NAAC Grade 'A"

R. C. PATEL
INSTITUTE OF TECHNOLOGY
An Autonomous Institute

MANTHAN

Department of Computer Science & Engineering (Data Science)

ISSUE YEAR
2025-26

"Experimental observations are only experience carefully planned in advance, and designed to form a secure basis of new knowledge."

Ronald Fisher

Father of Modern Statistics

Vision-Mission

Institute Vision

To become a leading Institute in Technical education fostering innovation, research, ethical values, and sustainable development for the betterment of society.

Institute Mission

To impart high quality Technical Education through :

- Innovative and Interactive learning process and high quality, internationally recognized instructional programs.
- Fostering a scientific temper among students by the means of a liaison with the Academia, Industries and Government.
- Preparing students from diverse backgrounds to have aptitude for research and spirit of Professionalism.
- To contribute to the nation’s sustainable development.

Vision-Mission

Department Vision

To provide cutting-edge Computer Engineering education in Data Science while instilling socio-moral values.

Department Mission

M1: To deliver state-of-the-art, ICT-enabled teaching and learning to achieve excellence in Data Science education.

M2: To develop professionally competent Data Science Engineers, meeting evolving industrial and societal needs.

M3: To prepare employable professionals with ethical values and a commitment to professional and social responsibility.

Visionary Insights

The Director

Prof. Dr. Jayantrao Bhaurao Patil

At RCPIT, we blend academic rigour with real-world exposure to create future-ready professionals. Emphasizing research, interdisciplinary learning, and soft skills, our hands-on approach empowers students to become ethical leaders, innovators, and entrepreneurs prepared to succeed in a rapidly evolving global landscape.

Visionary Insights

The Head of Department

Prof. Dr. Ujwala Patil

Established in 2020 at RCPIT, Shirpur, the Computer Science & Engineering (Data Science) program trains students in algorithmic thinking and statistical modeling. This interdisciplinary field blends computer science, mathematics, and business analytics. Driven by massive industry demand, it is widely recognized as a highly dynamic and lucrative global career path.

Editor’s Message

Dr. K. D. Chaudhari
Editor (Faculty)

DEAR READERS,

We are pleased to introduce the newest edition of our Department Magazine. This publication serves as a medium to exchange knowledge, thoughts, and accomplishments in the fields of Computer Science & Engineering (Data Science). It includes informative articles, career support from the Training and Placement Cell, and select of students' research publications. We hope these pages stimulate curiosity, promote innovation, and encourage every student to pursue continuous learning. May this magazine enhance the connection between academia and industry while helping our students build successful and fulfilling careers.

Editorial Board

Ms. Riya Deshmukh

Mr. Jinendra Gavit

Mr. Vaishnav Patil

Contents

Section-I: Technical Articles

Section-II: Career Horizon

Section-III: Selected Student Articles

Technical Article

Building Data Products: Beyond Dashboards

Prof. K. D. Deore

For most of the last decade, data teams were measured by the dashboards they built. That model is changing fast. The most valuable data organisations in 2025 build data products instead.

A data product is a well-defined, maintained, and owned data asset. It serves a clear purpose. It comes with quality guarantees. It evolves based on user feedback. Think of it like a software product, but the output is data itself.

Dashboards versus Data Products

A dashboard shows numbers. Someone built it, shared it, and moved on. When the data source changes, the dashboard breaks silently. No one owns it anymore. This happens in every organisation with a mature analytics team.

A data product is different. It has an owner. It has a version. It has documented assumptions and known limitations. Users can rely on it the same way they rely on a production API. Reliability is the defining quality.

