



Scalable Pipelines for Integrating Structured and Unstructured Clinical Data

¹FNU Sudhakar Abhijeet

¹Independent Researcher,

¹Northeastern University, Boston

Abstract: The sheer growth rate of the size and complexity of healthcare data, including structured data in the form of laboratory results, diagnostic codes, etc., in addition to unstructured data in the form of clinical notes, radiology reports, and more informal free-form narratives, presents major challenges when it comes to integration and analysis. The heterogeneity of these multimodal inputs, their scale, and velocity may exceed the capabilities of traditional data management frameworks, and thus, it may be difficult to leverage such underlying data management frameworks in clinical decision support, population health management, or research. Cloud-native data pipelines, the use of distributed computing frameworks, and continuous learning using modern natural language processing models have become a possible answer to consolidate disparate data feeds into a unified, queryable data source. These pipelines allow the construction of longitudinal patient records, real-time analytics, and secure cross-institutional collaboration, with adaptive orchestration and cost-aware optimization being included to stay operational. The architectures, operationalization, and research directions of such pipelines are discussed in this article in terms of resilience, compliance, and efficient management of large-scale clinical data environments.

IndexTerms - Clinical Data Integration, Scalable Pipelines, Structured and Unstructured Data, Natural Language Processing, Healthcare Analytics

I. INTRODUCTION

Clinical data has been exponentially increasing based on its breakout into structured fields such as lab results, diagnostic codes, and medication orders, and unstructured data such as physician notes, imaging reports, and discharge summaries, and it is realigning healthcare analytics. Healthcare systems must decide how to take advantage of this explosion of multimodal information and how to meet its challenge. To make better decisions, to promote personalized medicine, and to facilitate large-scale research efforts utilizing clinical data, integration across these various forms of data is key, but conventional data management practices have not been able to keep pace with the heterogeneity, scale, and speed of the more contemporary clinical information [1-3]. Specifically, the structured data with its limited schema and relational systems compatibility compares favourably to free-text accounts and image-based derived reports, which are not uniform and can only be digested after considerable processing to become meaningful [4, 5]. Much of the potential value of clinical data has been locked away and is unavailable to stakeholders in the field of healthcare in the absence of scalable pipelines that are capable of uniting such formats in a seamless manner [6-8].

The past few years have seen the introduction of distributed systems, cloud-native frameworks, and deep learning-based natural language processing, which have enabled the formation of coherent, queryable, and accessible resources across structured and unstructured clinical datasets. Multimodal inputs could now be ingested, processed, and aligned at scale due to technologies like scalable extract-transform-load (ETL) frameworks, semantic mapping engines, and transformer-based natural language understanding models in healthcare systems [9-11]. When combined with a modern pipeline, such tools allow generating unified patient histories and population-wide data that can drive the decision support, predictive analytics, and clinical research without making it necessary to violate privacy and security requirements [12-13].

This article discusses the infrastructures, architectures, and operational factors of a scalable pipeline aimed at blending unstructured and structured clinical data. The paper starts with an outline of the difficulties presented by multimodal healthcare integration, shifts to pipeline architectures and the technologies supporting them, and talks through their use in practice and research, and finishes with how they are likely to proceed in the years to come in the development of the next generation of healthcare data ecosystems. One section leads to another, and there is consistency of knowledge transfer in reading about how the old issues become the new solutions.

II. CHALLENGES OF INTEGRATING STRUCTURED AND UNSTRUCTURED CLINICAL DATA

The difficulty of mixing structured as well as unstructured clinical information has to do with the variability in how such types are presented, in context, and the processing they demand. Structured data, e.g., lab values, vitals, coded diagnoses, is in standardized formats, and thus it is fairly easy to store and query in a relational database. On the other hand, the content that is not structured, such as clinical narratives, radiology impressions, pathology notes, even voice transcriptions all display the features of being variable in language, style, and enumerativeness, frequently including domain specific abbreviations, synonyms, and expressions that themselves are sensitive to the context and therefore intractable to traditional parsing [14-16]. Such differences are problematic in three key dimensions: semantic fusion, computational upper bounding, and legal conformity. Semantic harmonization is the mapping of unstructured content to structured forms, usually through the extraction of key concepts and their reference to standardized ontologies such as SNOMED CT, LOINC, or ICD. This is going to change as it necessitates relatively advanced natural language processing where context is disambiguated, e.g., to distinguish between negated conditions and verified diagnosis [17, 18]. Limited computational scalability is also a challenge as healthcare organizations have to handle terabytes of text and imagery with billions of structured records, and thus require distributed processing frameworks as well as efficient storage systems [19, 20]. Such challenges are exacerbated by the need to become compliant with local or international regulatory requirements, such as privacy and patient confidentiality laws, e.g., the HIPAA and GDPR, which require sensitive patient data to be handled securely, audited, and de-identified with free-text data too. Such difficulties explain why it is important to integrate pipelines capable of addressing both the problems of semantic complexity and the scale of healthcare data. As we will see in the following section, the next generation pipe architecture can overcome these barriers through the use of cloud-native technologies, distributed systems, and sophisticated machine learning, shifting the focus from abstract challenges to real-world solutions.

