

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING IN CLINICAL DATA MANAGEMENT: WORKFLOW AUTOMATION AND QUALITY ENHANCEMENT

Keerthi Manne

Plot-No.391,Usha Nilayam,Bhagath Singh Nagar, Hyderabad
Pin:500085,Telangana,India

Received: 24/07/2026

Revised: 17/08/2026

Accepted: 25/09/2026

ABSTRACT:

Clinical data management (CDM) is the backbone that converts raw trial and real-world observations into a validated, analysis-ready dataset, and its quality directly determines the reliability of clinical and regulatory decisions. Traditional CDM workflows, however, remain heavily manual, relying on rule-based edit checks, page-by-page source data verification, and human medical coding, all of which are labour intensive, error prone, and slow relative to the pace of modern trials and the growing volume of electronic health record (EHR), registry, wearable, and multi-omics data. This paper examines how artificial intelligence (AI) and machine learning (ML) are reshaping CDM by automating repetitive workflow steps and elevating data quality beyond what manual processes can achieve. We review the principal AI/ML techniques applied in this domain, including natural language processing (NLP) for unstructured clinical text, supervised and unsupervised ML for anomaly and outlier detection, predictive models for query triage and risk-based monitoring, and robotic process automation (RPA) for system-to-system data reconciliation. A conceptual framework is presented that links data ingestion, model-assisted cleaning, human-in-the-loop review, and continuous learning, and this is discussed alongside emerging regulatory expectations for AI/ML-based tools, including the U.S. Food and Drug Administration's Total Product Life Cycle (TPLC) approach and Good Machine Learning Practice (GMLP) principles. We further discuss quality enhancement mechanisms such as automated query generation, intelligent medical coding, real-time anomaly flags, and risk-based monitoring, and we consider the operational, ethical, and validation challenges that must be addressed before such systems can be trusted at scale. The paper concludes that AI/ML, when embedded within a governed, human-supervised workflow, can substantially reduce cycle times and error rates in CDM while introducing new categories of risk that require proactive, lifecycle-based oversight.

Keywords: *Clinical Data Management, Artificial Intelligence, Machine Learning, Workflow Automation, Data Quality, Natural Language Processing, Electronic Health Records, Good Machine Learning Practice, Risk-Based Monitoring, Medical Coding*

INTRODUCTION

Clinical data management is the discipline responsible for collecting, cleaning, validating, and preparing the data generated during clinical trials, registries, and routine care so that it can be analyzed with confidence. A single multi-site trial can generate hundreds of thousands of individual data points spanning case report forms (CRFs), laboratory results, adverse event narratives, concomitant medication logs, and increasingly, data streamed from wearables and connected devices. Every one of these data points must be checked for completeness, consistency, plausibility, and protocol compliance before it can support a regulatory submission or a clinical decision. Historically this work has depended on manual query generation, fixed-rule edit checks embedded in electronic data capture (EDC) systems, and extensive source data verification (SDV) performed by clinical research associates.

This manual paradigm is increasingly strained. Trials have grown larger and more geographically dispersed, protocols have become more complex, and the volume of unstructured data — physician notes, discharge summaries, imaging reports, patient-reported outcomes captured in free text — has expanded far faster than the capacity of manual review teams. At the same time, sponsors and regulators are asking for shorter database lock timelines and near-real-time visibility into data quality, particularly for adaptive and decentralized trial designs. These pressures have created strong incentive to apply AI and ML not merely as an analytics layer sitting

downstream of CDM, but as an operational component embedded directly within the data management workflow itself.

AI and ML offer three broad capabilities that map naturally onto the pain points of CDM. First, pattern recognition allows models to learn what a common but incorrect entry looks like — for instance, an implausible vital sign combination or a duplicate adverse event — far faster than a rule engine that must be manually maintained. Second, natural language understanding allows unstructured narrative fields to be parsed into structured, codable concepts, closing a long-standing gap between the volume of free text collected and the capacity of human coders to process it. Third, predictive modelling allows data management teams to move from reactive, retrospective query cycles to proactive, risk-based monitoring in which sites and forms most likely to contain errors are prioritized before they compound downstream.

This paper synthesizes the current state of AI/ML application in CDM around two organizing themes: workflow automation, meaning the replacement or augmentation of manual, repetitive steps with model-assisted or fully automated processes, and quality enhancement, meaning the measurable improvement in data accuracy, consistency, and timeliness that results from those automated processes. Section II situates this work relative to existing literature on healthcare AI and data management. Section III outlines the traditional CDM workflow and its principal bottlenecks. Section IV surveys the AI/ML techniques most relevant to CDM. Section V proposes a conceptual framework for embedding these techniques into an operational workflow. Section VI details specific quality enhancement mechanisms. Section VII addresses the regulatory and governance considerations that accompany the use of AI/ML in a regulated clinical environment. Sections VIII through X discuss illustrative applications, limitations, and future directions, before Section XI concludes.

BACKGROUND AND RELATED WORK

The application of AI to healthcare data has been documented extensively in the biomedical informatics literature, most prominently in diagnostic image analysis, clinical decision support, and predictive risk scoring. A smaller but rapidly growing body of work addresses AI applied specifically to the operational layer of clinical research — that is, to the processes of data capture, cleaning, coding, and monitoring rather than to the clinical inference drawn from the data once it is clean. This distinction matters because operational CDM tasks tend to have well-defined, auditable ground truth (a laboratory value either matches the source document or it does not), which makes them comparatively tractable for supervised learning even where clinical diagnostic tasks remain difficult to validate.

A useful way to frame this body of work is to separate the data pipeline into an acquisition layer, a structuring layer, and an analytics layer. Figure 1 illustrates this pipeline in the context of a learning health system: electronic medical record (EMR) data together with imaging, genetic, and electrophysiological data form the raw substrate; natural language processing converts clinical notes written in ordinary human language into codified data; machine learning consumes this structured output, together with directly structured EMR fields, to support clinical activities such as screening, diagnosis, and treatment; and the outputs of those clinical activities in turn generate new notes and new data, closing a continuous improvement loop. The same architecture underlies most modern CDM automation platforms, with the difference that the 'clinical activities' box is replaced by data management activities — query resolution, coding, and monitoring — that feed back into progressively better-trained models.

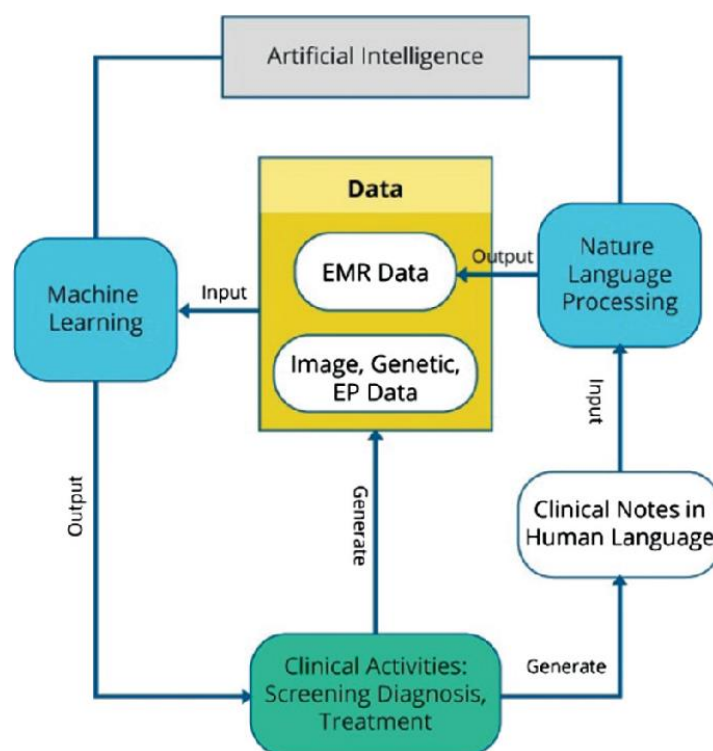


Figure 1. Conceptual data–AI cycle underlying learning health systems and AI-enabled clinical data management. Structured EMR and multimodal data are combined with NLP-derived concepts from clinical notes to drive machine learning models, whose outputs feed back into clinical and data-management activity.

