International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(Review Article)

Check for updates

Hybrid Approaches for NER in Noisy OCR Medical Records

FNU Sudhakar Abhijeet *

Northeastern University, Boston.

International Journal of Science and Research Archive, 2025, 16(03), 082-089

Publication history: Received on 20 July 2025; revised on 26 August 2025; accepted on 30 August 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.16.3.2499

Abstract

Named entity recognition of digitized medical records accessed through Optical Character Recognition (OCR) poses significant problems, since the character distortions, uneven formatting, and domain-specific acronyms render it quite difficult. Such artifacts worsen the quality of rule-based or machine learning models that do not perform well under such noisy conditions by retaining consistent entity extraction. The use of hybrid techniques, i.e., the combination of deterministic rule-based modules with neural convolutional models like transformer-based models, is a stable remedy to these problems. Hybrid systems show better tolerance to OCR-induced noise by combining lexicon-based rules with contextual embeddings and error correction mechanisms, and ensemble strategies to maximize precision and achieve higher recall in clinical entity extraction (diagnoses, medications, and time-related entities). The piece is an analysis of the issues related to processing OCR-generated medical text, the implementation and development of hybrid NER pipelines, their institution-agnostic scalability, and research directions, such as multimodal learning, self-supervised pretraining on noisy data, and the orchestration of large-scale healthcare systems with the help of AI.

Keywords: OCR Medical Records; Named Entity Recognition; Hybrid NLP Approaches; Clinical Text Processing; Noise-Robust Models

1. Introduction

The use of medical records digitized through Optical Character Recognition (OCR) poses specific challenges to the specific kind of natural language processing (NLP) named Language Context (NER) of application. OCR-based clinical documents are also known to be full of artifacts like incorrect recognition of characters, uneven line spacing, line and paragraph formatting mistakes, and inconsistent use of abbreviations, which are counterproductive to improved accuracy of automatic text comprehension. Although the NLP has made an abundance of progress through deep learning, the traditional methods continue to cede ground to deep learning in keeping entity extractions consistent, especially under noisy conditions that are characteristic of scanned records, handwritten notes, and other source materials that are of poor quality [1-3]. It is essential that such records can be mined to capture structured information (e.g., diagnoses, medications, procedures, and temporal expressions) and such information can be used in downstream applications such as clinical decision support, billing, and population health analytics [4, 5]. To overcome such limitations hybrid methodology of combining the strengths of rule-based techniques with the flexibility of machine learning and deep neural networks is a possible solution [6-8].

Hybrid NER systems take advantage of the compensatory power of the symbolic and the statistical paradigms to overcome OCR-related follies. In contrast, rule-based systems work well when the relationships they need to be aware of are deterministic, be it specific drug dosage patterns, laboratory values format, or an organizational structure section title, whereas machine learning models and especially transformer-based training frameworks can be used to gain better contextual insight on illegible and inconsistent text blocks [9-10]. These hybrid methods have enhanced resistance to character-level noise by applying handcrafted rules, domain-specialised lexicons, and context-based

* Corresponding author: FNU Sudhakar Abhijeet

embeddings, and the resulting ability to generalise adaptively to a wide range of clinical texts. In addition, error correction modules and confidence scoring systems also help to increase reliability in extractions and minimize OCR errors being carried downstream [11-13].

This paper focuses on how hybrid NER systems have grown, their architecture, and their effective application toward noisy medical records OCR. It starts with a discussion on the inherent challenges of the data that is generated by the OCR and goes on to discuss the design principles of hybrid models, assesses how well they perform in clinical information extraction, and finally discusses future research directions that would improve the scalability, robustness, and integration of systems into the larger healthcare systems. The various sections develop the previous discussion to provide a coherent story starting with the assumptions that form the reasons why the hybridization should occur in the first place, all the way to the operational strategies and innovations.

2. Challenges of NER in OCR-Derived Clinical Documents

Reliable NER of OCR-processed clinical records is a rather difficult task due to an intricate combination of linguistic idiosyncrasies, noise artifacts, and variability occurring within the domain. Commonly, OCR engines especially when working with handwritten notes, poor-quality scans or old documents, add character-level errors like replacement of one letter with another one ("0" for "O", "1" for "l"), authors do not know where to split the line, and they merge and split tokens [14,15]. This will cause problems with tokenization, a key initial process of most NLP pipelines, and cause compounded errors down the chain in other tasks, such as entity boundary detection and classification.

