International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Optimizing Healthcare Data Extraction With Ai: A Path To Improved Patient Care

Arun Kumar Ramachandran Sumangala Devi

Architect II- Software Testing UST Global Inc, Glen Allen, Virginia, USA
akumarrs@gmail.com

ABSTRACT

This study analyses the role of AI in enhancing the healthcare system by improving data extraction techniques to address various challenges in the healthcare sector. The roles of artificial intelligence (AI), like natural language processing (NLP) and machine learning (ML) in healthcare, allow the automation of the extraction of unstructured clinical information from records. This results in quicker diagnosis, enhanced treatment outcomes, and reduced paperwork. It has allowed for reduced burnout among healthcare professionals who would otherwise extract this data manually, allowing them to focus on providing better quality care to patients. While AI has been implemented in this sector, much must be done to achieve its full potential.

KEYWORDS: Data extraction. Artificial Intelligence (AI), patient care

1. INTRODUCTION

Healthcare systems worldwide face significant challenges in meeting four key goals of healthcare provision: promoting population health, increasing patient treatment outcomes, enriching caregiver experiences, and reducing overall healthcare expenditures. However, achieving these goals presents a challenge due to the increasing number of patients being treated daily. This means that a large amount of data is being generated daily about various conditions [1,2]. Gathering and processing of this information processing interferes and costs a lot of money for the healthcare systems [3]. Another challenge is working with unstructured textual information in EHRs and how patient data can be protected and kept confidential simultaneously.

The use of artificial intelligence (AI) in the health sector has excellent potential in the future to transform medicine. Some of the AI techniques that are being used in healthcare are machine learning and Natural Language Processing (Figure 1). Machine Learning (ML) is concerned with pattern analysis and is used to interpret ECGs, diagnose medical images and analyze patient risk [4,3].

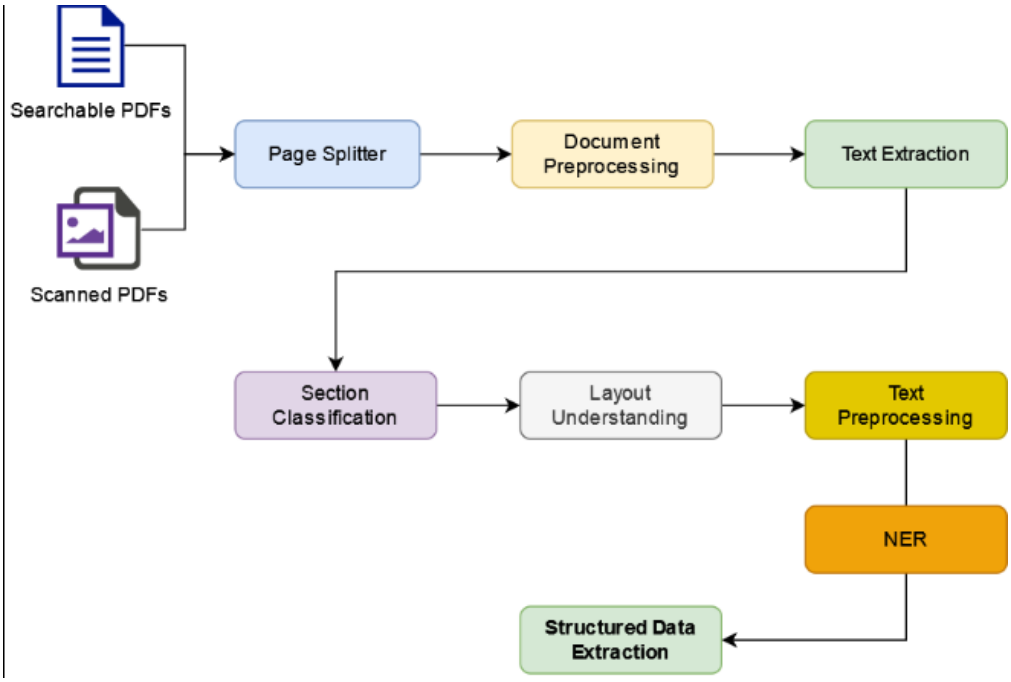

Figure 1: An example of how patient names are extracted using Named Entity Recognition (NER), a type of NLP that is used to identify and classify patient names in unstructured text [7]