Early automation efforts in CDM concentrated on deterministic rule engines: hard-coded range checks, cross-form consistency checks, and required-field checks built into EDC platforms such as those widely used across the pharmaceutical industry. These systems remain the backbone of most trials today, but they are brittle — every new check must be hand-authored, and rules cannot easily capture the kind of contextual plausibility judgement a trained data manager applies intuitively (for example, recognizing that a reported laboratory value is technically within range but inconsistent with a patient's other results on the same visit). ML-based anomaly detection was introduced to complement rather than replace these rules, learning statistical regularities across the dataset that static thresholds cannot express.

In parallel, the broader medical NLP literature has matured considerably, moving from rule-based and dictionary look-up systems toward transformer-based language models pretrained on biomedical corpora. Domain-adapted language models trained on large volumes of clinical or biomedical text have demonstrated substantially better performance on named-entity recognition, relation extraction, and concept normalization tasks than general-purpose language models, and these advances underlie current-generation tools for automated medical coding and adverse-event narrative structuring. This progress has, however, been accompanied by a steep rise in the computational resources required to train and operate the underlying models, a trend that has direct implications for how such tools can realistically be deployed within a CDM environment, as discussed in Section V.

Finally, a growing regulatory and standards literature — spanning good clinical data management practice guidance, ICH E6(R2) good clinical practice, ICH E8(R1) on quality by design in trials, and the FDA's evolving framework for AI/ML-based Software as a Medical Device — has begun to articulate expectations for how automated and adaptive systems should be validated, monitored, and governed when they touch regulated data. This regulatory dimension is treated separately and in depth in Section VII, because it materially shapes which AI/ML techniques are feasible to deploy in practice, independent of their technical performance.

TRADITIONAL CLINICAL DATA MANAGEMENT: WORKFLOW AND BOTTLENECKS

A conventional CDM workflow proceeds through a broadly linear sequence: protocol and CRF design, database build and edit-check specification, data entry or capture at site level, automated and manual edit-check execution, query generation and resolution, medical coding of adverse events and concomitant medications using standardized dictionaries such as MedDRA and WHODrug, source data verification, serious adverse event reconciliation, external data reconciliation (laboratory, imaging, ECG vendors), and finally database lock and extraction for statistical analysis. Each of these steps has traditionally required substantial manual effort, and each is a potential source of delay or error.

Several bottlenecks recur across this pipeline irrespective of therapeutic area or trial phase. Query volume scales with sample size and protocol complexity, and each query requires a site to investigate, respond to, and a data manager to review and close, a cycle that commonly takes days to weeks per query and can materially extend the time between last-patient-last-visit and database lock. Medical coding is similarly labour-intensive: coders must map free-text verbatim terms to standardized dictionary entries, a task that is conceptually simple for common terms but time-consuming at scale and error-prone for ambiguous or misspelled verbatims. Source data verification, historically performed on 100% of records for pivotal trials, consumes a disproportionate share of monitoring budgets while, according to extensive industry experience, catching a comparatively small fraction of the errors that ultimately affect analysis-critical fields. Finally, unstructured data — physician progress notes, discharge summaries, radiology and pathology reports — is frequently excluded from systematic review altogether simply because manual abstraction does not scale, leaving potentially decision-relevant information uncaptured.

These bottlenecks share a common structural cause: each depends on a human reviewer applying pattern recognition and judgement to a large, repetitive dataset, which is precisely the task machine learning systems are best suited to accelerate, provided output quality can be governed appropriately. Sections IV through VI describe how specific AI/ML techniques address each of these bottlenecks in turn.

A further structural feature of the traditional workflow compounds these bottlenecks: most of its steps are sequential and gated rather than parallel and continuous. Edit checks are frequently executed in scheduled batches rather than at the moment of entry, so an error made early in a trial may not be surfaced until a later scheduled run, by which point the same erroneous pattern may already have propagated across multiple subsequent visits or even across sites following a shared, mistaken interpretation of a CRF instruction. Coding is similarly batched, often queued until a sufficient volume of new verbatim terms has accumulated to justify a coder's review session. This batching behaviour was originally a reasonable response to the fixed cost of running a review session, but it introduces latency that compounds as trials grow in scale and as sponsors seek to compress the interval between last-patient-last-visit and database lock. Continuous, model-assisted review, discussed in Section V, directly targets this latency by moving detection closer to the point of data entry.

The cost structure of the traditional workflow is also worth making explicit, because it explains much of the commercial incentive behind AI/ML adoption in this space. Industry cost analyses have consistently identified data management and monitoring, including SDV, coding, and query resolution, as among the largest controllable cost categories in trial conduct, second only to site payments and comparable in scale to the cost of the investigational product itself in many programs. Because these costs scale roughly with data volume and with the number of queries generated, even modest reductions in query volume or in the proportion of records requiring full manual SDV translate into materially lower total trial cost, which is the primary commercial driver behind sponsor and vendor investment in CDM automation.

AI/ML TECHNIQUES APPLIED TO CLINICAL DATA MANAGEMENT

A. Natural Language Processing for Unstructured Clinical Text

NLP pipelines applied to CDM typically combine several sub-tasks: sentence segmentation and tokenization tuned to clinical writing conventions (heavy abbreviation, non-standard punctuation, templated headers); named-entity recognition to locate mentions of symptoms, diagnoses, medications, dosages, and temporal expressions; negation and assertion detection to distinguish a documented symptom from one explicitly ruled out; relation extraction to link an adverse event to a causal medication or a temporal onset; and concept normalization to map extracted mentions onto standard terminologies such as MedDRA, SNOMED CT, LOINC, or WHODrug. Contemporary

implementations increasingly rely on transformer-based language models pretrained on biomedical or clinical corpora and fine-tuned on smaller, task-specific labelled datasets, an approach that has substantially reduced the volume of manually annotated training data required relative to earlier statistical NLP methods.

Within CDM specifically, NLP is used to pre-populate structured adverse-event fields from investigator narrative text, to flag discrepancies between a structured CRF entry and the supporting narrative documented in source notes, and to surface safety signals embedded in free-text fields that might otherwise be missed by structured-field-only review. The output of these models is generally treated as a suggestion routed to a human coder or data manager rather than as an autonomous determination, reflecting both the imperfect precision of current models on rare or ambiguous terms and the regulatory expectation that clinically significant determinations retain human oversight.

B. Machine Learning for Data Cleaning and Anomaly Detection

Supervised classification models, trained on historical query data, can predict the likelihood that a given field, form, or visit will generate a query, allowing review effort to be prioritized toward high-risk records rather than applied uniformly. Unsupervised and semi-supervised anomaly detection methods — including clustering, isolation forests, and autoencoder-based reconstruction error — identify records that deviate from the statistical distribution of the surrounding dataset without requiring a labelled example of every possible error type, which is valuable because rare error patterns are, by definition, underrepresented in historical query logs. Time-series and longitudinal models can additionally flag implausible trajectories, such as a laboratory value that changes too abruptly between consecutive visits given the physiological plausibility of the parameter in question, a check that is difficult to encode as a simple static range rule.

