Anshul Gupta

Bristol | ans2852000@gmail.com | +44 790 945 24 01

Education

University of Bristol (UK), MSc Data Science — Sept 2024 – Sept 2025

- **Coursework:** Statistical Computing, Large-Scale Data Engineering, AI and Text Analytics, Visual Analytics

Amity University Noida (India), BSc Information Technology — Jul 2019 – Jun 2022

- CGPA: 8.82/10.0 (Gold Medalist in Academics)

Experience

Data Engineer, Deloitte, Bengaluru — Feb 2023 – Jun 2024

- Built and deployed a daily ETL pipeline processing 5–20M records through Kafka in under an hour, cutting data-load time by 1–2 hours.
- Wrote Python scripts for scheduling jobs and implementing advanced pattern matching to streamline workflows and enhance system monitoring, saving the team hours of manual checks each week.
- Automated Nokia router testing for the EtherSAM project, designing a framework that reduced manual effort, testing time, and operational costs for U.S. Cellular.
- Diagnosed and resolved over 5 critical bugs in the production pipeline, then refactored the EtherSAM codebase to follow SOLID principles, resulting in a much cleaner, more maintainable codebase.

Projects

Automated Polymer Property Extraction from Research Articles (Jun 2025–Sept 2025)

- Constructed two complementary corpora (3,000+ varied full texts and 43 polystyrene papers) with a unified seven-entity schema and relation links. Built a regex- and ontology-driven annotation pipeline to improve coverage, consistency, and extraction accuracy at scale.
- Developed a character-aware BiLSTM+CRF neural network, benchmarked against multiple BiLSTM and SciBERT models, achieving 96% macro F1 (30% improvement over baselines) while reducing computation time and resource usage by 80%.
- Demonstrated that model generalizability is driven by diverse patterns, full-text coverage, high-quality annotation, and schema enrichment rather than balancing techniques, with insights transferable to NLP in healthcare, finance, and legal domains.
- Delivered an end-to-end deployable solution by integrating data harvesting, automated annotation, model inference, and structured database storage into a FastAPI + Streamlit application.

Stratifying Tumors and Normal Tissues via MIR100HG-Centered Multi-Omics Signatures
Group Administrator, Industry Project with Nottingham Trent University, Jan 2025–Apr 2025

- Distilled a >56,000 multi-omics feature set into a 10–20-feature MIR100HG signature, allowing clear subtype identification, using differential expression analysis (FDR < 0.01) and network analysis.
- Revealed clear cancer subtype clusters by removing outliers with Isolation Forest (2% contamination) and applying PCA (PC1 explained 42–76% variance); validated features using a Random Forest classifier (F1 > 90% across all tumour types) and literature cross-checks.
- Demonstrated clear prognostic utility by stratifying pancreatic and lung cancer patients into high- and low-risk groups using top PCs in penalized Cox regression and Kaplan–Meier curves (log-rank p = 0.0116 for PAAD, p = 0.0003 for LUAD).

Skills

Programming: Python, C/C++, Java, JavaScript

Data Science & ML: Pandas, NumPy, Scikit-learn, PyTorch, HuggingFace Transformers, Tableau

Databases & Data Engineering: MySQL, PostgreSQL, Kafka, ETL

Cloud & DevOps: AWS, Docker, Jenkins, Linux, GitHub