Also, terminology is not usually standardized in clinical records; you will tend to find inconsistent use of terminology, high amounts of abbreviations, and inconsistent formatting. As an example, the dose formulations may differ extensively (e.g., 5mg bid, 5 milligrams twice daily), and though abbreviated or shortened notations or symbols commonly used by clinicians may additionally obfuscate creation of automated parsers [16, 17]. Ambiguity through context is also a challenging barrier; objects like lead could be a chemical element, an ECG electrode, or even a verb, which has to be interpreted through contextual means. OCR noise helps make these complexities worse because it may corrupt not just surface forms but also context clues, and corrupts the performance of purely statistical or purely rule-based models [18, 19].

This mix of syntactic variability, semantic ambiguity, and text degradation highlights why single-approach methods necessitate additional adequate approaches to be used in the NER task of OCR. To overcome the above, alternating architecture combos capable of integrating the pattern-recognition strength of rules and context sensitivity of deep learning are becoming more and more necessary. In the next section, instead of focusing on addressing such underlying challenges individually, we focus on how hybrid methods are gradually taking shape to solve these challenges, which is through unifications between structured heuristics and advanced neural-based architectures.

3. Design Principles of Hybrid NER Architectures

Hybrid NER systems are constructed based on the hypothesis that deterministic rules and machine learning models cover two types of issues in noisy environments, namely complementary aspects of the entity extraction task. Rule based components take advantage of lexicons, regular expressions and ontology-based mappings that can capture predictable instances of things like numeric lab values, date patterns and standardized section headings ("Impression", "Diagnosis") that are fairly semantically stable in the presence of OCR artifacts [20, 21] The rules achieve a high degree of precision particularly with well-formed or formulaic instances and form a basis of narrowing down the search space of statistical procedures.

Machine learning components, in turn, like the transformer-based models with the most prominent examples being BERT derivatives pretrained on biomedical corpora, are characterized by strong processing of irregular or context-sensitive entities, such as medications, symptoms, or procedures, the surface forms of which differ by a broad margin. Contextual embeddings used in these models permit the meaning of a text to be inferred despite character-level distortions because the self-attention mechanisms can down-weight noisy tokens and up-weight useful context [22-24]. Ensemble mechanisms usually combine the results of rule-based and neural elements by applying voting schemes or by weighting output based on the exuded confidence with the aim of achieving maximum recall, without compromising their precision [25].

There are also error correction modules to build a hybrid architecture. Recognition errors may be normalized by convolutional or recurrent modelling at the character level or by training transformer decoders in noisy synthetic OCR, prior to NER processing. To alleviate the possibility of tracking unravelled errors to downstream clinical uses, confidence calibration mechanisms detect low-respectability entities to be reviewed or re-certified by a human operator [26, 27]. These deficiencies of one paradigm or another can be surmounted through these architectural strategies that help the hybrid systems to address the drawbacks of one paradigm or another. After discussing their design, it is time to discuss their evaluation against hybrid ones in practical clinical NER challenges, where one is interested in their comparative output and robustness against monolithic options.

In addition to the architecture description of hybrid systems, it is also helpful to compare the behaviours of the rule-based, neural, and hybrid systems and their performance on various OCR-induced distortions. The table below makes the strengths and weaknesses of their performances in noise-heavy clinical settings stand out

Table 1 Comparative Strengths of Rule-Based, Neural, and Hybrid NER Approaches in Noisy OCR Clinical Text

NER Approach	Strength in OCR Context	Weakness in OCR Context	Best Use Case in Clinical Records
Rule-Based Systems	High precision for fixed formats (dates, lab values) and standardized headers	Poor recall for irregular phrasing; brittle to unseen abbreviations or artifacts	Extracting lab values, units, and structured sections in scanned reports
Neural Models (Transformers)	Robust contextual inference; tolerates moderate distortion through embeddings and self-attention	Susceptible to cascading errors when tokenization fails; can overfit to artifact patterns	Identifying variable entities like diagnoses, symptoms, and drug mentions
Hybrid Architectures	Combines precision of rules with adaptability of neural models; supports ensemble weighting	Higher computational and maintenance overhead; requires orchestration	Comprehensive extraction in noisy multi-institutional datasets

This comparative perspective clarifies why hybrid architectures, despite their complexity, are uniquely suited for OCR-derived medical records. These strengths become evident when evaluating their real-world performance, as discussed in the next section.

