Reconocimiento Óptico de Caracteres de Dominio Clínico: Análisis comparativo y Arquitectura para su transformación al estándar HL7

ITI. Irving Jesús Ramírez Alacio A., Dr. José Luis Sánchez Cervantes B., Dr. Isaac Machorro Cano C., MCE. Beatriz Alejandra Olivares Zepahua D., Dr. Giner Alor Hernández E.

[a] TecNM/Instituto Tecnológico de Orizaba. 94320, irving.jesus.ramirez.alacio@hotmail.com, Orizaba, Veracruz México.
[b] CONACYT/Instituto Tecnológico de Orizaba. 94320, jose.sc@orizaba.tecnm.mx, Orizaba, Veracruz México.
[c] Universidad del Papaloapan. 68301, imachorro@unpa.edu.mx, Tuxtepec, Oaxaca México.
[d] TecNM/Instituto Tecnológico de Orizaba. 94320, beatriz.oz@orizaba.tecnm.mx, Orizaba, Veracruz México.
[e] TecNM/Instituto Tecnológico de Orizaba. 94320, giner.ah@orizaba.tecnm.mx, Orizaba, Veracruz México.

Resumen

En el sector salud, la necesidad del intercambio de información es cada vez más necesaria e indispensable, puesto que permite al personal de salud tener acceso al historial clínico del paciente con el objetivo de prestar la atención médica de una forma más oportuna y eficiente. Sin embargo, ante la nueva representación de gestión de salud basada en el valor al que se acercan los sistemas de salud en todo el mundo, la interoperabilidad requiere ampliar su alcance y garantizar el intercambio de información.

Ante este contexto, surge la necesidad de contar con una herramienta que contribuya a realizar el proceso de interoperabilidad entre sistemas de información clínica, por tal motivo el presente trabajo describe un análisis comparativo de algoritmos de OCR (*Optical Character Recognition*) con la finalidad de seleccionar el más óptimo para el diseño de la arquitectura de un módulo de reconocimiento óptico que permita extraer información de diferentes documentos clínicos para su interpretación al estándar HL7 (por sus siglas en inglés *Health Level Seven*) utilizando técnicas de NLP (*Natural language processing*) con el objetivo de esquematizar la información clínica bajo un mismo formato y facilitar la interoperabilidad entre expedientes clínico-electrónico, y su acceso al personal responsable del sector salud.

Palabras clave— Análisis clínicos, *E-health*, Expediente clínico-electrónico, Interoperabilidad, Reconocimiento óptico de caracteres.

Abstract

In the healthcare sector, the need for information exchange is increasingly necessary and indispensable, since it allows healthcare personnel to have access to the patient's clinical history in order to provide medical care in a more timely and efficient manner. However, given the new representation of value-based healthcare management that healthcare systems around the world are approaching, interoperability requires broadening its scope and ensuring the exchange of information.

In this context, the need arises for a tool that contributes to the process of interoperability between clinical information systems, For this reason, this initiative describes a comparative analysis of OCR (Optical Character Recognition) algorithms in order to select the most optimal one for the design of the architecture of an optical recognition module that allows extracting information from different clinical documents for its interpretation to the HL7 (Health Level Seven) standard using NLP (Natural language processing) techniques with the objective of schematizing clinical information under the same format and facilitating the interoperability among clinical-records and its access to the responsible personnel of the health sector.

Keywords— *Clinical analysis, E-health, Electronic health record, Interoperability, Optical character recognition*.

1. INTRODUCCIÓN

En los últimos años las Tecnologías de la Información y Comunicación (TICs) tienen miles de aplicaciones en diferentes sectores empresariales, educativos y salud. Las tecnologías de la información son un conjunto de herramientas que permiten el procesamiento y transferencia de información.

Un área de las TICs, son las herramientas de software que ofrecen soluciones a determinados problemas, dentro de estas, se encuentran las aplicaciones Web, que tienen un papel importante dentro del reciente paradigma de gestión de salud basada en estándares de *E-health* utilizados en todo el mundo, la interoperabilidad requiere ampliar su alcance y garantizar el intercambio de información. Por tal motivo, es necesario un sistema de intercambio de información entre los actores participantes en la gestión y el cuidado de la salud tales como hospitales, pacientes, laboratorios clínicos, doctores, entre otros.

