International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 04 | April 2022 www.irjet.net p-ISSN: 2395-0072

Review on Medical Reports Extraction and Storage in EHR systems

Anushka Darade, Nandana Prabhu, Utsav Parekh, Tirth Thaker, Dhairya Umrania

[1]*Student, Dept. of Information Technology, K. J. Somaiya College of Engineering, Maharashtra, India*
[2]*Assistant Professor, Dept. of Information Technology, K. J. Somaiya College of Engineering, Maharashtra, India*
[3]*Student, Dept. of Information Technology, K. J. Somaiya College of Engineering, Maharashtra, India*
[4]*Student, Dept. of Information Technology, K. J. Somaiya College of Engineering, Maharashtra, India*
[5]*Student, Dept. of Information Technology, K. J. Somaiya College of Engineering, Maharashtra, India*

Abstract - *Medical reports can be difficult to manage, and they are also difficult to maintain in physical form. They suffer from physical conditions as well as improper storage, which can lead to unavailability of necessary information when it is most needed for the patients' well-being. Digitization of the health-related data can be a solution to above mentioned issues. In this paper, we have reviewed different tools, techniques and systems which can be used for data extraction from medical report images or Portable Document Format (PDF) documents, and then how that medical report data can be stored in databases.*

Key Words: Medical report data extraction, medical report data storage, OCR (Optical Character Recognition), EHR (Electronic Health Records) , EMR (Electronic Medical Records), PHR (Personal Health Records), semi-structured document extraction.

1. INTRODUCTION

Revolving around today's era, there is a lot of medical research surfacing new diseases, new symptoms as well as new medical reports. Layman faces several problems in maintaining their medical data efficiently and updating it time to time without loss of any reports. These reports are not only required for medical purposes but in the pandemic, such reports hold a lot of significance for daily activities. In such a situation, having an alternative to store medical reports, and related data eases several aspects for people. Some of the commonly noticed problems that have been discovered are:

- Frequent expensive consultations
- Loss of medical data due to inefficient maintenance
- Irregular updating of data
- Difficult access of data by patients and doctors
- Improper interpretation of report data
- Insecure storage of data

To tackle these issues we aim to build a system which can address all the points. The proposed system will require three main components: a data extraction component which can extract data from medical reports in the form of images or PDFs, a data storage component which can safely store the data extracted from the data extraction component, and a data analysis component, which can retrieve the data of the user and interpret the data of the reports for the user in a simple and convenient manner.

This paper is structured as follows: a literature review of methods to extract data from medical reports is conducted, which contain mainly two types of papers: papers covering medical data extraction from OCRs or PDFs and noting down their metrics and papers which have proposed systems. We have also prepared a dataset description from the papers that have used OCRs to extract data from medical reports. Next we have a review of medical data based database systems. Finally we cover the papers which include data analysis and visualizations of medical report data.

2. PAPERS BASED ON MEDICAL REPORT DATA EXTRACTION

In this section, we have reviewed papers which have researched Optical Character Recognition (OCR) on medical reports, and noted their results. First, we have the dataset description of all the papers which have used a medical report dataset. Then, each paper is mentioned with its method and approach and at the end of each paper, the results with the metrics used by that paper are given.

2.1 DATASET DESCRIPTION

For OCR purposes, datasets are difficult to procure and hence it can be difficult to train OCRs. Also since there are a vast number of image types where OCR is used, finding a dataset in the specific domain of your study can be elusive. The domain of our paper is Healthcare Reports, for which finding patient reports can prove to be a strenuous activity.

To obtain medical reports, one way to procure them is through pathology labs or hospitals, [8] have collected 100 medical reports from hospitals in China, [10], have collected sleep study reports from University of Texas Medical Branch, which has around 990 reports collected between 2015 and 2018 and [12] have collected cancer reports from a pathology lab, with annotations on the reports from the pathologists, it contains a total of 3632 reports.

But, a lot of times it is not possible to get medical report data from hospitals or labs because patient data is sensitive and needs to be confidential. Hence, most of the studies involving OCR for medical reports, have collected their personal medical reports from friends and family, such as [5] which have collected their personal medical reports and in total gathered 36 images, one of each of the following types: Physical Examination, Blood Chemistry, CBC, and Urinalysis. Another notable paper which has collected their own medical reports from friends and family is [6], which gathered a total of 119 reports, which came to be known as the Chinese Medical Document Dataset (CMDD).

Generally, to train and test OCRs, the more number of images there are, the better the performance will be, so to increase the dataset, many papers have performed preprocessing on their dataset, such as adding illuminations and rotations to each image, so that the OCR can be tested meticulously in several different conditions. This was done by [6], on their CMDD, for which, the illumination conditions were changed to different locations, such as, top right, bottom left, etc. and for the rotations, the images were rotated by small angles, such as, 5 degrees and 10 degrees. In total, they increased their corpus from 119 images to 357 images. The CMDD dataset was made publicly available and it was also utilized by [9] for their OCR study. Similarly, to add further testing to the OCR, the photos of reports can be clicked from varying distances, which will increase the corpus as well as help in testing the OCR thoroughly, as done by [5]. The authors of [10] performed post-annotation processing which increased their corpus from 990 images to 2995 scanned images. Taking different images from different devices is also done to increase the corpus, as done by [8], the images of their reports were taken from several different mobile devices such as iPhone, Nexus, Xiaomi, etc. In total, the dataset grew from 100 images to 1400 images. Some studies have also incorporated noise into their dataset such as [3], who have created 43 copies of each image which had different types of noise embedded in them such as blurs, watermarks and scribbles.