4. Evaluating Hybrid Approaches in Clinical Information Extraction

The best way to view the performance of hybrid NER systems is how well they perform on a variety of OCR-created clinical datasets, where precision and recall are both important to subsequent utility, as demonstrated in Figure 1. Benchmarks have always shown that hybrid systems have a much better performance than rule-based or neural systems alone, especially in conditions of intense noise and differently structured documents [28, 29]. Rule-based modules help here, as with predictable accuracy they identify standardized entities, whereas the neural components allow better recall as they extract semantically rich but structurally varied (and usually distorted by OCR) entities like symptoms and diagnoses [30]. The generalizability of hybrid frameworks is also better. Applied to collections of records composed of datasets of varying contexts (records scanned in different institutions with vastly different scanning technologies, OCR engines, and documentation practices) hybrid models perform relatively consistently, as opposed to standalone neural models driven to overfit to patterns and specific artifacts, and rule-based models where the limited ability to generalize fails when faced with unseen types of abbreviations or layouts [23-25]. Notably, the ensemble approaches applied to the hybrid architectures enable regular balancing of the input contributions of all parts, dependent on the nature of documents, enhancing the stability of the output quality.

In spite of those benefits, hybrid methods have high computational and maintenance overhead as compared to simple models. The rules have to be periodically updated in order to consider emergent terminologies, and the neural elements consume a lot of resources during training and inference. These trade-offs are, nevertheless, explicable through the critical advances in reliability and elasticity taken in difficult OCR environments. Now that the effectiveness of these systems has been established, it has been the discussion of how these systems can be implemented on a large scale, as it pertains to deployment approaches and how these systems may be integrated within a clinical setting.

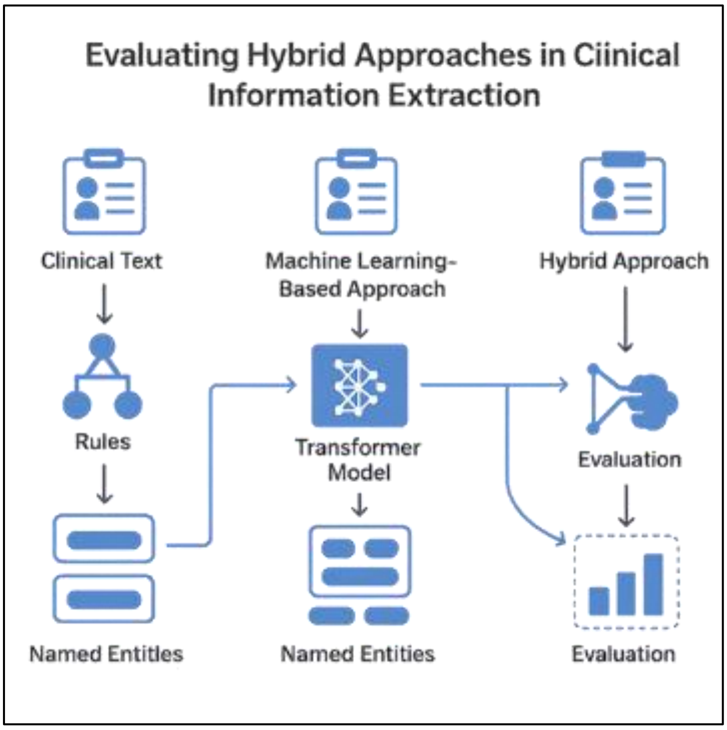

Figure 1 Evaluation of hybrid approaches in clinical information extraction, illustrating the comparison between rule-based, machine learning-based, and combined hybrid pipelines, leading to named entity recognition and performance evaluation

5. Deployment of Hybrid NER Pipelines in Clinical Environments

The implementation of hybrid NER systems in production to enable OCR-based clinical record reading would imply a pipeline capable of dealing with the scale, as well as the over- and under-representation of the data in the healthcare domain. These pipelines generally start off with preprocessing layers that normalize OCR text by fixing typical artifacts with character-level models, as well as dictionary-based substitutions. After normalization, the text is tokenized and is subjected to parallel processing of rule-based and machine learning modules. Rules quickly recognize a predictable pattern, and neural networks concentrate on entities that are contextually ambiguous or those that are semantically rich. The results of these two elements are combined with ensemble algorithms to create a cohesive set of labelled entities that can be used by the rest of the system [20,23].

