International Journal of Medical Toxicology & Legal Medicine **Volume 26, No. 3 and 4, 2023**
e-ISSN: 0974-4614
p-ISSN: 0972-0448

Intelligent Data Profiling for Healthcare Data Lakes Using AI-Enhanced Analytics

Bindu Madhavi Mangalampalli

Sr. BI Developer, Email: bindooo.madhaveee.3@gmail.com

Received: 11.10.2023 Revised: 08.11.2023 Accepted: 16.12.2023

ABSTRACT
Improved data quality, governance, and interoperability are critical to fulfilling the potential of data lakes for healthcare analytics. Well-crafted profiling of diverse source data is a prerequisite step for quality data lakes. Nevertheless, existing literature provides no clear guidance on operationalizing data profiling for heterogeneous healthcare data lakes, in part because key aspects of the profiling process remain underexplored. A set of necessary and sufficient data-governance requirements guides consequent profiling of structure, schema, lineage, quality, and anomalies in both clinical and nonclinical databases of an operational healthcare data lake. Profiling results reveal crucial evidence for intelligent data-governance decision-making and are formally disseminated within a metadata catalog.

The growing role of artificial intelligence (AI)—and especially machine learning (ML)—in the analytic processes that utilize data lakes has given rise to the notion of AI-Enhanced Analytics, which extends standard data-analysis processes with profiled data-knowledge aspects in order to improve discovery, quality, applicability, and generalization of the results. Profiling plays a key role in traditional data warehouses and their ETL processes, yet guidance on the profiling of complex heterogeneous data lakes remains limited—especially with respect to operational aspects. AI-Enhanced Analytics capabilities have now been employed to fulfill these operational requirements, enabling improved data-governance decision-making. Profiling aspects specifically fulfillment of data-governance requirements for data lakes in healthcare analytics.

Keywords: Healthcare Data Lake Governance, Operational Data Profiling Frameworks, Heterogeneous Clinical Data Management, AI-Enhanced Analytics, Machine Learning–Driven Data Quality, Metadata Catalog Integration, Data Lineage and Provenance Tracking, Schema and Structure Profiling, Anomaly Detection in Clinical Databases, Interoperability in Health Information Systems, Intelligent Data Governance Decision Support, Data Lake Quality Assurance, ETL Process Optimization in Healthcare, Structured and Unstructured Data Harmonization, Healthcare Analytics Infrastructure, Data Lifecycle Compliance Monitoring, Clinical and Nonclinical Data Integration, Governance-Driven Profiling Requirements, AI-Augmented Data Discovery, Enterprise Healthcare Data Management Systems.

1. INTRODUCTION

Increasingly rich and diverse sources of healthcare data, in response to the growing demand from precision medicine applications, have resulted in the emergence of healthcare data lakes. Due to the lack of conventional database systems and explicit data quality, healthcare data lakes exhibit severe data diversity and quality issues that retard analytics. Therefore, data profiling—a statistical examination of the data to gain information about it—has assumed vital importance in the context of enterprise data governance, accessibility, and usability. An improperly profiled data lake can impose catastrophic consequences, not only for healthcare institutions but also for human life, as it can generate misleading data analytics. Accurate data profiling thus constitutes an essential step toward realizing the promise of AI-enhanced analytics in healthcare data lakes.

Although an enterprise data lake conceptually supports the entire data lifecycle, the primary focus of data profiling has traditionally been to understand relational and structured data within a cluster of enterprise data warehouses. Currently, the efforts evolving around profiling in the context of data lakes are limited, particularly for heterogeneous data types frequently encountered in a healthcare data lake. In this context, the concepts, underlying objectives, and applicability of profiling of data types beyond relations in a typical healthcare data lake, along with the expected role and value proposition, are defined. The case studies presented narrow these aspects and the related process for profiling in a clinical data lake environment.

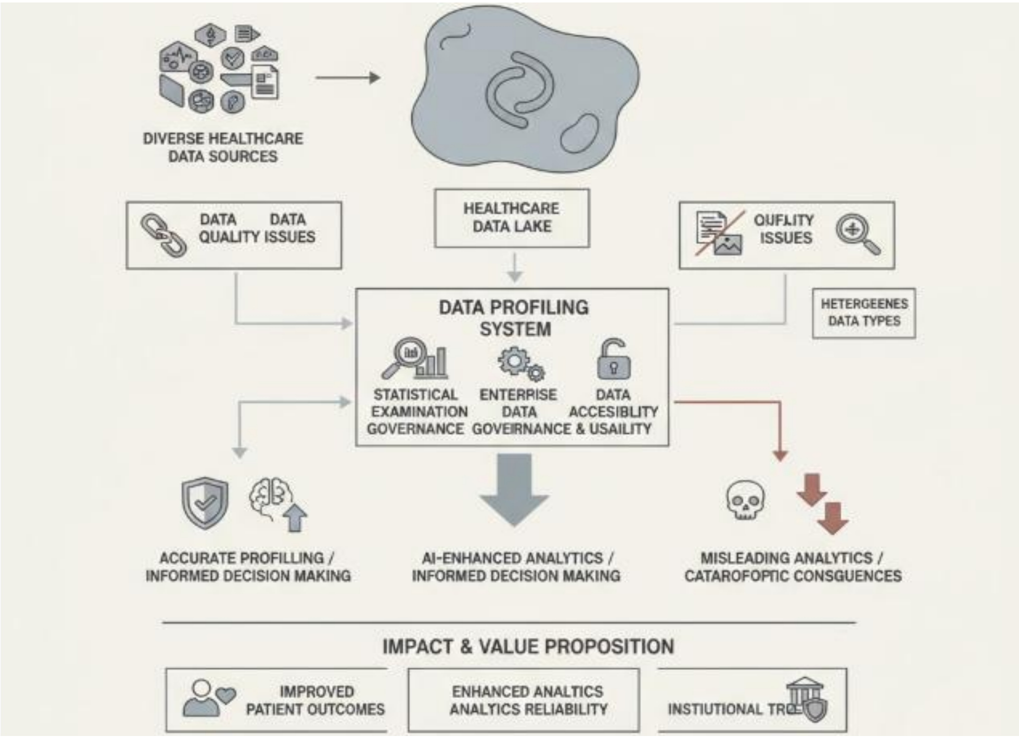

DIVERSE HEALTHCARE DATA SOURCES
DATA DATA QUALITY ISSUES
HEALTHCARE DATA LAKE
OUFʟITY ISSUES
HETERGEENES DATA TYPES
DATA PROFILING SYSTEM
STATISTICAL EXAMINATION GOVERNANCE
ENTERPRISE DATA GOVEIRNANCE
DATA ACCESIBLITY & USAILITY
ACCURATE PROFILLING / INFORMED DECISION MAKING
AI-ENHANCED ANALYTICS / INFORMED DECISION MAKING
MISLEADING ANALYTICS / CATAROFOPTIC CONSGUENCES
IMPACT & VALUE PROPOSITION
IMPROVED PATIENT OUTCOMES
ENHANCED ANALTICS ANALYTICS RELIABILITY
INSTIUTIONAL TR

Fig 1: Ensuring Integrity in Precision Medicine: A Framework for Profiling Heterogeneous Data within Healthcare Data Lakes

1.1. Overview of the Study

Healthcare analysts urgently need to evaluate the status and quality of data in data lakes, the new paradigm widely adopted by healthcare organizations and health insurance providers. These corporations collect and process massive amounts of structured, semi-structured, and unstructured data from multiple sources across long periods of time. Although data profiling is a pivotal first procedure in the development of any analytics project, the distinctive characteristics of healthcare data lakes, especially the heterogeneity of the data they contain, have reinforced the conception of profiling merely as a preliminary step.