El presente trabajo describe un análisis comparativo de algoritmos de OCR donde se examinan sus principales características tales como la velocidad de extracción y nivel de precisión por cada documento clínico procesado. De igual forma, se presenta una arquitectura de software que permite utilizar el reconocimiento óptico de caracteres en la interpretación de documentos clínicos al estándar HL7 utilizando el algoritmo OCR que mejores resultados ofreció.

En la siguiente sección se presenta el estado del arte integrado por trabajados que implementaron algoritmos de OCR y el estándar HL7 con el objetivo de diferenciar las iniciativas similares respecto a este trabajo y enfatizar nuestra contribución principal. De igual forma la se muestra la metodología utilizada para realizar el análisis comparativo de algoritmos de OCR, se presenta la arquitectura para un módulo de OCR para la interpretación de documentos clínicos al estándar HL7 utilizando técnicas de NLP. Posteriormente en la sección tres se presentan las conclusiones y el trabajo a futuro.

2. CONTENIDO

En la siguiente sección, se presenta el contenido de mayor relevancia del presente trabajo.

2.1 Estado del arte

A continuación, se presente el estado del arte, el cual fue identificado de acuerdo a dos aspectos principales: 1) Procesamiento de información en el estándar HL7 y, 2) Reconocimiento Óptico de Caracteres en el área de la salud.

2.1.1 Procesamiento de información en el estándar HL7

En [1] realizaron la propuesta de una arquitectura genérica para el desarrollo de servicios de salud móvil estandarizados y seguros con ayuda de las redes sociales. Como prueba de concepto, los autores realizaron una prueba mediante el desarrollo de dos aplicaciones Android y eligieron como red social Twitter®, como cubierta de seguridad openPGP (*open Pretty Good Privacy*), el estándar internacional HL7 y un algoritmo de incorporación de información. En cuanto a las pruebas realizadas a la aplicación, una prueba incluyo un escenario a pequeña escala y la otra prueba un escenario de límites. De igual forma en [2], se utilizó la tecnología de GraphQL y el estándar HL7 FHIR (*Health Level 7 Fast Healthcare Interoperability Resource*s) para HIE (*Health Information Exchang*) y se presentó un algoritmo para mapear recursos HL7 FHIR a un esquema GraphQL. Adicionalmente, se desarrolló un prototipo de implementa-ción el cual fue comparado con un enfoque RESTful. El resultado obtenido demostró que la combinación de GraphQL y las API (*Application Programming Interface*) Webs basadas en HL7 FHIR para HIE satisfacen los requisitos de los clientes Web y móviles. Así mismo, en [3] se presentó un mecanismo de interoperabilidad de recursos de salud que consistió en la conversión de datos de salud a estructura HL7 FHIR. La meta que se planteó fue desarrollar ontologías de datos, las cuales fueron almacenadas en un *triplestore*. Después para cada ontología desarrollada se calculó la semejanza sintáctica y semántica con las diferentes ontologías de HL7 FHIR con ayuda de la distancia Levenshtein y sus huellas semánticas correspondientes. Posterior a la incorporación de los resultados, se realizó la correspondencia con HL7 FHIR, traduciendo los datos clínicos a un estándar médico. En el mismo contexto V. Kilintzis [4] y sus colaboradores presentaron un marco de gestión de datos de telemedicina, que tuvo por objetivo el ayudar en los servicios de atención médica en pacientes crónicos. El marco de trabajo se apoyó en una ontología OWL (*Web Ontology Language*), desarrollada con recursos HL7 FHIR para almacenar y representar información HCE (Historia Clínica Electrónica) la cual fue semánticamente mejorada usando como guía los principios de *Linked Data*.

