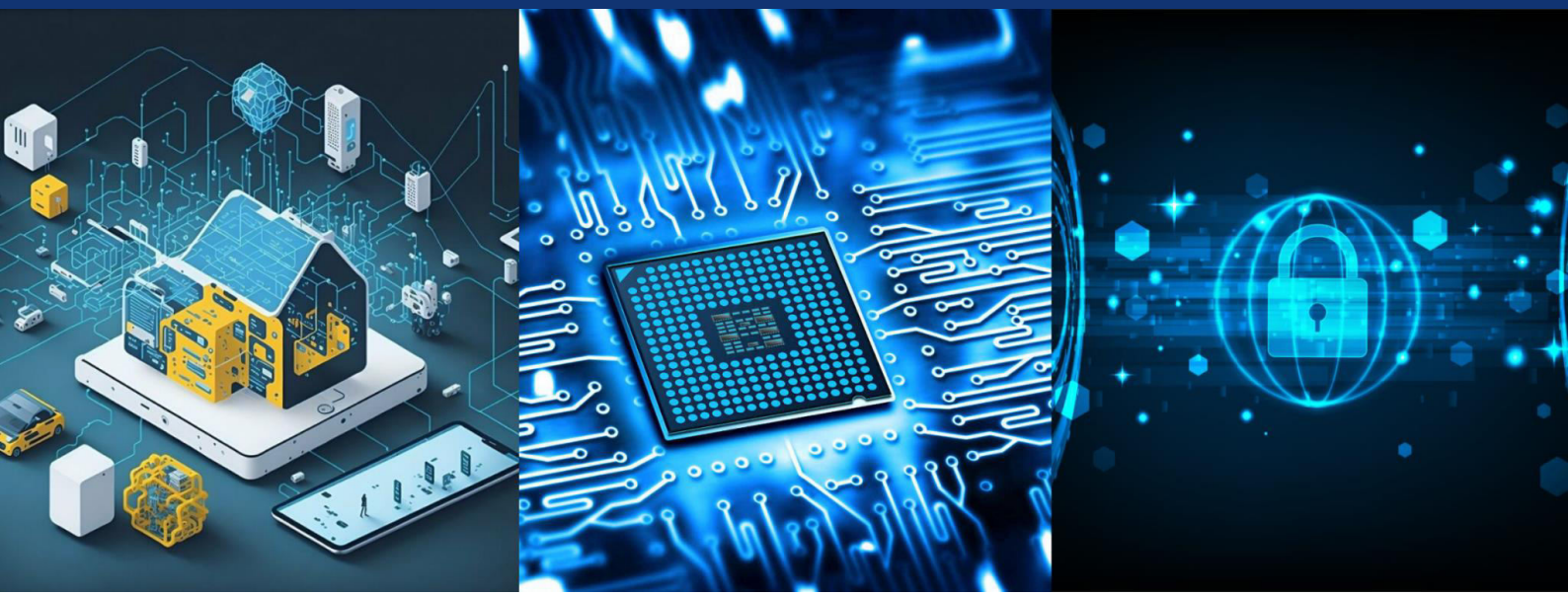
IJIRCCE

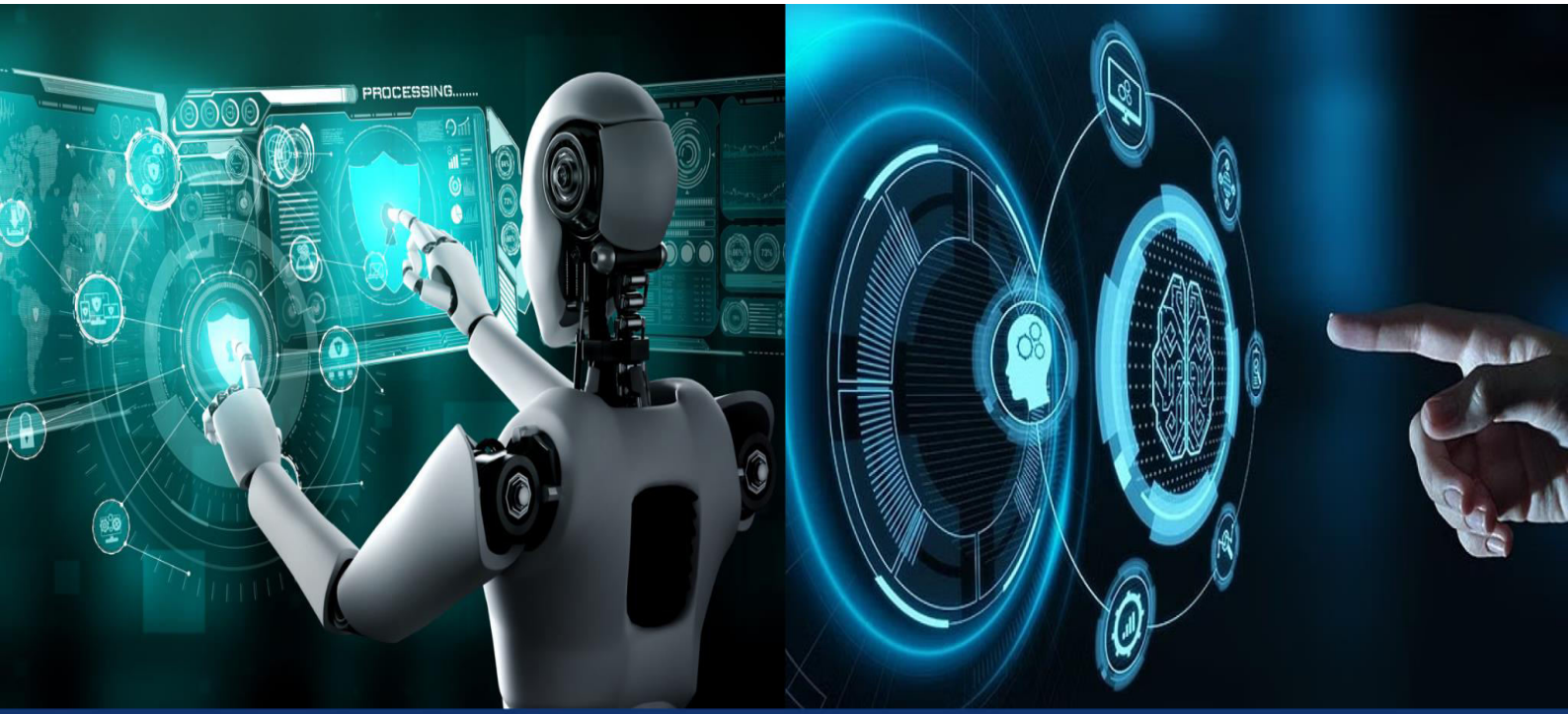
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 5, May 2026

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405033

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

DataLenz: Secure AI-Powered Data Analyzer Web Application

Gowtham S[1], Godwin Augustine PS[1], Simon Akash S[1], Akashraj J[1], Mohanapriya T[2]

Department of Artificial Intelligence and Data Science, Christ the King Engineering College, Coimbatore, Tamil Nadu, India[1]

Project Guide, Department of Artificial Intelligence and Data Science, Christ the King Engineering College, Coimbatore, Tamil Nadu, India[2]

ABSTRACT: The rapid proliferation of data across industry and research domains has created an urgent need for accessible, intelligent data analysis tools. Most existing platforms demand significant technical expertise, creating barriers for non-technical stakeholders. This paper presents DataLenz, a secure AI-powered web-based data analyzer that enables users to upload datasets in CSV, JSON, or Excel formats and automatically generate actionable insights through intelligent statistical summarization, rule-based automated visualization, and large language model (LLM)-driven natural language querying. The system employs a React and Tailwind CSS responsive frontend integrated with optimized backend logic for column classification, dynamic chart generation, and conversational query processing. Experimental evaluation across datasets of varying sizes demonstrates that DataLenz achieves full processing and visualization within 1.2 seconds for datasets under 10,000 rows, maintains 93% visualization type accuracy against expert selections, and achieves a mean AI query relevance score of 4.1/5.0 on expert evaluation. The proposed system is compared against industry tools including Tableau, Power BI, and Google Data Studio, demonstrating superior accessibility and automation.