These models are most effective when combined with the deterministic edit checks already embedded in EDC systems rather than deployed as a replacement for them: rule engines remain well suited to unambiguous protocol-defined constraints (for example, a visit date that must fall within a specified window), while ML-based anomaly detection is better suited to the softer, context-dependent judgements that previously required an experienced data manager's intuition.

C. Predictive Analytics for Risk-Based Quality Management

Predictive models are increasingly used to support risk-based monitoring (RBM) and risk-based quality management (RBQM) approaches endorsed by ICH E6(R2) and regulatory guidance more broadly. Rather than allocating monitoring and SDV resources uniformly across all sites and records, RBM approaches use statistical and ML-derived key risk indicators — query rates, protocol deviation frequency, enrollment velocity, data entry lag, and similar metrics — to identify the subset of sites or forms most likely to harbour material data quality issues, allowing centralized statistical monitoring to substitute for a portion of on-site SDV. This shift is one of the more mature and widely adopted applications of predictive analytics in CDM, in part because its outputs (a prioritized list of sites for review) preserve a human decision-maker in the loop and are readily auditable.

D. Robotic Process Automation and System Integration

Robotic process automation complements ML-based methods by automating the mechanical, rule-based steps that surround data cleaning — reconciling patient identifiers across EDC, laboratory, and safety database systems; triggering standardized query text when a defined discrepancy condition is met; and generating routine status reports. RPA does not itself involve statistical learning, but it is frequently packaged alongside ML components in commercial CDM automation platforms because the two approaches are complementary: RPA executes deterministic, high-volume integration tasks reliably, while ML handles the residual, judgement-dependent tasks that RPA cannot be hand-coded to resolve.

E. Ensemble and Hybrid Architectures

In practice, production CDM automation systems rarely rely on a single model type. A representative architecture might combine a lightweight gradient-boosted classifier trained on structured field metadata to produce an initial query-likelihood score, a fine-tuned transformer-based NLP model reserved for free-text fields flagged as high-priority by that initial score, and a deterministic rule layer that enforces protocol-defined hard constraints regardless of model output. This layered, ensemble approach is generally preferred over a single end-to-end model for two reasons: it allows each component to be validated and audited independently against its own well-defined task, which is considerably easier than validating a single opaque model against the full breadth of CDM decisions; and it allows the more computationally expensive components to be invoked selectively, containing the

infrastructure cost discussed in Section V. Ensemble designs also provide a natural mechanism for confidence calibration, since agreement or disagreement among component models can itself be used as an additional signal of how much human scrutiny a given record warrants.

A FRAMEWORK FOR AI-DRIVEN WORKFLOW AUTOMATION

Integrating the techniques described in Section IV into a coherent operational workflow requires more than deploying individual models; it requires a framework that governs how model output enters the human review process, how models are retrained as new data accumulates, and how the resulting system is monitored once deployed. We propose a four-stage framework consisting of data ingestion and harmonization, model-assisted cleaning and coding, human-in-the-loop adjudication, and continuous monitoring with periodic retraining.

In the ingestion and harmonization stage, data arriving from EDC systems, EHR extracts, laboratory feeds, and device streams is mapped onto a common data model — commonly an implementation of CDISC standards such as CDASH at the collection stage and SDTM at the submission stage — so that downstream models operate on a consistent schema regardless of source system. In the model-assisted cleaning and coding stage, the NLP, anomaly-detection, and predictive-triage models described above operate continuously against incoming data, producing a prioritized queue of suggested queries, coding assignments, and risk flags rather than final determinations. In the human-in-the-loop adjudication stage, data managers and medical coders review model output through an interface that surfaces the model's confidence and supporting evidence, accepting, rejecting, or correcting each suggestion; these adjudications are logged and become the labelled training data used in the final, continuous monitoring and retraining stage, closing the loop illustrated conceptually in Figure 1.

A practical constraint on this framework is computational: the transformer-based NLP models that provide the strongest performance on clinical text are substantially more resource-intensive than the classical statistical and rule-based methods they are replacing, and industry-wide benchmarking indicates that the computational cost of both leading medical-image and biomedical-text models has grown considerably faster than the corresponding growth in cost-efficient GPU hardware over the past decade. Figure 2 illustrates this divergence using published benchmarking of AI/ML model performance density and performance cost relative to the trajectory of state-of-the-art GPU hardware between 2015 and 2022: models used in biomedical information processing (including large clinical language models) show a markedly steeper growth trajectory in both computational demand and cost than the hardware performance curve against which they are plotted, while models used in medical image segmentation show a comparatively shallower trajectory. This has direct operational implications for CDM automation programs: sponsors and technology vendors must budget not only for initial model development but for the escalating infrastructure cost of retraining and operating large language models as they are updated, and should favor efficient, task-specific fine-tuned models over the largest available general-purpose model wherever task performance allows.

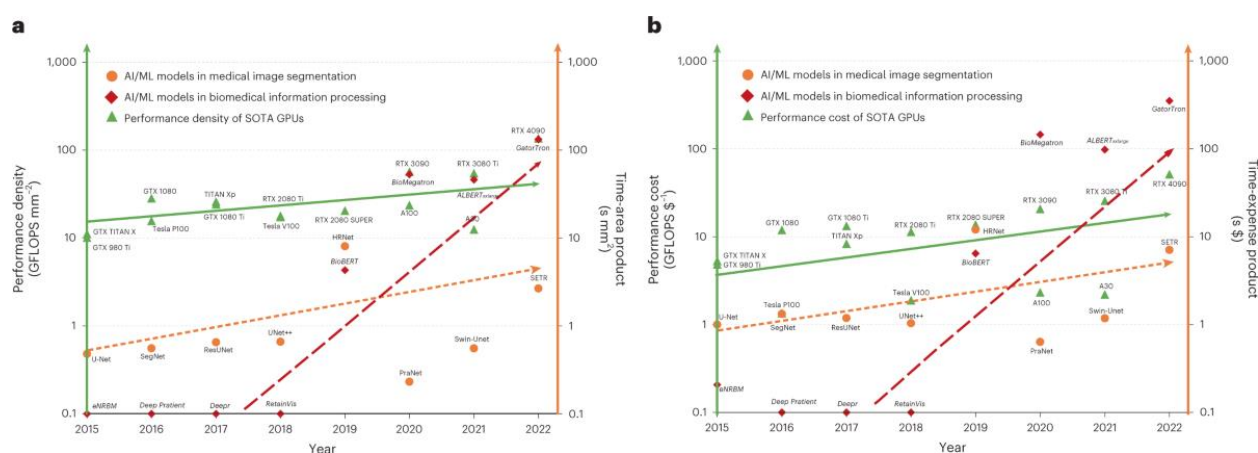


Figure 2. Growth in performance density (a) and performance cost (b) of AI/ML models used in medical image segmentation and biomedical information processing relative to state-of-the-art GPU hardware, 2015–2022. Biomedical information-processing models — the category most relevant to clinical-text NLP used in CDM — show the steepest year-over-year growth in both computational demand and cost.