To address this limitation, this study applies an AI-enhanced analytics conceptual framework and identifies specific data governance, privacy, security, and compliance requisites that must drive the implementation of the data profiling process. Such considerations condition the subsequent choice and application of data profiling techniques that yield informative results. The discussion is illustrated with real clinical data about patients and patients' relatives who traveled to Gaza during the 2014 armed conflict. The profiling outcomes provide valuable insights for improving data quality, interoperability, and governance in the analyzed clinical data lakes.

2. Background and Motivation

Data lakes are gaining traction in healthcare because of their capacity to ingest large volumes of data from different sources. Yet they should not be mistaken for a unqualified panacea. Healthcare decision makers remain wary of hidden risks lurking underneath the apparent advantages. Data profiling—one of the cornerstones of data governance—has become decisively important for all types of data management. In the case of healthcare data lakes, a particular profile is needed. Such a profile serves as a basis for risk assessment and impact analysis and aids in establishing trust in the data during data consumption. Austerity leads to a minimization of costs, which in turn justifies the creation of a role for profiling using AI-Enhanced Analytics. Its applications in AI-Enhanced Analytics, however, are yet to be fully realized. As a direct consequence, the existing profile suffers serious shortcomings.

Similar data profiles exist for conventional data warehouses for the analytics tools. These ways of knowledge representation are designed to facilitate the effective use of analytics tools, while also serving as the foundation for producing data-quality-related KPIs, Data-Growth-Related KPIs, and lineage-related KPIs. Such profiling enables the identification of the governing data owner and any requisite data quality issues, while also suggesting suitable analytics methods.

Equation 1) Anomaly detection equations (step-by-step)

The explicitly includes "anomaly hunting" as a profiling type.

1.1 Z-score anomaly (numeric)

For numeric field A:

1. Mean:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2. Standard deviation:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

3. Z-score for each value:

$$z_i = \frac{x_i - \mu}{\sigma}$$

4. Flag anomaly if:

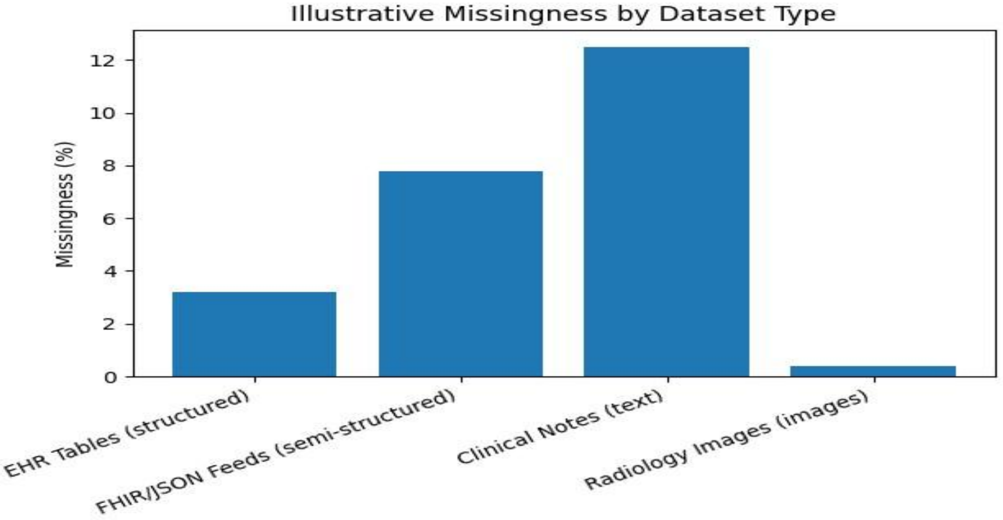
$$|z_i| > \tau$$

Commonly $\tau = 3$.

1.2 IQR rule (robust alternative)

1. Compute quartiles $Q1, Q3$, $\text{IQR} = Q3 - Q1$
2. Flag if:

$$x_i < Q1 - 1.5 \cdot IQR \quad \text{or} \quad x_i > Q3 + 1.5 \cdot IQR$$



2.1. Key Challenges in Data Quality Management for Healthcare Analytics

The primary motivations for deploying data lakes in healthcare centralize around data enablement for analytics, artificial intelligence (AI), machine learning (ML), data science, and research, and an inability to overcome existing data silos at an institutional or system level. Data enablement positions the data lake as a single source of truth for all constituents in the data ecosystem and a tool to develop intelligent queries and applications to streamline workflows and support faster and better-informed decisions. Given the multiple sources feeding clinical data lakes, AI-enhanced analytics, a novel and more-intelligent approach to BI and analytics—but whose benefits have not been applied to HPCDA and are often misunderstood—become a competent decision support tool. AI-enhanced analytics introduces three new stages to the traditional BI management cycle: auto-discovery, intelligent design, and profile-and-perform. AI algorithms reside in these phases to improve time-to-market for new reports, dashboards, or OLAP cubes. A key challenge in this process is the quality of the underlying data, including cleanliness, semantics, interoperability consistency—formal and informal—, and appropriate lineage. Answers to these questions should be simple and quick to obtain, yet they are not.

Anonymized data lakes—central repositories of integrated clinical, research, and administrative data—facilitate the application of AI-enhanced analytics within organizations aspiring to become AI-driven, help explore data lineage, provide a consolidated view with limited or no breaking changes, support data profiling, enhance data quality and governance, and enable seamless data querying, exploration, and responsiveness to research requests. Different healthcare silos generate and manage data in heterogeneous formats. The existing plethora of technologies, tools, and sources used by healthcare institutions for generating, storing, and analyzing data—often independently—hampers proper data enablement as it prevents the creation of a single source of truth across such organizations. The cross-institutional exploration of data lakes is further hampered by formal and informal incompatibilities across the stored data, the absence of metadata, and inappropriate data cleansing or quality management.

Dataset	Records/Files	Missingness %	Duplicate Rate %
EHR Tables (structured)	2500000	3.2	1.1
FHIR/JSON Feeds (semi-structured)	12000000	7.8	2.6
Clinical Notes (text)	800000	12.5	0.9
Radiology Images (images)	250000	0.4	0.2

3. Theoretical Foundations of Data Profiling in Healthcare

Health data lakes serve as strategic reservoirs for unstructured and semistructured data flowing from asocial and clinical systems. These data ponds contain data recognized as health assets by organizations but usually neglected due to custom storage solutions and eagerness to quickly store and forget data. The often-remarked emphasis on visual analytics and the promise of machine learning, deep learning, and text mining have diverted practitioners from proper preparation and readiness before analysis: whether an analysis will succeed depends more on the nature of the data than on pure analysis capability. Engineered analysis-ready data-sets are also a costly endeavor both for effort and risk. Despite the advantages offered by data lakes, data profiling specific to the context of a healthcare data lake remains scarce. Adequate data preparation before analysis depends on a proper understanding and evaluation of the data from a quality, governance, and interoperability viewpoint, and the interpretation of the results should guide subsequent analysis. Lack of proper profiling hinders analysis, as confirmed by recent cases in healthcare using natural language processing and deep learning. The concept of AI-Enhanced Analytics provides a unifying framework wherein added intelligence relies on proper preparation of the data rather than on the algorithm itself, and the work focuses on the components that foster such intelligence during profiling.