III. ARCHITECTURES OF SCALABLE INTEGRATION PIPELINES

Modular, distributed, and scalable pipelines to combine structured and unstructured clinical data rely on an elastic, fault-tolerant architecture that is designed to be compliant. Such pipelines are usually decoupled using event-driven or microservices-based architecture, where the ingestion, processing, and storage are independent of each other, scaling independently [21-23]. Structured systems (such as HL7/FHIR APIs) and streaming platforms (like Kafka or Pulsar) are used in ingestion layers to capture both structured data (in the form of APIs) and unstructured data (to capture free-text notes, imaging metadata, and audio transcripts, etc.). With an eye on big data, processing layers now use hybrid strategies that blend conventional ETL-style processing on structured fields with natural language processing and machine learning of unstructured sources. Pretrained language models in the form of transformers are generally combined to be used in entity recognition, negation detection, and relation extraction so that free-text narratives may be converted into structured data units in relation to usual terminologies [24-26]. Regarding images, pipelines will search computer vision models that will digitize metadata or textual observations, e.g., lesion size measurements or severity ratings, that can be associated with text reports and laboratory results to form complete reports [27]. These pipelines are based on a distributed computing platform, such as Spark or Flink, to handle the scales of data involved in such systems (both batch and streaming pipelines), as well as cloud-based object storage with healthcare compliance (e.g., AWS S3 with encryption and audit logging). The unification of access to joined data is achieved by metadata catalogs and semantic layers through which a researcher and clinician can make queries to multimodal patient history readily [28, 29]. With the hybrid of these technologies, these pipelines can assist in not only real-time clinical decision-making but also at scale, retrospective research, as well. The technical underpinnings of these systems have now been set, and the following discussion turns to the operationalization of these systems in healthcare, their practical utility, and factors to consider in the deployment of these pipelines within live clinical and research systems. While modular architectures provide the structural backbone for integrating structured and unstructured data, the selection of processing frameworks and their trade-offs also influences scalability and cost-effectiveness. The following table outlines commonly used frameworks for different pipeline layers, highlighting their scalability properties and constraints.

Table 1: Frameworks and Tools Supporting Scalable Clinical Data Pipelines

Pipeline Layer	Example Frameworks	Scalability Strengths	Limitations in Clinical Context
Ingestion & Streaming	Apache Kafka, Apache Pulsar	Handles millions of records per second; supports real-time ingestion	Requires tuning for large unstructured data bursts and secure transport layers
Batch Transformation	Apache Spark, Hadoop	Distributed ETL for structured datasets; fault-tolerant	Less efficient for sub-second latency requirements
Text and NLP Processing	Hugging Face Transformers, spaCy, BioBERT-based models	Efficient processing of large unstructured corpora with GPU acceleration	Requires careful tuning to comply with de-identification and privacy rules
Metadata & Query Layer	Presto, Trino, Data Catalog APIs	Low-latency query over integrated datasets; supports multi-cloud	Complex schema mapping needed for heterogeneous systems

The interplay between these frameworks allows pipelines to operate seamlessly under varying workloads, creating a foundation for operationalization in clinical and research environments, as discussed next.

