Shelke Rutvik Santosh, 2026, 14:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Artificial Intelligence In Data Analysis: A Review Of Automated Data Preparation, Predictive Analytics, Natural-Language Querying And Explainability

Shelke Rutvik Santosh[1]; Prof. S. R. Borhade[2]; Prof. M. C. Kolse[3]; Prof. S. K. Said[4]
Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India

Abstract—The volume, speed and variety of data produced by digital systems, sensors and online platforms have outgrown traditional analytical methods. Artificial intelligence enables organizations to extract insights from large, complex and unstructured data with greater speed and accuracy. This paper reviews how AI techniques, particularly machine learning, deep learning and natural language processing, are applied across the data analysis pipeline: automated data cleaning and preparation, automated machine learning for model and feature selection, pattern recognition and predictive modelling with ensemble methods, anomaly detection in transactions and sensor streams, and natural-language interfaces that translate questions into database queries so that non-technical users can explore data conversationally. Applications in healthcare, finance and retail are discussed. The review also examines explainable AI methods that attribute predictions to input features, and the problems of data quality, bias and fairness. A five-layer AI-assisted analytics architecture comprising data ingestion, preparation, modelling, explanation and conversational insight layers is proposed, together with research gaps, objectives, a comparison of traditional and AI-driven analysis and challenges such as privacy, compute cost and skill shortages. The review concludes that AI augments rather than replaces human analysts, freeing them for interpretation and strategic decisions.

Index Terms—Artificial intelligence, data analysis, machine learning, AutoML, anomaly detection, text-to-SQL, explainable AI, big data analytics

I. INTRODUCTION

A typical analyst spends much of the working day finding, cleaning and joining data before any real analysis begins. As data volumes grow from transactions, sensors, logs and social media, this manual approach becomes a bottleneck. Gandomi and Haider [1] described big data by its volume, velocity and variety and argued that analytics, not storage, is where value is created.

Artificial intelligence can automate many steps of the pipeline, from detecting dirty records to building and explaining predictive models, and even answering questions asked in plain language. Najafabadi *et al.* [2] reviewed how deep learning extracts representations from large and unstructured datasets.

This paper reviews AI across the data analysis pipeline and proposes a layered, human-centred analytics architecture.

II. BACKGROUND AND RELATED WORK

1. Types Of Analytics

Fig. 1 organizes analytics into descriptive (what happened), diagnostic (why it happened), predictive (what will happen) and prescriptive (what to do), and shows the AI techniques used at each level.

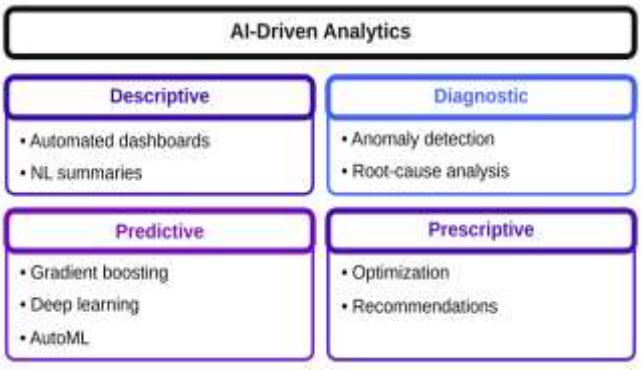


Fig. 1. Levels of analytics and supporting AI techniques.

© 2026 Shelke Rutvik Santosh This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

2. Data Analysis Pipeline

A pipeline consists of data collection, cleaning, transformation, modelling, evaluation, interpretation and communication. AI can assist at every stage, but human judgement remains essential in framing questions and acting on results.

3. Review Of Related Studies

Data preparation. Krishnan *et al.* [3] proposed ActiveClean, which prioritizes the records most likely to affect a model and cleans them interactively, showing that targeted cleaning improves model accuracy with less effort.

Automated machine learning. Feurer *et al.* [4] presented auto-sklearn, which uses Bayesian optimization and meta-learning to select algorithms and hyperparameters automatically. He *et al.* [5] surveyed AutoML across data preparation, feature engineering, model generation and evaluation.