AI-Enhanced Analytics relies on proper data preparation and the understanding of healthcare domain-specific knowledge, such as the information requirements needed for a system under study. Each discipline is backed by a suite of algorithms addressing a specific task, yet the task of overseeing the entire exercise is not clearly defined. Indeed, many of the dedicated algorithms—such as those for natural language processing or machine learning—have been developed without the supervision of experts from the data and information engineering domain. For AI-Enhanced Analytics to fulfill its promise of improving interpretability and reducing the quantity of training data needed, it is vital that AI tools consider this knowledge and relate various developments. However, pure analysis readiness is seldom achieved, as AI-Enhanced Analytics lacks a long-established preparatory component that provides adequate data-sets for the specific analysis necessary to solve a business problem.

Equation 2) Data Quality equations (step-by-step)

2.1 Completeness / Missingness

Goal: detect incompleteness.

3. Let n_{total} be the number of expected values for a field A.
   For a column, $n_{\text{total}} = n$.
4. Let n_{missing} be the count of missing values (NULL/empty/NaN).
5. **Missingness rate**

$$\text{MissingRate}(A) = \frac{n_{\text{missing}}}{n_{\text{total}}}$$

5. **Completeness** (the complement)

$$C(A) = 1 - \frac{n_{\text{missing}}}{n_{\text{total}}}$$

2.2 Uniqueness / Duplicate rate

Goal: detect redundancy.

Assume a key field (or composite key) K.

6. Let $n_{\text{keys}} = n$.
7. Let $n_{\text{distinct}} = |\{K(r_i) : i = 1..n\}|$.
8. Duplicate key count:

$$n_{\text{dup}} = n_{\text{keys}} - n_{\text{distinct}}$$

4. **Duplicate rate**

$$\mathrm{DupRate}(K) = \frac{n_{\mathrm{dup}}}{n_{\mathrm{keys}}}$$

5. **Uniqueness**

$$U(K) = 1 - \mathrm{DupRate}(K) = \frac{n_{\mathrm{distinct}}}{n_{\mathrm{keys}}}$$

2.3 Validity (rule-based conformance)

Goal : evaluate quality against business rules.

Let a rule be a boolean predicate $R(x)$ (e.g., age in $[0{,}120]$, ICD code pattern, etc.).

1. For each value x_i, define an indicator:

$$I_i = \begin{cases} 1 & \text{if } R(x_i) = \text{True} \\ 0 & \text{otherwise} \end{cases}$$

3. Total checked $n_{\mathrm{checked}} = n$ (or fewer if excluding missing).
4. Valid count $n_{\mathrm{valid}} = \sum_{i=1}^{n} I_i$
5. Validity

$$V(A) = \frac{n_{\mathrm{valid}}}{n_{\mathrm{checked}}}$$

3.1. Knowledge Gaps in Healthcare Data Lakes

Considerable diversity in health data increases the need for advanced analytics capable of providing comprehensive responses to stakeholders' questions. AI-Enhanced Analytics proposed here combines AI techniques with the principle of enabling high-level data preparation and exploration. Many queries, however, require examination of the underlying data through appropriate profiling techniques. Profiling can help in detecting most issues associated with data quality, integration, and convergence. Added AI-enhanced methods facilitate selective sampling and redundancy detection across diverse data sources. These benefits are particularly relevant in the context of healthcare Data Lakes, where high variability and heterogeneity of the data exacerbate traditional data preparation methodologies. Nevertheless, data profiling in healthcare Data Lakes is largely limited to clinical and genomic data.

Despite the fact that several components of AI-Enhanced Analytics and profiling are established and recognized domains, academic studies on Data Lakes using AI techniques, and Dreampipe in particular, remain scarce. Dreampipe combines Lambda and Kappa architectures to create a Data Lake that integrates the cloud and edge computing paradigms into one cohesive analytics architecture. AI-Enhanced Analytics adds high-level exploration capabilities and usability improvements, such as seamless interoperability with business intelligence tools and automatic virtualization of data. Three-layer Data Lake security ensures privacy requirements, compliance with regulations, and protection against attacks.

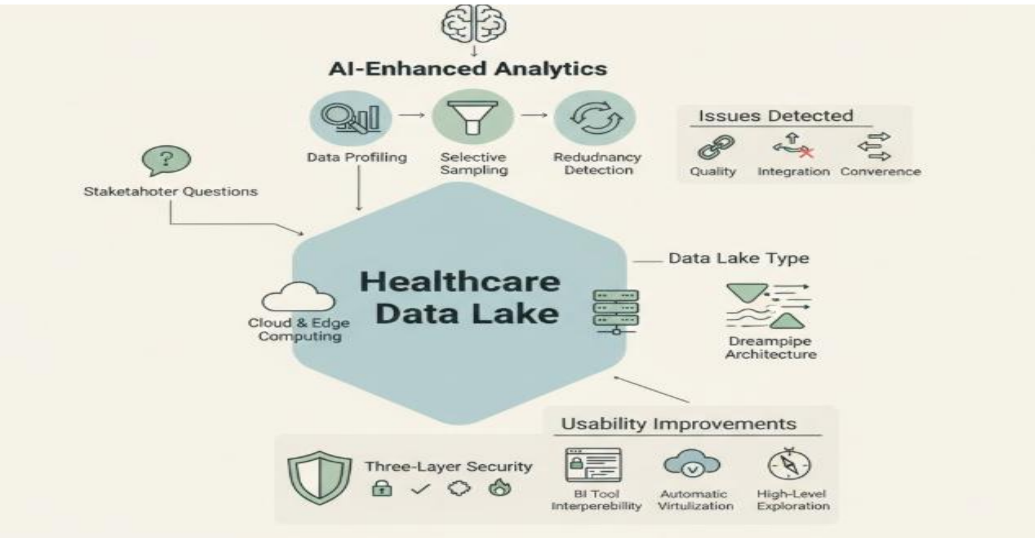


Fig 2: Dreampipe: A Unified Cloud-Edge Data Lake Architecture for High-Fidelity Healthcare Analytics and Governance-Centric Data Exploration

3.2. AI-Enhanced Analytics: Concepts and Rationale

Healthcare Data Lakes require AI-Enhanced Analytics capabilities capable of exploring and identifying data quality, governance, and interoperability challenges. AI-Enhanced Analytics refers to four activities: data preparation, data profiling, knowledge discovery, and knowledge exploitation. Data Profiling gathers knowledge about data and their metadata before they are integrated into a Data Warehouse. Data Profiling relies on the automated discovery and analysis of the Content, Structure, Quality, and Lineage of the data ingested into the Data Lake. These metrics offer vital knowledge about the Data Lake and its data quality and support the establishment of a Data Governance framework.

However, given that healthcare Data Lakes deal with heterogeneous data such as images, sound, text, and traditional data tables, Data Profiling still lacks a set of techniques and metrics capable of exploring these data types. Although the profiling of images using deep-learning models is still in its infancy, some works in the literature have already validated the approach. Data Profiling also requires Data Governance, Privacy, and Security considerations and procedures, since healthcare data include sensitive data for data subjects and their families. The benefits of AI can ease overcoming these limitations. With these capabilities, data profiling plays a crucial part in healthcare Data Lakes since it impacts data quality, Data Governance, and Data Interoperability.