IV. OPERATIONALIZING PIPELINES IN CLINICAL AND RESEARCH SETTINGS

Healthcare ecosystems require integration pipelines to be scaled, and there is a trade-off between performance, usability, and compliance. Pipelines also allow the generation of longitudinal patient records in clinical settings that bring together structured entries in the EHR and the analysis of narratives, radiology reports, and pathology results to allow clinicians access to a holistic picture of a patient's history in one place [30]. Decision Support Systems are based on these integrated records that have rules engines or predictive models that use unified data to establish care gaps, prescribe therapies, or prompt alerts to high-risk situations. Scalable pipelines allow the building of population-level data sets that contain both coded and narrative data across research settings. Connecting unstructured captures of comorbidities, adverse events, or lifestyle factors with the known structured data, such as lab results and prescriptions, researchers achieve higher-quality data in an observational study and predictive analysis. Cohort identification is simplified as well by pipelines since automatically cohort searches in both structured and unstructured data are possible to find patients whose patient records satisfy complex inclusion criteria, a task that used to be accomplished by manual examination of the charts [13,17]. This is to guarantee that there is a compliant deployment in these deployments. Strict governance is applied through the use of access controls and audit trails of personally identifiable information (PII) being removed prior to integration through de-identification modules. The data is stored and transmitted in a secure and encrypted manner, and federated learning systems enable collaboration across multiple institutions without any need to share patient data in any centralized fashion or even at all, which also prevents leakage of privacy [19,26]. However, all these practical advantages can only be realized through highly advantageous observability and optimization frameworks that support the maintenance of pipelines as performing, cost-effective, and reliable at scale. The second bit expounds upon how these kinds of frameworks permit long-term stability and efficacy in manufacturing.

V. OBSERVABILITY AND PERFORMANCE OPTIMIZATION IN INTEGRATED PIPELINES

The ability to ensure observability and improve performance is crucial in maintaining reliability and activity continuation in pipelines that have been made scalable with a combination of structured and unstructured clinical data, as illustrated in Figure 1. The first line of observability is ingestion, transformation, and storage, health and efficiency monitoring over time via distributed telemetry to monitor pipeline throughput, latency, and error rate. Framework-based dashboards (i.e., Grafana and Prometheus) can deliver timely views of the bottlenecks, both within textual processing backlogs, failures with ETL jobs, or network saturation in massive data-moving operations [22, 24]. This optimization of performance is especially important when there are pipelines that manipulate petabyte-level data, or when it is able to manipulate clinical narratives in near-real time to enable decision support. Stream processing layers complete the batch processing footwork to allow a role in achieving a balance between historic analysis and time-sensitive applications, to give example, an alert to a clinician with abnormal findings within minutes of a diagnostic report being produced. Cluster control utilizes autoscaling rules that are developed through historical workload behavior and reactive rules to dynamically scale compute resources in a manner that is cost-effective and does not impact service within a cluster [26, 27]. Security assurance and compliance observability are built into these observability constructs. Policy validation tools validate on a continuous basis that every component of the pipeline is in compliance with the healthcare policies, such as encryption standards, completeness of audit logs, and role-based access controls. For runtime, the policy drift can be detected to identify deviations of the baselines, like the occurrence of unauthorized access permission of containers in NLP processing pods, which automatically roll back and dispatch alerts to the security teams [29]. Observability and optimization, having been developed and identified as key to operational trust, the next question arises as to how these pipelines are able to deliver elasticity and scalability to the varying requirements in a variety of institutions and geographical locations.

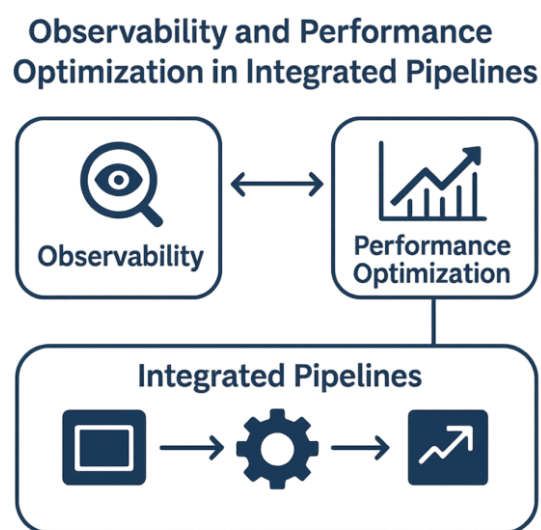


Figure 1: Overview of Observability and Performance Optimization in Integrated Pipelines. The diagram illustrates how observability (monitoring and visibility) and performance optimization (efficiency and scaling) interact to enhance the stages of integrated data or process pipelines.