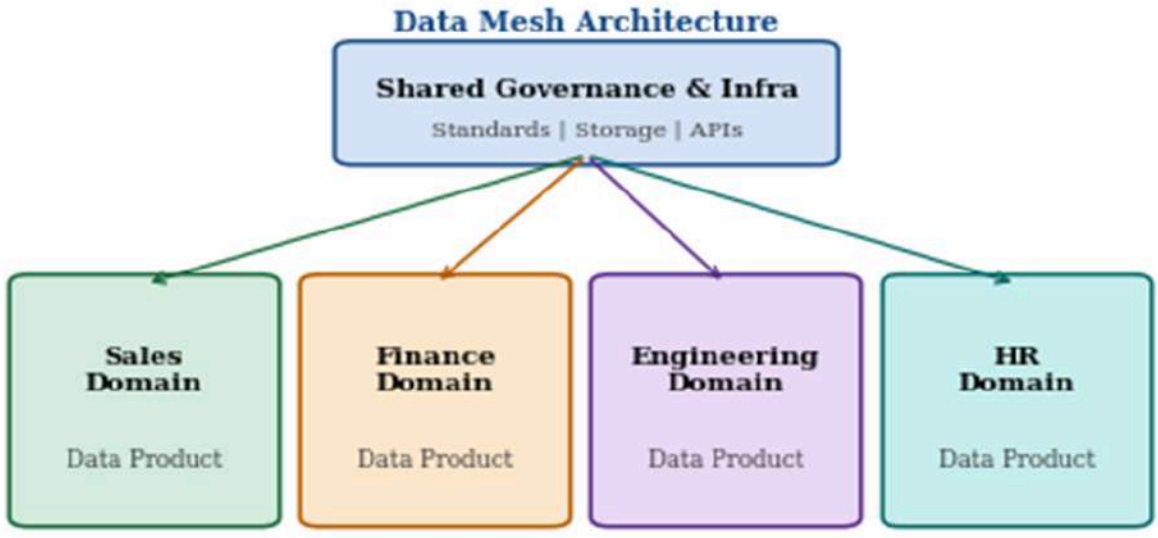

Fig. 1 — Data Mesh: each domain owns its data product; all governed by shared infrastructure and standards

Technical Article

Building Data Products: Beyond Dashboards

Data Mesh: Organising Around Products

Data Mesh is an architectural approach that distributes data ownership to the teams who know the data best. Each business domain owns and operates its own data products. A shared governance layer sets standards for the whole organisation.

The central data team no longer becomes a bottleneck. Domain teams in sales, finance, engineering, and HR each serve their own data to the rest of the organisation. This scales far better than centralised data warehouses managed by one small team.

Feature Stores: The ML Data Product

For machine learning applications, the feature store is the most important data product. It stores, versions, and serves features used by ML models across the organisation. Instead of every team engineering the same features independently, everyone consumes them from one shared source.

Tools like Feast, Tecton, and Hopsworks provide production-grade feature store capabilities. The feature store eliminates training-serving skew. The exact same features used during training are served at inference time. This mismatch is one of the most common causes of ML failures in production.

Real-Time Analytics: From Batch to Streaming

Overnight batch jobs are no longer enough for time-sensitive decisions. Fraud must be detected in milliseconds. Personalisation must respond to events as they happen. Ride-hailing platforms need real-time driver allocation data.

Apache Kafka and Apache Flink are the standard tools for real-time data products. Kafka buffers and routes high-throughput event streams. Flink performs stateful computations on those streams with sub-second latency.

Technical Article

Building Data Products: Beyond Dashboards

The Semantic Layer

Raw warehouse tables are not data products. A business user cannot interpret a table named fact_order_line_v3. A semantic layer translates raw structures into business-friendly terms: revenue, churn rate, active users, lifetime value.

Tools like dbt, Cube, and Looker provide semantic layer capabilities. When a semantic layer exists, business users can explore data independently. They no longer need a data analyst to translate every single request. This democratises data access across the organisation.

What This Means for You

Students who understand data product thinking will stand out in any data interview. Interviewers in 2025 ask not just what models you built, but who owns the data, how it is versioned, and what happens when the source changes.

Start thinking like a product owner, not just a model builder. The data products mindset is the skill that separates senior data professionals from those who only know how to run Jupyter notebooks.

Technical Article

Synthetic Data: The New Frontier for Model Training

Prof. S. P. Salunkhe

Machine learning needs data. More data almost always means a better model. But real-world data comes with serious constraints. Privacy laws limit what patient or financial records can be shared. Rare events like industrial accidents are infrequent by design. Labelling large datasets is expensive and slow.

Synthetic data addresses all three constraints at once. It is artificially generated data that preserves the statistical properties of real data. In 2025, it is one of the most rapidly evolving areas in applied machine learning.

What Makes Synthetic Data Useful?

Synthetic data is only useful if it is statistically faithful. It must preserve the correlations, distributions, and relationships that matter for the task. A synthetic medical dataset must reflect realistic correlations between symptoms and diagnoses. A synthetic transaction dataset must preserve the patterns that distinguish fraud from legitimate activity.

Fidelity is the key quality criterion. A high-fidelity synthetic dataset trains models that generalise to real data. A low-fidelity one produces models that fail the moment they meet the real world.

