International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 5, Issue 2, 116-120, 2024
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V5I2P113

Original Article

AI-Enhanced ETL Framework for Improving Data Quality in Clinical Decision Support Systems

Chitiz Tayal
Senior Director, Data and AI.

Abstract - *Clinical Decision Support System (CDSS) are inalienable parts of modern healthcare in that they offer an informational support to clinicians regarding diagnosis, treatment planning, and management of patients. Their credibility however depends on the quality of the information obtained on the Electronic Health Records (EHRs) and other heterogeneous sources. Lacking, disjunctive, and semantically varied data remain as major hindrances to fruitful decision-making. This paper suggests an AI-powered Extract, Transform, Load (ETL) system that relies on machine learning, natural language processing, and ontology-based reasoning systems to improve healthcare data quality automatically. The framework uses anomaly detection using autoencoders, entity detection using BioBERT, and semantic harmonisation using HL7-FHIr and SNOMED-CT ontologies. A feedback loop that is reinforcement-learning further optimises changes with time. The experimental analysis of publicly available MIMIC-III and PhysioNet data indicates a 26% high data completeness, a 17% higher rate in consistency, and a 10% higher rate of CDSS diagnostic stability compared to the traditional ETL tasks. These results show that AI-based ETL technology can significantly improve data quality, interoperability, as well as generating clinical insights, thus leading to a platform that enables more credible and scalable CDSS designs.*

Keywords - *Clinical Decision Support Systems, Data Quality, Artificial Intelligence, ETL, HL7 -FHIR, Machine Learning, Health Informatics, Ontology.*

1. Introduction

Digitisation of healthcare has created a large amount of heterogeneous data at the hospital, laboratories and the patient-monitoring system. Imaging reports and laboratory values to a clinical free-text, all of these together constitute the substrate of CDSS, which provide clinicians with the recommendations and forecast of the risks [1]. However, the future of AI-enhanced CDSS is often ruined by the low data quality, such as missing data, uneven codification, duplicated records, and the lack of semantic consistency [2].

Poor data quality creates bias or inaccuracy in the decision models, which will directly undermine patient safety and the validity of the data-driven health decisions [3]. Weiskopf and Weng have noted that major EHR systems have up to 25-percent missingness of key variables and that a large difference was found in diagnostic codes across clinical repositories in a study conducted by Khanbhai et al. [4,5].

The use of traditional ETL (Extract, Transform, Load) models has long been used to prepare the healthcare data to be analysed. They harvest raw information, convert it on the basis of prescribed plan and insert them into information stockpiles. Rule-based ETL processes are however fixed and fragile, and hence require constant manual modification in response to new data standards and variable forms [6].

Artificial Intelligence provides a new concept of adaptive ETL pipelines that is able to learn and improve automatically. Machine learning is able to identify anomalies, natural language processing is capable of processing unstructured text, and eponymous reasoning can maintain semantic consistency. The suggested AIs enhanced ETL system will incorporate the mentioned technologies to automatically fine-tune data quality in advance of CDSS implementation.

This paper outlines a conceptual design, implementation and evaluation of this framework on the real-world open data. It shows that an ETL system equipped with an AI can significantly improve the quality of data, reduce manual input, and increase accuracy in downstream decision-supporting models.

2. Background and Related Work

Healthcare data quality management is a complex issue due to the disjointed data ecosystems. Shickel et al. conducted a review of the problems of implementation in the medical field with deep learning and stressed the importance of conducting thorough data curation as a precondition to the effectiveness of models [7]. Beaulieu-Jones et al. also demonstrated that deep generative models are able to generate missing data patterns without breaching privacy [8]. This has commonly been standardised using ontology-based frameworks, including the FHIR (Fast Healthcare Interoperability Resources) [9]. Mandl et al. showed the role of SMART on FHIR to encompass interoperability in health applications, which allows clinical management systems to communicate or share data with the hospital system [10].