NLP applies ML techniques to enable computers to understand and communicate with human language. These two, among other techniques, are being used to optimize data extraction in healthcare and address the challenges related to the manual extraction of large volumes of data. Automated methods based on AI enable rapid, accurate, and completely anonymous data extraction for further processing [5]. This study aims to contribute to existing knowledge on how AI is optimizing data extraction techniques in healthcare and processing large volumes of raw and unstructured clinical data from various sources in the healthcare sector. The following research questions guide the study: i) How does AI improve the efficiency of healthcare data extraction? ii) How does optimized data extraction impact patient outcomes and care quality?

2. SOLUTION

This article is based on a comprehensive literature review of how artificial intelligence (AI) can optimize healthcare data extraction and improve patient care. The goal was to gather insights into how AI-driven systems are used for data extraction and what benefits they offer the healthcare system. Existing studies on AI tools, including natural language processing (NLP), machine learning, and deep learning, were examined to understand how these technologies are applied in extracting data from electronic health records (EHRs), medical images, and clinical notes. The study then focused on how AI-optimized data extraction techniques improved patient outcomes.

3. APPLICATIONS OF THE SOLUTION

3.1.1 The use of NLP for parsing clinical notes

While clinical notes are a rich source of medical data, the information available is underutilized as these notes entail subjective data most of the time. Prior to EHRs, data extraction from these notes was time-consuming, expensive, and cumbersome [6]. An example of how data is extracted using NLP is illustrated

in Figure 1 below. Despite the volume and continuing growth in the availability of healthcare data, [7] argued that more than 80% of text, images, signals, and other data continue to be unstructured and untapped. [8,9,10] identified that NLP can independently identify the family history of a patient from clinical unstructured textual content. They employed an ICD-10 diagnostic code and word embeddings in a Convolutional Neural Network (CNN). Compared to the traditional methods of data extraction, they got better results with almost no need for data preprocessing.

