International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0471
e ISSN: 2584-2854
Volume: 02
Issue: 10 October 2024
Page No: 3186-3190

Data Integration and Predictive Analysis for Healthcare: Survey

Gaurav Choughule[1], Yash Sodaye[2], Tejas Shelke[3], Sahil Wani[4], Nilesh Bhelkar[5]

[1,2,3,4]UG Scholar, Dept. of AIDS, Rajiv Gandhi Institute of Tech., Andheri, Maharashtra, India.

[5]Assistant Professor, Dept. of AIDS, Rajiv Gandhi Institute of Tech., Andheri, Maharashtra, India.

Emails: *gauravchoughule7@gmail.com[1] , yashsodaye13@gmail.com[2] , tejas.shelkee@gmail.com[3] , sahilwani31103@gmail.com[4], nilesh.bhelkar@mctrgit.ac.in[5]*

Abstract

In modern healthcare, handling of medical data is often divided, with sensitive information stored in various formats in documents like prescriptions, lab reports and patient summaries. The lack of standardization and management of such data can lead to errors and inefficiency in patient care. Various techniques like HCR, OCR, machine learning algorithms are for the purpose of data integration and predictive analysis. The aim of this paper is to conduct the survey of various techniques used in healthcare systems and domain for data handling and performing operations like predictive analysis on that data. We have reviewed and shortlisted 7 papers and the methods used in those studies. This survey is beneficial to those who wish to manage and store data of healthcare and use the same for their respective purposes.

Keywords: *Data Integration, Predictive Analysis, Machine Learning, Healthcare Systems.*

1. Introduction

In today's healthcare systems, critical information from prescriptions, lab reports, and patient summaries often exists in disparate formats, making it difficult for healthcare providers to access and analyze data efficiently. It is necessary for appropriate data management and integration for most productive and appropriate use of it. And since medicine being such a vital sector, it is necessary to be precise and responsible while doing so. With the bloom of Artificial Intelligence & Machine Learning, their application in various sectors have widely increased. Machine learning methods are highly used for predictive analysis. Domains like construction, botany, education, social service & medicine are the ones where predictive analysis is frequently used [2]. Out of these, medicine or healthcare is the most prominent or vital field of application. For any machine learning methods to be applied, data storing, manipulation & management forms the core step. Data is received in large volume, varying structures and at a high velocity in these times. This is called as Big Data. Such unstructured, voluminous data is necessary to be handled appropriately and accurately. Especially in the field of medicine, where data exists in various formats like prescriptions, lab reports and patient summary, and in physical format, it is essential to administer this data carefully. Only after standardizing and digitizing the data into a uniform format, it can be used for predictive analysis or any other required purpose. This paper includes a study of 7 of other papers, where the process of data digitization and standardization is executed using various techniques and further, in few other papers, predictive analysis is applied on the processed data for required output, specifically focusing on healthcare sector. The rest of the paper deals with the literature survey, results and conclusion of the various techniques and methods used for data standardization & digitization and predictive analysis.

2. Methods

OCR: Optical Character Recognition (OCR) is a technique that creates machine-readable data or rather digitizes the same from text pictures,

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0471
e ISSN: 2584-2854
Volume: 02
Issue: 10 October 2024
Page No: 3186-3190

including scanned documents or photos. OCR is significant to the healthcare industry because it digitizes and makes searchable medical documents, including test results and prescriptions, into a standard format which is easier to handle. It extracts and standardizes text using methods like machine learning, character recognition, and picture preparation. However, normal OCR still has drawbacks, such as managing different handwriting styles. This can be overcome by using advanced OCR technique that support handwritten texts as well.

HCR: Handwritten Character Recognition (HCR) is a specialization of OCR that aims to turn handwritten text into machine-readable data. It uses complex machine learning algorithms and neural networks to recognize and understand different types of handwriting, and convert them into digitized format. HCR increases efficiency in healthcare by automating data entry from handwritten papers, but there are possibilities that the model may not be completely accurate because of variety of handwriting styles. However, with continual developments, HCR improves accuracy and integration in document digitalization operations.

Predictive Analysis: Predictive analysis is to use historical data and based on it use statistical machine learning methods to predict future outcomes. In context of healthcare sector, this technology can be used to detect the possible diseases or issues a patient might suffer in the future with respect to his previous health records and reports. However, good quality of data and precise models are essential for effective outputs of predictive analysis.