An alternative to procuring a dataset if medical reports are not available, is to use several open source datasets such as the English Language Dataset, as done by [3]. The dataset comprises 322 color page scans from books printed between 1853 and 1920. There is also the MiBio dataset, which is a dataset which is created for the purpose of testing OCR for post-processing evaluation, which contains 211 scanned pages from the book "Birds of Great Britain and Ireland", as used by [11]. Datasets can also be used in combination, as done by [11], who have used the MiBio dataset and they have included 100 images of medical reports which they procured from the NHS (National Health Service of Britain).

2.2 EXTRACTION

For data extraction from medical reports, the technique depends on the type of data available. If the reports are in the form of images, Optical Character Recognition (OCR) is widely used and if the medical report is in the form of a PDF, then instead of an OCR, a PDF parser is required. The PDF file will be passed to a PDF parser, which will extract the raw text from it. After that, the text will be passed to a line picker which will iterate through the text line by line, and on each line there will be a boundary extractor which will extract the boundaries, and a pattern matcher such as Regular Expressions will be used to locate text that is to be extracted. After that, the extracted text will be entered into the database, from which it will be available for analysis [1].

2.2.1 BASIC WORKING OF OCRS

Extraction can be done in two broad manners, they are hand coded/ learning based and rule-based/statistical. The paper [2] focuses on extraction of data from images. Since there is a large amount of data stored in images, it is also crucial to note that there is variation in size, alignment, and orientation that makes it complex. Images, not only contain text but also semantic and perpetual contents. Before extraction, it is necessary for the image to have appropriate geometrical orientation which benefits the extraction process. The following is done in terms of Shape, Size, Alignment, Color and Texture.

There are several problems that are faced in Text Information Extraction namely, Detection, Localization, Track, Extraction and Identification (OCR). Any system performing text information extraction consists of the following steps: the images undergo text tracking, which consists of text localization and text detection. This is followed by text enhancement and extraction. Then the recognition using an OCR is done, after which the text is obtained.

Initially, text detection is performed where the presence of text in the image is determined. Next text extraction is performed which is different for different sets of images. For text document images, a canny edge detector is used which leads to dilation operation that in turn creates character string. To remove the non-text components, analysis of standard deviation from connected components and computing is used. For Scene Text images, there are techniques, one for detection and extraction from commercially taken screenshot images. It is a combination of two methods as blob extraction method (edge based method and connected component labeling method). The result of this was 94.66% success on a complex background. The second approach was the Boundary growing method (BGM) and the multi-color difference (MCD) of handwritten scenes of text. BGM was used for fixing boundaries in separated clusters while MCD was used for gaps between text and non-text pixels.

A method was used for multi-oriented graphics and scene text in video images for Heterogeneous Text Images. The method uses the laplacian operator for highlighting transitions in the background of the image for the text, then K-means for classification of text and non-text regions. Lastly, a morphological operation was used to separate the artifacts from the text cluster. The other approach was to use a text extraction algorithm that was unaffected by noise skewness, text orientation, color, or intensity of any type of heterogeneous document image.

In the case of Edge-based Method, it is focusing on contrast between the text and background. The input image will be filtered separately with 3 by 3 horizontal to image and perform threshold search for vertical edges. To find shape and texture in characters, the small edges and adjacent edges are removed, and the text intensity histogram of each cluster is used. The second method uses the canny operator for edge detection. Morphological dilation is used to connect edges in any cluster, the horizontal and vertical aspects are used to find the non-text clusters.

There are computational complexities involved in Texture-based methods and thus, it increases processing time. One solution makes use of frequency data such as the number of edges in pixels and the number of horizontal and vertical lines in the picture. The text areas are considered to be rectangular, with edges detected using the Hough transform. The Wavelet Transform is used for text localization. The variation is achieved by filtering geometric parameters such as size and ratio. Text regions detected as merged for final result. Another method is the combination of FFT and neural networks for reduction of processing time. The output obtained was 32 features.

Lastly the centralized process of Optical Character Recognition (OCR), where specialized software is used for converting scanned images of text into digital format. OCR engines are used for optimization and development of the extraction of data. Frequent provision of data can prove to be noisy but can be improved by using any algorithm which again depends on the complexity and accuracy in terms of time consumed.

OCR technology replaces the whole process of rewriting the printed document into digital format. The recognition time can vary from minutes to seconds based on hardware and software configuration. There are several steps to perform the OCR operation which includes identifying the text direction, dimension of text, deciding the baseline position, tokenizing lines into single characters and running the tokens by unknown characters [2].