VI. ACHIEVING ELASTICITY AND SCALABILITY ACROSS INSTITUTIONS

The fact that healthcare organizations carry out their work in distributed environments, such as hospital networks, research collaborations, and cloud-based infrastructure, requires very elastic and high-scaling pipelines, as illustrated in Figure 2. Elasticity gives pipelines the flexibility to adjust to peak loads at certain intervals, like seasonal patients or mass retrospective studies, and down to minimal activity through resource adjustment to manage the cost [21,25]. Horizontal scaling is often used, and it allows processing structured records and unstructured stories simultaneously, whereas vertical scaling is used on compute-intensive elements, such as huge transformer models, to comprehend the meaning of the text. The cross-institutional deployments are based on the federated data processing model that involves models or analytics queries being performed per individual organization in its environment, instead of moving sensitive patient data to a central place. This will both reduce the risk to privacy and give worldwide insights based on aggregation and de-identification of outputs [26, 27]. In order to scale to the vagaries of the network, edge-compute nodes perform initial text and structured data processing at the site of creation and transmit only key features to centralized systems to do more complicated analytics. The hybrid model reduces latency, ensures centralized governance, and provides the ability to audit.

Similarly, multi-cloud or hybrid-cloud strategies are also used to scale pipelines across regions, so that the institutions can take advantage of vendor-specific capabilities in machine learning acceleration or geographical compliance certification. The architectures also have resilience to failovers that are automated, giving these architectures the ability to continue operating even when there is an outage in the regions where clouds operate [23]. These environments have become more and more complex, and, as a result, smarter methods of resource management require investigation, which is the topic of the next section: adaptive optimization and AI-driven orchestration exploration.

Achieving Elasticity and Scalability Across Institutions

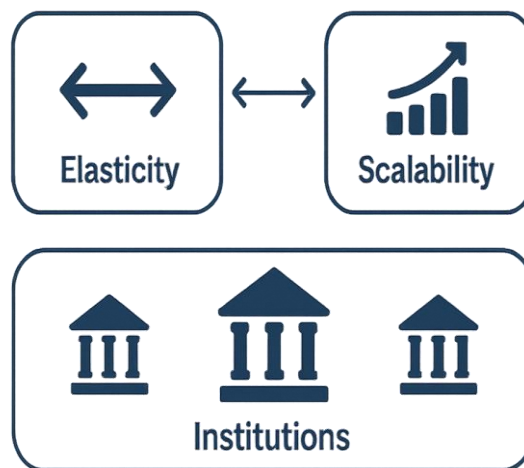


Figure 2: Diagram illustrating Achieving Elasticity and Scalability Across Institutions. The figure highlights how elasticity (the ability to rapidly adjust to changing demands) and scalability (capacity expansion) work together to support the growth and adaptability of institutions such as universities and organizations.

VII. ADAPTIVE ORCHESTRATION AND COST-AWARE AUTOMATION

Much like the clinical data pipelines grow in size and the types of work they have to handle, the cost of the operations and the efficiency thereof have to be maintained. Adaptive orchestration frameworks solve this problem through dynamic prioritizing of tasks and the allocation of resources in real-time according to demand, criticality, and cost restrictions. Auto rocket science traditionally, the policies to drive autoscaling are either specified manually or based on empirical equations of past usage volumes to predict future usage demand [28]. Those systems enable latency-sensitive workloads, like real-time diagnostic alerting, to keep allocated resources and allocate less expensive compute nodes, such as Spot Instances, to standard batch jobs. The techniques of model efficiency, such as pruning and quantization, also minimize the infrastructure costs. Through inference compression of transformer models, it is possible to run edge deployments on more precise NLP tasks, such as named entity recognition or clinical summarization, with very little accuracy loss [25]. Pipelines, when used together with tiered storage systems, which are commonly referred to as datasets that are often accessed and integrated in high-performance storage levels and are migrated to inexpensive cold storage, can help attain highly reduced costs with the potential to scale within a short period of time when necessary [30]. Importantly, any kind of optimization strategy is to be carried out in a compliance-sensitive framework. Workload orchestration systems may include policy enforcement layers to prevent the cost-saving actions, like those of moving workloads to lower-cost compute environments, from running afoul of the relevant regulations of data residency in various regions or with typical healthcare privacy requirements. These safeguards not only make adaptive orchestration a means to control cost, but also a means to achieve resilient and compliant scalability. With these operational optimizations becoming a benchmark, it is necessary to review the challenges that continue to inhibit the potential of fully integrated clinical data pipelines, and this is outlined in the next section. While adaptive orchestration governs real-time workload management, specific cost-control techniques determine whether such pipelines remain economically viable at scale. The table below summarizes distinct optimization techniques and their practical impacts on healthcare data processing costs.