Predictive modelling. Random forests [6] and gradient-boosted trees such as XGBoost [7] remain strong for tabular data.

Anomaly detection. Chandola *et al.* [8] surveyed anomaly detection techniques, and Isolation Forest [9] isolates anomalies efficiently with random partitioning, useful for fraud and sensor fault detection.

Natural-language querying. Zhong *et al.* [10] generated SQL queries from natural-language questions, and Yu *et al.* [11] released the Spider benchmark for complex cross-domain text-to-SQL, a key step toward conversational analytics.

Explainability and fairness. Lundberg and Lee [12] proposed SHAP for consistent feature attribution, and Mehrabi *et al.* [13] surveyed sources of bias in machine learning and fairness definitions. Table I summarizes these works.

TABLE I Summary of Literature Survey

Ref.	Author(s) / Year	Technique / Approach	Key Contribution	Limitation / Gap
[9]	Liu et al., 2008	Isolation Forest	Efficient anomaly detection	Global anomalies
[8]	Chandola et al., 2009	Anomaly detection survey	Taxonomy of methods	Pre-deep-learning
[4]	Feurer et al., 2015	Auto-sklearn	Automated model selection	Tabular data only
[3]	Krishnan et al., 2016	ActiveClean	Targeted data cleaning	Convex models
[7]	Chen and Guestrin, 2016	XGBoost	Scalable boosting	Needs tuning
[12]	Lundberg and Lee, 2017	SHAP explanations	Consistent attributions	Compute cost
[11]	Yu et al., 2018	Spider text-to-SQL	Complex query benchmark	Academic schemas
[13]	Mehrabi et al., 2021	Bias and fairness survey	Sources and definitions	Review only

III. RESEARCH GAP AND OBJECTIVES

1. Identified Research Gaps

1. 1) Automated cleaning tools still need domain knowledge to judge errors.

2. 2) Text-to-SQL systems may produce plausible but wrong queries.

3. 3) AutoML can hide important modelling assumptions from users.

4. 4) Explanations are often not understood by business users.

5. 5) Bias in historical data can be amplified by predictive models.

2. Objectives Of The Review

- **O1:** To review AI techniques across the data analysis pipeline.
- **O2:** To examine AutoML, anomaly detection and natural-language querying.
- **O3:** To study explainability and fairness in AI-driven analysis.

- **O4:** To propose a five-layer AI-assisted analytics architecture.
- **O5:** To identify challenges and future directions.

IV. DESIGN OF THE PROPOSED SYSTEM

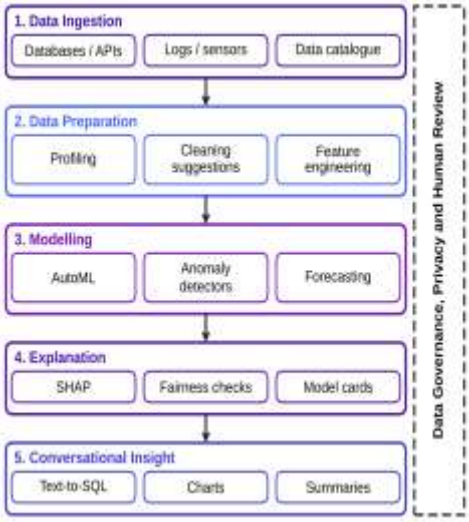


Fig. 2 shows the proposed architecture.

Fig. 2. Proposed five-layer AI-assisted analytics architecture.

1. Data Ingestion Layer

Connectors bring data from databases, spreadsheets, APIs, logs and sensors into a governed store with a catalogue and access control.

2. Data Preparation Layer

ML-assisted profiling detects missing values, duplicates, outliers and type errors and suggests fixes for analyst approval.

3. Modelling Layer

AutoML searches over algorithms such as gradient boosting and neural networks; anomaly detectors monitor streams.

4. Explanation Layer

SHAP values, partial dependence and fairness checks explain predictions and detect bias.

5. Conversational Insight Layer

A natural-language interface converts questions into SQL or code, shows the generated query for verification and returns charts and summaries.