This is generally the approach taken by OCRs, but every OCR can give different results based on how it's trained.

2.2.2 DIFFERENT OCRS

There are many different OCRs available, the 3 prominent ones are Tesseract, Amazon Textract and Google Document AI. [3] Compared the 3 on images of English and Arabic document images. The dataset comprises 322 color page scans from books printed between 1853 and 1920, which are part of the English Language dataset. For the Arabic dataset, 4587 articles were printed from Wikipedia and scanned to pdf. From this 100 randomly selected pages were chosen. The authors also discussed how important it is to have noise in the dataset so that it reflects real world scenarios more accurately. So to implement this, the authors created 43 copies of each image which had different types of noise embedded in them such as blurs, watermarks, scribbles, etc. In total this generated a total corpus of over 14000 English documents and 4400 Arabic documents.

For testing metrics of each OCR, the accuracy was measured using the ISRI tool, known as Ocreval. It calculates the Levenshtein distance which is then used to calculate the Word Accuracy.

The paper concluded that Document AI had the lowest error, followed by Textract, followed by Tesseract. The error rate went up for all the OCR engines when noise was implemented in the dataset. Another observation was that the engines performed worse for Arabic documents than English documents.

But for different datasets, it is possible that the results may change. Hence it is an empirical process to find the optimal OCR that works best for a particular dataset. Another thing that affects how good an OCR performs, is the preprocessing of the images. In the dataset description we saw the different ways in which the datasets are processed, such as, changing angles of images, changing illumination direction, altering distance between camera lens and reports, and so on. And one more thing that may affect the accuracy of an OCR, is the domain of the images. For example, since this is a review for medical reports, in medical report images, there will be domain specific terms, such as "Erythrocytes" or "Platelets" which some OCRs may not be aware of and would interpret them as different words.

2.2.3 OCRS FOR MEDICAL REPORTS

For effective medical care and precise diagnosis, reports are required to be latest and complete. Paper files are inconvenient over a long term preservation. A lot of clinical data in regards to the archival data of a patient is in the form of clinical reports (CR) which are inserted into medical records. The texts describe the patient details, histories and findings during procedures. An efficient system is required for extracting information in a structured way so that it can be analyzed accurately.

Medical reports are a form of unstructured documents. (Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence) does an in-depth review of various techniques to extract data from unstructured documents. For this literature review, we will focus on the OCR part of the paper. The paper provides a general pipeline on processing data from unstructured documents, and explores each area of the pipeline extensively. The pipeline begins with parsing the dataset through the OCR, after which there is preprocessing of the raw text done, followed by segmentation and feature extraction before classification and post processing.

In preprocessing, the authors have mentioned various techniques which can help the OCR to detect words. The techniques stated are: Binarization, to convert an image to black and white followed by Noise reduction, cleaning of image by removal of unwanted noise pixels, Skew Correction, and Slant removal. Segmentation is to divide the image into parts, which will be processed further. The methods of segmentation include text line detection and word extraction methods followed by Feature Extraction. The raw data is extracted into the desired form of data using methods such as statistical feature extraction and structural feature extraction.

OCR is trained to learn and classify the extracted words correctly using K Nearest Neighbors, Naive Bayes, Neural Network, Support Vector machine. Post-processing detects and corrects any misspelled words, if present. To evaluate the OCR models Word Error rate (WER) and Character error rate (CER) is used. In contrast to unstructured documents as a whole, a general methodology pipeline for medical documents is given by [4], which utilizes the Tesseract OCR engine, whose engine consists of 3 main components, Connected Component Analysis, Line and Region Analysis, and Word Recognition, which works together to extract text from an image.

This word recognition module begins the recognition process by first identifying how a word should be separated into characters. The results from line analysis are classified first. The non-fixed pitch text is then processed by the rest of the word recognition. Then the Linguistic Analysis unit is called on.

Overall the paper presents a method for reading a camera-based document image by utilizing Tesseract OCR Engine. They have investigated character fonts and distance between the document and the camera on the smartphone. The key takeaways from the experiment were that Calibri and Tahoma are the suitable fonts, optimal distance from document to smartphone is 12 to 15 cm for a report block size of about 21 cm x 3 cm and a reflective (shiny) background is very difficult to threshold and isolate the text from it.

For the pipeline to work properly, the OCR must produce good results. We'll take a look at the approaches and experimentation done by [5], [6], [7], [8] and [9] along with the results and metrics of their approaches and experimentation.

The authors of [5] have thus presented an image processing system to extract information from medical reports. The work mainly relates to Optical Character Recognition which has two main steps: text localization and recognition.

The work mainly relates to Optical Character Recognition which has two main steps: text localization and recognition. There are various elements in documents that help in expressing our minds like figures, formulas, etc. In the work, the authors collected 119 paper files of a medical laboratory to design the dataset that consists of three groups, scanned images, and images that are captured in conditions with various illuminations and rotations. The medical document has a clear structure, the first table located on top containing patients' private information which has been erased for privacy concerns. The second table has details of test names, results, units and other information. The third table at the bottom is where the doctor and examiner have their signatures. The system follows a top-down pipeline to extract text. It is divided into mainly segmentation and recognition.