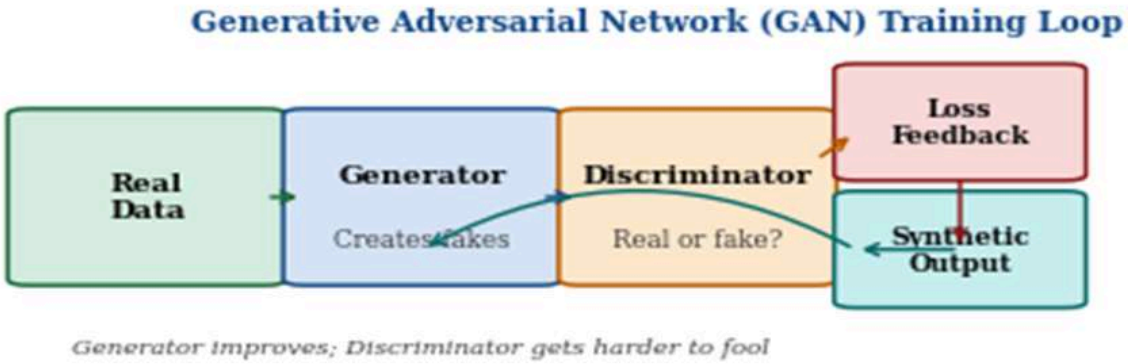

Fig. 2a — GAN training loop: the generator creates fake samples; the discriminator learns to detect them; both improve together

Generative Adversarial Networks were the first widely-used deep learning approach to data synthesis. A generator network produces synthetic samples. A discriminator network tries to distinguish synthetic from real. Both networks improve through competition until the generator fools the discriminator consistently.

Technical Article

Synthetic Data: The New Frontier for Model Training

Diffusion models now surpass GANs for image synthesis quality. They are also being adapted for tabular and time series data. Large language models are routinely used to generate synthetic text for training NLP systems.

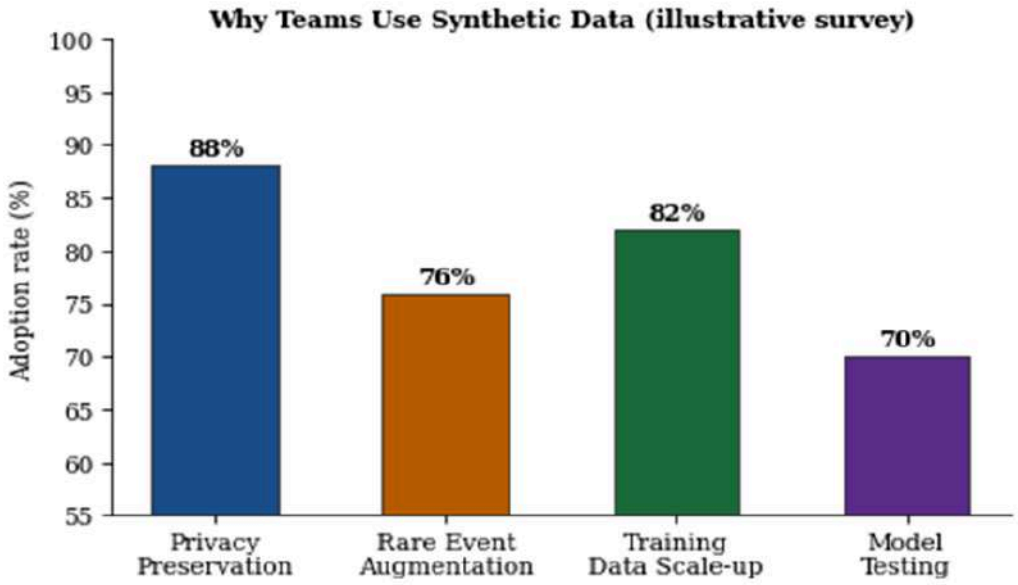

Fig. 2b — Why teams adopt synthetic data: illustrative survey across four primary use cases

Privacy-Preserving Synthesis

Synthetic data is often proposed as a privacy-safe alternative to real patient or customer records. The argument is appealing. But it is not automatic. A naive generative model can memorise and reproduce exact training samples. This is called memorisation, and it destroys the privacy benefit entirely.

Differential privacy provides a rigorous mathematical framework for guaranteeing that synthetic data cannot reveal individual records. Training generative models with differential privacy guarantees is the correct approach for sensitive data contexts.

Technical Article

Synthetic Data: The New Frontier for Model Training

Data Augmentation for Rare Events

Industrial safety models face a chronic data scarcity problem. A well-maintained factory may have only ten bearing failures per year. Training a predictive maintenance model on ten examples produces a model that cannot generalise.

