP), Prompt engineering, Tool/function calling, Embeddings (BGE, E5, Sentence Transformers)
Vector & Search Systems	Qdrant , FAISS, Elasticsearch, ChromaDB
Database	MySQL , PostgreSQL, MSSQL, Redis
Data Engineering	Apache Spark , Kafka, ETL/ELT pipelines, Dataset curation
ML & Data Tools	pandas , NumPy , joblib, CUDA , cuDNN, PyTorch DDP, Playwright
Frontend	React , Material-UI, React Router
Version Control	Git , Agile/Scrum

AI MODELS & ARCHITECTURES

Computer Vision	YOLOv7/YOLOv5, DBNet, Mask R-CNN, RCNN+CTC, LCNN
Document & Layout Understanding	LayoutLM, BROS, DocFormer
Language Models	LLaMA 3, Gemma, Mistral, Qwen, BERT, RoBERTa
Fine-tuning & Alignment	Instruction tuning/SFT, PEFT/LoRA, DPO
Retrieval & Embeddings	BGE, E5, Instructor, Sentence Transformers, BGE-reranker, Cross-encoder reranking

SYSTEM DESIGN & PRODUCTION AI STACK

Model Serving & Optimization	ONNX Runtime , OpenVINO, TensorRT, Quantization/pruning/distillation, FastAPI + gRPC inference systems
Distributed Systems & Backend	Microservices architecture, gRPC , Asyncio/concurrency patterns, Multiprocessing/thread pools
DevOps & Production Deployment	Docker, Docker Compose, Kubernetes (HPA, deployments, scaling), CI/CD (GitHub Actions, GitLab CI), Prometheus, Grafana, OpenTelemetry, AWS (EC2, ECS, S3, Lambda, CloudWatch, CodeCommit/ECR)
Data Engineering & Pipelines	Apache Spark, ETL/ELT pipelines, Streaming systems (Kafka, Redis Streams), Dataset versioning, Annotation/labeling pipelines

EDUCATION

Ahsanullah University of Science and Technology (AUST) <i>Bachelor of Science in Computer Science and Engineering — CGPA: 3.59</i>	<i>Dhaka, Bangladesh</i> <i>Apr. 2016 – Jan. 2021</i>
SOS Hermann Gmeiner School and College <i>Higher Secondary Certificate — GPA: 5.00</i>	<i>Dhaka, Bangladesh</i> <i>2013 – 2015</i>
Mirpur Bangla High School and College <i>Secondary School Certificate — GPA: 5.00</i>	<i>Dhaka, Bangladesh</i> <i>2003 – 2013</i>

PROJECTS

Agami Assistant — Voice-First Enterprise AI Assistant 🗣️

- ▷ Built a voice-first enterprise assistant that delivers grounded, spoken answers from product catalogs, policies, and IT records, with live transcripts and captions. Live demo 📄
- ▷ Designed intelligent agent workflows with knowledge retrieval, contextual follow-up support, and responsive answers for business and operational queries.

Tech: [FastAPI](#), [Next.js](#), [Pydantic AI](#), [Gemini Embeddings](#), [FAISS](#), [Groq](#), [SSE](#), [PyO3/Rust](#), [ASR/TTS](#), [Render](#)

Search Microservice — High-Performance Product Discovery 🗣️

- ▷ Built a typo-tolerant product discovery service for **1M products** and **~29K brands**, enabling fast, relevant search on constrained infrastructure.
- ▷ Achieved **179 QPS at 3.17 ms P95** for warm traffic, **13.1 ms P95** for unseen typo queries, and a validated peak capacity of **1,046 QPS** under concurrent load.

Tech: [Python](#), [FastAPI](#), [gRPC](#), [OpenSearch](#), [PostgreSQL](#), [Redis](#), [Nginx](#), [Docker Compose](#), [k6](#), [Prometheus](#)

Shorol Notes — AI-Powered Note-Taking 🗣️

- ▷ Voice notes with AI transcription/summarization; calendar sync; Bangla-first UX.
- ▷ Pluggable backends (OpenAI/Ollama/Hugging Face); responsive, offline-friendly.

Tech: [React](#), [TypeScript](#), [Vite](#), [Node.js/Express](#), [Tailwind CSS](#), [OpenAI API](#), [Hugging Face](#), [Ollama](#)

Japanese Lawyer Assistant 🗣️

- ▷ Legal Q&A over Japanese corpus using RAG (retriever + FAISS) with instruct LLM.
- ▷ Knowledge-base pipeline; concise answers with custom prompts.

Tech: [Python](#), [LangChain](#), [FAISS](#), [FastAPI](#), [ELYZA LLaMA-2](#), [Hugging Face](#)

RESEARCH EXPERIENCE

Bengali Intent Classification with Generative Adversarial BERT 📄

First Author, IEEE Xplore, 2023

Developed **GAN-BnBERT**, a generative adversarial [BERT](#)-based model for Bengali intent classification. Introduced the **BNIntent30** dataset with 30 intent classes, demonstrating superior performance over [Bi-LSTM](#) and baseline [BERT](#) models.

Design of an Arrhythmia Classification Algorithm Using 2-D Convolutional Neural Network 📄

First Author, Undergraduate Thesis, AUST, 2021

Proposed a [CNN](#)-based arrhythmia detection algorithm from [ECG](#) signals, comparing performance with [VGGNet](#) for improved beat classification accuracy.

ACADEMIC PROJECTS

Hospital Management 📄 | Web Application

Developed a full-stack hospital management system using [ASP.NET MVC](#), [C#](#), [Razor](#).

Tolet-Dhaka 📄 | Web Application

Built a flat-renting website with [HTML5](#), [CSS](#), [PHP](#); enhanced interactivity with [JavaScript](#).

Bank Management System 📄 | Database Application

Implemented transactional features with [Java](#), [Java GUI](#), [MySQL](#).

Dhaka City Bus Routes 📄 | Android Application

Created a route-finder app using [Java](#), [Google Maps API](#), [JSON](#).

Railway Management System  | *Database Application*
Developed client-server components in [PL/SQL](#).

NON-TECHNICAL SKILLS

- ▷ Stakeholder communication (PM/business)
- ▷ Work Breakdown Structure (WBS)
- ▷ SRS & documentation
- ▷ Versioned runbooks
- ▷ Technical presentations

LANGUAGE SKILLS

Bangla (Native); English (Professional Working Proficiency).