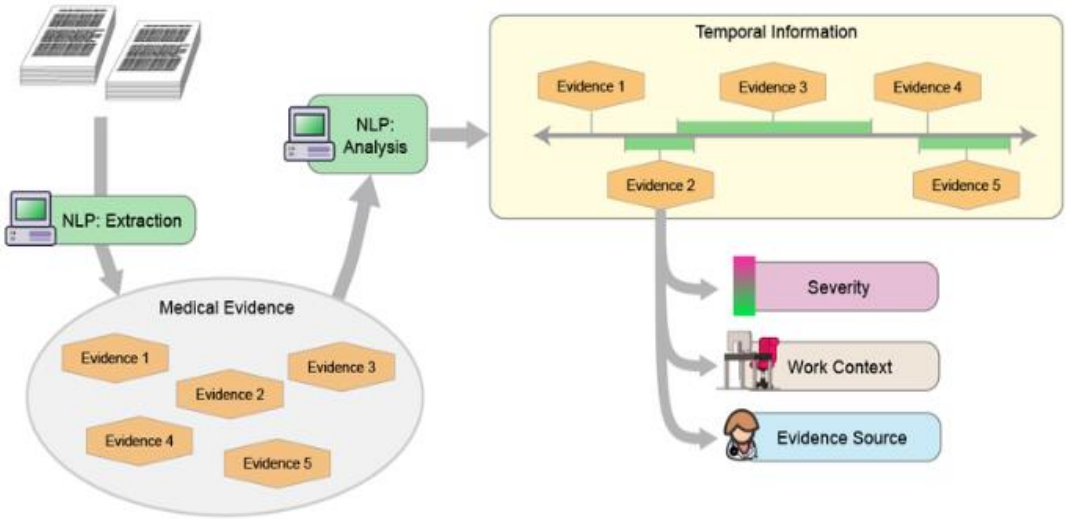

Figure 2: Process of data extraction using NLP

Various studies have shown that patients with both acute and chronic conditions could be diagnosed and treated through the application of NLP techniques. For example, [11] used annual per capita healthcare expenditures on pneumonia, influenza, acute bronchitis, and upper respiratory illnesses. [12] used early data pertaining to COVID-19 to build and validate a machine-learning decision tree model in order to identify actual positive RT-PCR tests. In their cross-sectional study, [13] used a combination of conventional epidemiology and NLP and ML predictive analysis to find features that would predict that COVID-19 patients would be admitted to the ICU. [14] applied a simple-tree XGBoost model to identify high-risk COVID-19 cases. [10] also applied NLP for the automated extraction of data from free text documents in patients' records, with particular attention to ischemic stroke patients who underwent reperfusion therapies. The outcomes demonstrated that NLP techniques enhanced efficiency for model learning for the vast majority of comorbidities.

3.1.2 Machine learning application in data extraction

Machine learning (ML) has been consistently applied to collect and analyze various forms of patient data from various sources. An example of feature extraction from a given dataset using ML is illustrated in Figure 2 below. [15] developed ML methods for pathological Complete Response (pCR) to neoadjuvant chemotherapy and survival probabilities in breast cancer using multiparametric magnetic resonance imaging (mpMRI) data. Using material from 38 female breast cancer patients, [15] ranked features associated with residual cancer burden (RCB), recurrence-free survival (RFS), and disease-specific survival (DSS) using eight classifiers, L-SVR, logistic regression, and XGBoost. Their results showed that XGBoost offered the best accuracy in RCB and DSS compared to any other models, while logistic regression offered the best results in RFS. In another study, [16] used a multilevel perception model to

estimate the survival rates of non-small cell lung cancer patients with two-year data collected from 559 patients. They used the RelifF feature selection method, and the multilayer neural network turned out to be the best prediction model, with an AUC of 0.75.

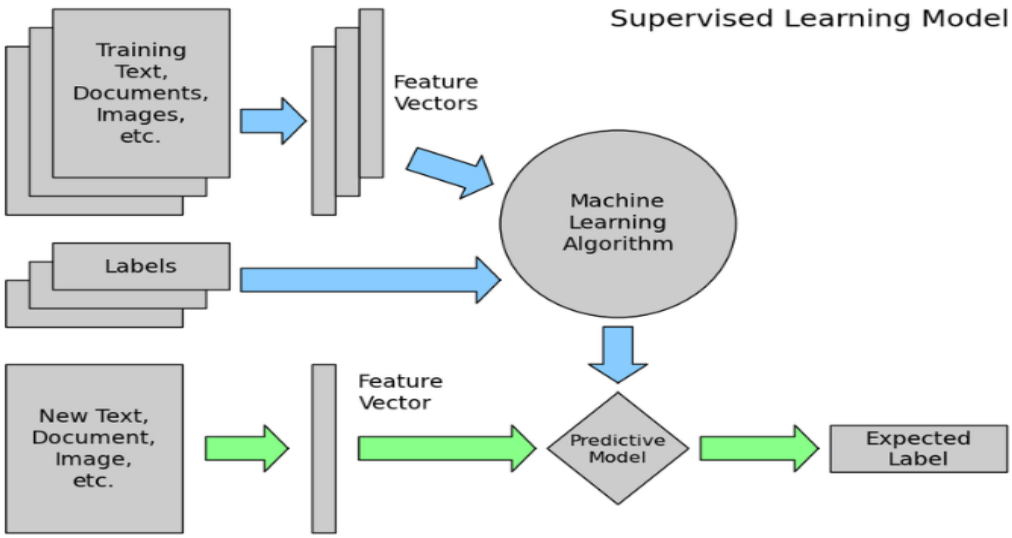

Figure 3: Feature extraction using ML

[17] proposed a framework to identify Type 2 Diabetes Mellitus (T2DM) patients from EHR data containing 300 samples. They extracted 114 features and used different classifiers, noting that Support Vector Machine (SVM) achieved the highest accuracy of 96%. [17] found that Support Vector Machine Recursive Feature Elimination (SVM-RFE) performed the highest in classifying the tumor as well as in grading the glioma with accuracies of 85% and 88%, respectively. In another study by [18] to predict high-risk surgical patients using various machine-learning techniques, ML techniques were trained on the Pythia dataset consisting of 194 clinical factors. From the analysis, [18] found that penalized logistic regression was the most effective, with an AUC of 0.924.