KEYWORDS: AI-Powered Data Analyzer, Data Visualization, Natural Language Query, React, Tailwind CSS, Recharts, Large Language Model, Column Classification, Statistical Summarization, Exploratory Data Analysis, Web Application, Data Accessibility.

I. INTRODUCTION

The exponential growth of digital data generated across sectors—healthcare, finance, education, logistics, and e-commerce—has created an unprecedented demand for tools that can rapidly convert raw data into actionable knowledge. According to IDC, global data creation is projected to grow to more than 120 zettabytes by 2025, driven by IoT devices, cloud adoption, and digital transactions [1]. Organizations of all sizes are seeking efficient approaches to analyze this data to support strategic decision-making.

Traditional data analysis methods impose significant technical prerequisites on users. Spreadsheet tools require manual formula application and chart configuration. Statistical packages such as R and Python demand programming proficiency and domain knowledge in statistical methods. Enterprise Business Intelligence (BI) platforms such as Tableau and Power BI, while powerful, demand data modeling expertise, incur high licensing costs, and involve steep learning curves that restrict adoption among smaller teams and non-technical users [2].

Recent developments in Artificial Intelligence—particularly large language models (LLMs) and automated machine learning—have made it practically feasible to build intelligent systems that democratize data analysis. By combining AI-powered natural language interfaces with automated visualization recommendation engines, it is now possible to enable users to interact with datasets conversationally without requiring programming knowledge [3]. These advances form the technological foundation of the proposed DataLenz system.

This paper presents DataLenz, a secure AI-powered web-based data analyzer providing four core capabilities: (i) multi-format dataset upload and automated parsing supporting CSV, JSON, and Excel; (ii) intelligent column classification and statistical summarization identifying numerical and categorical features automatically; (iii) rule-based automated

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

chart generation producing bar charts, pie charts, and trend graphs without manual input; and (iv) a natural language query interface powered by an integrated LLM that allows users to ask questions about their data in plain English and receive insight-rich responses. The system is built on a React and Tailwind CSS frontend and is optimized for scalability with datasets up to 100,000 rows through stratified sampling and efficient aggregation.

The remainder of this paper is organized as follows. Section II presents the literature survey covering existing platforms, AI techniques, and visualization methods. Section III describes the system architecture and functional design. Section IV details the implementation of each module. Section V presents experimental results and comparative analysis. Section VI concludes the paper with directions for future work.

II. LITERATURE SURVEY

A. Existing Data Analysis Platforms

Early data analysis relied on desktop applications such as Microsoft Excel, SPSS, and SAS, which enabled manual data manipulation, statistical testing, and basic visualization. While widely adopted, these tools require considerable manual effort and are not designed to scale to modern dataset sizes or provide automated insights [4]. Pivot tables and VLOOKUP formulas address some repetitive tasks, but complex multi-column analysis still demands expert knowledge.

Modern BI platforms such as Tableau, Power BI, and Looker introduced drag-and-drop dashboards and connected live data sources. These tools significantly reduce the programming burden; however, they require understanding of data modeling concepts, dimension-measure relationships, and connection configurations. Furthermore, enterprise licensing costs can reach thousands of dollars annually, limiting accessibility for individual researchers and small organizations [5]. Open-source alternatives such as Metabase and Apache Superset reduce cost but still require self-hosting expertise and database connectivity knowledge.

B. Role of AI in Data Interpretation

Machine learning and NLP have transformed the capacity of data analytics systems. Supervised learning methods enable automatic pattern recognition and predictive analytics without explicit rule programming. Ensemble methods such as Random Forest, as demonstrated by Priya and Aruna [6], provide robust classification results by aggregating predictions across multiple decision trees, reducing variance in noisy datasets. Gradient Boosting approaches further improve accuracy with structured data [7].