2.1.2 Reconocimiento Óptico de Caracteres en el área de la salud

En [5] se presentó la construcción de un sistema de extracción de información clínica apoyado en ontologías, denominado "OB-CIE". El sistema "OB-CIE" provee un método para extraer datos clínicos de las notas de texto que genera el médico y transforma las anotaciones clínicas no organizadas en información organizada a la cual se accede a través de las HCE (Historias Clínicas Electrónicas). Siguiendo la misma línea, en [6] se expuso el uso de un nuevo enfoque híbrido basado en NLP de los gráficos que son dilucidados con el OCR para extraer información clínica importante de los reportes escaneados de colonoscopia y patología. Además de lo anterior, [7] describió el diseño y la evaluación de un sistema para clasificar los documentos en categorías clínicamente importantes y no importantes. El propósito fue mostrar que los sistemas de clasificación de textos son capaces de clasificar con precisión los documentos clínicos escaneados del cual el texto se obtuvo utilizando el reconocimiento óptico de caracteres OCR. Además, en [8] se realizó un enfoque apoyado en el aprendizaje profundo para la extracción de datos textuales a partir de imágenes de informes de laboratorio clínico, el cual apoya al personal médico a solucionar el problema de compartir datos entre instituciones. Además, se describió la creación de dos módulos: detección y reconocimiento de textos. Para el caso de la detección de textos, se empleó una estrategia de entrenamiento basada en "parches". Por otro lado, en el reconocimiento de textos, se empleó una estructura de concatenación para ajustar las características de las capas superficiales y profundas de las redes neuronales. En las pruebas realizadas se mostró que el identificador de textos presentado en el enfoque mejora la precisión del reconocimiento de textos multilingües. La tabla 1 muestra una comparación de los trabajos relacionados con esta propuesta con el objetivo de identificar los elementos más importantes de cada uno de ellos y como se relacionan con la propuesta presentada.

Tabla 1 Comparativa entre los trabajos relacionados.

Trabajo	Problema	Contribución
[1]	Falta de seguridad y privacidad de la información que se comparte a través de sistemas de información.	Arquitectura genérica para el desarrollo de servicios de salud para dispositivos móviles.
[2]	Necesidad de interoperabilidad entre sistemas de salud.	Algoritmo para mapear recursos HL7 FHIR a un esquema GraphQL.
[3]	Necesidad de compartir datos clínicos y mecanizar el proceso de seguimiento del paciente.	Mecanismo de interoperabilidad de recursos de salud para la conversión de datos de salud a estructura HL7 FHIR.
[4]	Necesidad de mejorar los servicios de atención médica a pacientes crónicos	Marco de gestión de datos de telemedicina para ayudar a los servicios de atención médica.
[5]	La incorrecta interpretación y documentación de notas clínicas.	Método para extraer datos clínicos de las notas de texto que genera el médico.
[6]	En la actualidad una colonoscopia es uno de los procedimientos médicos para el cribado del cáncer colorrectal en Estados Unidos de América. Generalmente los informes se realizan en un formato no estandarizado y regularmente no están	Enfoque híbrido basado en NLP de los gráficos que son dilucidados con el OCR para extraer información clínica importante de los reportes escaneados de colonoscopia y patología.

Trabajo	Problema	Contribución
	agregados en los registros clínicos electrónicos.	
[7]	Por lo general las HCE (Historias Clínicas Electrónicas) están formadas documentos escaneados de diversas fuentes y tipos, entre los cuales se encuentran tarjetas de identificación, informes de radiología, correspondencia clínica.	Se desarrollaron y probaron múltiples modelos de aprendizaje automático de clasificación de textos, incluyendo tanto enfoques denominados como "bolsa de palabras", así como de aprendizaje profundo.
[8]	Generalmente no se dispone de registros clínicos completos durante el tratamiento por causa del problema funcional del sistema de HCE o a las diversas brechas de información.	Se realizó un enfoque apoyado en el aprendizaje profundo para la extracción de datos textuales a partir de imágenes de informes de laboratorio clínico, el cual apoya al personal médico a solucionar el problema de compartir datos entre instituciones.