Synthetic minority over-sampling generates additional rare-event examples by interpolating between existing ones. Physics-based simulation generates sensor readings for different failure modes at different stages of progression. The augmented dataset provides the diversity needed to train a robust model.

Limitations and Best Practice

Synthetic data is not a complete substitute for real data. Models trained only on synthetic data often fail when they meet real-world inputs. The synthetic distribution never perfectly captures the true complexity of the real one.

The best practice is to combine synthetic and real data strategically. Use synthetic data to fill specific gaps: rare classes, missing demographics, or privacy-restricted domains. Do not use it to replace real data entirely. Validate every synthetic-data-trained model on a real held-out test set before deployment.

Technical Article

End-to-End ML Pipelines with Airflow and MLflow

Prof. A. K. Patil

Building a machine learning model is only part of the job. The model must be trained on fresh data regularly. Its performance must be monitored continuously. It must retrain automatically when that performance degrades. All of this must happen without manual intervention.

Apache Airflow handles workflow orchestration. MLflow handles experiment tracking and model management. Together they form the backbone of a production ML pipeline.

Apache Airflow: Scheduling Workflows

Airflow lets you define data pipelines as Directed Acyclic Graphs, or DAGs, using Python code. Each node in the graph is a task. Airflow schedules tasks, manages dependencies between them, retries failed tasks, and provides a web UI for monitoring.

For ML pipelines, Airflow schedules nightly data ingestion, triggers preprocessing, kicks off model training, runs evaluation, and conditionally deploys the best model. The entire pipeline runs automatically. Data scientists can focus on improving models instead of babysitting workflows.

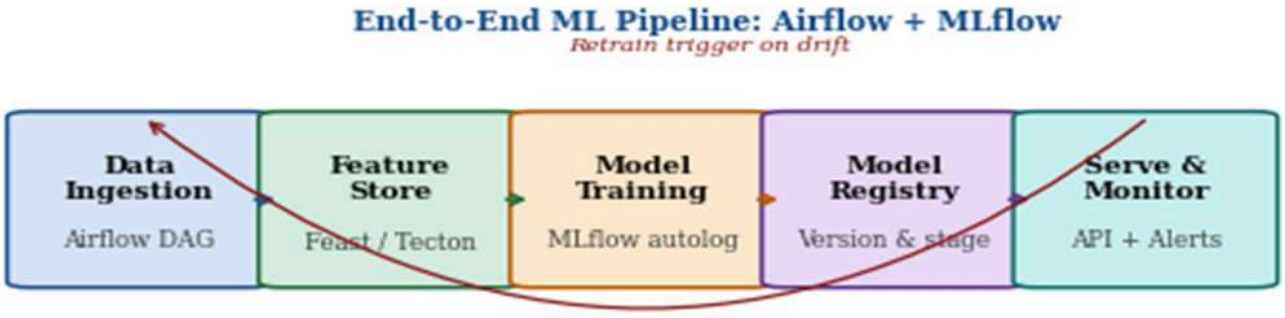

Fig. 3a — End-to-end ML pipeline: five stages orchestrated by Airflow, tracked by MLflow, with automatic retraining on drift

MLflow: Tracking Experiments and Models

MLflow tracks every training run automatically when you call mlflow.autolog(). It records model parameters, training metrics at each epoch, and the final model artefact. You can compare all past runs in the MLflow web UI with a few clicks.

Technical Article

End-to-End ML Pipelines with Airflow and MLflow

The MLflow Model Registry stores every model version with a lifecycle stage: Staging, Production, or Archived. Promotion from Staging to Production requires the new model to outperform the existing one on a held-out test set. This gate prevents regressions from reaching users.

Detecting and Responding to Drift

Model performance degrades over time as the world changes. You need to monitor two things. Data drift means the statistical distribution of input features has changed. Concept drift means the relationship between features and the target has changed.

Tools like Evidently AI generate drift detection reports that compare incoming production data against the training data distribution. When significant drift is detected, the Airflow pipeline triggers retraining automatically. No human intervention is needed for routine operation.