3. Related Work

In [1], Rathod and Sari (2020) have presented their design for digitization of health data. They have implemented Handwritten Character Recognition (HCR) for converting physical data like general medical records into digitized format. The objective was to automate the process of converting handwritten data into a digital format so that it can be stored and accessed easily in databases. The project successfully demonstrated the objective and using HCR they have improved efficiency and accuracy in managing health data. The authors in [2] have compiled brief research on predictive analysis using machine learning. The paper provides guidance to researchers interested in performing predictive analysis for choosing the most suitable machine learning method based on the field of application. Their research spreads across various fields such as medicine, social science, building, botany and education. In [3], the authors have proposed a system to integrate a 3D human body model for visualization of health data and to automate the process of digitizing paper-based medical records for enhancing the efficiency of managing an Electronic Health Record (EHR) system. The project utilizes the benefits of OCR for digitizing the medical health records. A success rate of 100% was achieved in digitizing medical records, tested over 200 medical reports. The researches in [4], Kadadi and Agrawal (2014), have analyzed the challenges of data integration and interoperability in big data. The objective is to explore the data integration techniques, address the complexity regarding big data and design an integration architecture to tackle these challenges. The authors propose potential solutions through various tools and technologies like Hadoop, KARMA and JNBridgePro for addressing these issues. In [5], authors have proposed and described a data model based on Fast Healthcare Interoperability Resource (FHIR) standard for an intelligent multimodal interface. They have demonstrated the application of FHIR standard for creating a multimodal user interface supporting production and research in context of medical applications. The FHIR-based architecture is insightful, including interaction between components like EEG and accelerometer data processing. In [6], Sinha and Jenckel (2019), authors have used Generative Adversarial Networks (GANs) for examining the OCR models in an unsupervised manner. The project demonstrates improved validation of OCR models by using generated text line images with the help of GANs which resemble the original input images from OCR output text, so that the model can be evaluated even when softmax layer is not accessible. The authors in [7], Saluja and Punjabi (2019), have developed a method to improve OCR error correction for Indic languages using sub-word embeddings to account

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0471
e ISSN: 2584-2854
Volume: 02
Issue: 10 October 2024
Page No: 3186-3190

for complex aspects of linguistic rules. The project uses sub-word embeddings, including n-grams to train a model capable of correcting OCR errors in Indic languages. The fastText-based model outperformed the baselines models in terms of F-scores and Word Error Rates (WER) demonstrating effectiveness in handling out-of-vocabulary words. All the above papers demonstrate the use variety of technologies like OCR, HCR, GANs, Hadoop, FHIR for data digitization & standardization and use of machine learning algorithms for predictive analysis.

4. Results AND DISCUSSION

The model proposed in [1] successfully demonstrated the ability to convert non-digitized data into digital format with the help of Handwritten Character Recognition (HCR). Figure 1 represents the final model proposed by the authors for data digitization.

However, HCR finds it difficult to deal with identification of characters in cursive script. This is the only flaw of the technique.