Después de presentar los trabajos relacionados con la propuesta del presente trabajo, se determinó que aunque comparte similitud con algunos trabajos presentados como es el caso de [7], esta propuesta se distingue de otras ya que se centra en la interpretación de formatos de análisis clínicos en idioma Español, y considerando que el sector salud es muy amplió se contemplan cuatro categorías de análisis: 1) Análisis clínicos para enfermedades crónico-degenerativas; 2) Análisis clínicos para enfermedades cardiovasculares; 3) Análisis clínicos para padecimiento de estrés y, 4) análisis clínicos generales. Además, la presente propuesta busca el generar documentos clínicos personalizados, es decir que el usuario seleccione la información más relevante de dos o más análisis clínicos con el objetivo de obtener un nuevo documento con lo más importante y posteriormente transformar este documento al estándar Hl7 como si se tratara de un documento tradicional.

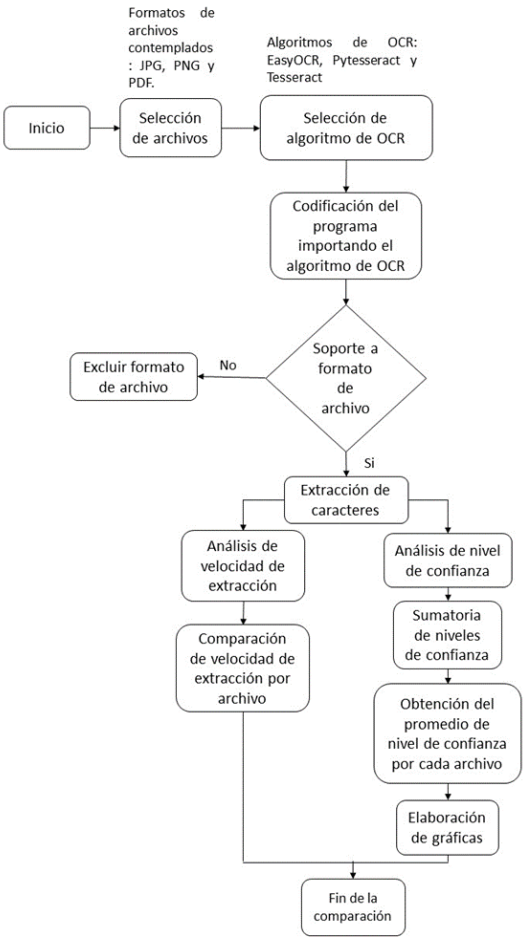
2.2 Metodología del Análisis Comparativo

En esta sección se presentan los pasos aplicados a cada uno de los algoritmos de OCR con el objetivo de determinar sus características de mayor importancia las cuales se utilizaron para realizar un análisis comparativo de las mismas.

De igual forma con el objetivo de obtener el resultado antes mencionado se utilizó el mismo conjunto de archivos para cada algoritmo buscando el detectar las variaciones en el procesamiento de cada archivo. Para lo cual el conjunto de archivos se conformó de cinco imágenes de análisis clínicos de los cuales dos correspondieron al formato JPG (por sus siglas del inglés *Joint Photographic Experts Group*) y las otras tres imágenes correspondieron al formato PNG (del inglés *Portable Network Graphics*) de igual forma se incluyó un formato de análisis clínico con extensión PDF (del inglés *Portable Document Format*).

Los algoritmos seleccionados para el presente análisis comparativo son los siguientes: EasyOCR, Pytesserac y Tesserac. En la figura 1 se presenta el flujo de los pasos aplicados a cada uno de los algoritmos.

Fig. 1 Diagrama del procedimiento de comparación.



2.3 Características a comparar

Una vez aplicada la metodología descrita anteriormente se compararon las características más importantes de cada algoritmo de OCR entre las cuales destacan el soporte a múltiples idiomas, velocidad de extracción por archivo, nivel de precisión por carácter y promedio de nivel de confianza por archivo.

En la tabla 2. se presenta un resumen de las características de los algoritmos de OCR (Soporte a múltiples idiomas, velocidad extracción y promedio del nivel de precisión), mientras que en las figuras de "Fig. 2 a la Fig. 6" corresponden a las imágenes analizadas con los algoritmos OCR.

Tabla 2. Comparación de velocidad de extracción y nivel de precisión.