Such pipelines have been constructed on distributed computing systems like Apache Spark or Flink to maintain performance on large throughput hospital systems and support batch and stream processing of millions of documents. Kubernetes managed containerized deployments offer the flexibility to scale easily, with heavier components (such as transformer models) on GPU-equipped nodes, and lighter ones (the rules) on CPU clusters [24, 25]. Interoperability interfaces to electronic health record (EHR) systems may be accomplished using standards like FHIR APIs, and structured output from the NER pipeline can directly feed clinical decision support and coding systems or analytics dashboards [21,26].

Such deployments help simplify workflows in business operations as well as increase interoperability between structured and unstructured data flows. Since hybrid systems are gradually expanded to most, or even all, university departments or institutions, fault tolerance and elasticity come to the forefront, of which one naturally speaks about how these pipelines are designed (or would need to be) concerning resilience and multi-site scalability.

6. Scalability and Multi-Institutional Integration

Scaling hybrid NER pipelines to support regional health networks or research consortia introduces significant architectural considerations, as shown in Figure 2. OCR-derived records vary widely by institution due to differences in scanning technology, document templates, and OCR engines, making standardization a primary challenge. To address this, federated learning frameworks allow machine learning components to train collaboratively across distributed datasets without centralizing sensitive patient data. Neural models are updated using gradient sharing, while rules remain localized, customized to each site's unique artifact patterns [19, 26]. Elastic scaling mechanisms ensure responsiveness under fluctuating workloads. Autoscaling clusters dynamically allocate resources based on real-time processing demand, using metrics such as queue depth and throughput to trigger scaling events. Horizontal scaling allows parallel document processing across nodes, while vertical scaling provisions GPU-accelerated nodes for transformer inference when entity recognition workloads spike [25, 28]. Multi-cloud or hybrid-cloud deployments distribute workloads across providers, leveraging vendor-specific strengths, such as AI accelerators or regional compliance certifications, while failover strategies ensure uninterrupted service during localized outages [23]. By embedding these scalability principles, hybrid NER pipelines can reliably support both clinical operations and large-scale research initiatives across heterogeneous environments. Yet scalability alone cannot mitigate the impact of OCR-specific distortions on data quality, which necessitates a deeper exploration of specialized error correction mechanisms.

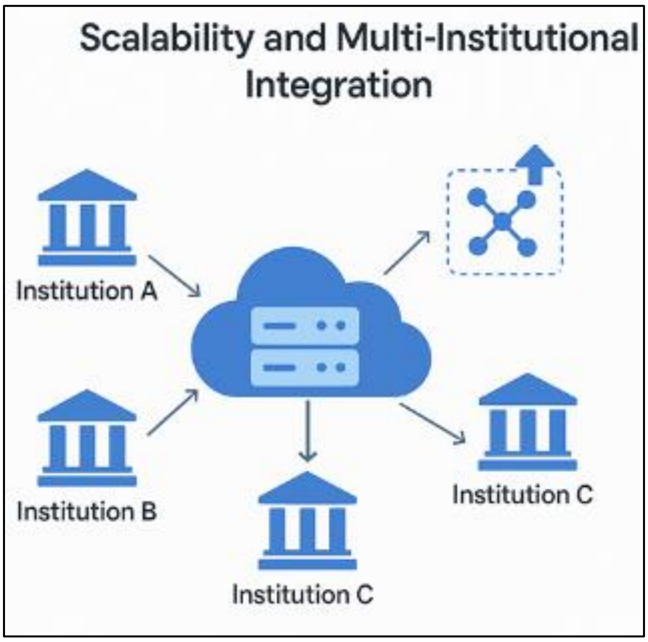

Figure 2 Scalability and multi-institutional integration, depicting a centralized cloud-based system connecting multiple institutions to enable seamless data sharing and scalable analytics