3.1.3 Optical Character Recognition (OCR)

Optical Character Recognition (OCR) technology transforms information from various documents—like patient forms, medical records, doctors' notes, prescription slips, and lab results—into machine-readable text. This technology helps computers edit, search, and process information more efficiently. It distinguishes characters from their backgrounds, recognizes the text, and compares it to databases to extract relevant data. OCR tools can capture and process data from documents in just 45-60 seconds and support batch processing, allowing multiple documents to be handled simultaneously, which saves valuable time [3]. This efficiency enables healthcare professionals to focus more on enhancing operational workflows and providing personalized patient care. Additionally, OCR delivers data extraction with over 95% accuracy, and advanced AI-powered solutions like Docsumo can achieve up to 99% accuracy [2].

4. BENEFITS OF THE SOLUTION

AI-based extraction techniques offer significant benefits that lead to improved quality of care provision.

4.1 Alleviate worker burnout

The healthcare industry faces a workforce shortage, leading to increased burnout among healthcare professionals who are often overwhelmed by administrative tasks, such as data entry. Traditionally,

unstructured data—like faxes, scanned images, and PDFs—requires manual input to be incorporated into electronic health records, adding to the workload of nurses and other staff members. AI can recognize key information—like patient demographics and diagnoses—from various documents, allowing quicker data processing without requiring extensive manual entry [19]. By using machine learning to understand and extract relevant fields from documents, AI reduces the time healthcare workers spend on administrative tasks, allowing them to focus more on patient care.

4.2. Improved diagnostic accuracy

Machine learning algorithms improve diagnostic speed and provide a deeper understanding of complex medical conditions by continuously learning from diverse databases. Moreover, AI and machine learning play a crucial role in disease prediction through predictive analytics, which examines historical and real-time healthcare data to identify disease patterns and risk factors [20]. For instance, in cardiovascular health, machine learning can evaluate a person's risk of heart disease by analyzing indicators like blood pressure and cholesterol levels.

4.3 Cost reduction

Investing in AI-based extraction techniques solutions can significantly reduce costs for healthcare organizations by automating repetitive tasks like data entry, which minimizes the need for manual labor. This automation streamlines operations and enhances accuracy, leading to error-free billing and claims processing that helps avoid financial losses. A prime example is Access Healthcare, which automated the processing of Explanation of Benefits (EOB) documents using Echopay's OCR engines. As a result, they achieved a remarkable 50% reduction in operational costs.

4.4 Compliance and Security

AI-powered data extraction enhances compliance and security by automating the processes of data extraction, validation, and compliance audits. These systems help ensure adherence to important regulations and laws, including HIPAA, GDPR, the HITECH Act, CCPA, and the 21st Century Cures Act. They also come equipped with standardized security features like access controls, cloud storage, and encryption, which protect sensitive patient information from unauthorized access [21]. This helps prevent fraud and mitigates the risk of legal issues that could lead to significant fines.

5. Conclusion

Healthcare practitioners can improve patient outcomes by utilizing cutting-edge technology like Optical Character Recognition (OCR), machine learning (ML), and Natural Language Processing (NLP) to handle data more accurately and efficiently. This study shows how AI-driven solutions can save operational costs, improve diagnostic accuracy, guarantee regulatory compliance, and lessen the administrative strain on healthcare staff. In the end, applying AI to optimize data extraction methods not only solves present healthcare issues but also opens the door to a more effective, patient-centered, and efficient method of providing healthcare.