Nombre de la imagen (Img.)	Dimensiones por cada imagen (Ancho por Alto)		Tamaño por cada imagen	Velocidad de extracción	Promedio de precisión alcanzado por imagen
ALGORITMO EasyOCR					
Img1.jpg	1200px	1600px	343KB	51.55 Seg.	0.534158247
Img2.png	1600px	1200px	253KB	45.37 Seg.	0.830197166
Img3.jpg	1200px	1600px	326KB	36.24 Seg.	0.842506857
Img4.png	1275px	1650px	2 MB	30.56 Seg.	0.777081683
Img5.png	1200px	1600px	343KB	55.29 Seg.	0.903403476
ALGORITMO Pytesseract					
Img1.jpg	1200px	1600px	343KB	09.33 Seg.	49.29762059

Nombre de la imagen (Img.)	Dimensiones por cada imagen (Ancho por Alto)		Tamaño por cada imagen	Velocidad de extracción	Promedio de precisión alcanzado por imagen
Img2.png	1600px	1200px	253KB	05.41 Seg.	75.45926225
Img3.jpg	1200px	1600px	326KB	04.61 Seg.	87.63577598
Img4.png	1275px	1650px	2 MB	05.01 Seg.	82.82350248
Img5.png	1200px	1600px	343KB	05.53 Seg.	91.35961277
ALGORITMO Tesseract					
Img1.jpg	1200px	1600px	343KB	35.40 Seg.	39.64965449
Img2.png	1600px	1200px	253KB	1.45.05 Min.	76.04048154
Img3.jpg	1200px	1600px	326KB	N/A	N/A
Img4.png	1275px	1650px	2 MB	1.56 Min	83.09301211
Img5.png	1200px	1600px	343KB	38.82 Min	81.94425444

Nomenclatura: Seg. = Segundos; Min. = Minutos; Esp. = Español; Ing. = Inglés; Img. = Nombre de la Imagen; px = Pixeles; KB = *kilobyte*; MB = *Megabyte*.
Soporte a múltiples idiomas (Español e Inglés):
EasyOCR = Sí
Pytesseract = Sí
Tesseract = Sí

Fig. 2 "Img1".

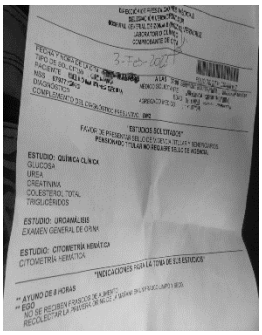


Fig. 3 "Img2".

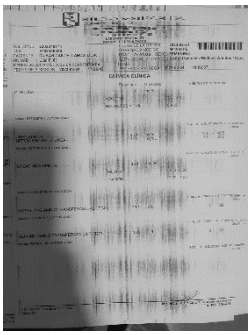


Fig. 4 "Img3".

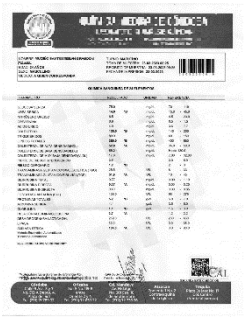


Fig. 5 "Img4".

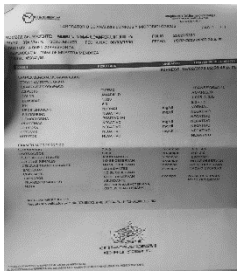
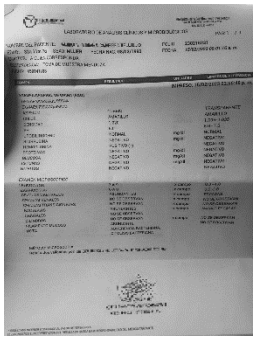


Fig. 6 "Img5".