The first major step is extracting table areas from the document image. In each report, content is listed column wise, each column segmented by a projection method. Table lines are detected by Line Segment Detector in the input image. The table lines are separated from unwanted lines by two steps, the first step includes counting the lengths and slopes for a loose selection, where they can be filtered using the fixed threshold method. Every line gets a new coordinate after the first step. The lines are filtered again such that the angle of inclination is not more than 2 degrees. Two lines are obtained which are regarded as the table outlines. Like this, there are different extraction methods like the Yen algorithm which is used for mobile captured images. In this method, the image will be separated into subareas and a gray histogram is calculated. T denotes the table area in the document.

Once T is located, columns have to be separated. The holes around the characters in T are filled by erosion operation. It is very common that text is unevenly distributed, because of which, the top half of eroded T is considered for calculating features for projection analysis.

Here featCol, meanCol and stdDevCol are feature value, mean value and standard deviation of the i^{th} column in pixel level. Candidate columns where characters may exist, can be selected according to threshold calculated by adaptive threshold algorithms. Median filter is also used for feature vector featCol to avoid gaps between characters.

In terms of recognition, the accuracy is improved significantly by the use of LSTM. CRNN is directly applied, where the image height has to be resized to 32 pixels. Several convolutional layers are applied to extract features and this is fed to a deep bidirectional LSTM network with Connectionist Temporal Classification (CTC) layer to output the prediction. Some texts appear several times in the dataset, thus the FreeType library is used to generate a million training data.

The authors have considered two different thresholds which are used for the evaluation of the current methods and the proposed method has better metric evaluation in comparison to the current methods.

Similarly, the authors of [6] have firstly worked on the process of extraction of information. CRs have been collected and analysis has been performed, these CR contain observations from certain examinations, measurements, etc. The authors assembled a corpus of hundred CRs wherein 50 reports were training corpus, and the rest for the test phase. The major goal was the automatic structuring of the clinical report text into sections that were defined earlier to serve as a pre-processing step for the entity extraction. The report begins with patient identification details, history, text giving information about symptoms, findings and evolution of the patient, etc. A rule based algorithm was developed for the segmentation of the report into sections. The linguistic characteristics that are a part of the grammatical categories of the nominal phrases (NP) serve as a source for the recognition. The authors first focus on tokenizing and tag words using the TreeTagger tool. The tagged medical report is parsed to detect nominal phrases. A filter is applied to favor the longest NP over others.

The template for Medix includes Patient details like Name, Surname, age, etc. and disease details. It includes the symptom list, clinical reviews and medication details.

Process extraction is based on the rule approach and context surrounding the information. The authors have incorporated several dictionaries from several resources for tagging the NP components. There are four features decided to encode for each NP. The first one represents part of speech information of each token present in NP that can be extended with a second feature. The third feature is the NP length and fourth being the contextual information. Privacy is a major concern and so the patient data is anonymized before using it. In the research, a new anonymous identifier is generated by the Standard Hash Algorithm for each patient, in the structured database, the generated code is used for identification.

As a result, 50 reports were processed and out of 651 entities and properties, MedIX accurately matched 450 and missed 201, identified 3 entities erroneously. This gave a precision of 98.9% and a recall of 68.8%. Which is close to the results achieved by [5].

Similar to [4], [7] proposes a pipeline to extract text from scanned medical records. The Dataset Corpus consists of 100 medical report documents collected from hospitals in China. The images of these images were taken from several different mobile devices such as iPhone, Nexus, Xiaomi, etc. In total, the dataset consists of over 1400 images.

The pipeline starts with clicking a picture, then, there is the image pre-processing stage. Here the images are denoised, binarized and deskewed. After that, the OCR is run on the images. The OCR engine used is the IBM Datacap Taskmaster Capture 8.1. After this, it's the post-processing stage, where word resegmentation and Multi-engine Synthesis occurs. After this the annotations are done, where each entity, such as numbers and labels are annotated. Finally the Personal Health Record (PHR) is created where a confidence threshold is measured, and if it passes it, the document is constructed and stored in the database.

For evaluation, the precision, recall and F-measure was calculated. The results were calculated after each step in the pipeline for all the different categories. They achieved a result of F1 score of 0.918 for Diagnosis name detection, 0.845 for medication name detection and 0.922 for test item name detection for entity detection.

Deep learning approaches have been improving the results that OCRs have achieved previously. For example, [8] describes a deep-learning method for extracting textual information from Medical record. For text detection it achieves 99.5% recall in the experiments. For text recognition, a concatenation structure is developed. This approach is useful for integrating historical health records and patient medical records. In this paper they performed experiment as: Given a photo of a medical laboratory report they detect a feature detection algorithm. Each detected textual object is cropped from the source image and fed into a text recognition algorithm. Algorithm then with help of proposed concatenation structure and trained on synthetic data.

The paper adopts a two-stage architecture originally developed for generic object detection. A patch-based approach is applied to this architecture for text detection.
Network Architecture.