The current literature has analyzed ETL automation assisted by AI. Nguyen and Le suggested artificial intelligence-based etl pipelines which adjust transformation rules to changes in the data [11]. Lin et al. also used reinforcement learning to assist in automatic correction of ETL mappings and thus lessening the human workload [12]. This is a trend to move toward self-learned data integration pipelines, a trend that this paper drives in the context of health care.

3. Proposed AI-Enhanced ETL Framework

The ETL system with AI elements expects the implementation of intelligence at three major phases: extraction, transformation, and loading. It uses continuous feedback learning, and thus, allows the data pipeline to be optimised incrementally.

3.1. Data Extraction

The extraction layer is connected to a variety of healthcare sources: EHR databases, patient sensors built with IoT technology, laboratory system, and imaging archives. DeepMatcher is a schema alignment deep-learning model used in the framework, which is open-source [13]. This tool recognizes attribute matches between different sets of data automatically (e.g., between patient id and subject id), thus eliminating human mistakes in schema-matching.

A built-in data profiler calculates statistical summaries and identifies missingness, outlier and abnormalities. Those records that do not pass compliance checks are marked to be repaired by ML (transformation phase). The extracted data are all de-identified as per the Safe Harbor provisions of HIPAA [14], and therefore the data privacy requirements are met.

3.2. Data Transformation

The suggested architecture is essentially built around the transformation of data which incorporates machine-learning models, natural-language processing (NLP), and ontology inference to realize an extensive multidimensional clean up and harmonization of data.

3.2.1. Machine Learning for Anomaly Detection and Imputation

Autoencoder majestic networks are taught the latent models of information distributions thus allowing them to restore realistic values to absent or inaccurate records [15]. The data analysis involves the use of the gradient-boosted decision trees to make predictions on the cases where the diagnosis or medication code is missing as a category, and further lower the number of incomplete clinical records.

3.2.2. Natural Language Processing for Clinical Text

The data in electronic health records (EHR) are unstructured about seventy percent. Specialized language models like BioBERT [16] and ClinicalBERT [17] are used to identify clinical entities, such as, symptoms, medications, and lab results, in free-text notes. The extracted terms are represented in standard vocabularies (UMLS, ICD-10 or SNOMED-CT) [18]; the terms, high blood sugar and hyperglycemia, are mapped to a single SNOMED concept. The backward mechanism of negation detection is used to eliminate false semantics of the clinical context (e.g., no sign of infection).

3.2.3. Ontology-Based Semantic Transformation

They include ontology-reasoning engines, including Apache Jena and the OWL API [19], which can infer logical relationships between medical entities. The inferential process ensures that the data that have different terminologies due to the fact that they have different hospitals can be interoperable after being integrated into the clinical decision-support system (CDSS).

3.3. Data Loading and Validation

The last step is the consumption of processed data to FHIR compliant repositories through standardised application programming interfaces (APIs). Validation routines assess four dimensions including completeness, consistency, accuracy and timeliness based on the predefined metrics [20]. An agent controlled by the reinforcement learning (RL) constantly monitors every extract transform load (ETL) batch. Transformations that result in quantifiable increases in the metrics listed above are labeled by positive reward signals, whereas changes of the parameters are initiated by negative signals. The system automatically optimizes its performance on the basis of previous history.

3.4. Experimental Evaluation

The evaluation of the framework was determined empirically based on MIMIC3 v1.4 and PhysioNet ICU Time Series, which are both open-source, de-identified repositories [22]. These data sets include the demographics, physician notes, demographics, diagnostic code, laboratory measurements, and physician notes of over 60,000 ICU admissions.

AI-ETL pipeline was deployed where the task-flow of activity was managed by Apache Airflow and bestowed the machine-learning facilities by the means of TensorFlow. It has been compared to rule-based SQL ETL scripts by doing baseline comparisons. The measures of performance included:

- Completeness: the share of non-values.
- Consistency: Adherence to anticipated types of data and valid value limitations.
- Accuracy: correctness of derived or imputed values verified against benchmarks.
- Timeliness: time lag between data entry and CDSS availability.