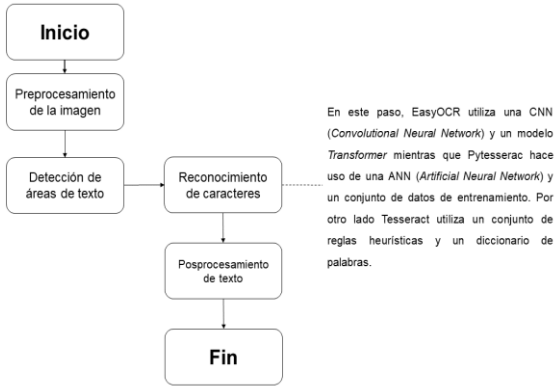
* Por solicitud expresa de los laboratorios clínicos (públicos y privados) algunos datos de las imágenes utilizadas estan difuninados.

Como se observa en la tabla 2 los tres algoritmos ofrecen soporte a múltiples idiomas. Sin embargo, EasyOCR y Pytesseract presentan mayor velocidad en la extracción en comparación con Tesseract siendo Pytesserac el algoritmo que presenta mayor velocidad de extracción en las cinco imágenes utilizadas como prueba.

Además, durante las pruebas realizadas EasyOCR no fue capaz de soportar el formato de archivo PDF, de igual forma Tesseract no tuvo la capacidad de procesar la imagen número tres. Por otro lado, Pytesseract procesó las cinco imágenes designadas para prueba, además del archivo PDF.

Finalmente, la figura 7 presenta los pasos realizados por los algoritmos de OCR para llevar a cabo la extracción, siendo el paso de "Reconocimiento de caracteres" donde cada algoritmo implementa diferentes técnicas para realizar este paso.

Fig. 7 Diagrama del proceso realizado por los algoritmos de OCR para la extracción de caracteres.

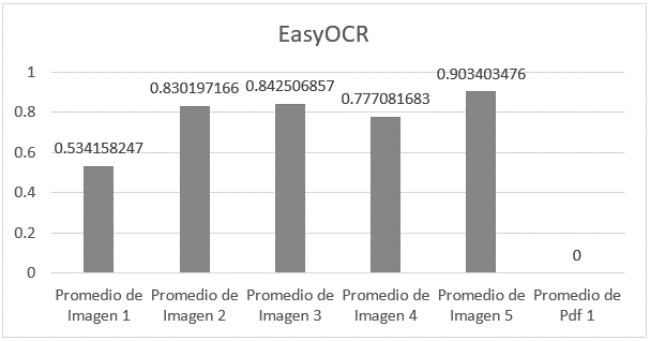


2.4 Resultados

Una vez comparadas las características, se seleccionó el nivel de confianza de cada carácter extraído para realizar una sumatoria del conjunto de cada archivo con el fin de obtener el promedio del nivel de confianza por cada imagen.

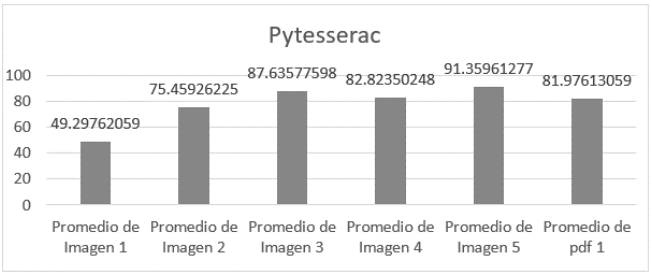
En la figura 8 se presenta el promedio de nivel de precisión obtenido para cada archivo utilizando EasyOCR.

Fig. 8 Gráfica del promedio de nivel de precisión para cada archivo procesado con EasyOCR



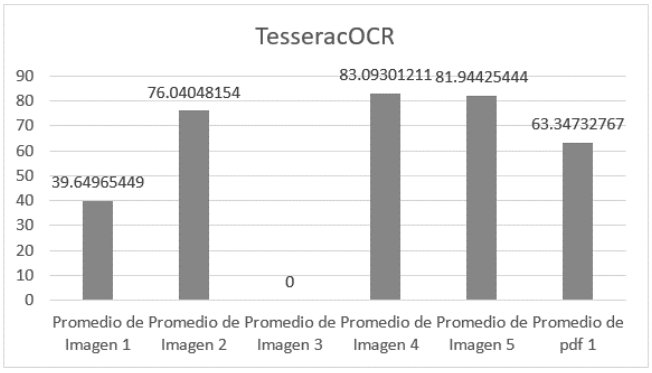
De igual forma en la figura 9 se muestran los promedios de nivel de precisión obtenidos con Pytesserac.