1) Image Goes through VGG16 network to extract a group of feature maps.
2) RPN takes feature maps as inputs and proposes axis-aligned bounding boxes that have more overlap areas with the ground-truth boxes.
3) The locations of proposals, region-of-interest (ROI) pooling extracts the features from the previous feature maps and transforms them into fixed size (7 × 7 in our experiments).
4) ROI features are flattened and pass through two fully connected layers

5) Output layer, connected with Fully Connected Layers and then calculates the loss of text/non-text classification and bounding box regression.

Similar [8], [9] proposes a Deep learning-NLP based approach to extract data from scanned electronic health records. They have evaluated this model on sleep study reports, which is a type of medical report but this model can also be applied to other types of reports such as urine reports or blood reports. They have utilized the Tesseract OCR engine, 7 bags of words models and 3 deep learning based models. They also evaluated different combinations of image preprocessing to improve the results.

For the data source, the authors collected sleep study reports from University of Texas Medical Branch, which has around 990 reports collected between 2015 and 2018. Then the authors performed post-annotation processing which increased their corpus to 2995 scanned images. The data was split into a 70-30 Test-train split.

Then image preprocessing was performed where images were first converted to gray scale, then dilate and erode process was used and finally the contrast was increased by 20%.

For the OCR, the Tesseract OCR engine was used in python under the library pytesseract. After that, for text segmentation, regex was used to search numerical data. For the text classification to determine what kind of numeric data is extracted, the position indicators of the numeric values, page number and floating point representation of numeric value was collected and this data was then trained in NLP models. The approaches of NLP models used were Bag of words and deep learning-based sequential models.
The bag of words model consists of classifiers such as Logistic Regression, Ridge Regression, Lasso Regression, Support Vector Machine, k-Nearest neighbors, Naive Bayes and Random Forest. A 5-fold cross validation was performed during training with a 6:1 training-to-validation ratio. From the bag-of-words model, the best accuracy was given by Random Forest which was 93.71%.

The sequential deep learning based models consist of Bidirectional Long short-term memory (BiLSTM), Bidirectional Encoder Representations from Transformers (BERT), and pretrained BERT with EHR data, Clinical BERT. All the three sequence models outperformed the bag-of-words models, with BERT giving the highest accuracy of 95.1%.

Most of the studies have implemented preprocessing, which has improved the results, but in [10] the authors have discussed how the lack of medical terms, specific to the domain, while training OCR's can cause inaccuracies in the OCR. An OCR is trained on general words used in the English language, so particular medical terms may not get recognized. This paper suggests a Post Processing method, to counter this problem.

The paper uses OCR Tesseract Engine, and uses the RoBERTa language model to find wrongly predicted words. For evaluation, the dataset on MiBio is used, and a few medical reports collected from the NHS are used and to measure the quality of the OCR the Word Error Rate (WER) and Character Error Rate (CER) are calculated.

After running experiments, the paper found that the WER and CER were reduced by implementing post processing using their proposed method. Before post processing the Average insertions, substitutions and deletions were 0.93, 0.812 and 0.858 respectively. After post processing it reduced to 0.813, 0.645 and 0.711 respectively.

As we saw in [9] and [8], Recurrent Neural Networks are also prevalent for developing OCRs. [11] have implemented RNNs as well. The paper is successful to extract information from descriptive pathology reports to structured information. Using structured pathology reports will help in easy understanding and detection of cancer. The structured pathology report can make it conducive to subsequent knowledge extraction and analysis, and the construction of pathology ontology and knowledge map.

The future scope of the paper is to improve the accuracy of their model in extracting data and mapping pathological images to pathology reports.

The data source of the paper is from the TCGA (the cancer Genome atlas) project, they have selected four types of cancer, which are kidney cancer, lung cancer, breast cancer, and prostate cancer as the cancer types for study. They have encoded the sentence with a recurrent neural network and implemented the RNN as a bi-directional LSTM (Bi-LSTM). From the results we can see that the GCN-based neural network performs better on most tasks.

Another way to deploy the proposed technique to end-users such as pathologists oncologist and clinicians is by integrating the model with digital pathology report to the existing digital pathology platform such as OpenHI user interface, thus allowing pathologists to upload the diagnosed pathological information to the database so that clinicians can see the structured pathology report extracted with automatic prediction information.

After the data has been extracted, the next part of the pipeline is to store the medical report data in a database.

3. PAPERS BASED ON MEDICAL DATABASES

Databases are a crucial part of any system, it is used for storing the different values that are required in a system. The databases have different schemas and organizations that differentiate their usage in particular use cases. There are

several types of databases which have different storages, and their particular way of querying and retrieving information. Storage of medical reports can be a challenging tasks because medical data is sensitive and needs to be confidential, at the same time patients need a convenient way to access their medical report data, so to address these needs, we have reviewed papers related to EHR systems, medical report storage systems, etc. to provide the continuation from the previous stage of data extraction to the next stage which is data storing.