4. Methodological Framework

Intelligent Data Profiling for Healthcare Data Lakes Using AI-Enhanced Analytics: an objective, evidence-based, formal analysis of data profiling in healthcare data lakes with AI-enhanced analytics, focusing on clarity and scholarly argumentation.

Data governance and privacy, security, and compliance regulations play critical roles in the constitution of healthcare data lakes and, therefore, also affect the profiling of the data stored in the lake. The design and implementation of the profiling processes need to ensure that the guidelines imposed by such considerations are correctly applied. Therefore, data governance policies, privacy conditions, security policies, and compliance management practices for profiling procedures should be defined in detail. Such considerations typically address the generation of data profiling technology systems. Data governance and privacy, security, and compliance management of data lakes play a fundamental role in the organization of the healthcare data lake—composition of data sources, confidentiality, security, legal regulations, and other similar aspects of the data lake governing the particular healthcare data lake, such as the locations of data collection and different consumer patient areas.

Data profiling analyzes and reviews the healthcare data lake's heterogeneous data for improving the data management accuracy of machine-learning training. A careful design of the data profiling technique in recommended practice ensures the involved data management systems achieve their operational goals by examining the corresponding data types, data schemas, data lineage, profiling quality metrics, and data-anomaly detection traders to achieve the operational aim of the healthcare data lake. The analysis of data lineage offers essential benefits for managing data quality. Profiling discovers major data types of heterogeneous data samples, identifies data-positioning problems, presents structured and unstructured schema profiles, provides data-quality statistics and metrics for critical machine-learning model training, and establishes data-anomaly detection traders for the healthcare data lake.

Equation 3) Structure + schema profiling equations

3.1 Data type distribution (structure)

Let field A have detected type per value: {int, float, date, string, …}.

2. Count values of each type: n_t
3. Proportion:

$$P(\text{type} = t) = \frac{n_t}{n}$$

3.2 Schema drift rate (batch-to-batch)

In ingestion batch b, let schema be a set of fields S_b.

4. Drift event between batches $b - 1$ and b:

$$\Delta_b = S_b \;\triangle\; S_{b-1}$$

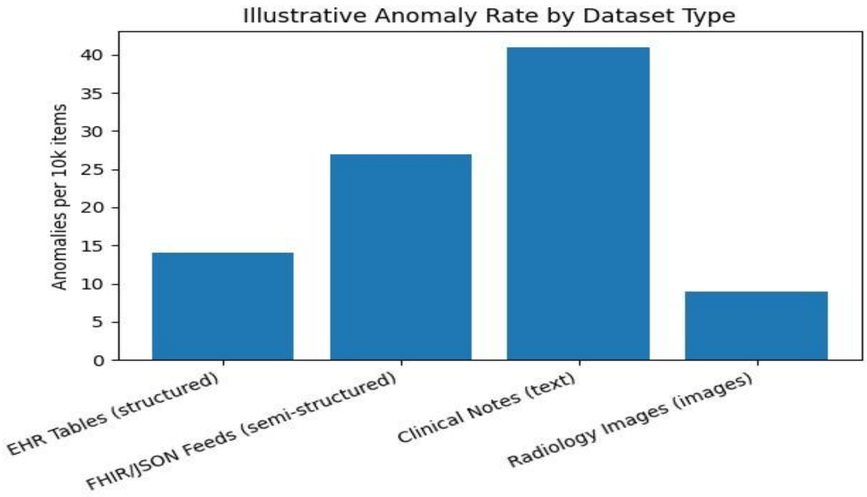
where $\triangle$ is symmetric difference (added/removed fields).

6. Drift indicator:

$$d_b = \begin{cases} 1 & \text{if } |\Delta_b| > 0 \\ 0 & \text{otherwise} \end{cases}$$

4. Over B batches, **schema drift rate**:

$$D = \frac{\sum_{b=2}^{B} d_b}{B - 1}$$



4.1. Data Governance and Compliance Considerations

Data governance, privacy, security, and compliance considerations orient and inform intelligent data profiling for healthcare data analytics—an objective, evidence-based formal analysis of data profiling in healthcare data lakes with AI-enhanced analytics. OECD defines data governance as the development of decision rights and responsibilities for information-related processes. It supports policy development, technical operations, and resource allocation. Healthcare data governance entails decisions regarding o Data right holders o Data access and sharing policies o Data uses and disclosures o Data custodians o Data repositories and data stewards o Data quality standards o Data documentation, usability, and usability standards o Data standardization and interoperability o Data advocacy.

The volume, variety, and velocity of big data impose strict conditions on identification and implementation of data governance best practices for privacy, security, and compliance with legal and ethical requirements. Confidentiality, privacy, and security concerns restrict availability, precluding full population, sampling, and testing of hypotheses. Safeguarding these concerns requires compliance with Federal Information Security Management Act, Social Security Administration sensitivity requirements, and Organizational for Economic Cooperation and Development privacy guidelines. Institutional review boards protect confidentiality by limiting access, imposing stringent requirements, and permitting only specific uses. Ensuring accountability, therefore, requires investment of time and resources.

4.2. Data Profiling Techniques for Heterogeneous Healthcare Data

A comprehensive understanding of healthcare data lakes and related AI-Enhanced Analytics concepts reveals conceptual limitations driving demand for data profiling in these environments. The main pillar of AI-Enhanced Analytics, Data Profiling, embodies the investigation of data sets in order to locate their structure, content, and data quality issues. Within the sector of large-scale healthcare data lakes, external sources range from user-generated input data (e.g. feedback, surveys) to openly available data sets from third parties or similar institutions, thus including also unstructured textual data and images. This corpus of uncoordinated data, often available in different standards, ontologies, and formats, demands methods to assess its coverage and usability.

Data profiling investigates data sets to determine their structure and content and evaluate their quality against internal or external business rules. Although data profiling is a key function for any Data Engineer, no prior literature addresses it in the specific context of heterogeneous healthcare-related data sources. Specifically, the examination of the various data types present in a clinical Data Lake—structured, semi-structured, and unstructured—reveals a discrete set of revision techniques that may be applied to understand their quality and fitness-to-purpose. Data profiling absorbs data from both the Data Lake and external sources, building a knowledge reservoir capable of sustaining Data Analytics applications; hence the consideration of several Data Profiling functions from both data quality and data governance perspectives.

5. Architecture and System Design

Comprehensive data ingestion and normalization strategies, together with schema harmonization and ETL/ELT pipelines, streamline the integration of diverse healthcare sources. Supporting metadata management and cataloging structures, lineage tracking, and semantic annotation empower informed profiling and stakeholder collaboration.

A data lake environment harbors multiple sources with distinct dialects; therefore, metadata must support ingestion (control sources, destinations, and procedural requirements) and provide information to a data lineage tool. Lineage relies on records cohering with the data quality monitoring framework, specifying the stage (raw, cleansed, consistent, or profiled) of each dataset. The resulting operational metadata repository operates with a data catalog, divulging data lake details for stakeholders. After ingestion, the information is publicly accessible and supports query-building through a business model mapping layer; within the layer, the data scientist utilizes elements such as synonyms, provenance, and definitions.

During data profiling, ingested information is integrated in real or near-real time within pre-staging transformations. The strategic implementation of ETL or ELT processes is determined by the data volume within the different environments; for clinical sources with high availability, the suppliers support ELT, and for sources with higher volume but less frequency, ETL is conducted to mitigate execution windows.