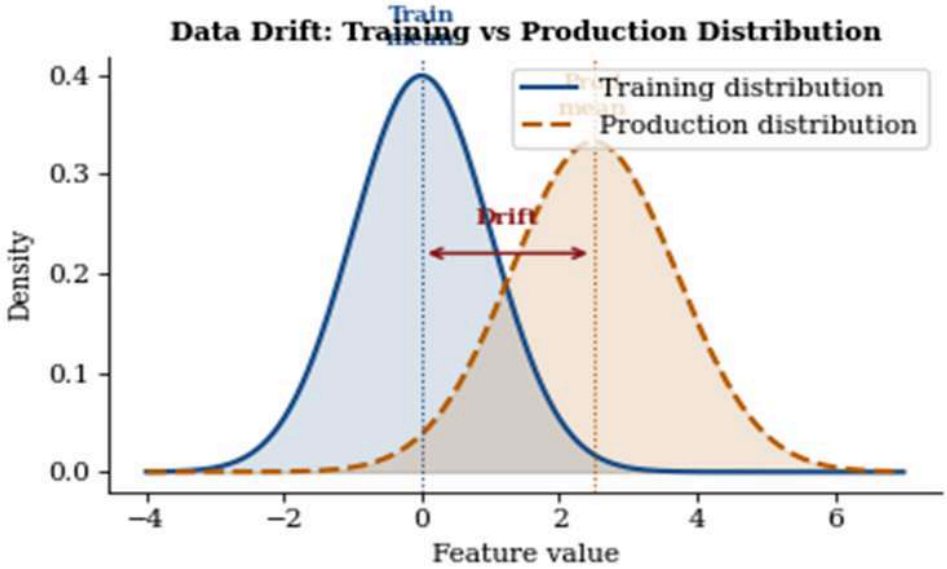

Fig. 3b — Data drift: the production distribution shifts away from the training distribution, degrading model accuracy

Building the Pipeline Step by Step

Start with data ingestion. Write an Airflow PythonOperator that reads fresh data, validates it, and writes cleaned features to the feature store. Schedule it nightly.

Technical Article

End-to-End ML Pipelines with Airflow and MLflow

The training task reads from the feature store and trains the model with MLflow autologging enabled. The evaluation task compares the new model against the current production model on a held-out test set. If the new model wins by more than a threshold, it is promoted automatically to Production in the MLflow Registry.

The monitoring task runs daily. It compares the distribution of production inputs against the training baseline. When drift exceeds a configured threshold, it fires an alert and triggers an immediate retraining run instead of waiting for the nightly schedule.

Why This Architecture Matters

This pipeline requires no human intervention for routine operation. The model self-heals when its environment changes. Engineers are alerted only when human judgment is genuinely required. This is the standard for production ML in 2025.

Implement this architecture even for your college projects. A project that includes a monitoring and retraining loop demonstrates a level of engineering maturity that most candidates never show in interviews.

Technical Article

Vector Databases and Semantic Search for Engineers

Prof. P. D. Lanjewar

When you search Google for "how to fix a leaking pipe," you get results about plumbing repair even if those results never use that exact phrase. The system matches by meaning, not by keyword. This is semantic search. The technology behind it is vector embeddings and vector databases.

In 2025, vector databases underpin every Retrieval-Augmented Generation system, every intelligent document search tool, and every AI-powered knowledge base. Understanding them is an essential skill for data scientists building AI applications.

What Is a Vector Embedding?

A vector embedding is a numerical representation of a piece of content as a list of floating-point numbers. These numbers are produced by a neural network trained to map similar content to similar vectors.