Given this cost trajectory, most production CDM automation deployments adopt a tiered model strategy: lightweight, rule-augmented statistical models handle high-volume, low-ambiguity tasks (duplicate detection, simple range-plausibility checks), while larger fine-tuned language models are reserved for the smaller subset of genuinely ambiguous unstructured-text tasks where their added accuracy justifies the additional computational and validation cost.

QUALITY ENHANCEMENT MECHANISMS

The workflow automation described above translates into measurable data quality improvements through several specific mechanisms.

A. Automated and Intelligent Query Generation

Rather than waiting for a scheduled batch edit-check run, model-assisted systems can flag a likely discrepancy at or near the point of data entry, shortening the interval between an error being made and being detected. Because ML-derived triage models rank queries by predicted likelihood of representing a genuine, analysis-relevant discrepancy rather than a benign or explainable variation, reviewer effort concentrates on the queries most likely to matter, which tends to reduce both the total query burden on sites and the proportion of queries ultimately closed as 'no error found.'

B. Consistency and Duplicate Detection Across Sources

Record-linkage and similarity-based ML methods can identify duplicate patient records, duplicate adverse event entries, or inconsistent demographic information across systems that would otherwise require manual cross-referencing, a task that becomes increasingly important as trials integrate data from multiple EHR systems, registries, and wearable-device platforms with heterogeneous identifier schemes.

C. Automated Medical Coding with Human Verification

NLP-assisted coding tools propose MedDRA or WHODrug codes for verbatim terms with an associated confidence score, allowing a human coder to auto-accept high-confidence, unambiguous matches while focusing manual review on low-confidence or novel terms. This selective review model preserves coder oversight over clinically significant judgements while substantially reducing the volume of terms requiring full manual lookup, and it improves coding consistency across large coding teams by reducing the inter-coder variability that arises when different coders interpret an ambiguous verbatim term differently.

D. Risk-Based Monitoring and Reduced Source Data Verification Burden

As discussed in Section IV-C, predictive key-risk-indicator models allow monitoring resources to be redirected from uniform, resource-intensive on-site SDV toward centralized statistical monitoring focused on the sites and data domains where model output indicates elevated risk. Properly implemented and governed, this shift has been associated across the industry with maintained or improved data quality at reduced monitoring cost, because centralized statistical review can, in aggregate, examine patterns across the full dataset that individual site-level SDV cannot detect.

E. Continuous, Closed-Loop Quality Monitoring

Because the framework in Section V logs every human adjudication of a model suggestion, aggregate model performance (precision, recall, and calibration against human decisions) can be tracked continuously rather than assessed only at the point of initial validation. Systematic drift in model performance — for example, a fall in precision following a protocol amendment that changes the meaning of a previously well-modelled field — can be detected and addressed through retraining or rule adjustment before it materially affects data quality, rather than being discovered retrospectively at database lock.

REGULATORY, GOVERNANCE, AND LIFECYCLE CONSIDERATIONS

Because CDM operates within a regulated clinical research environment, the introduction of AI/ML components is subject to a different and in some respects more demanding set of expectations than AI deployed in lower-stakes commercial settings. Regulatory guidance relevant to good clinical practice, including ICH E6(R2)'s emphasis on a risk-based, quality-by-design approach to trial oversight and ICH E8(R1)'s articulation of critical-to-quality factors, provides the foundational framework within which any automated data-quality tool must be validated: the sponsor remains accountable for data integrity regardless of which tool — manual or automated — was used to achieve it.

A second, increasingly relevant body of guidance concerns AI/ML as a regulated technology in its own right. The FDA's evolving framework for AI/ML-based Software as a Medical Device (SaMD) articulates a Total Product Life Cycle (TPLC) approach: rather than validating a model once at a fixed point in time, as is conventional for locked, deterministic software, the TPLC approach recognizes that ML models can be updated as new data becomes available and asks sponsors to build a governed process — spanning premarket assurance of safety and effectiveness, a predefined change protocol describing how and when the algorithm may be modified, and real-world performance monitoring after deployment — around that capacity for change. Figure 3 depicts this TPLC concept schematically: data selection and model training feed a validation loop assessing performance and clinical evaluation, all embedded within a broader culture of quality and organizational excellence; premarket assurance and a defined algorithm-change protocol govern the transition of a model into a deployed state; and once deployed, new live data is continuously logged and evaluated through real-world performance monitoring, with insights feeding back into retraining.

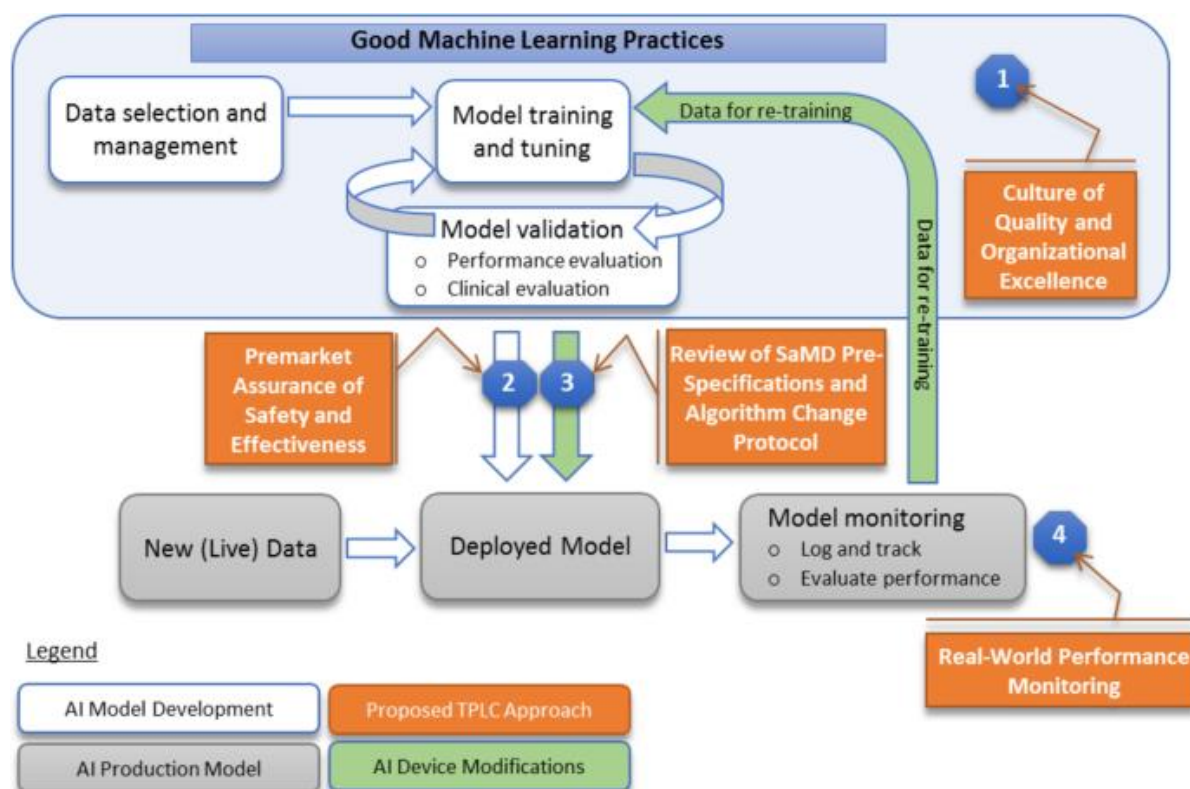


Figure 3. Total Product Life Cycle (TPLC) approach to AI/ML-based software, integrating Good Machine Learning Practices, premarket assurance, algorithm change protocols, and real-world performance monitoring within a continuous improvement loop.