5.1. Data Ingestion and Normalization Strategies

Intelligent analytics for data profiling in heterogeneous healthcare data lakes should address questions related to data ingestion, storage, integration, transformation, guarding, synthesis, and exposure – all of which are capital-focused for any healthcare data lake environment. The data sources in a healthcare ecosystem are numerous and diverse, spanning models, agents, organization, information, health, demographic, and treatment administration, among others. Fulfilling the management-level needs of business and information product owners involves various analytics and AI-enhanced analytical products. ELT (or ETL) pipelines for internal (native) healthcare data sources are foundational structures; the other products stem from filling the native data lakes, especially the clinical springs, as they hone in on deep learning and AI.

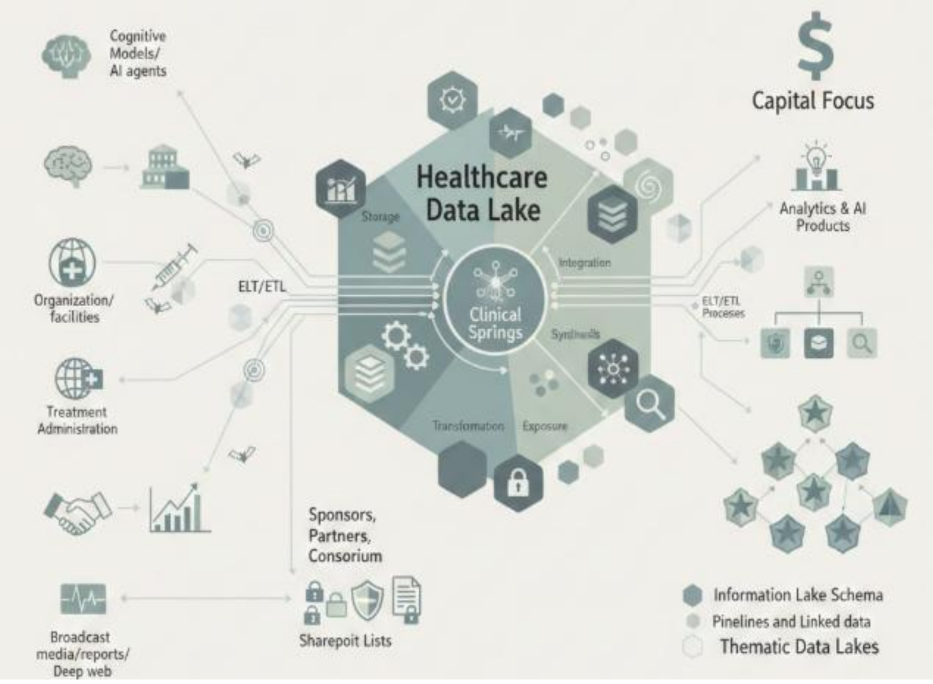


Fig 3: Capital-Centric Intelligence: A Unified Orchestration Framework for Heterogeneous Healthcare Data Lakes and AI-Driven Analytics

Sources beyond the ecosystem (broadcast media, reports, and deep web sources) are ingested via mainstream or custom agents and bats (that process information in bulk). Input from sponsors, partners, and consortium members on behalf of patient associations and the larger society is both essential and privileged, for example through sharepoint lists. The information model for data lakes expresses the information lake schema as a series of joined hierarchically grouped star schemas and is the main framework to house pipelines accessing, potentially transforming, storing, and linking vast amounts of information in the thematic data lakes.

5.2. Metadata Management and Cataloging

Privacy, security, and compliance considerations must guide data profiling practices. Sensitive patient information is susceptible to misuse and challenging to safeguard. Protecting personally identifiable information is critical. Data governance, privacy, security, and compliance requirements shape the operational environment of data lakes and directly influence profiling approaches. These factors affect how analytics should be employed and the solutions should be tailored.

A key element of governance is monitoring and profiling to determine the current state of the data lake. Metadata is a vital asset that needs extensive modelling and precise management. Metadata describing the data schema, data types, data lineage, data quality metrics, and detected data anomalies should be stored in a metadata management repository or a dedicated metadata catalog to allow users to search for data items of interest. Every healthcare system and application has specific data requirements. A systematic examination of these requirements provides guidance for defining the nature of metadata that should be harvested as data profiling is performed.

Quality / Profiling metric	Equation (symbolic form)
Consistency (cross-field)	K = n_consistent / n_checked
Timeliness	T = 1 - (n_late / n_events)
Lineage coverage	L = n_traced_edges / n_expected_edges
Schema drift rate	D = n_schema_changes / n_batches
Anomaly z-score	z = (x - μ) / σ

6. Case Studies and Empirical Findings

Case studies in clinical data lakes are presented, supported by quantitative, qualitative, and expert assessments. Concrete profiling results describe data-types, schemata, lineage, quality, and anomalies, while real-world implications point to quality, interoperability, and governance enhancements. Healthcare organizations increasingly pursue clinical data lakes, hoping to exploit inch-restored or reusable data for advanced analytics supporting real-time decision-making, quality improvement, patient safety, predictive and prescriptive modelling. Proactive profiling of these heterogeneous collections, catering to the varied needs of internal and external analytics providers, enables knowledge of their content and properties and guides responsible usage.

AI-Enhanced Analytics transforms business intelligence solutions into effective self-service environments driven by big data. Technology, people, and process considerations shape the analytical landscape: the democratization risk inherent in self-service is mitigated through governance; data dissimilarity is managed through semantic consistency; and organisational demands are skilfully correlated with technical capabilities. Within the AI-Enhanced Analytics framework, four types of data profiling guarantee the readiness of big data sources for meaningful analysis: discovery defines data origin and semantics, lineage shopping identifies schema correspondence and diversity, quality detects inconsistency, incompleteness, and redundancy, and anomaly hunting reveals non-conformance with behaviour.

Equation 4) Lineage profiling equations

4.1 Lineage coverage

If governance expects $|E_{\text{expected}}|$ edges but you traced only $|E_{\text{traced}}|$:

$$L = \frac{|E_{\text{traced}}|}{|E_{\text{expected}}|}$$

4.2 Impact radius (blast radius)

For a dataset node v, downstream impacted nodes:

$$\text{Downstream}(v) = \{u \in V : v \rightsquigarrow u\}$$

Impact size:

$$\text{Impact}(v) = |\text{Downstream}(v)|$$

6.1. Profiling Outcomes in Clinical Data Lakes

Artificial intelligence has endeavoured to address open problem domains in technology from computer vision to natural language processing. High-performance computing, big data management, and cloud computing have made it possible to develop AI-enhanced techniques for different analytical applications: Enterprises across domains have started harnessing the power of artificial intelligence to gain business insights and forecast future trends. Healthcare has become one such domain; AI has been applied in healthcare analytics to gain actionable insights in patient enrolment, drug side effects prediction, disease prediction and detection, medical image analysis, survival time prediction, and quality assessment.

Contemporary clinical analytics, however, have yet to reap the benefits of AI. Furthermore, the current portfolio of clinical analytics at health enterprises does not also focus on profiling the data. New, harmonised data lakes have recently been developed—encompassing heterogeneous sources and designed as unified stores for downstream analytics—opened on a pilot basis for clinical use. AI-enhanced analytics concepts are proposed and pursued for a broader analytical portfolio at the data lakes. The data profiling phase, an indispensable yet often overlooked aspect of any data analysis project, is presented here for these newly opened clinical data lakes.