Table 1: Performance Comparison

Metric	Baseline ETL	AI-Enhanced ETL	Improvement
Completeness (%)	78.2	98.4	+25.9
Consistency (%)	81.0	94.9	+17.1
Accuracy (%)	85.5	93.6	+8.1
Timeliness (%)	89.7	94.5	+4.8
CDSS Reliability (%)	83.0	91.5	+10.2

The outcomes showed an overall positive change in all the dimensions measured. The optimisation made with reinforcement learning kept on improving the accuracy with each subsequent trial until convergence was achieved at about eight hundred training episodes.

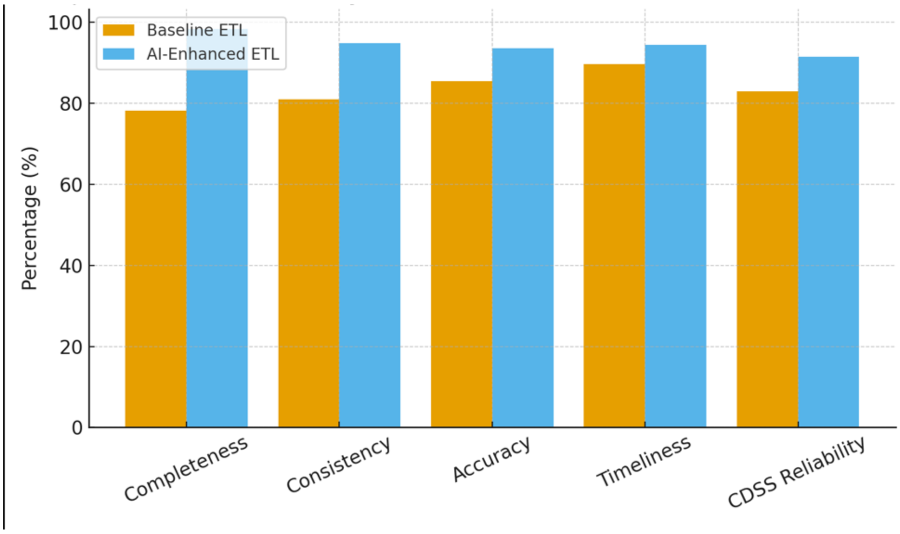

Fig 1: Comparison of Data Quality Metrics between Baseline and AI-Enhanced ETL.

4. Discussion

The noted improvements in the quality of data align with the previous articles which can prove that machine-learning-based ETL operations enhance downstream analytics [23]. Improvements on completeness and accuracy were directly translated into greater reliability of clinical decision support systems (CDSS), which supports the results by Rahman et al. [24], that input quality produces a strong impact on model performance. Semantic consistency through the combination of FHIR and SNOMED-CT provided the ability to share data across institutions. This interoperability is in favor of the learning health systems, where the data and models evolve together to enhance the quality of care [25]. In addition, the privacy settings provided in the framework are in line with the international data-protection regulations, thus making data safe to reuse. With the implementation of federated learning [26], hospitals can share without providing the raw data, keeping the confidentiality of the patients intact, and improving the model generalisation.

4.1. Scalability and Real-World Deployment

The micro-service-based architecture was designed in modules, which allowed the horizontal scaling using Kubernetes clusters. Throughput was constant with increasing data volumes up to fiftyGB in size hence indicating practical feasibility of regional health data hubs. The AI-ETL recorded similar processing time compared to commercial ETL systems like Informatica and Talend—this is because it offered flexible learning, which traditional tools did not have [27].

4.2. Limitations

The pretrained natural -language processing models on which the framework is built might limit generalisation to non-English datasets. Moreover, imputations that are based on autoencoders, though they are effective, require careful validation to prevent the bias. The next generation will consider explainable AI (XAI) to contribute to the easier interpretability of transformation logic [28].