Although most CDM automation tools are deployed as internal operational software rather than as a regulated SaMD product, the underlying TPLC logic is directly transferable and is increasingly adopted voluntarily by sponsors and technology vendors as a governance model: models used to flag queries, suggest codes, or prioritize monitoring should have a documented premarket validation against a representative, labelled dataset; any planned mechanism for retraining or updating the model should be pre-specified rather than made ad hoc; and post-deployment performance should be actively logged and periodically reviewed against the original validation benchmarks, with material performance drift triggering re-validation before continued use. This lifecycle discipline is what allows an adaptive, learning system to be reconciled with the static, audit-trail-based expectations of good clinical data management practice.

Good Machine Learning Practice (GMLP) principles — spanning multidisciplinary expertise in model development, representative and well-characterized training and test data, appropriate reference-standard definitions, and rigorous, clinically meaningful performance evaluation prior to and after deployment — provide

a practical checklist against which a CDM automation vendor's model development process can be assessed during sponsor due diligence, independent of whether the tool is formally regulated as a medical device.

Data privacy and security considerations run alongside these quality-focused frameworks. Because CDM systems handle protected health information, any AI/ML component must operate within the applicable regulatory perimeter — for example, HIPAA in the United States or the GDPR in the European Union — including provisions governing de-identification, data minimization, cross-border transfer, and the auditability of automated decisions that affect a data subject's records. Model training pipelines built on real clinical data should incorporate privacy-preserving techniques, such as federated learning or synthetic-data augmentation, where feasible, to reduce the volume of identifiable data that must be centralized for model development.

Audit trail requirements deserve particular emphasis because they are frequently underestimated during early-stage tool selection. Regulatory inspection of a trial's data management practices routinely requires a sponsor to reconstruct, for any given data point, who or what changed it, when, and why. For a deterministic edit-check rule this reconstruction is straightforward, since the rule's logic is static and documented in the database specification. For an ML-derived suggestion, the equivalent reconstruction must additionally capture which model version produced the suggestion, what confidence score it carried, what evidence supported it, and what human reviewer accepted, rejected, or modified it, together with a durable record of the training data and validation results associated with that model version at the time it was in production. CDM automation platforms intended for use in regulated trials should therefore be evaluated not only on model accuracy but on the completeness and immutability of this model-decision audit trail, since a highly accurate model that cannot produce an inspectable decision history is of limited practical value in a GxP environment.

Vendor and sponsor responsibilities also require clear delineation. Where a CDM automation capability is supplied by a third-party technology vendor, the sponsor retains ultimate accountability for data integrity under applicable good clinical practice obligations, and sponsor due diligence should accordingly extend to the vendor's model development process, training data provenance, validation methodology, and change-control procedures, typically documented through a combination of vendor audits, service-level agreements specifying model performance thresholds, and contractual notification requirements whenever a materially retrained or reconfigured model version is deployed into a live trial.

ILLUSTRATIVE APPLICATIONS

Several categories of application illustrate how the framework described above operates in practice across the CDM lifecycle. In adverse event processing, NLP pipelines extract structured event, seriousness, causality, and outcome information directly from investigator narrative fields, pre-populating the safety database entry that a pharmacovigilance reviewer then confirms or amends, shortening the interval between event documentation and case processing. In protocol deviation detection, ML models trained on historical deviation patterns flag visits or forms exhibiting a statistical signature associated with past deviations — for example, a cluster of out-of-window visits at a specific site — well before a scheduled monitoring visit would have surfaced the same pattern. In external data reconciliation, record-linkage models automatically match laboratory or ECG vendor records against EDC entries using a combination of patient identifiers and visit metadata, flagging only the subset of records that fail to match with high confidence for manual reconciliation rather than requiring every record to be manually cross-checked.

In decentralized and hybrid trials, where participants may generate data through home-based devices, patient-reported outcome applications, and telehealth visits in addition to conventional site visits, automated data harmonization and anomaly detection become particularly important because the volume and heterogeneity of incoming data streams exceed what a conventional, visit-scheduled manual review cadence can absorb. AI-assisted triage in this setting allows data management teams to maintain continuous data quality oversight across an asynchronous, high-frequency data stream rather than reviewing data in scheduled batches.

Registry and real-world-evidence programs present a related but distinct application. Because registries typically draw data directly from routine EHR extracts rather than from purpose-built CRFs, the proportion of clinically relevant information held only in unstructured text is considerably higher than in a conventional interventional trial, and NLP-based abstraction is correspondingly more central to achieving an analysis-ready dataset. Programs of this kind have used NLP pipelines to abstract comorbidity, treatment-line, and outcome information at a scale

that would be commercially impractical using purely manual chart abstraction, while retaining a sampled manual audit process to estimate and report the abstraction error rate, a quality-control pattern that is likely to generalize well to other high-volume, unstructured-data-heavy CDM applications.

CHALLENGES AND LIMITATIONS

Despite its demonstrated benefits, AI/ML-enabled CDM faces several substantive challenges. Model performance depends heavily on the quality and representativeness of the historical query and coding data used for training; because this data reflects the judgements, conventions, and occasional errors of the human reviewers who generated it, a poorly curated training set can encode and perpetuate systematic biases — for example, under-flagging error types that historically went unqueried because reviewers considered them low-priority rather than because they were genuinely benign.

Explainability remains a persistent limitation, particularly for transformer-based NLP models whose internal decision process is not readily interpretable by a human reviewer. Because data managers and auditors are expected to understand and justify why a given record was flagged, cleaned, or coded in a particular way, opaque model output can complicate both routine adjudication and retrospective audit, motivating continued preference in regulated settings for models that can surface supporting evidence (the specific text span or feature values driving a prediction) alongside their output rather than a bare classification.

Validation itself is nontrivial: unlike a deterministic edit-check rule, whose behaviour can be exhaustively tested against a finite set of conditions, an ML model's behaviour must be validated statistically against a representative test set, and its performance can degrade silently as the characteristics of incoming data drift from those of the training set — for instance, following a protocol amendment, the introduction of a new site population, or a change in an upstream EHR system's documentation conventions. Sustained governance, including the retraining and monitoring discipline described in Section VII, is required to manage this risk, and that governance itself carries an ongoing resource cost that must be weighed against the efficiency gains automation provides.

Finally, organizational and workforce considerations should not be understated. Effective deployment requires data managers and medical coders to develop new skills centred on model output adjudication rather than exclusively manual review, and change-management effort is required to build appropriate trust in model suggestions — trust that is neither excessive, leading to insufficiently critical acceptance of incorrect suggestions, nor insufficient, leading to reviewers manually re-verifying every model output and thereby forfeiting the efficiency gains the system was intended to provide.

A related limitation concerns generalizability across therapeutic areas and protocol types. A model trained predominantly on oncology trial data, for instance, may encode assumptions about visit schedules, adverse-event terminology usage, and laboratory panels that do not transfer cleanly to a cardiovascular outcomes trial or a rare-disease natural-history study with a small and heterogeneous patient population. Sponsors and vendors should therefore treat cross-indication transfer as an empirical question to be validated for each new therapeutic area rather than assumed, and should be particularly cautious when deploying a model in a population that is materially smaller or more heterogeneous than the population on which it was originally trained and validated, since model confidence estimates themselves can become unreliable outside the training distribution.