The authors in [12] have talked about the scenario of the Indian healthcare system. The current state of information management is an unstructured method of keeping records on paper. The Electronic Medical Record (EMR) is an electronic version of patient-related health data collected from a single healthcare service provider, which is typically utilized at the secondary and tertiary levels. Today, the vast majority of hospitals handle their patients' health records in hybrid forms. Organizations employ the EMR system to gather patient data in electronic format, but comprehensive health data is not available to healthcare practitioners at the time of service. The 'Digital Health Mission' of the government resulted in the creation of new procedures and instruments for preserving patients' digital data. The Government of India took the initiative to formulate and publish an EHR standard back in September 2013 and has constantly revised and published new versions for the same. The adoption of these terminologies and standard coding systems in Health Information System (HIS) ensures that data is unambiguous. The Union Cabinet in 2019 accepted the personal data protection bill that says the sensitive personal data can only be processed by explicit consent of an individual and prevents other organizations from access without the same. The authors visited the government primary healthcare centers where questionnaires were used for identifying issues to research building the EHR system. The PHC patient's case sheet was the sole record that provided specifics regarding health information, and it was paper-based, putting it at risk of damage and deterioration over time. Medical documentation at the primary health care level is mostly paper-based and book-based. The ongoing necessity for high-speed Internet access is a fundamental impediment to implementing IT application infrastructure in rural healthcare systems. At the secondary and tertiary levels, patient data was recorded in a hybrid record structure that was both electronic and paper. During the evaluation, it was discovered that the system aids in the documentation of patient information but does not facilitate the sharing of health information between different levels of healthcare institutions. The fundamental goal of the HIS is to provide lifetime clinical care at all times; therefore, data syntactic and semantic interoperability must be maintained at all times. The suggested concept links HIS users to a central point and distributes information across the network. The Administrative system module, which records patient registration, admission data, and so on, is the system's initial component. The nursing module, which comprises height, weight, blood pressure, and BMI information, is the second component. The laboratory data, which is vital in the clinical process for the healthcare service provider, is the third component. The clinical documentation module, which collects the patient's clinical data such as diagnosis, procedures, prescriptions, and so on, is a significant component of EHR. The EHR design requirement analysis phase is the phase where the requirements are collected and specified. The Design phase provided the solution to the problem with iterative steps. Design basically is the blueprint or a plan of the solution of the problem by the system. The EHR system aids in capturing the patient's past medical history and present symptoms. The details of major events or illnesses are better recorded for diagnosis of current illness. The details include physical examination, observations, diagnosis, lab investigation, encounter type, demographics and patient history. The EHR is a private communication between the patient and the healthcare staff and should maintain confidentiality. The framework developed mainly focuses on Authentication, Authorization and Attribute based encryption. Testing is needed for the EHR system to validate and verify all the functionalities, interoperability, compliance and performance.

In [13], methods of medical data extraction from databases/data warehouses using OLAP tools, and then generating a report from that data is mentioned. The authors of this paper have created their own Custom made Medical report generation software, which does the above said things and generates a report for the same. OLAP stands for Online Analytical Processing and it is used in various complex forms of data analysis such as reporting, summaries, KPI's, analysis of trends, time series forecasting and top down hierarchical analysis. The OLAP server has proven to be the most important component, it is present between the client and the database management systems. OLAP servers understand the types of data structures present in the databases and have special functions and methods which can be used for analysis of those data structures. Electronic Health Records (EHR), are used for storage of patient health records and tracking of patient medical health. For that a database is required, in this paper, the authors have proposed a EHR metadata model which consists of 6 tables, named: 'profile', 'Tables', 'MeasurementUnit', 'ProfileItem', 'Fields' and ranges, in the form of a relational schema. Lastly, the authors propose a report generating tool which will retrieve patient data and display it to the user in the form of tables and charts, which the user can download for reference, in the form of XML files or PDF files.

Authors of [14] propose a cloud based Electronic Health Record (EHR) system which facilitates a storage for patient medical report data and since it is on the cloud, it provides a key generation system with the Key Attributes Method. This also allows the patient to control access points to their data, which means they can grant access to their doctors or relatives as well.

The authors have also provided a framework for the cloud which will host the EHR, which includes the EHR System Doc, the Hosts and the PHR interacting with the EHR system.

The authors of (Customization of Medical Report Data) mention how the current state of storage of medical reports is and what are the subsequent problems of the same. It mentions how to store medical data in a structured fashion and why it is better than conventional methods.

If report data is stored in a structured format, it has many advantages such as: user customized data, historical data retrieval, interpretation analysis and structured medical data can also be used to practice evidence based medicine. This paper proposes a standardized structure for report data, and mentions that it is an absolute prerequisite for structured medical data. It will also ease the process of data mining, data retrieval and data output and make it quicker and more versatile for different processes and requirements. This paper proposes a table like timeline for patient data with fixed parameters. The table contains parameters such as: Date, Modality, Anatomy, Findings, Clinical Significance, Interval change and Follow-up recommendations. It is tracked through time and since it is in a table format, which is similar to DBMS relational tables, the data can be filtered as well as per need. This paper believes that standardizing structured medical data across every platform and institution can improve efficiency of data retrieval, facilitate data driven analysis, improve workflow, and many other aspects which will ultimately improve clinical outcomes and patient safety.