7. Error Correction and Noise-Robust Entity Recognition

Correction of mistakes lies at the heart of ensuring consistent results of NER: due to the very characteristics of OCR-generated medical language, it is noisy. The hybrid pipelines combine several noise reduction layers where the text normalization modules, which employ character-level convolutional or transformer decoders trained on synthetic noise, address typical recognition errors, most commonly errors in the reading of the letters slurred together nowhere, denomination ("1" vs. "l"), and omission ("0" vs. O "zero"), but also gaps in punctuation. They have no downstream impact on rule-based or neural components, as they standardize the inputs before tokenization [16, 26]. Not just normalization, correction strategies of hybrid systems happen in a context-dependent way. With the attention mechanisms used in neural networks, likely corrections can be deduced using the context around the error positions and neutralizing noisy tokens, and enhancing pure signals when disambiguating entity boundaries. Confidence score is also involved with manually reviewed extractions/secondary validation against knowledge bases and lexicons, given to the low confidence checks [17, 29]. Rules aid these processes by imposing structure uniformity, such as requiring the units to be valid in a laboratory value or requiring drug doses to follow an expected series of numerals [20]. The introduction of such layered correction mechanisms is not only more accurate when it comes to extraction but also increases the downstream trustworthiness by making sure structured outputs do not actually transfer OCR artifacts into clinical processes. Now that error correction has been covered, the remaining section will review the support of security, compliance, and observability in healthcare-grade deployments via the routes of these pipelines.

Although the hybrid systems have combined the neural and symbolic modules, error correction layers are important in increasing their robustness. As shown in the table below, the significant noise-mitigation methods and their effects or quantitative effects on the downstream entity recognition performance were summarized.

Table 2 Impact of Noise-Correction Layers on NER Performance in OCR-Derived Clinical Text

Noise-Correction Technique	Description	Typical Performance Gain (F1 Score)	Practical Considerations for Deployment
Character-Level Transformer Decoders	Pretrained to normalize OCR-specific artifacts before tokenization	+8–12%	Requires synthetic noise datasets for training
Context-Aware Confidence Calibration	Flags low-confidence entities for review or secondary validation	+4–6%	Needs integration with manual or semi-automated review workflows
Lexicon-Guided Normalization	Matches and corrects distorted tokens against curated clinical lexicons	+5–7%	Periodic updates to lexicons are needed as terms evolve
Ensemble Consistency Checking	Cross-validates outputs from rules and neural models to reduce false positives	+6–10%	Adds computational cost but enhances trustworthiness

These layered strategies not only elevate extraction accuracy but also ensure that OCR-induced artifacts do not propagate into downstream analytics, reinforcing the pipeline's reliability before compliance and governance considerations are addressed in the following section.

8. Compliance, Observability, and Trust in Hybrid Systems

When used in the public health domain, hybrid NER pipeline solutions would come subject to strict regulations and security procedures, as well as high transparency among end-users. All processing elements support encryption both at rest and in motion, and access to such data was based on a role-based policy that limited data manipulation to authorized individuals. De-identification modules strip or obscure personally identifiable information (PII) before data use to train or redistribute data between institutional boundaries of HIPAA, GDPR, and other regional privacy laws [21, 22]. This observability is inculcated through telemetry structures that monitor the latency, throughput, drift on models, and policy adherence across the stages of the pipeline. Dashboards offer operational visibility, and automated scanners are used in round-the-clock verification of system settings with regard to healthcare security baselines. In another example, a policy drift identification will detect the violation, as in the case of unauthorized privileges within the NLP processing containers, to generate a rollback and alert administrators of this possible security threat [29]. Clinicians are also extended transparency since, in many cases, they need justifications regarding automated extractions when results are used in diagnostic or billing processes. Explainability tools show the weight of the model's attention and point out the rule triggers, which enable the user to audit how certain entities were identified. All these measures are enough to maintain trust in hybrid NER systems, and nonetheless, some issues have not been solved yet to a great extent, which will be discussed in the next section.