Fig. 9 Gráfica del promedio de nivel de precisión para cada archivo procesado con Pytesserac.



Además, en la figura 10 se presentan los promedios de nivel de precisión obtenidos con Tesseract.

Fig. 10 Gráfica del promedio de nivel de precisión para cada archivo procesado con Tesserac.



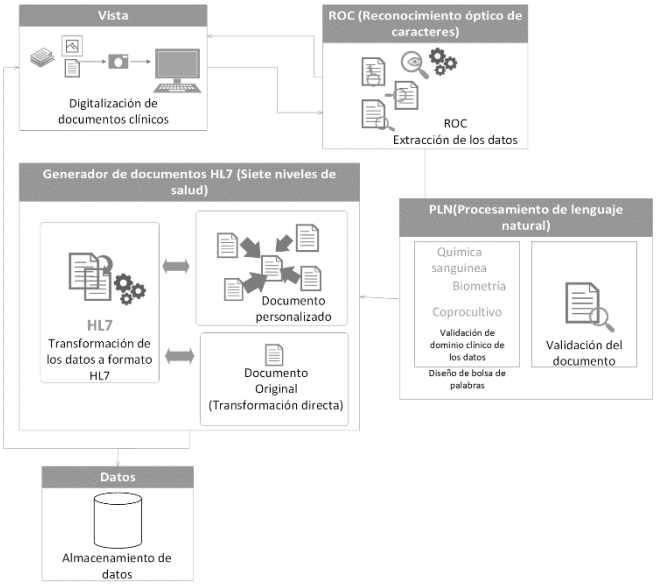
De acuerdo a los resultados obtenidos en las diferentes pruebas realizadas a los tres algoritmos de OCR, se encontró que el mayor promedio de nivel de confianza alcanzado fue por Pytesseract con un promedio de nivel confianza de 91.35961277 para la imagen cinco. Por otro lado, EasyOCR tuvo 0.903403476 como mejor pro-medio de nivel de confianza siento este para la imagen cinco. En cambio, Tesseract tuvo un mejor desempeño con la imagen cuatro obteniendo el promedio de nivel de confianza de 83.09301211.

2.5 Arquitectura

En esta sección se presenta la arquitectura para desarrollar un módulo de reconocimiento óptico de caracteres para interpretar documentos clínicos al estándar HL7 utilizando técnicas de procesamiento del lenguaje natural. La arquitectura (figura 11) presenta un diseño basado en capas que facilitan su mantenimiento, cada capa incluye los componentes que intervienen para su funcionamiento y son brevemente descritos a continuación.

Capa de vista: Esta capa muestra la interface grafica del módulo la cual permite la interacción del usuario con el sistema como es la manipulación de documentos digitales. Para lo cual es necesario enviar a procesar los documentos en alguno de los siguientes formatos: JPG, PNG o PDF.

Fig. 11 Arquitectura del módulo de reconocimiento óptico de caracteres para la interpretación de documentos clínicos al estándar HL7 utilizando técnicas de procesamiento del lenguaje natural.



Capa de vista: Esta capa muestra la interface grafica del módulo la cual permite la interacción del usuario con el sistema como es la manipulación de documentos digitales. Para lo cual es necesario enviar a procesar los documentos en alguno de los siguientes formatos: JPG, PNG o PDF.

Capa de ROC (Reconocimiento óptico de caracteres): Esta capa contiene los componentes que realizan el reconocimiento óptico de caracteres el cual consiste en los siguientes pasos: escaneo óptico, segmentación de la ubicación, preprocesamiento, segmentación, representación, extracción de características, entrenamiento y reconocimiento, posprocesamiento y el texto de salida. Cabe resaltar que el proceso de ROC se aplica a una o varias secciones del archivo especificado con el objetivo de obtener la información más relevante de cada documento.