The authors in [15] provide a relational model for storing hospital report data which can then be accessed for research work, and other analytics. The paper provides a very basic relational schema and the paper mainly focuses on how to convert the patient data which was stored in the form of .csv files into tables which can be stored in a relational database. The tools mentioned in the paper for extracting csv data such as MaxSplitter, and Microsoft Access for the relational database are now deprecated but the concept is still viable to convert .csv medical report data into a relational schema. The concept is to create tables based on column headings of the.csv files, identifying primary and foreign keys, and then splitting the data in csv files and storing them in their respective tables.

In [16], the authors mention an object-relational model with entity attribute value with classes and relationships for an EHR system. The paper has first explained a conventional model which has 3 main columns: patient identity, test attribute name, and the value for the said test attribute. Along with these 3 columns, the relational schema also includes a temporal field in the form of date and the primary key, patient ID. So the conventional relational schema has a total of 5 columns and then more for other testing attributes. The paper proposes an Entity-Value-Attribute (EAV) schema which will have only 3 main columns: the date, the attribute name, including the patient identity and lastly the attribute value. This table will have 4 columns, because the patient identity and attribute names have been cascaded into one column. This EAV schema has a few disadvantages as this schema is still in 1st Normal form which means the database will have poor data display, increased query complexity and poor constraint checking. To counter this the authors have suggested a metadata model which would logically be on top of the EAV model, which will act as a description and contain all the relations contained in the EAV model. When this metadata model is combined with the EAV model, it leads to better querying, smaller storage, as there will be lesser null values and easier understanding of patient health data. It is a good alternative to EHRs which have a lot of varying test attributes and are prone to a lot of null variables.

The visualization of data that is stored and used, is a necessary step for better understanding and realization of data. It is further discussed as the next step in the pipeline.

4. PAPERS BASED ON MEDICAL DATA VISUALIZATIONS

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Traditionally, medical visualization has focused on single-subject data, such as applying visualization techniques to an MRI dataset of a single patient for diagnosis or to a single study participant to explore their brain connectivity.

The authors of [17] discuss medical institutions that store a lot of medical data that is complex, while more and more data is being collected to support the growing medical informatics. Good record keeping helps in smooth communication and better understanding between patients, doctors and other related professionals. Electronic Health Records (EHR) is introduced for sharing this information, making it available and securely to authorized users. Health Data Visualization is a way to gain insights derived from EHR. People can understand and act upon information by seeing the results in a visual context.

The deformable human body model is built using a three step statistical analysis method. The Statistical shape model uses the Generalized Procrustes Analysis to analyze real differentiation of human body data as well as changes of parameters such as height and weight of the mode. The Principal Component Analysis divides the samples into male and female data. The internal deformation parameters are extracted between these by PCA. The interpolation algorithm is used on height, weight and other parameters. The customized 3D model of the human body is able to be generated when the real data is provided. The user can generate their 3D digital human body with organs, muscles,

etc. after the parameters are entered and the model is generated on their smartphone. The procedure has 4 main parts. The image preprocessing is performed so that the irrelevant data can be eliminated, the useful information can be saved and enhanced. OCR is used for digitizing the medical records. The third step is the OCR results analysis, page ranking is implemented on the OCR results, the page style is recognized and the page is digitally recreated with the obtained information. The header and footer content is recognized when the contents are in the same position.

Lastly, the OCR results and the medical database double check method is applied, to ensure the correctness of the data. The database is built by completing the schema with the help of professionals including entities and relationships. Filling the database with entities, relationships and related properties to complete the relation and filling the database after data scrubbing by medical professionals.

5. CONCLUSIONS

In this paper, we saw many papers which use tools and techniques to extract data from medical reports and also how different EHR/EMR/PHR store medical data. We also reviewed systems which cover similar topics and explored other related works for the same. Some systems have also created their own pipelines which do extraction and storage of medical report data from reports. Most of the papers used OCRs for data extraction, utilizing different datasets of medical reports.

For datasets, it was observed that most of the papers have procured the dataset on their own, with their personal reports because as mentioned earlier, obtaining a lot of medical reports can be a difficult task. The papers which did not gather their medical reports, either got them from a pathology lab or they used few of the sample OCR datasets that are available online such as the MiBio dataset. Almost all the papers have increased the size of the corpus by adding natural data preprocessing methods like changing the angle of the image, changing the illumination conditions or altering the distance from the camera to the image. This helps in increasing the data to test the OCRs and also it ensures that the OCR is being thoroughly tested in several varied circumstances.

After the data is being extracted, to detect the words and correct them if necessary, most papers used Machine learning algorithms for correction of OCR text and then utilized Regex extracting crucial information from raw OCR texts, which can then be stored in databases.