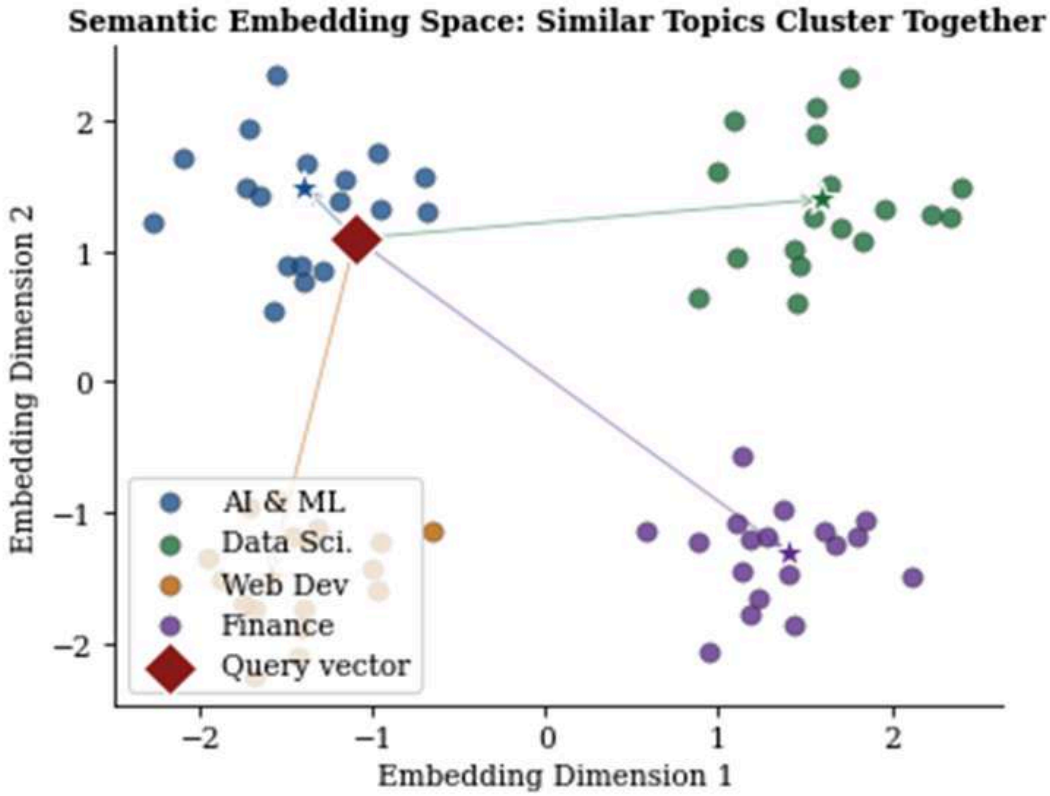

Fig. 4a — Semantic embedding space: similar topics cluster together; a query vector (diamond) finds its nearest neighbours

Technical Article

Vector Databases and Semantic Search for Engineers

Two sentences with the same meaning will have very similar vectors, even if they share no words. Two sentences with opposite meanings will have very different vectors. The similarity between two vectors is measured using cosine similarity.

How Vector Databases Work

A vector database stores high-dimensional vectors and answers nearest-neighbour queries: which stored vectors are most similar to this query vector? Exact search requires comparing the query against every stored vector. This is too slow for large collections.

Vector databases use Approximate Nearest Neighbour algorithms, or ANN, to find near-exact results in milliseconds at scale. Common algorithms include HNSW (Hierarchical Navigable Small World) and IVF (Inverted File Index). Chroma, Qdrant, Weaviate, and Pinecone are widely used vector database systems.

Building a Semantic Search System

Choose an embedding model first. For English text, the sentence-transformers library on Hugging Face offers many strong options. The model "all-MiniLM-L6-v2" produces 384-dimensional embeddings and runs efficiently on a CPU.

Embed all your documents using the chosen model. Store the resulting vectors in Chroma, which runs locally with no server setup needed. When a user submits a query, embed the query with the same model. Retrieve the top-k most similar document vectors. Present the matching documents to the user.

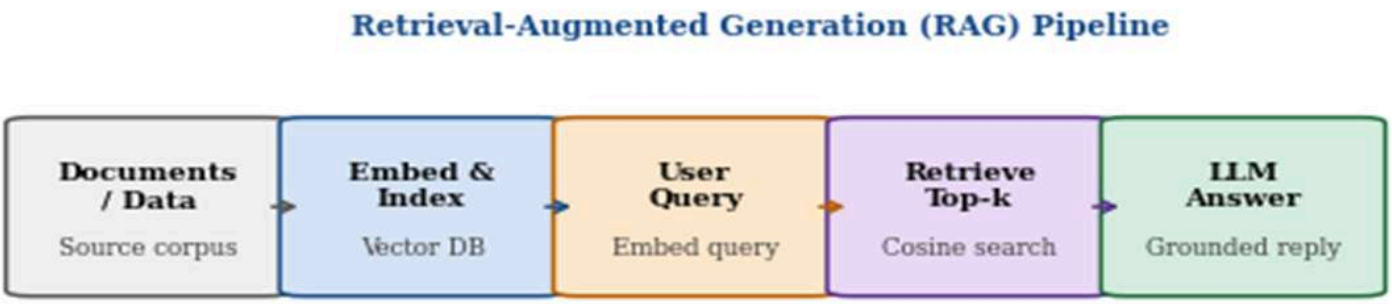

Fig. 4b — RAG pipeline: documents are embedded and indexed; at query time the top-k chunks are retrieved and sent to the LLM

Technical Article

Vector Databases and Semantic Search for Engineers

Retrieval-Augmented Generation

RAG combines vector search with a language model. The user asks a question. The system embeds the question and retrieves the most relevant document chunks from the vector database. Those chunks are included in the prompt sent to the language model.

The language model generates an answer grounded in the retrieved content. It is not asked to recall facts from training memory. Hallucination rates drop dramatically. This is the foundation of every enterprise knowledge assistant and intelligent document search system in 2025.