9. Challenges and Research Frontiers

Hybrid NER systems of OCR-based medical text have some remaining issues despite their promise. The reliance on hand-made rules is rife and requires upkeep as medical terminologies, abbreviations, and documentation procedures change. Whereas machine learning parts acquire new patterns as a whole during fine-tuning, rules have to be later updated manually to maintain complete coverage, and this becomes a long-term maintenance cost [17,20]. The other stumbling block is the computational and energy requirements of transformer models, especially in cases where the models have to be implemented on a large scale and in multi-institutional networks. Whereas the methods of model compression, including pruning and quantization of the model, mitigate some of these issues, sparse and low-rank transformer structures should be further explored to balance accuracy with minimization of the computing cost [25]. Moreover, the problem of interpretability is persistent, and despite the existing explainability modules, a lower level at which ML connections find applicable clinical usage remains present, restricting end-user faith [18, 27]. Lastly, there are no common benchmarks that can be used to evaluate noisy OCR clinical datasets objectively across competing systems comparatively. Although there are some synthetic noise datasets, they are usually far less diverse than reality, and there is a need to share and release de-identified benchmark corpora on which to evaluate the heterogeneity of healthcare documentation [15, 23]. Challenges give the opportunity to be creative, and the second section covers that by citing the rising research directions that might find a solution to these obstacles.

10. Future Directions and Opportunities

The development of hybrid NER systems on noisy OCR medical records is on the verge of gaining pace with a number of research and technology advancements. An improvement in entity recognition through the cross-checking of visual and contextual information through multimodal learning pipelines, which are able to ingest OCR text, related imaging, and structured metadata concurrently, would consequently cut down on how much the system looks at noisy textual data. A second promising trend is the use of self-supervised pretraining on synthetic OCR noise data so that neural models can derive their own character-level noise invariance without the need for many manually annotated examples. This pretraining may be complemented with agents of reinforcement learning, which adaptively adjust the rule-based and neural parts balance depending on the features of a document, maximizing the precision and recall in real time. At the operational level, cost-restricted orchestration and price-sensitive scaling, which is feasible with AI, will play a vital role in the process of scaling up hybrid NER pipelines. Agents that combine aspects of predictive autoscaling learn the historical rate of processing and pattern prices based on historical patterns of processing, and may optimally assign and decommission computations across large-scale research and clinical workloads. On top of this, multi-region failover benchmarking will be done so that the systems ensure continuity in the event of cloud or network interruptions, a feature indispensable to health sector activities. Hybrid NER architectures will also become the norm of OCR-derived clinical document processing as these innovations mature, allowing highly disorganized medical data of an analogue origin to flow into structured digital information ecosystems vital to the demands of contemporary healthcare analytics and decision support.

References

[1] Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., ... & Liu, H. (2018). Clinical information extraction applications: a literature review. *Journal of Biomedical Informatics*, 77, 34–49.

[2] Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1), 18.

[3] Solares, J. R. A., Raimondi, F. E. D., Zhu, Y., Rahimian, F., Canoy, D., Tran, J., ... & Salimi-Khorshidi, G. (2020). Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of Biomedical Informatics*, 101, 103337.

[4] Cui, S., & Mitra, P. (2024). Automated multi-task learning for joint disease prediction on electronic health records. *Advances in Neural Information Processing Systems*, 37, 129187–129208.

[5] Banerjee, I., Chen, M. C., Lungren, M. P., & Rubin, D. L. (2018). Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort. *Journal of Biomedical Informatics*, 77, 11–20.

[6] Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, 17(01), 128–144.

[7] Kalusivalingam, A. K., Sharma, A., Patel, N., & Singh, V. (2021). Leveraging BERT and LSTM for enhanced natural language processing in clinical data analysis. *International Journal of AI and ML*, 2(3).

[8] Chiu, C. C., Wu, C. M., Chien, T. N., Kao, L. J., Li, C., & Chu, C. M. (2023). Integrating structured and unstructured EHR data for predicting mortality by machine learning and the latent Dirichlet allocation method. *International Journal of Environmental Research and Public Health*, 20(5), 4340.

