Aayush Malla

Kathmandu, Nepal mallaaayush59@gmail.com (+977) 9851203045 Aayush Malla

Research Interests

AI/ML, optimization, human-centered systems, behavioral analytics, geospatial intelligence, trustworthy and safe AI, security and privacy, multimodal models, human–AI interaction, AI for health and climate, scalable ML systems

Education

Kathmandu University, Nepal *Bachelor of Computer Engineering* *2016 – 2020*

- GPA: 3.60/4.00
- **Relevant Coursework:** Machine Learning, Algorithms & Complexity, Data Structures, Database Management Systems, Artificial Intelligence, Computer Architecture & Organization, Computer Graphics, Microprocessor and Assembly Language, Compiler Design,Human Computer Interaction, Statistics & Probability
- **Capstone Projects:** Nepali License Plate Recognition — **(Best Final Year Project)**

Experience

Data Engineer *Kathmandu, Nepal*
Fusemachines *March 2022 – Present*

- **Quantum ERP Data Integration:** Built modular ETL pipelines in Azure ADF to integrate ERP data across HR, Finance, and Procurement systems. Developed PySpark and Spark SQL scripts in Databricks for data cleaning, transformation, and modeling, enabling structured ingestion into Azure Data Lake. Automated workflows and CI/CD using ADF, Databricks, and Azure DevOps, reducing manual effort by 50%, minimizing release errors, and ensuring over 95% data consistency.
- **Internal Analytics:** Designed ETL pipelines in Azure ADF to extract organizational data, including food operations, attendance, project metrics, and Jira logs. Integrated data into Azure Data Lake and built Power BI dashboards for leadership reporting. Leveraged Azure Synapse, Monitor, and Logic Apps to streamline scheduling, monitoring, and alerting.
- **Instructor & Mentor:** Delivered training sessions on SQL, PySpark, Spark SQL, Databricks, and Azure technologies. Mentored junior engineers in ETL development, CI/CD automation, and best practices for scalable data engineering.

Software Engineer *Kathmandu, Nepal*
COTIVITI *Feb 2021 – 2022 Feb*

- **ETL (Extract, Transform Load) System for Health Care:** Developed PL/SQL ETL workflows for 20+ insurance clients, enabling custom data processing and automated claim validation. Built reusable ODI pipelines to normalize and centralize raw health insurance data from 20+ providers. Analyzed healthcare claims data and created Oracle Apex dashboards to support fraud detection and cost optimization. Automated routine ETL and data quality tasks with PL/SQL scripts, reducing manual effort and accelerating claim validation processes.
- **Data Hackathon:** Participated in a Data Hackathon and automated dynamic CRUD operations across multiple tables for onboarding new healthcare clients, improving efficiency and reducing manual effort.

Teaching Experience

Teaching Assistant –National Institute of Science and Technology, Nepal *May 2022 – March 2023*

- Taught Object-Oriented Programming and Database Management Systems to undergraduate BCA students.
- Led lab sessions and tutoring, providing hands-on guidance in problem-solving and SQL programming.
- Contributed to assignment design and evaluation, fostering an interactive and collaborative learning environment.

Projects & Research Experience

ML Tool Suite — Q&A, Summaries, Keyphrases, English Translation, Sentiment (MCP-enabled) *2025*

- Finds answers directly from the source text, writes short summaries, and pulls out key terms.
- Detects the input language and translates to English; works offline or online with automatic fallback.
- Rates sentiment (positive/neutral/negative) using a small ML + rule-based mix; includes a simple web app and MCP integration.
- **Technologies:** Python; scikitlearn; Hugging Face Transformers / MarianMT (translation model); NLTK, spaCy (text processing); VADER AFINN (lexiconbased sentiment); Streamlit; MCP (Model Context Protocol).

AirAware — $PM_{2.5}$ Air Quality Forecasts with Health Guidance *2025*

- Predicts fine particulate pollution — $PM_{2.5}$ (particles 2.5 µm that can reach deep into the lungs) — for the next 6–24 hours, with live nowcasts and uncertainty bands.
- Combines multiple models for better accuracy, uses public air sensor + weather data, and explains results; provides simple health tips and alerts.
- Provides a web app and an API for realtime forecasts, "whatif" checks, and notifications; shipped with testing, monitoring, and containers.
- **Technologies:** Python; FastAPI; Streamlit; PyTorch; Prophet (timeseries forecasting); ARIMA (timeseries model) ; scikitlearn; SHAP (explainability); MLflow (experiment tracking); Docker.

NepaliGov—Retrieval-Augmented Q&A for Nepali Government Documents (Nepali/English) *2025*

- End to end system that turns scanned Nepali government PDFs into searchable text and answers questions with citations (Nepali & English).
- Uses staged OCR with fallback, builds a vector index for fast search, and routes answers in the user's preferred language.
- Provides a simple web app and REST API to ask questions; includes quality checks, caching, and basic monitoring for reliability.
- **Technologies:** Python; FastAPI/Flask; FAISS (vector search); SentenceTransformers (text embeddings); PaddleOCR / Tesseract (OCR); MarianMT (translation); Redis (caching/queues); SQLite / Parquet; Docker; MLflow (experiment tracking).

NEPSE Time Series Forecast *2024*

- Predicts the next day's NEPSE closing price with an uncertainty range.
- Runs a daily pipeline that pulls live/historical data and handles gaps or bad values.
- Trains with timeseries crossvalidation and tuning for stable, repeatable results.
- Uses practical signals (lags, returns, RSI, MACD, ATR, Bollinger Bands, volatility) to improve forecasts.
- **Technologies:** Python; pandas/NumPy; scikitlearn; LightGBM (forecast model); PyTorch; Optuna (tuning);Docker.

Nepali License Plate Recognition *2020*

- Developed a machine learning-based License Plate Recognition system to detect and extract Nepali license plate numbers from images with 95% accuracy.
- Created custom datasets by collecting and labeling thousands of Nepali character images for supervised model training.
- Built and trained an SVM model in OpenCV to improve recognition accuracy and feature extraction.
- **Technologies**: Python, OpenCV, SVM, Image Processing

Achievements and Awards

First Runner up – Data Hackathon:	Awarded for developing "Content Generator", an AI-driven tool that automated dynamic content creation using NLP and generative models.
Data Engineering Trainer:	Taught PySpark, Python, Flask, SQL, and Azure to data engineering trainees through hands-on sessions.
Merit-Based Scholarship:	Awarded a merit-based scholarship in the 7th semester for outstanding GPA performance.
Outstanding Project Award:	Recognized for developing a machine learning-based Nepali License Plate Recognition system using SVM and OpenCV, achieving 95% accuracy.

Skills

Programming:	SQL (5+ yrs), Python (4+ yrs)
Data Engines & Libs:	PySpark, pandas, NumPy
Core ML:	scikitlearn, PyTorch, LightGBM
NLP / IR:	RAG (BM25, FAISS), semantic search, TextRank, TFIDF, translation
Databases & Storage:	PostgreSQL, SQL Server, MySQL, SQLite, Parquet
Cloud & Platforms:	AWS (S3, Glue, Athena, MWAA), Azure (ADF, ADLS Gen2, Synapse, Databricks)
Orchestration:	Apache Airflow, Databricks Workflows, Delta Live Tables, ADF Pipelines
APIs & Apps:	FastAPI, Flask, Streamlit, MCP
MLOps & Dev:	MLflow, Docker, Git/GitHub
OCR:	PaddleOCR, Tesseract
Languages:	Nepali, Hindi, English