Engineering Considerations

Chunking strategy matters enormously. Split documents into chunks of 200 to 500 tokens with 20 percent overlap between adjacent chunks. Too large a chunk reduces retrieval precision. Too small a chunk loses the surrounding context that makes an answer coherent.

Evaluate your system quantitatively. For each test question, check whether the correct chunk appears in the top-k results. This retrieval recall metric is the foundation of RAG quality. If retrieval fails, the best language model cannot compensate. Invest time in tuning chunking and retrieval before optimising the generation step.

Career Horizon

The Data Science Job Market in 2025: Where Is It Heading?

Prof. K. D. Deore

The data science job market in 2025 has matured significantly from the gold-rush years of 2018 to 2022. The title "data scientist" no longer commands a premium for anyone with basic Python and ML skills. For those with deep applied expertise, however, opportunities are larger and more diverse than ever before.

The fundamental shift is this: companies have moved from exploring data science to operationalising it. Exploratory, research-oriented roles have been partially replaced by engineering-focused roles that build and maintain production ML systems at scale.

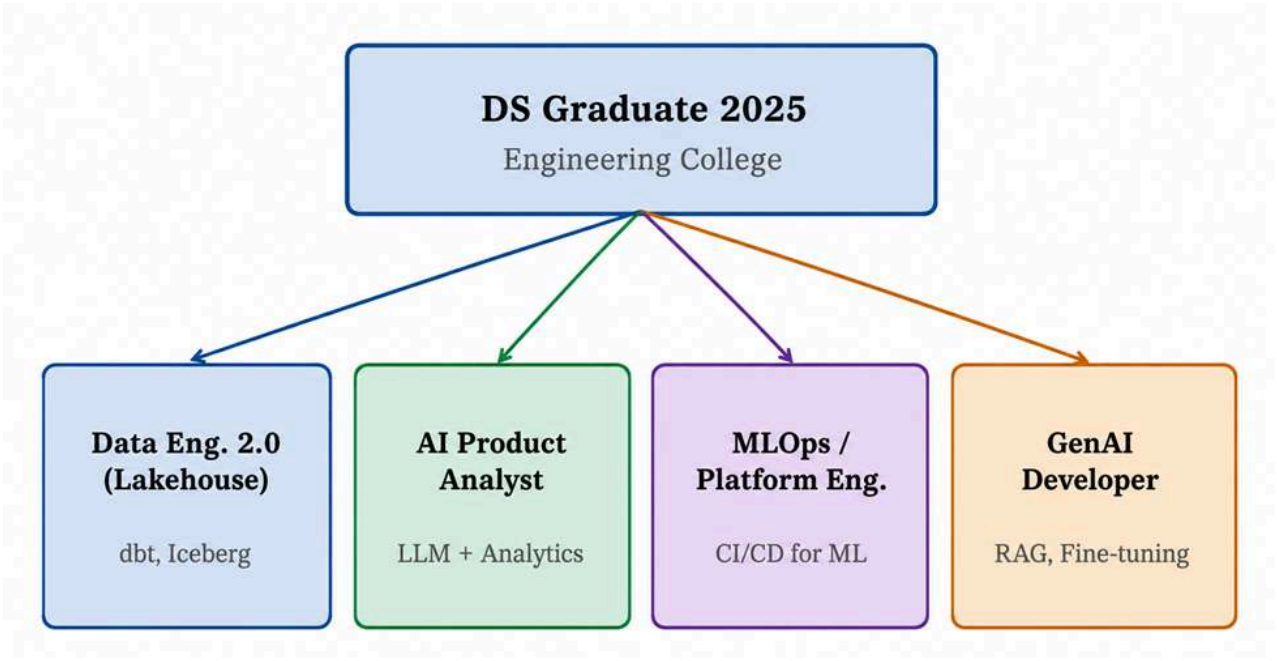

Fig. 5 — Hot data science roles in 2025: four high-demand specialisations for graduating engineers

Data Engineer 2.0

Modern data engineers are not just pipeline builders. They work with the lakehouse architecture: a hybrid of data lakes and data warehouses using formats like Delta Lake, Apache Iceberg, and Apache Hudi. They build pipelines that include LLM-powered processing steps alongside traditional transformations.

Career Horizon

The Data Science Job Market in 2025: Where Is It Heading?

AI Product Analyst

This role sits between data analyst and product manager. The AI product analyst understands both the technical capabilities of AI systems and the business context they operate in. They define success metrics for AI features, design evaluation studies, and translate findings into product decisions.