[9] Velupillai, S., Mowery, D., South, B. R., Kvist, M., & Dalianis, H. (2015). Recent advances in clinical natural language processing in support of semantic analysis. *Yearbook of Medical Informatics*, 24(01), 183–193.

[10] Tsanousa, A., Bektsis, E., Kyriakopoulos, C., González, A. G., Leturiondo, U., Gialampoukidis, I., ... & Kompatsiaris, I. (2022). A review of multisensor data fusion solutions in smart manufacturing: Systems and trends. *Sensors*, 22(5), 1734.

[11] Chava, K. (2022). Harnessing Artificial Intelligence and Big Data for Transformative Healthcare Delivery. *International Journal on Recent and Innovation Trends in Computing and Communication*, 10(12), 502–520.

[12] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1), 119.

[13] Sun, H., Depraetere, K., De Roo, J., Mels, G., De Vloed, B., Twagirumukiza, M., & Colaert, D. (2015). Semantic processing of EHR data for clinical research. *Journal of Biomedical Informatics*, 58, 247–259.

[14] Neamatullah, I., Douglass, M. M., Lehman, L. W. H., Reisner, A., Villarroel, M., Long, W. J., ... & Clifford, G. D. (2008). Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1), 32.

[15] Soltanmohammadi, E., & Hikmet, N. (2024). Optimizing healthcare big data processing with containerized PySpark and parallel computing: a study on ETL pipeline efficiency. *Journal of Data Analysis and Information Processing*, 12(4), 544–565.

[16] Liu, H., Carini, S., Chen, Z., Hey, S. P., Sim, I., & Weng, C. (2022). Ontology-based categorization of clinical studies by their conditions. *Journal of Biomedical Informatics*, 135, 104235.

[17] Zhang, X., Wang, Y., Jiang, Y., Pacella, C. B., & Zhang, W. (2024). Integrating structured and unstructured data for predicting emergency severity: an association and predictive study using transformer-based natural language processing models. *BMC Medical Informatics and Decision Making*, 24(1), 372.

[18] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., ... & Bakas, S. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1), 12598.

[19] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240.

[20] Imran, S., Mahmood, T., Morshed, A., & Sellis, T. (2020). Big data analytics in healthcare− A systematic literature review and roadmap for practical implementation. *IEEE/CAA Journal of Automatica Sinica*, 8(1), 1–22.

[21] Boda, V. V. R., & Immaneni, J. (2021). Healthcare in the Fast Lane: How Kubernetes and Microservices Are Making It Happen. *International Journal of Emerging Research in Engineering and Technology*, 2(3), 33–42.

[22] Tahir, A., Chen, F., Khan, H. U., Ming, Z., Ahmad, A., Nazir, S., & Shafiq, M. (2020). A systematic review on cloud storage mechanisms concerning e-healthcare systems. *Sensors*, 20(18), 5392.

[23] Gaw, N., Schwedt, T. J., Chong, C. D., Wu, T., & Li, J. (2018). A clinical decision support system using multi-modality imaging data for disease diagnosis. *IISE Transactions on Healthcare Systems Engineering*, 8(1), 36–46.

[24] Fan, Z., Yan, Z., & Wen, S. (2023). Deep learning and artificial intelligence in sustainability: a review of SDGs, renewable energy, and environmental health. *Sustainability*, 15(18), 13493.

[25] Teo, Z. L., Jin, L., Liu, N., Li, S., Miao, D., Zhang, X., ... & Ting, D. S. W. (2024). Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. *Cell Reports Medicine*, *5*(2).

[26] Wu, Z., Dadu, A., Tustison, N., Avants, B., Nalls, M., Sun, J., & Faghri, F. (2024). Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*.

[27] Dabral, S., & Mohana, R. (2023). Healthcare Data Pipeline.

[28] Paiva, B., Gonçalves, M. A., da Rocha, L. C. D., Marcolino, M. S., Lana, F. C. B., Souza-Silva, M. V. R., ... & Morais, V. C. (2024). 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 Medical Informatics*, 12(1), e54246.

[29] Dasari, K. K. (2023). Cross-Cloud Continuity: A Scalable Framework for Resilient and Regulated Digital Infrastructure.

[30] Shurrab, S., & Duwairi, R. (2022). Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8, e1045.