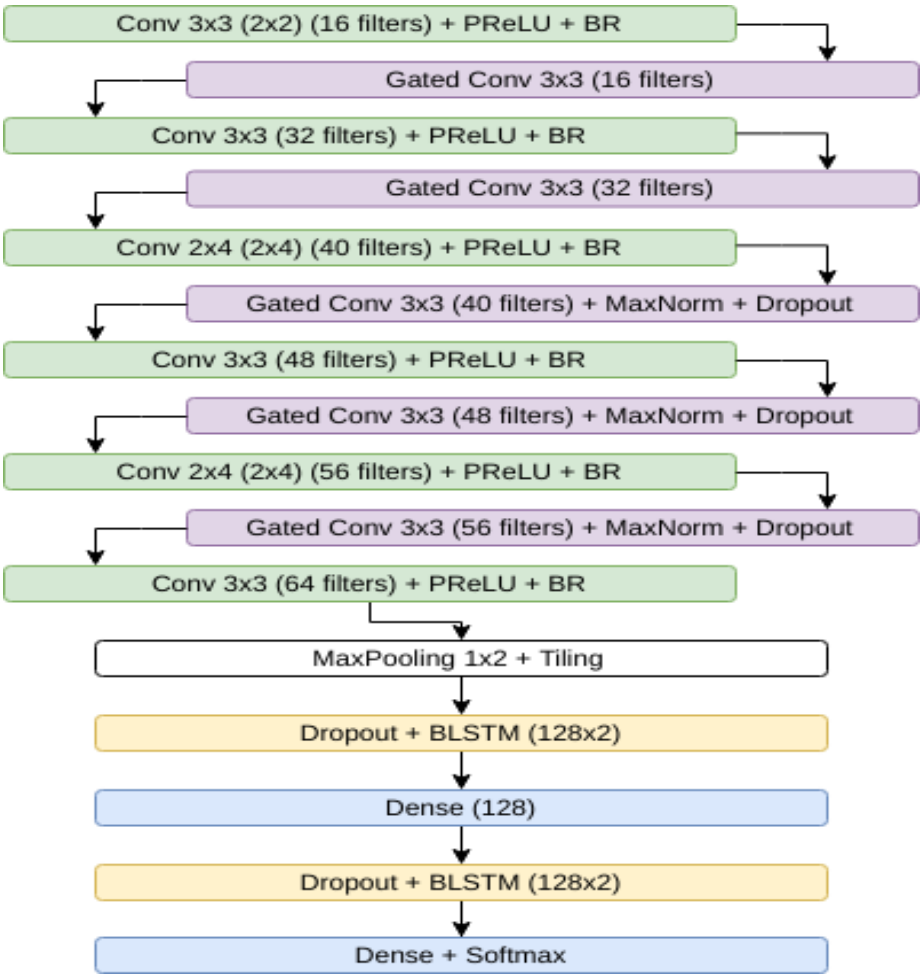

Figure 1 Model Architecture [1]

In [2], the conclusion was achieved that 56.7% application of Machine Learning methods in predictive analysis is in the field of medicine. Also, 70% of the methods used fall under supervised learning, among which, Random Forest (RF) is the most commonly used method. In [3], 100% success rate was achieved in digitizing over 200 medical reports. Figure 2 represents the design flow of the OCR model used by the authors. Although, the testing of the model was only done on Chinese-script medical records.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0471
e ISSN: 2584-2854
Volume: 02
Issue: 10 October 2024
Page No: 3186-3190

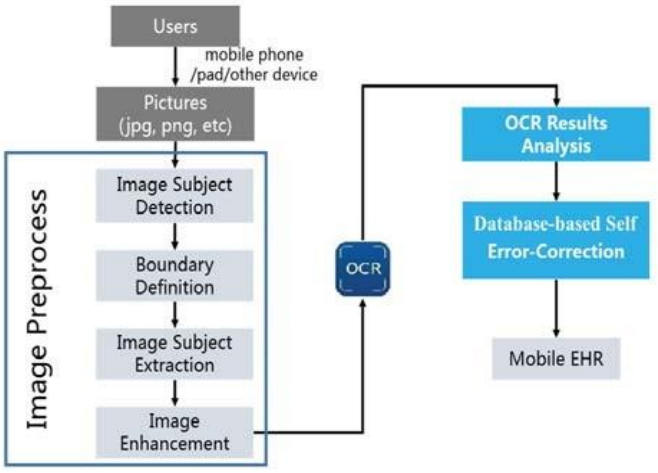

Figure 2 Design Flow for Auto Digitization [3]

In [4], key challenges in big data integration have been identified and further, using tools and methods like KARMA, JNBridgePro, Hadoop and distributed query optimizations has been suggested. Figure 3 represents the technique used by the authors to integrate various forms of data together. However, the effectiveness of the proposed solutions is uncertain due to lack of real-world case studies.

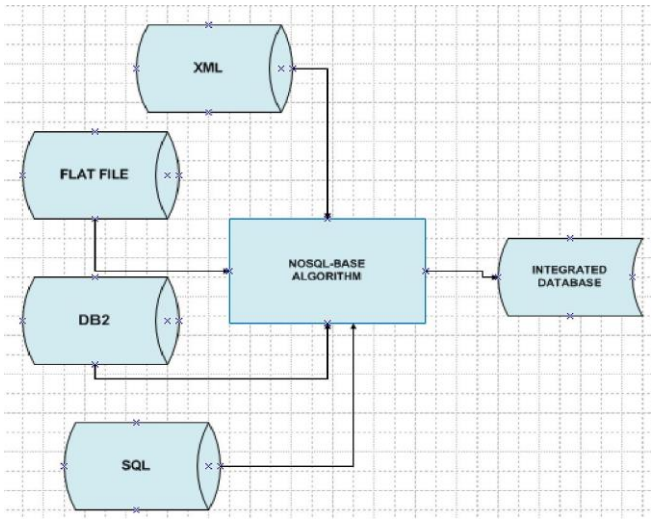

Figure 3 Integration of Different Types of Data Using NOSQL Algorithm [4]

In [5], the FHIR-based architecture is very detailed and insightful, which illustrates use cases and diagrams showcasing the functionality in both research and production modes. Figure 4 shows one of the use cases of the architecture proposed by the authors.