6. Methodology

Fig. 3 describes the workflow. Model quality is measured on held-out data, for example with the root mean squared error for regression,

$$RMSE = \sqrt{(1/n)\,\Sigma\,(y_i - \hat{y}_i)^2} \qquad (1)$$

and each prediction is explained by additive feature contributions,

$$f(x) = \varphi_0 + \Sigma_j\, \varphi_j \quad (2)$$

where φ_j is the SHAP value of feature j.

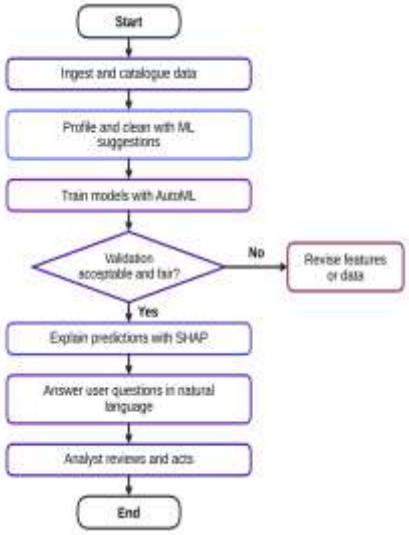


Fig. 3. AI-assisted data analysis workflow.

V. DISCUSSION AND COMPARISON

Table II compares traditional and AI-driven data analysis. AI is faster and handles unstructured data, but requires governance and explanation to be trusted.

Aspect	Traditional	AI-Driven
Data preparation	Manual scripts	ML-suggested cleaning
Model building	Hand-tuned	AutoML search
Unstructured data	Limited	Text, images, logs
Access for non-experts	Low	Natural-language queries
Transparency	High	Needs XAI tools

TABLE II Traditional vs. AI-Driven Data Analysis

VI. CHALLENGES AND FUTURE DIRECTIONS

Table III summarizes challenges. Future directions include LLM-based analytical agents with verification, privacy-preserving analytics, real-time streaming insights and data literacy programmes.

Challenge	Mitigation
Poor data quality	Profiling; validation rules
Bias	Fairness audits; diverse data
Wrong generated queries	Show query; human verification
Privacy	Access control; anonymization
Skill shortage	Training; guided tools

TABLE III Challenges and Mitigation

VII. CONCLUSIONS

AI is transforming data analysis by automating preparation, modelling and querying and by making insights accessible to non-experts. The reviewed literature shows mature methods for prediction and anomaly detection and rapid progress in AutoML and natural-language interfaces. Explainability, data quality and fairness must accompany these tools so that AI augments human analysts and supports reliable, evidence-based decisions.

1. Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

REFERENCES

1. A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," Int. J. Inf. Manage., vol. 35, no. 2, pp. 137–144, 2015.
2. M. M. Najafabadi et al., "Deep learning applications and challenges in big data analytics," J. Big Data, vol. 2, Art. no. 1, 2015.
3. S. Krishnan, J. Wang, E. Wu, M. J. Franklin, and K. Goldberg, "ActiveClean: Interactive data cleaning for statistical modeling," Proc. VLDB Endow., vol. 9, no. 12, pp. 948–959, 2016.
4. M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in Proc. NeurIPS, 2015, pp. 2962–2970.
5. X. He, K. Zhao, and X. Chu, "AutoML: A survey of the state-of-the-art," Knowl.-Based Syst., vol. 212, Art. no. 106622, 2021.
6. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
7. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794.
8. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Surv., vol. 41, no. 3, Art. no. 15, 2009.
9. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in Proc. 8th IEEE Int. Conf. Data Min. (ICDM), 2008, pp. 413–422.
10. V. Zhong, C. Xiong, and R. Socher, "Seq2SQL: Generating structured queries from natural language using reinforcement learning," arXiv:1709.00103, 2017.
11. T. Yu et al., "Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task," in Proc. EMNLP, 2018, pp. 3911–3921.
12. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, vol. 30, 2017.
13. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Comput. Surv., vol. 54, no. 6, Art. no. 115, 2021.