Statistical literacy, clear causal thinking, and the ability to communicate quantitative findings to non-technical audiences are the core competencies. Students who are passionate about data but also drawn to product and user experience thinking should target this path.

MLOps and Platform Engineer

As more companies move from experimental ML to production ML, the demand for MLOps engineers has grown sharply. MLOps stands for Machine Learning Operations. It is the practice of deploying, monitoring, and maintaining ML models in production systems reliably and efficiently.

MLOps engineers use Airflow for workflow orchestration, MLflow for experiment tracking, and Kubernetes for container management. They build monitoring dashboards and automated retraining systems. Salaries for this profile have increased 30 to 40 percent over the past two years in India.

GenAI Developer

The GenAI developer builds applications powered by large language models. This includes designing RAG pipelines, fine-tuning models on domain-specific data, building evaluation frameworks, and managing the lifecycle of LLM-powered features through production.

Required skills include Python, at least one LLM API, vector database experience, and familiarity with LangChain or LlamaIndex. Companies place heavy emphasis on demonstrated project experience over academic credentials for this role.

Salaries and Location

Bangalore accounts for approximately 45 percent of all open data science positions in India. Hyderabad and Pune each contribute around 20 percent. Remote work has expanded significantly: roughly 25 percent of data roles are now fully or partially remote.

Career Horizon

The Data Science Job Market in 2025: Where Is It Heading?

Entry-level salaries for strong candidates with a portfolio typically range from 6 to 12 LPA in Bangalore for standard DS and ML engineer roles. GenAI developer roles at product companies offer 10 to 18 LPA at entry level. MLOps engineers with one year of experience command 14 to 22 LPA.

Your Action Plan

Identify which of these four profiles matches your strengths and interests. Build your portfolio specifically for that profile rather than spreading effort across all areas. One deep, well-documented project beats five shallow ones in every interview.

Apply for internships in your chosen specialisation from your third year. Start building publicly on GitHub from now. The students who succeed in the 2025 and 2026 market are those who show something real that works, not just a notebook with a trained model. Set that as your standard today.

Selected Student Articles

Development Of AI/ML Based Solution for Detection of Face-Swap Based Deep Fake Videos

Abstract

Deepfake technology, especially face-swap manipulation, has emerged as a major threat because of its potential misuse in misinformation, identity fraud, and digital deception. This paper describes an AI-driven deepfake detection framework that incorporates hybrid CNN-LSTM architecture. Using the benchmark FaceForensics++ dataset, the model learns both spatial artifacts and temporal inconsistencies between original and manipulated videos. The proposed system attained an accuracy of 90.2 % and an AUC of 0.95, showing superior performance to state-of-the-art frame-based CNN methods.

Experimental results evidence that spatial and temporal learning combined brings significant improvements in robustness against compression and unseen manipulation.

Keywords – Deepfake Detection, Face-Swap Manipulation, Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), Deep Learning, Video Forensics, FaceForensics++, Spatial–Temporal Feature Learning, AI-based Media Authentication.

Rushikesh R. Rajput, Vaibhav R. Bhadane, Yashashri R. Borse, Suvarnsing P. Rajput, and Dr. Ujwala M. Patil.

Selected Student Articles

Student Dropout Prediction

Abstract

In India, student dropout is a major issue, especially in the case of students attending private schools, which are not affordable for the majority of the population. This affects not only the distribution of education but also the whole country's development. Therefore, the paper suggests the use of a machine learning-based system for early warning that would help in predicting which students might drop out by using several indicators such as academic performance, attendance, behavior, socio-economic status, and digital engagement. A diverse dataset was created by combining institutional student records with open-access educational datasets, and the data was preprocessed via normalization, encoding, and feature engineering techniques. A number of predictive models such as Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting, and XGBoost were trained and their performance evaluated with respect to accuracy, precision, recall, and Fl-score. The results indicated XGBoost as the best model, achieving an accuracy of 93%, a precision of 91%, a recall of 92%, and an F1-score of 91%, which is higher than that of the other baseline models. The system proposed makes it possible to detect these students at risk of failing in a proactive manner, which in turn facilitates timely academic, financial, and emotional support through the existing institution's mechanisms. The results prove that the use of ensemble learning techniques is very productive when it comes to dealing with complex educational data and can greatly improve the effectiveness of data-driven student retention strategies in educational institutions.

Keywords – Student Dropout Prediction, Educational Data Mining, Parental Engagement, Financial Support Systems, Student Engagement, Academic Performance Prediction

Purva Sonaje, Swamini Patil, Vaishnavi Borase, Jagruti Desale and Dr. P. S. Sanjekar