For databases of medical report data, there were several approaches that were observed, most of the older techniques involve usage of OLAP tools and techniques, and with the development of blockchain technology, many blockchain based databases have also been developed which provide an alternative to the traditional methods. Cloud based systems have also grown in popularity recently, but most of them act as hosts/servers which can house the aforementioned databases which can act as access control providers and also a secure place to keep medical report data of patients. Medical reports are possible to be mapped onto a relational schema, which makes it convenient and easy to store and access, hence making them the most used technique for storage of medical reports.

For future work, we propose a full-fledged system which can extract and store medical report data which can combine and apply the tools and techniques reviewed in this paper, and after that use that data for analysis and interpretation for the patients and also aid doctors by giving them access to their patient records.

REFERENCES

[1] Prashant M Ahire, Anil P Gagare, Yogesh B Pawar, Savan S Vidhate, “EXTRACT AND ANALYSIS OF SEMI STRUCTURED DATA FROM WEBSITES AND DOCUMENTS”, www.ijcsmc.com, Vol 4, Issue 2

[2] Shanti Pragnya, Swayanshu. (2017). Study of Information Extraction and Optical Character Recognition. buiwww.ijcst.com. 8. 22-26.

[3] Hegghammer, T. OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment. J Comput Soc Sc (2021). https://doi.org/10.1007/s42001-021-00149-1

[4] A. Kongtaln, S. Minsakorn, L. Yodchaloemkul, S. Boontarak and S. Phongsuphap, "Medical document reader on Android smartphone," 2014 Third ICT International Student Project Conference (ICT-ISPC), 2014, pp. 65-68, doi: 10.1109/ICT-ISPC.2014.6923219.

[5] W. Xue, Q. Li, Z. Zhang, Y. Zhao and H. Wang, "Table Analysis and Information Extraction for Medical Laboratory Reports," 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2018, pp. 193-199, doi: 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00043.

[6] B. Fatiha, B. Bouziane and A. Baghdad, "MedIX: A named entity extraction tool from patient clinical reports," 2011 International Conference on Communications, Computing and Control Applications (CCCA), 2011, pp. 1-6, doi: 10.1109/CCCA.2011.6031494.

[7] Li X, Hu G, Teng X, Xie G. Building Structured Personal Health Records from Photographs of Printed Medical Records. AMIA Annu Symp Proc. 2015 Nov 5;2015:833-42. PMID: 26958219; PMCID: PMC4765700.

[8] W. Xue, Q. Li and Q. Xue, "Text Detection and Recognition for Images of Medical Laboratory Reports With a Deep Learning Approach," in IEEE Access, vol. 8, pp. 407-416, 2020, doi: 10.1109/ACCESS.2019.2961964.

[9] Deep learning-based NLP Data Pipeline for EHR Scanned Document Information ExtractionEnshuo Hsu (1, 3, and 4), Ioannis Malagaris (1), Yong-Fang Kuo (1), Rizwana Sultana (2), Kirk Roberts (3) ((1) Office of Biostatistics, (2) Division of Pulmonary, Critical Care and Sleep Medicine, Department of Internal Medicine, University of Texas Medical Branch, Galveston, Texas, USA. (3) School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, Texas, USA. (4) Center for Outcomes Research, Houston Methodist, Houston, TX, USA.)

[10] S. Karthikeyan, A. G. Seco de Herrera, F. Doctor and A. Mirza, "An OCR Post-correction Approach using Deep Learning for Processing Medical Reports," in IEEE Transactions on Circuits and Systems for Video Technology, doi: 10.1109/TCSVT.2021.3087641.

[11] J. Wu, K. Tang, H. Zhang, C. Wang and C. Li, "Structured Information Extraction of Pathology Reports with Attention-based Graph Convolutional Network," 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020, pp. 2395-2402, doi: 10.1109/BIBM49941.2020.9313347.

[12] Pai, M.M.M., Ganiga, R., Pai, R.M. et al. Standard electronic health record (EHR) framework for Indian healthcare system. Health Serv Outcomes Res Method 21, 339–362 (2021). https://doi.org/10.1007/s10742-020-00238-0

[13] P. J. Rajkovic and D. S. Jankovic, "Custom made medical data reporting tool," 2009 9th International Conference on Telecommunication in Modern Satellite, Cable, and Broadcasting Services, 2009, pp. 306-309, doi: 10.1109/TELSKS.2009.5339518.

[14] "Impact of Cloud Database in Medical Healthcare Records based on Secure Access." International Journal of Engineering and Advanced Technology (2019): n. pag.

[15] Cutting, A.C. & Goldberg, Gerson & Jervis, Kathryn. (2007). An easier method to extract data from large databases: The medicare hospital cost report database. 11. 27-34.

[16] El-Sappagh, Shaker & El-Masri, Samir & Riad, Alaa el-din & Elmogy, Mohammed. (2012). Electronic Health Record Data Model Optimized for Knowledge Discovery. IJCSI International Journal of Computer Science Issues. 9. 329-338.

[17] N. Liu et al., "A New Data Visualization and Digitization Method for Building Electronic Health Record," 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020, pp. 2980-2982, doi: 10.1109/BIBM49941.2020.9313116.