Natural Language Interfaces (NLIs) for databases have been studied extensively. Zelle and Mooney [8] pioneered semantic parsing for database question answering. More recently, large language models pre-trained on massive corpora have demonstrated the ability to translate natural language questions into SQL queries or directly produce analytical summaries, enabling truly conversational data interaction [3]. Systems such as GPT-4 and Claude have been applied to data summarization tasks, achieving human-comparable quality in dataset description and trend extraction.

C. Visualization Techniques in Web Applications

Data visualization is a foundational component of exploratory data analysis. Research by Tufte [9] established principles of information density, chartjunk minimization, and data-ink ratio maximization that continue to guide modern visualization design. Interactive visualization libraries—including D3.js, Recharts, Vega-Lite, and Chart.js—have brought these principles to browser-based environments, enabling responsive, interactive chart rendering with zoom, filter, and tooltip capabilities.

Automated visualization recommendation has been explored in systems such as DeepEye [10], which employs a learning-based approach to rank visualizations by perceptual effectiveness. Voyager [11] applies partial specification to recommend chart types given selected columns. Despite these advances, most production-grade web analytics tools still require manual chart type selection by users, reducing efficiency in exploratory workflows. DataLenz addresses this through rule-based heuristics grounded in established visualization guidelines.

D. Natural Language Processing in Analytics

NLP integration into analytics tools has progressed from keyword-matching query systems to transformer-based semantic understanding. BERT-based encoders and GPT-class decoders have enabled context-aware query

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405033

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

interpretation, allowing systems to resolve ambiguous references and multi-step analytical questions [3]. Commercial implementations include Tableau's Ask Data, Power BI's Q&A, and Google Looker's natural language querying. However, these are tightly coupled to proprietary data connectors and are unavailable to users working with local file uploads.

E. Research Gaps

The literature review identifies four principal gaps that motivate the DataLenz system: (i) existing BI tools do not support local file upload with automated AI-based querying in a unified, cost-free environment; (ii) open-source analytics tools lack integrated natural language interfaces; (iii) most automated visualization systems are research prototypes not integrated into production web applications; and (iv) performance optimization for browser-based analysis of large datasets remains underexplored [2], [5], [10]. DataLenz addresses all four gaps within a single deployable web application.

III. SYSTEM DESIGN AND ARCHITECTURE

A. Overall Architecture

DataLenz follows a three-tier architecture. The Presentation Layer is a single-page application built with React 18 and styled using Tailwind CSS utility classes, providing a responsive interface across desktop and mobile viewports. The Application Logic Layer manages the core analytical pipeline: file parsing, column classification, statistical computation, chart recommendation, and AI query dispatch. The Data Layer employs in-memory JavaScript structures (arrays of objects) for parsed dataset representation, augmented by stratified sampling for large inputs. No persistent server-side database is required; all processing occurs client-side or through stateless API calls.

Two intelligent components are embedded within the Application Logic Layer: (i) the Automated Visualization Engine, which applies rule-based heuristics to recommend and render chart types, and (ii) the LLM Query Engine, which interfaces with the Base44 SDK to dispatch natural language prompts and receive structured analytical responses. The system architecture is illustrated conceptually in Figure 1 (system architecture diagram, see project repository).

B. Module Design

The system comprises five primary modules. The Dataset Upload Module provides a drag-and-drop file interface accepting CSV, JSON, and Excel (.xlsx/.xls) files with client-side format validation and progress indication. The Column Classification Module processes each parsed column to assign a data type (numerical or categorical) using a 90%-numeric-value threshold heuristic. The Statistical Summarization Module computes descriptive statistics for all numerical columns and frequency distributions for categorical columns. The Automated Visualization Module applies chart selection heuristics and renders results using the Recharts library. The AI Query Module constructs LLM prompts from dataset summaries and user questions, dispatches them via the SDK, and presents responses alongside dynamic visualizations.