Table 2: Cost Optimization Techniques in Clinical Data Pipelines

Optimization Strategy	Typical Application Scenario	Estimated Cost Reduction (%)	Potential Risks and Mitigation
Tiered Storage (Hot/Warm/Cold Tiers)	Archival and retrieval of integrated patient datasets	30–45%	Retrieval latency in cold tiers is mitigated by caching policies
Quantized Transformer Inference (8-bit/16-bit)	Near-real-time NLP tasks in edge deployments	40–55%	Slight model accuracy drop offset by retraining or calibration
Spot Instance Utilization for Batch Jobs	Historical analytics on unstructured notes and EHR data	50–70%	Requires job checkpointing and pod eviction handling
RL-Driven Autoscaling of Compute Clusters	Real-time alert systems and multimodal analytics	35–50%	Risk of latency spikes mitigated by minimum-capacity baselines

These strategies, while effective in reducing costs, must be carefully integrated with compliance frameworks and resilience mechanisms to maintain clinical-grade reliability, a concern echoed in the challenges that follow.

VIII. CHALLENGES LIMITING PIPELINE EFFECTIVENESS

Although technological advances have paved the way for scaling up integration of structured and unstructured clinical data, there are a few challenges that remain unsolved. The challenge of data quality is also ubiquitous; lack of agreement in coding, missing data points, and lack of clarity in free-form text narratives continue to compromise subsequent analyses, despite powerful NLP methods being introduced [14,18]. Although to alleviate these issues, ontological mapping and data validation frameworks have been developed, manual oversight and curation are commonly needed in order to provide a reinforcement of reliance upon datasets to support high-stakes clinical use cases. The other obstacle is privacy and safety. Although there is less risk of data being lost in case of federated learning and secure multi-party because of decentralized data aggregation, it creates more computation overheads and needs strong governance structures so that the data is not leaked in case of federated updates [19,26]. In addition, it is claimed there is a barrier to interpretability of NLP-driven extraction where clinicians tend to insist on transparent and auditable results, especially when the results are used in making life-changing decisions or in regulatory submission [20,27].

Lastly, the sizeable number of computations required in handling petabyte-scale information, and particularly petabyte-scale data with complex models, can demand even cloud-based systems. On the one hand, optimization methodologies minimize costs, but on the other hand, there is still a demand for energy-efficient processing solutions, especially as the idea of sustainability increases its popularity in the world of healthcare [28]. The challenges mentioned show that there is room to improve further through innovation, and that is what will be discussed in the following section as we look into the future of scalable clinical data pipelines.

IX. FUTURE DIRECTIONS AND RESEARCH OPPORTUNITIES

The advancement of scalable clinical data integration pipelines will be further enhanced via the increased and a few upcoming lines of research. Multimodal learning, i.e., learning with concurrent processing of text, structured data, and imaging or genomic inputs, is a significant frontier. Pipelines that natively support the combination of so many and contrastingly varied modalities will make it possible to include in the patient representation much more information, and consequently, it will increase predictive value as well as support the clinical decision [24, 30]. Orchestration based on AI is also expected to grow in complexity, and cost-aware autoscaling agents will use real-time pricing, workload criticality measurements, and predictive scheduling to make on-the-fly decisions about the allocation of resources [28]. In addition to cost-efficiency, such systems will provide resiliency to automatically redistribute workloads when there is a regional outage or a disruption with a cloud provider, based on the learning in multi-region failover benchmarking work that quantifies the effects of downtime on clinical workflows [30].

The opportunities for cross-institutional cooperation will be widened as well, thanks to the development of privacy-preserving analytics. Advanced federated learning schemes, in conjunction with homomorphic encryption schemes, will enable training of more complex models on decentralized data distributions without having to share sensitive patient data, resolving one of the fundamental challenges to the deployment of multi-institutional pipelines [26]. Lastly, to ensure both the financial and environmental sustainability of large-scale clinical NLP and analytics, specialized architectures based on energy-efficient models, e.g., sparse and distilled transformers, will minimize their environmental and financial footprint and make pipeline scalability sustainable for both production and sustainability goals [25]. These innovations hold the potential to change the way structured and unstructured clinical data are brought together, so that in the future, all healthcare systems may benefit on a large scale with the information richness of these innovations, at the same time maintaining cost-effectiveness, security, and clinical trustworthiness.