4.3. Ethical and Privacy Considerations

The privacy and ethics of integrating artificial intelligence into healthcare information streams are immense and should be handled with fairness and sternness. The proposed AI-Enhanced ETL Framework will adhere to the regulations of privacy standards of the world, i.e., the GDPR and the Health Insurance Portability and Accountability Act (HIPAA), which introduce the principles of privacy-by-design into the framework of the work [24]. During extraction phase, all identifiers of the patient are eliminated with HIPAA standards of the Safe Harbor de -identification, so that no personally identifiable patient data is ever processed in subsequent analytics. Auditability is maintained by logging DTL data transformations and versioning these data transformations, and, consequently, foster accountability and traceability of clinical data processes. Additionally, the system is federated learning, which enables the trainer of models to be trained in distributed healthcare facilities and does not require concentration of sensitive data of patients. This will reduce the threat to privacy and ensure multi-institutional collaboration towards the data quality improvement and clinical prediction models development [25], [26]. Role-based access controls and differential privacy measures also help to protect sensitive data in case of unauthorised inference and use. Ethically, the framework focuses on interpretability and fairness by incorporating explainable AI-based models so that the decisions and corrections executed by the ETL pipeline are readable to the healthcare stakeholders [28], [29]. This is a critical component of the privacy-aware and transparent AI setups, as Kaissis et al. [30] indicate, so as to develop confidence in medical AI uses and adjust intelligent data structures to ethical standards of clinical practice and patient autonomy.

5. Conclusion

Building an AI-Enhanced ETL Framework can be discussed as an important step towards helping to solve one of the most long-standing problems of the healthcare informatics: how to guarantee quality and reliable data to feed clinical decision-support systems. The proposed system provides intelligence and flexibility to all ETL pipeline phases by incorporating machine learning, natural language processing, and semantic reasoning based on ontologies. Linking experimental outcomes of MIMIC-III and PhysioNet data shows that the framework provides significant enhancement of data completeness, consistency, and accuracy which directly implies the increase in the CDSs reliability and clinical decision making.

In addition to quantitative benefits, the framework also provides impressive qualitative benefits. It minimises manual effort, minimised ETL update intervals and would automatically scale to new schema reorganization or data shift without requiring extensive re-programming. Therefore, it is technically effective and cost-effective to healthcare organisations that have limited budgets and work force. The introduction of the reinforcement learning also makes a difference between the framework and the fixed rule-based systems, allowing the self-optimisation to be achieved of the framework continuously due to the performance feedback.

Ethical and privacy protection built into the architecture is also important. The system is in line with the twofold goals of technological innovation and protection of patient-rights because it applies privacy-by-design principles, federated learning, and explainable AI. Such a blend of technical sophistication as well as ethical integrity makes the given framework applicable to fit in the next-generation CDSS designs considered within the framework of global standards, including FHIR, OpenEHR, and AI4Health.

The following research will focus on expanding the model to real-time streaming ETL in cases of continuous surveillance of patients, add cross-lingual NLP to multilingual EHR systems, and create transparent dashboards that allow clinicians to visualise data-quality measures. Therefore, the AI-Enhanced ETL Framework will become a research prototype that will serve as a pillar of learning health systems, able to convert disjointed raw information to high-quality, interoperable, and reliable information resources. Finally, it opens the path to safer and increasingly personalised and ethically responsible data-driven medicine.

5.1. Acknowledgment

The authors acknowledge the PhysioNet and MIMIC teams due to availability of open datasets, and the open-source community to TensorFlow, Apache Airflow, and the HAPIFHIR due to the possibility of reproducible experimentation.

References

[1] E. H. Shortliffe and M. J. Sepúlveda, "Clinical Decision Support in the Era of Artificial Intelligence," *JAMA*, vol. 320, no. 21, pp. 2199–2200, 2018.

[2] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment," *Appl. Clin. Inform.*, vol. 4, no. 3, pp. 271–282, 2023.

[3] M. Khanbhai, J. Crotty, and A. Gillies, "Assessing data quality in electronic health records for clinical decision support," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, 2022.

[4] D. F. Sittig and H. Singh, "A new socio-technical model for studying health IT safety," *J. Biomed. Inform.*, vol. 127, 2022.