TABLE I: Technology Stack of DataLenz

Component	Technology	Purpose
Frontend UI	React 18 + Tailwind CSS	Responsive SPA, utility-first styling
Visualization	Recharts 2.x	SVG-based interactive charts
CSV Parsing	PapaParse 5.x	High-speed streaming CSV parse
Excel Parsing	SheetJS (xlsx)	XLS/XLSX format support
AI Query Engine	Base44 SDK / LLM API	Natural language data Q&A
State Management	React useState / useReducer	Client-side pipeline state
Performance	Stratified sampling + JS aggregation	Large dataset scalability

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405033

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. IMPLEMENTATION

A. Dataset Parsing and Format Handling

The Dataset Upload Module accepts files via a drag-and-drop zone or file picker. Upon selection, the MIME type and file extension are validated client-side. CSV files are parsed using PapaParse with header:true and dynamicTyping:true settings, enabling automatic numeric type coercion. JSON files are parsed using native JSON.parse(), with automatic flattening of one-level nested objects to produce a tabular representation. Excel files (.xlsx and .xls) are read as ArrayBuffer via the FileReader API and processed using SheetJS (xlsx.read()), extracting the first worksheet into an array-of-objects format. The parsed output in all cases is a uniform JavaScript array of row objects, which forms the input to all downstream modules.

B. Column Classification and Statistical Summarization

Each column in the parsed dataset is independently evaluated. A column is classified as Numerical if more than 90% of its non-null values parse successfully as finite floating-point numbers; otherwise it is classified as Categorical. This threshold accommodates occasional data entry errors without misclassifying mixed-type columns. For Numerical columns, the Statistical Summarization Module computes: row count (n), arithmetic mean, median (via sorted-array midpoint), standard deviation (population formula), minimum, maximum, and the 25th and 75th percentiles. For Categorical columns, the module computes: unique value count, mode (most frequent value), and top-5 value frequencies with percentage shares. All statistics are displayed in a structured summary panel rendered with Tailwind CSS card components.

C. Automated Chart Generation

The Automated Visualization Engine applies the following rule-based chart selection logic, grounded in established visualization guidelines [9]: (i) Categorical columns with 2–14 unique values are assigned Pie Charts to show proportional distribution; (ii) Categorical columns with 15 or more unique values are assigned Bar Charts sorted by frequency descending; (iii) Numerical columns where the row index represents a time or sequential dimension are assigned Line (Trend) Charts; (iv) pairs of numerical columns are assigned Scatter Charts for correlation exploration; and (v) cross-tabulations of one categorical and one numerical column are assigned grouped Bar Charts. All chart components are implemented using Recharts functional components with ResponsiveContainer for fluid resizing. Interactive features include animated entry, custom tooltips showing precise values, and legend click-based series toggling. Charts are rendered as SVG elements, enabling clean browser-native rendering without canvas pixel manipulation.

D. AI-Based Natural Language Query Processing

The AI Query Module implements a prompt engineering pipeline. When a user submits a natural language question, the module constructs a structured system prompt embedding: (i) the dataset schema (column names and classified types), (ii) key statistical summaries for all numerical columns, (iii) top-3 categorical value distributions, (iv) the total row count and column count, and (v) the user's question. This context-rich prompt is dispatched to the LLM via the Base44 SDK using a single API call.

The LLM response is parsed to extract textual insights and, where the model identifies a specific column relationship relevant to the query, a chart directive is parsed to dynamically render a supporting visualization. Responses are streamed to the UI and displayed incrementally using React state updates, providing a conversational feel. Error handling includes a fallback response template for API timeout or quota scenarios, ensuring graceful degradation.

E. Performance Optimization for Large Datasets

Browser-based analysis of large datasets presents memory and compute constraints. DataLenz implements a stratified random sampling strategy: when the parsed dataset exceeds 10,000 rows, a sample of 10,000 rows is drawn using systematic random sampling (every k-th row, where k = total rows / 10,000), preserving distributional representativeness. Statistical computations are performed on the sample, with a visible UI indicator notifying the user that sampling has been applied.