FUTURE DIRECTIONS

Several directions are likely to shape the next phase of AI/ML application in CDM. Multimodal models capable of jointly reasoning over structured CRF fields, free-text narratives, and imaging or waveform data may allow more holistic consistency checking than current unimodal pipelines, for example cross-verifying a reported adverse event against both the narrative documentation and any associated diagnostic imaging in a single pass. Federated and privacy-preserving learning architectures are likely to see wider adoption as sponsors seek to train models across multiple institutions' data without centralizing protected health information, addressing both privacy regulation and the practical reluctance of individual sites to share raw data externally.

Standardization of model validation and reporting practices specific to CDM automation tools — potentially building on existing frameworks for AI/ML SaMD and on CDISC data standards — would materially ease sponsor due diligence and cross-vendor comparability, an area where the field currently lacks consensus. Finally, closer

integration between AI-enabled CDM systems and adaptive trial design platforms may allow data-quality signals generated during a trial to inform real-time adaptive decisions (for example, triggering additional monitoring of a site or arm) rather than being reviewed only retrospectively at scheduled milestones, moving CDM from a primarily retrospective quality-assurance function toward a more continuously integrated component of trial conduct.

CONCLUSION

Artificial intelligence and machine learning are shifting clinical data management from a largely manual, retrospective discipline toward a continuously monitored, model-assisted workflow capable of absorbing the growing volume and heterogeneity of modern clinical data. NLP-based extraction closes the long-standing gap between unstructured narrative data and structured analysis-ready fields; ML-based anomaly detection and predictive triage direct scarce human review effort toward the records most likely to matter; and risk-based monitoring frameworks, increasingly informed by predictive analytics, allow quality oversight to scale without a proportional increase in manual review burden. Realizing these benefits safely, however, depends on embedding AI/ML tools within a governed, human-in-the-loop framework rather than deploying them as autonomous replacements for human judgement, and on adopting a lifecycle-based approach to validation, change management, and post-deployment monitoring consistent with emerging regulatory expectations for AI/ML-based software. Organizations that combine strong technical implementation with this governance discipline are best positioned to realize the workflow efficiency and quality gains that AI/ML-enabled clinical data management makes possible, while organizations that treat automation as a purely technical upgrade risk introducing new, less visible categories of data quality and compliance risk in place of the manual errors the technology was intended to eliminate.

REFERENCES

1. **Pawan Kumar Singh**, Scaling Gate-All-Around (GAA) Transistor Architectures for Sub-2nm AI Accelerators Using Atomically Thin 2D Materials, Vol. 33 No. 08 (2024): JOCAAA, Journal Name: Journal of Computational Analysis and Applications, Journal's ISSN: 1521-1398 (Paper),1572-9206 (Online), <https://eudoxuspress.com/index.php/pub/article/view/5713>
2. **Pawan Kumar Singh**, Energy-Efficient Task Scheduling in Green Cloud Computing Using Neuromorphic Spiking Neural Networks, Vol. 33 No. 08 (2024): JOCAAA, Journal Name: Journal of Computational Analysis and Applications, Journal's ISSN: 1521-1398 (Paper),1572-9206 (Online), <https://eudoxuspress.com/index.php/pub/article/view/5712>
3. **Rangavinay Teja Royal Jagga**, 2026, DOI: 10.1109/GCWCN66157.2025.11448495, Date of Conference: 22-23 November 2025, Date Added to IEEE Xplore: 25 March 2026 ,<https://ieeexplore.ieee.org/document/11448495>
4. **Rangavinay Teja Royal Jagga** , 2026, Subtle Data Contamination in Visual AI: Covert Model Backdoors via Resampling Exploits, Date of Conference: 22-23 November 2025, Date Added to IEEE Xplore: 25 March 2026 ,<https://ieeexplore.ieee.org/document/11448514>
5. **Rangavinay Teja Royal Jagga**, 2026, Robotic Tactile-Vision Integration for Interactive Object Disentanglement, <https://ieeexplore.ieee.org/document/11448337>
6. **Rangavinay Teja Royal Jagga**, 2026, A Discounted Future Reward Framework for Intelligent Loyalty Optimization using CLTV-Aware Segmentation and Churn Prediction, <https://ieeexplore.ieee.org/document/11497560>
7. **Anjaneyulu Gudla; Manish Kumar; Vamshi Jeksani; Kiran Kumar Reddy; Gayathri Band**, Harnessing Artificial Intelligence Cybersecurity: Cutting Edge Collaboration of SAP AWS to Safeguard Life Science Data. 2026 <https://doi.org/10.1109/ai69164.2026.11497833>
8. **G. Mahender, Department of CSE(Data Science), Vardhaman College of Engineering ; Vamshi Jeksani; Manish Kumar; Koppuravuri Sri Lakshmi Sruthi; Kiran Kumar Reddy**, SAP Cloud

- System on AWS for Risk Detection and Integration Artificial Intelligence Scalable Cybersecurity 2026 <https://doi.org/10.1109/aiei69164.2026.11497435>
9. **Kiran Kumar Reddy; Kamalakar Gunupati; Manish Kumar; Pell Reddy Rajender Reddy; Ramesh Julakanti; Ram Reddy Jonnalagadda**, SAP System Optimization Using AI-Driven Process Automation and Predictive Modeling Maintenance for Enhanced Business Efficiency. 2025 <https://doi.org/10.1109/computingcon64838.2025.11376613>
 10. **Amilineni, M.V. 2025**. Audit-Safe Agentic RAG Framework for Regulated Enterprise Systems: A Trustworthy AI Architecture for SAP, Finance, Healthcare, and Government Workflows. INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES. (Jan. 2025), 34–46. DOI: <https://doi.org/10.29284/7fjzk883>
 11. **Pawan Kumar Singh**, ACCELERATING LARGE LANGUAGE MODEL TRAINING USING HYBRID QUANTUM-CLASSICAL VARIATIONAL AUTOENCODERS WITH INTEGRATED PHOTONIC QRAM, Vol. 54 No. 3 (2026): July-September 2026, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/501>, DOI: <https://doi.org/10.46121/pspc.54.3.45>
 12. **Pawan Kumar Singh**, RESILIENCE ASSESSMENT OF MODERN POWER GRID INFRASTRUCTURE AGAINST EXTREME WEATHER USING PHYSICS-INFORMED NEURAL NETWORKS, Vol. 54 No. 1 (2026): January-March 2026, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/502>, DOI: <https://doi.org/10.46121/pspc.54.1.64>
 13. **Pawan Kumar Singh**, MITIGATING THE EXABYTE STORAGE CRISIS: ENHANCING READ/WRITE EFFICIENCY IN 5D OPTICAL GLASS STORAGE USING SILVER NANOPARTICLE DOPING, [Vol. 38 No. 11s \(2025\)](#), International Journal of Applied Mathematics (IJAM), pISSN 1311-1728, eISSN 1314-8060, <https://ijamjournal.org/ijam/publication/index.php/ijam/article/view/2153>
 14. **Mr. Pawan Kumar Singh (Author), Miss. Pari Nidhi Singh (Author)**, Intelligent Bot, ASIN : B0DBGCQY8X, Publisher : Namya Press, Publication date : November 25, 2022, Language : English, Print length : 105 pages, ISBN-10 : 9390124867, ISBN-13 : 978-9390124862, Item Weight : 7.5 ounces, Dimensions : 6.24 x 0.43 x 9.24 inches, https://www.amazon.com/Intelligent-Bot-Pawan-Kumar-Singh/dp/B0DBGCQY8X/ref=sr_1_5?crid=2VAXN5BRN5SMS&dib=eyJ2IjojMSJ9.v3OCQvHBpy20RIjfmWh0qbOl-RLOIUJIC8Xa0Pjeqilvo2sbZ_suN3LbSCw_OA_ZD_4lgjRLMGlBwZQQ1ehZEfDpt_MJiOUfVUqqz_Mav0I5_mb6pN5DQog0U7SA826n1DQO3QAtWwopHy4lC5IfhZXttKj9faiyATwhdtDATSSV1-r7FmzM2VQyIA0jxcfq6E6hp9ECfH1fby0li9kUfGpLxMq4GktlzUIJO5ggrhsQ.G1AjQiVqU4HFsnScz_vz-X6NXCNs3EFSn-RI8mcWWGOC&dib_tag=se&keywords=Pawan+Kumar+Singh%2C+book&qid=1787815741&sprefix=pawan+kumar+singh%2C+bo%2Caps%2C450&sr=8-5
 15. **Mr. Pawan Kumar Singh (Author)**, STORAGE CRISES DUE TO RISE OF AI: A Comprehensive Guide to Managing the Storage Challenges in the Age of Artificial Intelligence, ASIN : B0HFK12ZCT Publisher : EB1 Profile, Accessibility : Learn more, Publication date : August 17, 2026, Edition : 1st Language : English, File size : 1.7 MB, Screen Reader : Supported, Enhanced typesetting : Enabled X-Ray : Not Enabled, Word Wise : Not Enabled, Print length : 261 pages, ISBN-13 : 979-8951479112, Page Flip : Enabled, https://www.amazon.com/STORAGE-CRISES-DUE-RISE-Comprehensive-ebook/dp/B0HFK12ZCT/ref=sr_1_1?crid=9MF6C4WDVJR5&dib=eyJ2IjojMSJ9.DQ8fesBch7ELmRnilDt-YA.NBSCHLmetvnbhkwIpOoj3s0IVsCV8y6-