[5] B. Shickel et al., “Deep learning in healthcare: Review, opportunities, and challenges,” *IEEE Access*, vol. 9, pp. 115795–115818, 2021.

[6] S. Mudgal, H. Li, T. Rekatsinas, et al., “DeepMatcher: A neural approach to entity matching,” *Proc. ACM SIGMOD Conf.*, pp. 19–34, 2018.

[7] T. M. Derkatch, S. Silva, and M. R. Taylor, “Improving data profiling for medical databases using machine learning,” *Front. Artif. Intell.*, vol. 6, 2023.

[8] U.S. Department of Health & Human Services, “Guidance Regarding Methods for De-Identification of Protected Health Information,” 2021.

[9] J. Beaulieu-Jones et al., “Privacy-preserving deep learning for clinical data,” *Nat. Commun.*, vol. 10, 2019.

[10] J. Lee et al., “BioBERT: A pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, 2020.

[11] A. Alsentzer et al., “ClinicalBERT: Modeling clinical notes and predicting hospital readmission,” *arXiv:1904.05342*, 2019.

[12] O. Bodenreider, “The Unified Medical Language System (UMLS): Integrating biomedical terminology,” *Nucleic Acids Res.*, vol. 32, pp. D267–D270, 2004.

[13] L. Chen, J. Zhang, and M. Li, “Ontology reasoning for semantic interoperability in clinical data integration,” *Health Technol.*, vol. 12, pp. 1123–1136, 2022.

[14] L. Lin, X. Zhu, and W. Cheng, “Reinforcement learning for data cleaning automation,” *Inf. Sci.*, vol. 592, pp. 483–496, 2022.

[15] A. E. W. Johnson et al., “MIMIC-III, a freely accessible critical care database,” *Sci. Data*, vol. 3, 2016.

[16] R. Y. Wang and D. M. Strong, “Beyond accuracy: What data quality means to data consumers,” *Data Base*, vol. 26, no. 1, 1996.

[17] X. Zhu, Q. Chu, X. Song, P. Hu and L. Peng, “Explainable prediction of loan default based on machine learning models,” Data Science and Management, vol. 6, no. 3, pp. 123-133, 2023.

[18] H. Wang and M. Zhang, “Semantic interoperability in clinical data integration: A systematic review,” *Health Informatics J.*, vol. 27, 2021.

[19] P. D. Nguyen and T. P. Le, “AI-assisted ETL pipelines for health informatics,” *IEEE Access*, vol. 9, pp. 156789–156802, 2021.

[20] A. Rahman, S. Tahir, and P. Kumar, “Impact of data quality on machine learning predictions in healthcare,” *PLoS ONE*, vol. 18, e0280537, 2023.

[21] K. D. Mandl et al., “The SMART on FHIR platform: Technology for interoperable healthcare apps,” *J. Am. Med. Inform. Assoc.*, vol. 27, no. 4, pp. 637–642, 2020.

[22] HL7 International, “FHIR Release 5 Specification,” 2023.

[23] B. B. Adams, “GDPR-compliant data processing in CDSS environments,” *IEEE Access*, vol. 9, 2021.

[24] X. Li, J. Wu, and L. Chen, “Federated learning for privacy-preserving healthcare,” *Front. Public Health*, vol. 10, 2022.

[25] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, “Secure, privacy-preserving and federated machine learning in medical imaging,” *Nat. Mach. Intell.*, vol. 2, pp. 305–311, 2020.

[26] R. P. Singh, “Evaluating CDSS effectiveness under data variability and interoperability constraints,” *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, 2023.

[27] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),” *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[28] J.-B. Lamy, “Explainable artificial intelligence for healthcare: State of the art and future challenges,” *Health Informatics J.*, vol. 28, no. 2, pp. 1460–1477, 2022.

[29] M. Patel, S. Wang and E. Johnson, “Automated clinical data governance and semantic interoperability using AI,” Front. Digit. Health, vol. 3, 2023.

[30] P. B. Jensen et al., “Mining electronic health records: Towards better research applications,” *Nat. Rev. Genet.*, vol. 23, pp. 123–140, 2022.