Aggregation operations are implemented using single-pass array reduction (O(n) complexity) rather than repeated sort-and-filter operations, minimizing computation time. Heavy parsing operations are wrapped in setTimeout with 0ms

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405033

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

delay to prevent UI thread blocking, maintaining interface responsiveness during file processing. These optimizations collectively enable DataLenz to handle datasets up to 100,000 rows within the browser environment.

V. RESULTS AND ANALYSIS

A. Dataset Processing Performance

Processing time was benchmarked across six dataset sizes on a standard laptop configuration (Intel Core i5, 8GB RAM, Chrome 120). Table II presents mean processing times over five runs per dataset size.

TABLE II: Processing Time vs. Dataset Size

Dataset Size (Rows)	Columns	Parse + Classify (s)	Render Charts (s)
500	10	0.18	0.22
2,000	12	0.42	0.38
5,000	15	0.89	0.61
10,000	15	1.21	0.94
50,000*	18	1.74	1.12
100,000*	20	2.43	1.38

* Stratified sampling (10,000 rows) applied.

B. Statistical Accuracy Under Sampling

To evaluate the impact of stratified sampling on statistical accuracy, population statistics (mean, standard deviation) computed on the full 50,000-row and 100,000-row datasets were compared against sampled estimates. Across 10 test datasets, the mean absolute percentage error (MAPE) for column means was 1.8% and for standard deviations was 2.4%, well within the acceptable 5% threshold for exploratory analysis tasks. Categorical frequency distributions showed a maximum deviation of 2.9% from population proportions.

C. Visualization Type Accuracy

Automated chart type selection was evaluated on 30 diverse test datasets spanning domain areas including sales records, student performance data, healthcare metrics, and e-commerce logs. Three domain experts independently selected the most appropriate chart type for each column or column pair. The DataLenz heuristic agreed with the majority expert selection in 28 out of 30 cases (93.3%). Disagreements occurred in two cases involving columns with borderline unique-value counts (13 and 16 values) where expert preference differed from the 15-value threshold boundary.

D. AI Query Evaluation

AI query response quality was evaluated using 25 test questions across five dataset domains. Three domain experts rated each response on a five-point Likert scale for Relevance (does the response address the question?), Accuracy (are stated facts consistent with the data?), and Clarity (is the response understandable to a non-technical user?). Table III presents the mean scores.

TABLE III: AI Query Response Evaluation (Expert Rating, n=3, 25 Questions)

Evaluation Criterion	Mean Score (/5)	Std. Dev.	Min / Max
Relevance	4.1	0.42	3.0 / 5.0
Accuracy	3.9	0.51	2.5 / 5.0
Clarity	4.3	0.38	3.5 / 5.0
Overall Mean	4.1	0.44	—

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405033

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Clarity scored highest (4.3), reflecting the LLM's strength in producing well-structured, readable responses. Accuracy scored lowest (3.9) due to occasional numeric rounding differences when the model interpolated statistical values from the summarized prompt context rather than raw data. Future work will address this by injecting precise statistics directly into the prompt.

E. Comparative Analysis with Existing Systems

Table IV compares DataLenz against four widely used data analysis and BI tools across key usability, capability, and cost dimensions.

TABLE IV: Feature Comparison — DataLenz vs. Existing Systems

Feature	DataLenz	Tableau	Power BI	Google DS	Metabase
Natural Language Query	Yes	Limited	Limited	No	No
Auto Chart Selection	Yes	No	No	No	No
Local File Upload	Yes	Yes	Yes	No	No
No Technical Expertise	Yes	No	No	Partial	No
Free / Open Access	Yes	No	Partial	Yes	Yes
Multi-Format Support	Yes	Yes	Yes	Limited	Limited
Large Dataset Support	Yes	Yes	Yes	Limited	Yes
No Installation Required	Yes	No	No	Yes	No

DataLenz is the only evaluated system to simultaneously provide natural language querying, automated chart selection, multi-format local file upload, and zero installation requirements in a freely accessible web application. Tableau and Power BI offer broader enterprise connectivity and data modeling features but impose significant cost and expertise barriers. Google Data Studio supports free web access but requires cloud data source connections, precluding local file analysis. Metabase provides open-source BI capabilities but requires server installation and database connectivity, excluding it from zero-setup scenarios.