- [ZCfCP6U11c&dib_tag=se&keywords=Pawan+Kumar+Singh%2C+book%2C+eb1+Profile&qid=1787816040&sprefix=pawan+kumar+singh%2C+book%2C+eb1+profile%2Caps%2C420&sr=8-1](https://doi.org/10.46121/pspc.53.2.34)
16. **Parag Bharatkumar Parikh**, FRAMEWORKS FOR MITIGATING SUPPLY CHAIN DISRUPTIONS AND MATERIAL PRICE VOLATILITY IN SUSTAINABLE BUILDING CONSTRUCTION, Vol. 53 No. 2 (2025): April-June 2025 , , Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/447>, DOI: <https://doi.org/10.46121/pspc.53.2.34>
 17. **Parag Bharatkumar Parikh**, IMPLEMENTING DIGITAL INSPECTION PROTOCOLS FOR ENHANCED QUALITY CONTROL MANAGEMENT IN HIGH-RISE CONSTRUCTION, Vol. 54 No. 3 (2026): July-September 2026, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/439>, DOI: <https://doi.org/10.46121/pspc.54.3.12>
 18. **Parag Bharatkumar Parikh** , EVALUATING THE FEASIBILITY AND EFFICIENCY OF 3D-PRINTED CONCRETE TECHNOLOGY IN SUSTAINABLE MODULAR CONSTRUCTION, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/issue/view/30>, DOI: <https://doi.org/10.46121/pspc.54.1.61>
 19. **Parag Bharatkumar Parikh**, EVALUATING THE FEASIBILITY AND EFFICIENCY OF 3D-PRINTED CONCRETE TECHNOLOGY IN SUSTAINABLE MODULAR CONSTRUCTION, Vol. 54 No. 1 (2026): January-March 2026, System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/440>, DOI: <https://doi.org/10.46121/pspc.54.1.61>
 20. **Ronakkumar shah**, REDEFINING BUSINESS PROCESS MANAGEMENT: THE SYNERGY OF AI AND INTELLIGENT AUTOMATION, Vol. 54 No. 3 (2026): July-September 2026, System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/448>, DOI: <https://doi.org/10.46121/pspc.54.3.14>
 21. **21: Ronakkumar shah**, OPERATIONAL ACCOUNTABILITY IN ACTION: METRICS AND MECHANISMS FOR SYSTEMIC SUCCESS, Vol. 53 No. 4 (2025): October-December 2025, System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/452>, DOI: <https://doi.org/10.46121/pspc.53.4.40>
 22. **Ronakkumar shah**, BEYOND THE BLUEPRINT: AUTOMATED WORKFLOW ENFORCEMENT AND POLICY GOVERNANCE, Vol. 53 No. 2 (2025): April-June 2025 , System Protection and Control, ISSN-1674-3415 ,<https://pspac.info/index.php/dlbh/article/view/451>, DOI: <https://doi.org/10.46121/pspc.53.2.35>
 23. **Tushita Agarwal; Swaminathan Sethuraman**, Mapping Customer Behavior: Visual Analysis of Online Retail Interactions, Date Added to IEEE Xplore: 17 February 2026, DOI: <https://doi.org/10.1109/IC2NC67409.2025.11376261>
 24. **Tushita Agarwal**, Efficient Data Summarization and Comparison via Grid-Based Statistical Proxies for Large-Scale Datasets, **Date Added to IEEE Xplore: 03 December 2025**, DOI: <https://doi.org/10.1109/ICBATS66542.2025.11258564>
 25. **Dip Bharatbhai Patel**, Building scalable business intelligence systems in the cloud: A technical approach, International Journal of Science and Research Archive, 2021, 02(01), 212-215., Received on 03 February 2021; revised on 17 April 2021; accepted on 19 April 2021, DOI: <https://doi.org/10.30574/ijrsra.2021.2.1.0047>
 26. **Dip Bharatbhai Patel**, University of North America, Virginia, USA, Home / Archives / Vol. 5 No. 02 (2025): Volume 05 Issue 02 / Articles, <https://www.academicpublishers.org/journals/index.php/ijdsml/article/view/5958>