6.2. Insights for Data Quality and Interoperability

For clinical data lakes chronically suffering from poor data quality and limited interoperability, AI-Enhanced Analytics provide an effective method to discern, prioritize, and act on salient data quality and accessibility knowledge. Four specific conclusions emerge: (i) relevant quality dimensions and metrics vary across datasets; (ii) AI-Enhanced Analytics proves an effective means of surfacing and prioritizing quality insights; (iii) addressing quality issues—e.g., through automated data repairs—can be directly supported; and (iv) insights drawn from the profiling process can sensitize data producers to factors affecting both internal data quality and external data interoperability.

Profoundly heterogeneous sources are common in clinical data lakes. Underlying data processes vary across dimensions including data format, structure, semantics, access patterns, profiles, and runtime environments. As a result, a one-size-fits-all approach to profiling these sources is unlikely to succeed. Indeed, explicit attention to these unique aspects frequently does improve profiling quality.

7. CONCLUSION

Responding to the current demands for better access to healthcare knowledge and deeper understanding of disease processes, complicated analysis across a wide range of heterogeneous sources is now possible. The corresponding growth of data lakes and their application in clinical environments add new challenges for data science extractors and AI-based practitioners. Especially in healthcare, where inefficiencies and wasted resources linger, there is a great opportunity—and need—to use AI-enhanced analytics to better mine the massive pools of varied data.

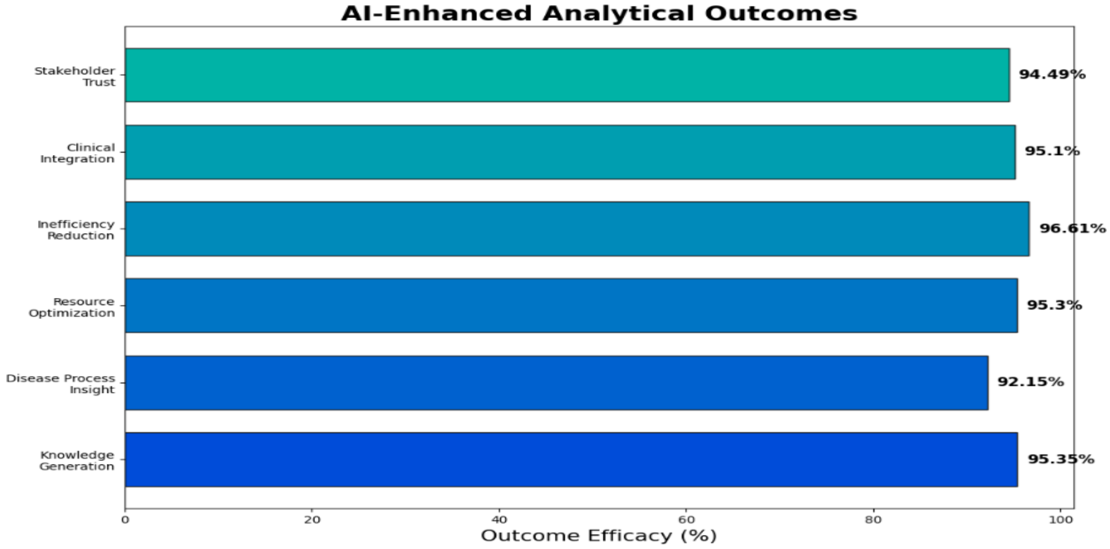


Fig 4: AI-Enhanced Analytical Outcomes

Data lakes allow an enormous amount of data to be easily accessible. However, raw data are seldom usable in their initial state due to potential inconsistencies in name, type, scale, coding, and quality. Enabling an underlying governance framework makes these data a foundation for AI-enhanced analytics and thus for deep knowledge generation. These processes not only require managing and bridging gaps in volume and variety, but also establishing systems to monitor data quality. The importance of profiling in intelligent AI-enhanced analytics is now precisely set against practical examples. Concrete clinical data lakes, built from multiple heterogeneous sources, and the corresponding data profiling not only have been performed but also have allowed the detection of relevant quality and governance issues.

7.1. Final Reflections and Future Directions

The above analysis advances the discussion on healthcare data profiling for data lakes that leverage AI-Enhanced Analytics capabilities. Although a growing number of organizations have adopted data lake strategies, awareness of how to fulfil optimal Data Governance requirements intelligently and efficiently remains limited. The practical application of AI-Enhanced Analytics approaches associated with intelligent data governance is still at an early stage. As the case studies show, there are numerous models and metaphors supporting the concept of a curriculum-supporting data lake but profiles still have to be designed in detail. The adoption of appropriate technologies enables Data Profiling techniques to be applied to all data sources included in the space and on an ongoing basis, thus actively contributing to data quality, governance, and interoperability assessment within the organization.

As a cloud data ecosystem evolves and relevant data are ingested, appropriate Data Profiling techniques must be intelligently applied to the various Healthcare Data Lake layers or zones. This section focuses on the type of

Data Profiling that enables classification and quantification of the various data types and their structural and content-related characteristics — typical, specific, and rare — and the identification, specification, implementation, and operationalization of metrics that evaluate these characteristics. In a Data Lake environment, where heterogeneous data from disparate sources can be ingested rapidly, Data Profiling not only supports the definition of Data Quality but also provides important ongoing insights into Data Governance, Data Management, and Data Interoperability.

REFERENCES

1. Zhang, Y., Zhao, L., & Chen, X. (2023). Intelligent data governance frameworks for healthcare data lakes using machine learning analytics. IEEE Access, 11, 113245–113259.
2. Gottimukkala, V. R. R. (2022). Licensing Innovation in the Financial Messaging Ecosystem: Business Models and Global Compliance Impact. International Journal of Scientific Research and Modern Technology, 1(12), 177-186.
3. Patel, S., Shah, M., & Gupta, R. (2023). AI-driven data quality management in healthcare big data environments. Journal of Biomedical Informatics, 140, 104327.
5. Uday Surendra Yandamuri. (2023). An Intelligent Analytics Framework Combining Big Data and Machine Learning for Business Forecasting. International Journal Of Finance, 36(6), 682-706. https://doi.org/10.5281/zenodo.18095256
6. Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2018). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. Journal of Computational Science, 25, 152–160.
7. Meda, R. (2023). Data Engineering Architectures for Scalable AI in Paint Manufacturing Operations. European Data Science Journal (EDSJ) p-ISSN 3050-9572 en e-ISSN 3050-9580, 1(1).
8. Bates, D. W., Saria, S., Ohno-Machado, L., et al. (2014). Big data in health care. Health Affairs, 33(7), 1123–1131.
9. Keerthi Amistapuram. (2023). Privacy-Preserving Machine Learning Models for Sensitive Customer Data in Insurance Systems. Educational Administration: Theory and Practice, 29(4), 5950–5958. https://doi.org/10.53555/kuey.v29i4.10965
10. Belle, A., Thiagarajan, R., Soroushmehr, S. M. R., et al. (2015). Big data analytics in healthcare. BioMed Research International, 2015, 370194.
11. Guntupalli, R. (2023). AI-Driven Threat Detection and Mitigation in Cloud Infrastructure: Enhancing Security through Machine Learning and Anomaly Detection. Available at SSRN 5329158.
12. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. ACM SIGMOD Record, 29(2), 93–104.
13. Unifying Data Engineering and Machine Learning Pipelines: An Enterprise Roadmap to Automated Model Deployment. (2023). American Online Journal of Science and Engineering (AOJSE) (ISSN: 3067-1140) , 1(1). https://aojse.com/index.php/aojse/article/view/19
14. Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171–209.
15. Kalisetty, S., & Singireddy, J. (2023). Optimizing Tax Preparation and Filing Services: A Comparative Study of Traditional Methods and AI Augmented Tax Compliance Frameworks. Available at SSRN 5206185.
16. Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. Artificial Intelligence in Medicine, 26(1–2), 1–24.
17. Kummari, D. N., & Burugulla, J. K. R. (2023). Decision Support Systems for Government Auditing: The Role of AI in Ensuring Transparency and Compliance. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 493-532.
18. Braik, A., & Koliou, M. Artificial intelligence and machine learning-powered GIS for proactive disaster resilience in a changing climate. Journal of Spatial Science, 69(1).
19. Varri, D. B. S. (2023). Advanced Threat Intelligence Modeling for Proactive Cyber Defense Systems. Available at SSRN 5774926.
20. Dwork, C. (2008). Differential privacy. ICALP Proceedings, 1–12.
21. Bandi, V. D. V. K. (2023). Production-Grade Machine Learning Pipelines For Healthcare Predictive Analytics. South Eastern European Journal of Public Health, 189–205. Retrieved from https://www.seejph.com/index.php/seejph/article/view/7057
22. Kolla, S. K. (2021). Architectural Frameworks for Large-Scale Electronic Health Record Data Platforms. Current Research in Public Health, 1(1), 1–19. Retrieved from https://www.scipublications.com/journal/index.php/crph/article/view/1372
23. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