References

1. Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., & Biancone, P. (2021). The role of artificial intelligence in healthcare: a structured literature review. BMC medical informatics and decision making, 21, 1-23.
2. Yanamala, A. K. Y. (2023). Data-driven and artificial intelligence (AI) approach for modelling and analyzing healthcare security practice: a systematic review. Revista de Inteligencia Artificial en Medi-

cina, 14(1), 54-83.
3. Panesar, A. (2019). Machine learning and AI for healthcare (pp. 1-73). Coventry, UK: Apress.
4. Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., ... & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Information Fusion, 96, 156-191.
5. Khan, Z. F., & Alotaibi, S. R. (2020). Applications of artificial intelligence and big data analytics in m-health: A healthcare system perspective. Journal of healthcare engineering, 2020(1), 8894694.
6. BuHamra, S. S., Almutairi, A. N., Buhamrah, A. K., Almadani, S. H., & Alibrahim, Y. A. (2022). An NLP tool for data extraction from electronic health records: COVID-19 mortalities and comorbidities. Frontiers in Public Health, 10, 1070870.
7. Kong, H. J. (2019). Managing unstructured big data in healthcare system. Healthcare informatics research, 25(1), 1-2.
8. John Lin, C. C., Yu, K., Hatcher, A., Huang, T. W., Lee, H. K., Carlson, J., ... & Deneen, B. (2017). Identification of diverse astrocyte populations and their malignant analogs. Nature neuroscience, 20(3), 396-405.
9. Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019). Natural language processing of clinical notes on chronic diseases: systematic review. JMIR medical informatics, 7(2), e12239.
10. Zhou, L., Lu, Y., Vitale, C. J., Mar, P. L., Chang, F., Dhopeshwarkar, N., & Rocha, R. A. (2014). Representation of information about family relatives as structured data in electronic health records. Applied clinical informatics, 5(02), 349-367.
11. DeCapprio, D., Gartner, J., McCall, C. J., Burgess, T., Kothari, S., & Sayed, S. (2020). Building a COVID-19 vulnerability index. MedRxiv, 1-12.
12. Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj digital medicine, 4(1), 1-5.
13. Izquierdo, J. L., Ancochea, J., Savana COVID-19 Research Group, & Soriano, J. B. (2020). Clinical characteristics and prognostic factors for intensive care unit admission of patients with COVID-19: retrospective study using machine learning and natural language processing. Journal of medical Internet research, 22(10), e21801.
14. Guan, X., Zhang, B., Fu, M., Li, M., Yuan, X., Zhu, Y., ... & Lu, Y. (2021). Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study. Annals of medicine, 53(1), 257-266.
15. Kanjo, E., Younis, E. M., & Ang, C. S. (2019). Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. Information Fusion, 49, 46-56.
16. Dagli, Y., Choksi, S., & Roy, S. (2019). Prediction of two year survival among patients of non-small cell lung cancer. In Computer Aided Intervention and Diagnostics in Clinical and Medical Images (pp. 169-177). Springer International Publishing.
17. Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., ... & Chen, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. International journal of medical informatics, 97, 120-127.
18. Bisaso, K. R., Anguzu, G. T., Karungi, S. A., Kiragga, A., & Castelnuovo, B. (2017). A survey of machine learning applications in HIV clinical research and care. Computers in biology and medicine, 91, 366-371

19. Hoppe, N., Härting, R. C., & Rahmel, A. (2023). Potential benefits of artificial intelligence in healthcare. In Artificial intelligence and machine learning for healthcare (pp. 225-249). Springer, Cham.
20. Moulaei, K., Yadegari, A., Baharestani, M., Farzanbakhsh, S., Sabet, B., & Afrash, M. R. (2024). Generative artificial intelligence in healthcare: A scoping review on benefits, challenges and applications. International Journal of Medical Informatics, 105474.
21. Kayal, C. K., Bagchi, S., Dhar, D., Maitra, T., & Chatterjee, S. (2019). Hepatocellular carcinoma survival prediction using deep neural network. In Proceedings of International Ethical Hacking Conference 2018: eHaCON 2018, Kolkata, India (pp. 349-358). Springer Singapore.