REFERENCES

- [1] Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. *J Biomed Inform.* 2018;77:34–49.
- [2] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med.* 2018;1(1):18.
- [3] Solares JRA, Raimondi FED, Zhu Y, Rahimian F, Canoy D, Tran J, et al. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J Biomed Inform.* 2020;101:103337.
- [4] Gibson J, Atkins DC, Creed TA, Imel Z, Georgiou P, Narayanan S. Multi-label multi-task deep learning for behavioral coding. *IEEE Trans Affect Comput.* 2019;13(1):508–18.
- [5] Banerjee I, Chen MC, Lungren MP, Rubin DL. Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort. *J Biomed Inform.* 2018;77:11–20.
- [6] Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform.* 2008;17(1):128–44.
- [7] Kalusivalingam AK, Sharma A, Patel N, Singh V. Leveraging BERT and LSTM for enhanced natural language processing in clinical data analysis. *Int J AI ML.* 2021;2(3).
- [8] Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data.* 2023;10(1):1.
- [9] Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent advances in clinical natural language processing in support of semantic analysis. *Yearb Med Inform.* 2015;24(1):183–93.
- [10] Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Med Inform Decis Mak.* 2020;20(1):280.
- [11] Kumar S, Rani S, Sharma S, Min H. Multimodality fusion aspects of medical diagnosis: A comprehensive review. *Bioengineering.* 2024;11(12):1233.
- [12] Chava K. Harnessing Artificial Intelligence and Big Data for Transformative Healthcare Delivery. *Int J Recent Innov Trends Comput Commun.* 2022;10(12):502–20.
- [13] Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *NPJ Digit Med.* 2020;3(1):119.
- [14] Kumar G, Basri S, Imam AA, Khowaja SA, Capretz LF, Balogun AO. Data harmonization for heterogeneous datasets: a systematic literature review. *Appl Sci.* 2021;11(17):8275.
- [15] Neamatullah I, Douglass MM, Lehman LWH, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak.* 2008;8(1):32.
- [16] Soltanmohammadi E, Hikmet N. Optimizing healthcare big data processing with containerized PySpark and parallel computing: a study on ETL pipeline efficiency. *J Data Anal Inf Process.* 2024;12(4):544–65.
- [17] Liu H, Carini S, Chen Z, Hey SP, Sim I, Weng C. Ontology-based categorization of clinical studies by their conditions. *J Biomed Inform.* 2022;135:104235.
- [18] Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Med Inform Decis Mak.* 2020;20(1):280.
- [19] Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep.* 2020;10(1):12598.
- [20] Wang B, Xie Q, Pei J, Chen Z, Tiwari P, Li Z, Fu J. Pre-trained language models in biomedical domain: A systematic survey. *ACM Comput Surv.* 2023;56(3):1–52.
- [21] Gao X, He P, Zhou Y, Qin X. Artificial intelligence applications in smart healthcare: a survey. *Future Internet.* 2024;16(9):308.
- [22] Boda VVR, Allam H. Scaling Kubernetes for Healthcare: Real Lessons from the Field. *Int J Emerg Res Eng Technol.* 2023;4(3):27–34.
- [23] Tahir A, Chen F, Khan HU, Ming Z, Ahmad A, Nazir S, et al. A systematic review on cloud storage mechanisms concerning e-healthcare systems. *Sensors.* 2020;20(18):5392.
- [24] Ardıc N, Dinc R. Emerging trends in multi-modal artificial intelligence for clinical decision support: A narrative review. *Health Inform J.* 2025;31(3):14604582251366141.
- [25] Kichuk T, Avalos JL. Shape Matters: The Utility and Analysis of Altered Yeast Mitochondrial Morphology in Health, Disease, and Biotechnology. *Int J Mol Sci.* 2025;26(5):2152.
- [26] Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep.* 2020;10(1):12598.
- [27] Asif M, Ur Rehman F, Rashid Z, Hussain A, Mirza A, Qureshi WS. An Insight on the Timely Diagnosis of Diabetic Retinopathy using Traditional and AI-Driven Approaches. *IEEE Access.* 2025.
- [28] Dabral S, Mohana R. Healthcare Data Pipeline.
- [29] Paiva B, Gonçalves MA, da Rocha LCD, Marcolino MS, Lana FCB, Souza-Silva MVR, et al. A New Natural Language Processing-Inspired Methodology (Detection, Initial Characterization, and Semantic Characterization) to Investigate Temporal Shifts (Drifts) in Health Care Data: Quantitative Study. *JMIR Med Inform.* 2024;12(1):e54246.
- [30] Somarapu SK. Cross-Region Consistency Models for Geo-Distributed Event Streams. 2025.