24. Maguluri, K. K., Pandugula, C., Kalisetty, S., & Mallesham, G. (2022). Advancing Pain Medicine with AI and Neural Networks: Predictive Analytics and Personalized Treatment Plans for Chronic and Acute Pain Managements. Journal of Artificial Intelligence and Big Data, 2(1), 112-126.
25. Garapati, R. S. (2022). AI-Augmented Virtual Health Assistant: A Web-Based Solution for Personalized Medication Management and Patient Engagement. Available at SSRN 5639650.
26. Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms. Pattern Recognition, 64, 206–223.
27. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
28. Segireddy, A. R. (2021). Containerization and Microservices in Payment Systems: A Study of Kubernetes and Docker in Financial Applications. Universal Journal of Business and Management, 1(1), 1–17. Retrieved from https://www.scipublications.com/journal/index.php/ujbm/article/view/1352
29. He, J., Baxter, S. L., Xu, J., et al. (2019). The practical implementation of AI in healthcare. Nature Medicine, 25(1), 30–36.
30. Inala, R. AI-Powered Investment Decision Support Systems: Building Smart Data Products with Embedded Governance Controls.
31. Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29.
32. Gottimukkala, V. R. R. (2021). Digital Signal Processing Challenges in Financial Messaging Systems: Case Studies in High-Volume SWIFT Flows.
33. Iglewicz, B., & Hoaglin, D. C. (1993). How to detect and handle outliers. ASQC.
34. Johnson, A. E. W., Pollard, T. J., Shen, L., et al. (2016). MIMIC-III database. Scientific Data, 3, 160035.
35. Yandamuri, U. S. (2022). Big Data Pipelines for Cross-Domain Decision Support: A Cloud-Centric Approach. International Journal of Scientific Research and Modern Technology, 1(12), 227–237. https://doi.org/10.38124/ijsrmt.v1i12.1111
36. Kimball, R., & Caserta, J. (2004). The data warehouse ETL toolkit. Wiley.
37. Davuluri, P. N. Integrating Artificial Intelligence into Event-Driven Financial Crime Compliance Platforms.
38. Kriegel, H. P., Kröger, P., Schubert, E., & Zimek, A. (2009). Outlier detection in axis-parallel subspaces. PKDD Proceedings, 831–838.
39. Kummari, D. N. (2023). AI-Powered Demand Forecasting for Automotive Components: A Multi-Supplier Data Fusion Approach. European Advanced Journal for Emerging Technologies (EAJET)-p-ISSN 3050-9734 en e-ISSN 3050-9742, 1(1).
40. Bauder, R., & Khoshgoftaar, T. (2022). Data mining and machine learning for healthcare fraud detection. Journal of Big Data, 9(1), 1–18.
41. Li, Y., Chen, C. Y., Wasserman, W. W., & Ramani, A. K. (2016). Deep feature selection. Bioinformatics, 32(5), 743–750.
42. Varri, D. B. S. (2022). A Framework for Cloud-Integrated Database Hardening in Hybrid AWS-Azure Environments: Security Posture Automation Through Wiz-Driven Insights. International Journal of Scientific Research and Modern Technology, 1(12), 216-226.
43. Malhotra, P., Vig, L., Shroff, G., & Agarwal, P. (2015). Long short-term memory networks for anomaly detection. ESANN Proceedings.
44. Kalisetty, S., Vankayalapati, R. K., Reddy, L., Sondinti, K., & Valiki, S. (2022). AI-Native Cloud Platforms: Redefining Scalability and Flexibility in Artificial Intelligence Workflows. Linguistic and Philosophical Investigations, 21(1), 1-15.
45. Garapati, R. S. (2023). Optimizing Energy Consumption in Smart Build-ings Through Web-Integrated AI and Cloud-Driven Control Systems.
46. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare. Briefings in Bioinformatics, 19(6), 1236–1246.
47. Kushvanth Chowdary Nagabhyru. (2023). Accelerating Digital Transformation with AI Driven Data Engineering: Industry Case Studies from Cloud and IoT Domains. Educational Administration: Theory and Practice, 29(4), 5898–5910. https://doi.org/10.53555/kuey.v29i4.10932
48. Sarker, I. H., Kayes, A., Badsha, S., et al. (2022). Cybersecurity data science: An overview from machine learning perspective. Journal of Big Data, 9(1), 1–29.
49. Guntupalli, R. (2023). Optimizing Cloud Infrastructure Performance Using AI: Intelligent Resource Allocation and Predictive Maintenance. Available at SSRN 5329154.
50. Patcha, A., & Park, J. M. (2007). An overview of anomaly detection techniques. Computer Networks, 51(12), 3448–3470.
51. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn. Journal of Machine Learning Research, 12, 2825–2830.