27. **Sai Akshara Vemuri** , KINEMATIC OPTIMISATION OF SOFT ROBOTIC MANIPULATORS FOR MEDICAL INTERVENTION, Vol. 54 No. 3 (2026): July-September 2026, System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/450>
28. **Sai Akshara Vemuri**, <https://www.periodicos.ulbra.org/index.php/acta/article/view/528>
29. **Aziz, A., Chandusha, K., Naga Kiran, B., Prasad, M. L. M., Golla, K., Rajeswari, S., & Katta, S. K. R.** (2026). Exploring Machine Learning Applications for Genomic Data Analysis in Personalized Medicine. International Journal of Drug Delivery Technology, 16(38s), 1166-1173.
<https://doi.org/10.25258/ijddt.16.38s.127>, <https://www.periodicos.ulbra.org/index.php/acta/article/view/687>
30. **Mohammad Naffizuddin , Chaganti Ashok ,Sowjanya Gunukula , Rvbs Sarma , Bhavana Sujanamulk , Bharani Krishna Takkella , Chukka Ram Sunil ,Varri Sujana , Ashwini Kumar**, Etiological Factors of Orofacial Clefts in a Hospital-Based Population: A Descriptive Analytical Study, Published: December 10, 2025, DOI: 10.7759/cureus.98868, DOI: <https://doi.org/10.7759/cureus.98868>
31. **Manoj Kumar Kagitha**, Automated Security Compliance in Ci/Cd: A Natural Language Processing (NLP) Approach to Detecting Vulnerabilities in Infrastructure-As-Code (Terraform/Bicep), Publication year 2021, Volume 2021, Issue 12, <https://computerfraudsecurity.com/index.php/journal/article/view/970>
32. **Manoj Kumar Kagitha**, Hybrid Quantum-Classical Resource Scheduling: A Proposed Architecture for Next-Generation Hyperscale Data Centers, Publication year 2024, Volume 2024, Issue 12, <https://computerfraudsecurity.com/index.php/journal/article/view/969>
33. **Manoj Kumar Kagitha**, Self-Healing Cloud Infrastructure: A Hybrid Reinforcement Learning Framework for Predicting Microservice Failures in Azure Kubernetes Service (AKS), Publication year 2023, Volume 2023, Issue 12, <https://computerfraudsecurity.com/index.php/journal/article/view/971>
34. **Manoj Kumar Kagitha**, GREENOPS IN THE CLOUD: AN AI-DRIVEN TELEMETRY ANALYSIS MODEL FOR REDUCING CARBON FOOTPRINT IN HIGH-AVAILABILITY CLUSTERS, Vol. 28 No. 6 (2020): JOCAAA, <https://www.eudoxuspress.com/index.php/pub/article/view/5054>
35. **Lakshmi Rahul Reddy Mareddy.** (2026). Security and Performance Assessment of Encryption Techniques for Cloud Platforms. International Journal of Communication Networks and Information Security (IJCNIS), 18(1), 68–80. Retrieved from <https://ijcnis.org/index.php/ijcnis/article/view/8735>
36. **Lakshmi Rahul Reddy Mareddy**, Context-Aware Learning Approaches for Improving Prediction Accuracy in Dynamic Systems, International Journal of Computational Mathematical Ideas (IJCMI), Vol.17 No.3(2025): IJCM Vol17 NO3 2025
<https://www.ijcmi.com/index.php/journal/article/view/15>
37. **Lakshmi Rahul Reddy , Mareddy.** (2026). Revisiting Autonomy in Artificial Intelligence through Agentic System Design. Journal of Computational Analysis and Applications (JoCAAA), 35(1), 707–713. <https://eudoxuspress.com/index.php/pub/article/view/4780>
38. **Mahendrakumar Kalal**, Lean-Driven SAP Production Planning Optimization Using Machine Learning for Inventory and Throughput Efficiency, Date of Conference: 19-20 March 2026, Date Added to IEEE Xplore: 08 April 2026, ISBN Information:, ISSN Information:

- DOI: 10.1109/ISDFS69419.2026.11459029, <https://ieeexplore.ieee.org/abstract/document/11459029>
39. **Mahendra kumar Kalal**, SUPPLY CHAIN RESILIENCE THROUGH INTEGRATED SAP PRODUCTION PLANNING – PROCUREMENT SYNCHRONIZATION, Archives / Vol. 54 No. 1 (2026): January-March 2026 / Articles, <https://pspac.info/index.php/dlbh/article/view/259>
40. **Dr Manoj Goel (December 2018)**. THE ATTITUDE OF DIFFERENT STATES STUDENTS TOWARDS PHYSICAL EDUCATION; STUDYING IN PHYSICAL EDUCATION COLLEGES OF MAHARSHTRA STATE, <https://ijeponline.com/index.php/journal/article/view/965/847>
41. **Ilakiya Ulaganathan, Tagore Engg. College, Anna University, Chennai, India, ; FNU Antara; Santhosh Chitraju Gopal Varma; Zokir Mamadiyarov; Ziyadulla Turayev**, Published in: 2025 International Conference on Sustainability, Innovation & Technology (ICSIT), Date of Conference: 22-23 August 2025, Date Added to IEEE Xplore: 19 December 2025, ISBN Information: <https://ieeexplore.ieee.org/document/11294748/references#references>
42. **Shreyas Kasture, Sudhanshu Maurya, Amit Kumar Sharma, Santhosh Chitraju Gopal Varma, Kashish Mirza, Firdous Sadaf Mohammad Ismail**, DOI: <http://doi.org/10.11591/ijai.v15.i2.pp1746-1761>, IAES International Journal of Artificial Intelligence (IJ-AI), ISSN/e-ISSN 2089-4872/2252-8938 <https://ijai.iaescore.com/index.php/IJAI/article/view/26932>
43. **Shiv Shankar Dwivedi, Mr. Santosh Kumar Sharma**, Future-Proofing Data Security with Quantum-Resistant Cryptography, Vol. 11 No. 3 (2025): IJCESEN, DOI: <https://doi.org/10.22399/ijcesen.3447>, <https://ijcesen.com/index.php/ijcesen/article/view/3447>
44. **Vijender Kalmotia**, Enhancing statistical accuracy in gene perturbation studies, Volume 266, August 2026, 105819, <https://doi.org/10.1016/j.biosystems.2026.105819>, <https://www.sciencedirect.com/science/article/pii/S0303264726001292>
45. **Naveen Undrathi**, Design and implementation of generative AI-driven data transformation pipelines using spring boot for scalable financial applications, eISSN: 2581-9615 || CODEN: WJARAI || Impact Factor 8.2 || CrossRef DOI, September 2026 (Volume 31, Issue 3), DOI url: <https://doi.org/10.30574/wjarr.2026.30.3.1629>,
46. **Naveen Undrathi**, Agentic generative AI-enabled microservices framework for intelligent financial data processing in java-based enterprise systems, September 2026 (Volume 31, Issue 3), eISSN: 2581-9615 || CODEN: WJARAI || Impact Factor 8.2 || <https://doi.org/10.30574/wjarr.2026.30.3.1628>
47. **Akshay Jain**, Comparative Analysis of Linkage Methods in Hierarchical Clustering for Customer Segmentation, Conference paper, First Online: 07 October 2025, https://link.springer.com/chapter/10.1007/978-981-96-7289-9_44
48. **Ismail Hossain**, Department of Computer Science, George Mason University, Fairfax, VA, USA, Mridul Banik, Department of Computer Science, Colorado State University, Fort Collins, CO, USA, <https://ieeexplore.ieee.org/abstract/document/11446534> DOI: 10.1109/FMLDS67896.2025.00106
49. **Ismail Hossain, Mridul Banik , Fahmid Al Farid , Jia Uddin , Hezerul bin Abdul Karim ,** Advancing Large Language Models for Low-Resource Languages: A Systematic Review of Pretraining, Adaptation, and Ethical Challenges, 10.32604/cmes.2026.075507, Date: 2026 <https://www.sciencedirect.com/org/science/article/pii/S152614922600322X>

50. **Ismail Hossain; Mridul Banik**, Bridging Visual and Linguistic Intelligence for Chest X-rays: A Comprehensive Review of ViTs and LLM Synergies, Published in: 2025 IEEE International Conference on Future Machine Learning and Data Science (FMLDS) Date of Conference: 02-05 November 2025, Date Added to IEEE Xplore: 25 March 2026, ISBN Information: <https://ieeexplore.ieee.org/abstract/document/11446534>
51. **Rifat Sarker Aoyon¹, Ismail Hossain², M. Abdullah-Al-Wadud³, Jia Uddin⁴**, A Secured and Continuously Developing Methodology for Breast Cancer Image Segmentation via U-Net Based Architecture and Distributed Data Training, <https://www.sciopen.com/article/10.32604/cmes.2025.060917>