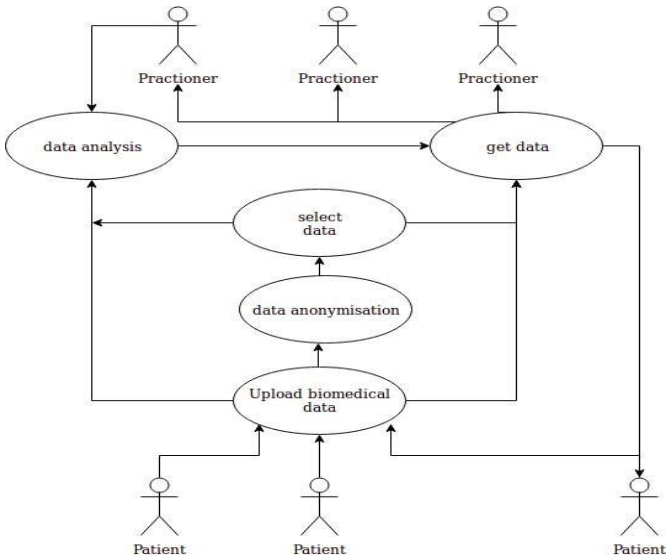

Figure 4 Use Case Diagram for Patient in Research Mode [5]

In [6], the use of GANs for generating fake images to evaluate the OCR model is very beneficial when softmax layer is not accessible. However, the generated images lacked details such as font style (e.g., bold or italic), which makes it difficult to compare these images with the original input when such styling is important. In [7], the proposed model outperformed various other baseline models in terms of F-scores and Word Error Rates (WER).

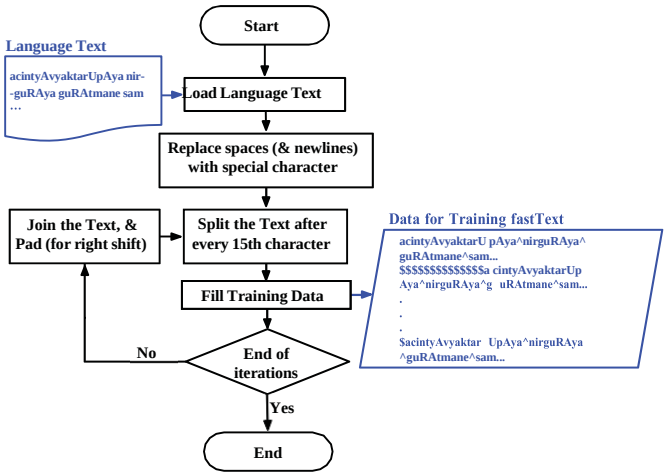

Figure 5 Flowchart for Transformation of Language Data [7]

Figure 5 is the flowchart of the system proposed by the authors for their model. Although, the results are particularly dependent upon availability of high-

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0471
e ISSN: 2584-2854
Volume: 02
Issue: 10 October 2024
Page No: 3186-3190

quality date. The method also shows increase in WER for some out-of-vocabulary words in Malayalam, due to differences between training and test sets.

Conclusion

This literature survey provides the compiled research of shortlisted 7 papers which consist of digitization and standardization of various forms of unstructured or physical data, especially in the context of healthcare sector and to further perform operations like predictive analysis on the same. On the basis of this research, we aim to create an Electronic Health Record (EHR) which will tackle the existing limitations in data standardization and further perform predictive analysis for the treatment of the patient.

References

[1]. Rathod & Sarita. (2020). Converting non-digitized health data to digital format. Asian Journal of Convergence in Technology, 6(1), 10-13.
[2]. Loola B., Khadija, O. T., & Souissi N. (2020). Predictive analysis using machine learning: Review of trends and methods. 2020 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), 1-6.
[3]. Liu, N., et al. (2020). A new data visualization and digitization method for building electronic health record. 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2980-2982.
[4]. Kadadi, A., Agrawal, R., Nyamful, C., & Atiq, R. (2014). Challenges of data integration and interoperability in big data. 2014 IEEE International Conference on Big Data (Big Data), 38-40.
[5]. Borisov, V., Minin, A., Basko, V., & Syskov, A. (2018). FHIR data model for intelligent multimodal interface. 2018 26th Telecommunications Forum (TELFOR), 420-425.
[6]. Sinha, A., Jenckel, M., Bukhari, S. S., & Dengel, A. (2019). Unsupervised OCR model evaluation using GAN. 2019 International Conference on Document Analysis and Recognition (ICDAR), 1256-1261.
[7]. Saluja, R., Punjabi, M., Carman, M., Ramakrishnan, G., & Chaudhuri, P. (2019). Sub-word embeddings for OCR corrections in highly fusional Indic languages. 2019 International Conference on Document Analysis and Recognition (ICDAR), 160-165.