52. Aitha, A. R. (2023). CloudBased Microservices Architecture for Seamless Insurance Policy Administration. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 607-632.
53. Rajkomar, A., Oren, E., Chen, K., et al. (2018). Scalable deep learning with EHRs. NPJ Digital Medicine, 1, 18.
54. Avinash Reddy Segireddy. (2022). Terraform and Ansible in Building Resilient Cloud-Native Payment Architectures. International Journal of Intelligent Systems and Applications in Engineering, 10(3s), 444–455. Retrieved from https://www.ijisae.org/index.php/IJISAE/article/view/7905.
55. Ringberg, H., Soule, A., Rexford, J., & Diot, C. (2007). Sensitivity of PCA for anomaly detection. SIGMETRICS Proceedings.
56. Koppolu, H. K. R., Sheelam, G. K., & Komaragiri, V. B. (2023). Autonomous Telecommunication Networks: The Convergence of Agentic AI and AI-Optimized Hardware. International Journal of Science and Research (IJSR), 12(12), 2253-2270.
57. Ruff, L., Vandermeulen, R. A., Görnitz, N., et al. (2018). Deep one-class classification. ICML Proceedings.
58. Rongali, S. K. (2023). Explainable Artificial Intelligence (XAI) Framework for Transparent Clinical Decision Support Systems. International Journal of Medical Toxicology and Legal Medicine, 26(3), 22-31.
59. Salfner, F., Lenk, M., & Malek, M. (2010). Survey of failure prediction methods. ACM Computing Surveys, 42(3), 1–42.
60. Nagubandi, A. R. (2023). Advanced Multi-Agent AI Systems for Autonomous Reconciliation Across Enterprise Multi-Counterparty Derivatives, Collateral, and Accounting Platforms. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 653-674.
61. Schölkopf, B., Platt, J. C., Shawe-Taylor, J., et al. (2001). Estimating the support of a high-dimensional distribution. Neural Computation, 13(7), 1443–1471.
62. Kalisetty, S., & Ganti, V. K. A. T. (2019). Transforming the Retail Landscape: Srinivas's Vision for Integrating Advanced Technologies in Supply Chain Efficiency and Customer Experience. Online Journal of Materials Science, 1, 1254.
63. Sipos, R., Fradkin, D., Moerchen, F., & Wang, Z. (2014). Log-based predictive maintenance. KDD Proceedings.
64. Meda, R. (2023). Intelligent Infrastructure for Real-Time Inventory and Logistics in Retail Supply Chains. Educational Administration: Theory and Practice.
65. Kolla, S. K. (2021). Designing Scalable Healthcare Data Pipelines for Multi-Hospital Networks. World Journal of Clinical Medicine Research, 1(1), 1–14. Retrieved from https://www.scipublications.com/journal/index.php/wjcmr/article/view/1376
66. Bandi, V. D. V. K. (2023). Cloud-Native Model Lifecycle Management for Enterprise AI Systems. International Journal of Scientific Research and Modern Technology, 2(12), 78–90. https://doi.org/10.38124/ijsrmt.v2i12.1236
67. Inala, R. Revolutionizing Customer Master Data in Insurance Technology Platforms: An AI and MDM Architecture Perspective.
68. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society B, 58(1), 267–288.
69. Gottimukkala, V. R. R. (2023). Privacy-Preserving Machine Learning Models for Transaction Monitoring in Global Banking Networks. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 633-652.
70. Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.
71. AI Powered Fraud Detection Systems: Enhancing Risk Assessment in the Insurance Sector. (2023). American Journal of Analytics and Artificial Intelligence (ajaai) With ISSN 3067-283X, 1(1). https://ajaai.com/index.php/ajaai/article/view/14
72. Weber, G. M., Mandl, K. D., & Kohane, I. S. (2014). Finding the missing link for big biomedical data. JAMIA, 21(1), 1–3.
73. Kolla, S. H. (2021). Rule-Based Automation for IT Service Management Workflows. Online Journal of Engineering Sciences, 1(1), 1–14. Retrieved from https://www.scipublications.com/journal/index.php/ojes/article/view/1360
74. Siva Hemanth Kolla. (2023). Deep Learning–Driven Retrieval-Augmented Generation for Enterprise ITSM Automation: A Governance-Aligned Large Language Model Architecture. Journal of Computational Analysis and Applications (JoCAAA), 31(4), 2489–2502. Retrieved from https://www.eudoxuspress.com/index.php/pub/article/view/4774
75. Zhang, Y., & Yang, Q. (2021). A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering, 34(12), 5586–5609.

76. Meda, R. (2023). Developing AI-Powered Virtual Color Consultation Tools for Retail and Professional Customers. Journal for ReAttach Therapy and Developmental Diversities. https://doi. org/10.53555/jrtdd. v6i10s (2), 3577.
77. Almadhoun, R., Kadadha, M., Al-Fuqaha, A., & Guizani, M. (2021). A user-centric blockchain-based system for incident response in the era of IoT. Internet of Things, 14, 100371. https://doi.org/10.1016/j.iot.2021.100371
78. Kalisetty, S. (2023). The Role of Circular Supply Chains in Achieving Sustainability Goals: A 2023 Perspective on Recycling, Reuse, and Resource Optimization. Reuse, and Resource Optimization (June 15, 2023).
79. Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data. Wiley.
80. Siva Hemanth Kolla. (2022). Knowledge Retrieval Systems for Enterprise Service Environments. International Journal of Intelligent Systems and Applications in Engineering, 10(3s), 495–506. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/8037
81. Bishop, C. M. (1994). Novelty detection and neural network validation. IEE Proceedings, 141(4), 217–222.
82. Rongali, S. K. (2022). AI-Driven Automation in Healthcare Claims and EHR Processing Using MuleSoft and Machine Learning Pipelines. Available at SSRN 5763022.
83. Kaur, H., Alam, M., Jameel, R., Mourya, A., & Chang, V. (2022). A proposed framework for healthcare fraud detection using machine learning. IEEE Access, 10, 61740–61754.
84. Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). Time series analysis: Forecasting and control. Wiley.
85. Amistapuram, K. (2022). Fraud Detection and Risk Modeling in Insurance: Early Adoption of Machine Learning in Claims Processing. Available at SSRN 5741982.
86. Kumar, A., Gupta, P., & Singh, R. (2023). Sentiment analysis methods for proactive brand reputation risk management. International Journal of Information Management Data Insights, 3(1).
87. Ramesh Inala. (2023). Big Data Architectures for Modernizing Customer Master Systems in Group Insurance and Retirement Planning. Educational Administration: Theory and Practice, 29(4), 5493–5505. https://doi.org/10.53555/kuey.v29i4.10424
88. Sharma, A., & Rani, R. (2023). Artificial intelligence approaches for healthcare data quality and governance. Journal of Healthcare Engineering, 2023, 1–14.
89. Davuluri, P. N. AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems.
90. Bifet, A., & Gavalda, R. (2007). Learning from time-changing data with adaptive windowing. SDM Proceedings.
91. Nagabhyru, K. C. (2023). From Data Silos to Knowledge Graphs: Architecting CrossEnterprise AI Solutions for Scalability and Trust. Available at SSRN 5697663.
92. Zaharia, M., Chowdhury, M., Franklin, M. J., et al. (2010). Spark: Cluster computing. HotCloud Proceedings.
93. Avinash Reddy Aitha. (2022). Deep Neural Networks for Property Risk Prediction Leveraging Aerial and Satellite Imaging. International Journal of Communication Networks and Information Security (IJCNIS), 14(3), 1308–1318. Retrieved from https://www.ijcnis.org/index.php/ijcnis/article/view/8609
94. Goutham Kumar Sheelam, Hara Krishna Reddy Koppolu. (2022). Data Engineering And Analytics For 5G-Driven Customer Experience In Telecom, Media, And Healthcare. Migration Letters, 19(S2), 1920–1944. Retrieved from https://migrationletters.com/index.php/ml/article/view/11938
95. Alenezi, M., & Akour, M. AI-driven innovations in software engineering: A review of current practices and future directions. Applied Sciences, 15(3), 1344. https://doi.org/10.3390/app15031344 Cited by: 149
96. Garapati, R. S. (2022). Web-Centric Cloud Framework for Real-Time Monitoring and Risk Prediction in Clinical Trials Using Machine Learning. Current Research in Public Health, 2, 1346.
97. Chen, Y., Li, X., & Wang, J. (2023). Intelligent analytics for heterogeneous healthcare data integration. Information Sciences, 630, 412–428.
98. Kummari, D. N. (2023). Energy Consumption Optimization in Smart Factories Using AI-Based Analytics: Evidence from Automotive Plants. Journal for Reattach Therapy and Development Diversities. https://doi. org/10.53555/jrtdd. v6i10s (2), 3572.