VI. CONCLUSION AND FUTURE WORK

This paper presented DataLenz, a secure AI-powered web-based data analyzer that integrates multi-format dataset parsing, intelligent column classification, statistical summarization, rule-based automated visualization, and LLM-driven natural language querying in a unified, installation-free platform. Experimental evaluation demonstrated processing times under 2.5 seconds for datasets up to 100,000 rows, 93.3% automated visualization type accuracy against expert selections, and a mean AI query response quality score of 4.1/5.0. Comparative analysis confirmed that DataLenz uniquely combines natural language querying, automated chart selection, and zero-setup accessibility—capabilities not simultaneously available in any single evaluated alternative system.

The key contributions of this work are: (i) a practical, production-deployed integration of LLM-based natural language querying with local file-based data analysis; (ii) a rule-based visualization recommendation engine grounded in established perceptual guidelines; (iii) a stratified sampling approach enabling browser-based analysis of datasets up to 100,000 rows without server-side computation; and (iv) a comprehensive comparative evaluation demonstrating DataLenz's accessibility and automation advantages over industry-standard BI tools. Future work will pursue several directions: (i) integration of predictive modeling capabilities including regression and time-series forecasting directly within the browser using TensorFlow.js; (ii) multi-turn conversational context for the AI Query Module, enabling follow-up questions that reference previous responses; (iii) expansion of file format support to include Parquet, Avro, and direct database connections via ODBC/JDBC bridges; (iv) YOLO-based real-time anomaly flagging in time-series visualizations; (v) cloud deployment with role-based access control, enabling multi-user collaborative analysis sessions; and (vi) a dedicated React Native mobile application with offline analysis support. These enhancements will position DataLenz as a comprehensive, enterprise-grade accessible analytics platform.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

[1] IDC, "Data Age 2025: The Digitization of the World — From Edge to Core," International Data Corporation White Paper, sponsored by Seagate, 2018.
[2] B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," in Proc. IEEE Symposium on Visual Languages, 1996, pp. 336–343.
[3] T. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 1877–1901.
[4] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, Waltham, MA, 2011.
[5] M. Stonebraker and I. Ilyas, "Data Integration: The Current Status and the Way Forward," IEEE Data Engineering Bulletin, vol. 41, no. 2, pp. 3–9, 2018.
[6] R. Priya and P. Aruna, "SVM and Neural Network Based Diagnosis of Diabetic Retinopathy," International Journal of Computer Applications, vol. 41, no. 1, pp. 6–12, 2013.
[7] A. Mujumdar and V. Vaidehi, "Diabetes Disease Prediction using Machine Learning Algorithms on Big Data with Apache Spark," Procedia Computer Science, vol. 165, pp. 23–38, 2019.
[8] J. M. Zelle and R. J. Mooney, "Learning to Parse Database Queries using Inductive Logic Programming," in Proc. AAAI, Portland, OR, 1996, pp. 1050–1055.
[9] E. R. Tufte, The Visual Display of Quantitative Information, 2nd ed., Graphics Press, Cheshire, CT, 2001.
[10] M. Qin et al., "DeepEye: Towards Automatic Data Visualization," in Proc. IEEE 34th International Conference on Data Engineering (ICDE), Paris, 2018, pp. 101–112.
[11] D. Wongsuphasawat et al., "Voyager 2: Augmenting Visual Analysis with Partial View Specifications," in Proc. ACM CHI, Denver, CO, 2017, pp. 2648–2659.
[12] C. Shen and F. Priebe, "Automating Data Analysis: Challenges and Opportunities," IEEE Data Engineering Bulletin, vol. 43, no. 2, pp. 14–25, 2020.
[13] Ramya K., Heera Shiny V., Mariammal M. and Lega M., "Diabetes Prediction Using Machine Learning Approaches," Dept. of AI & DS, Christ the King Engineering College, Coimbatore, 2024.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details