Capa PNL (Procesamiento del lenguaje natural): Esta implica la elaboración de una bolsa de palabras con términos clínicos. Adicionalmente, en esta capa se utilizan los datos extraídos mediante el ROC para validar con la bolsa de palabras que la terminología extraída con el ROC, semánticamente corresponda a términos clínicos.

Por otro lado, en esta capa también se lleva a cabo la validación del documento a través de técnicas de PNL con el fin de validar que este no haya sido alterado por el usuario.

Capa HL7: En esta capa se integran los componentes responsables de la interpretación de los términos presentes en la bolsa de palabras al estándar HL7 con el objetivo de facilitar el intercambio de información entre sistemas de salud.

Capa de datos: Después del proceso de interpretación de los términos a HL7 los datos clínicos se mantienen de manera persistente en un SGBD (Sistema Gestor de Base de Datos).

La arquitectura propuesta en la figura 5 es una contribución al sector salud, ya que se espera contribuya en el proceso de interoperabilidad entre sistemas de salud, ayudando al

personal especializado en la materia de salud a intercambiar información clínica del paciente de una manera más eficiente a través de la unificación los diferentes formatos clínicos bajo un estándar.

3. CONCLUSIONES Y RECOMENDACIONES

Al investigar el estado del arte se llegó a la conclusión que existe una gran oportunidad para aplicar estudios relacionados con la interoperabilidad entre sistemas de salud ya que es una necesidad cada vez más grande entre los sistemas de información. Del mismo modo, al analizar diferentes algoritmos OCR nos permitió definir una arquitectura para el reconocimiento óptico de caracteres para la interpretación de documentos clínicos al formato HL7 a través de tecnologías semánticas que se espera contribuya a generar diversos documentos clínicos bajo un formato para intercambiar o compartir información de forma segura.

Como trabajo futuro se espera que el presente trabajo contribuya a la integración de los sistemas y servicios de salud electrónicos para facilitar la disponibilidad de los datos de manera rápida, independientemente de la ubicación espacial de la información.

3.1 Agradecimientos

Agradecemos al COVEICYDET por apoyar este trabajo a través del proyecto Prevención y Detección Temprana de Enfermedades Cardiovasculares (Arritmias y Taquicardias) utilizando Técnicas de Aprendizaje Automático, Big Data e Internet de las Cosas (identificador número 12 1806). Los autores están muy agradecidos con el Tecnológico Nacional de México (TNM) por apoyar este trabajo. Asimismo, este trabajo de investigación fue patrocinado por el Consejo Nacional de Ciencia y Tecnología (CONACYT), así como por la Secretaría de Educación Pública (SEP) a través del PRODEP.

4. REFERENCIAS

[1] J. D. Trigo, Ó. J. Rubio, M. Martínez-Espronceda, Á. Alesanco, J. García, and L. Serrano-Arriezu, “Building Standardized and Secure Mobile Health Services Based on Social Media,” *Electronics* , vol. 9, no. 12. 2020. doi: 10.3390/electronics9122208.

[2] S. K. Mukhiya, F. Rabbi, V. K. I Pun, A. Rutle, and Y. Lamo, “A GraphQL approach to Healthcare Information Exchange with HL7 FHIR,” *Procedia Comput Sci*, vol. 160, pp. 338–345, 2019, doi: https://doi.org/10.1016/j.procs.2019.11.082.

[3] A. Kiourtis, S. Nifakos, A. Mavrogiorgou, and D. Kyriazis, “Aggregating the syntactic and semantic similarity of healthcare data towards their transformation to HL7 FHIR through ontology matching,” *Int J Med Inform*, vol. 132, p. 104002, 2019, doi: https://doi.org/10.1016/j.ijmedinf.2019.104002.

[4] V. Kilintzis, I. Chouvarda, N. Beredimas, P. Natsiavas, and N. Maglaveras, “Supporting integrated care with a flexible data management framework built upon Linked Data, HL7 FHIR and ontologies,” *J Biomed Inform*, vol. 94, p. 103179, 2019, doi: https://doi.org/10.1016/j.jbi.2019.103179.

[5] E. Yehia, H. Boshnak, S. AbdelGaber, A. Abdo, and D. S. Elzanfaly, “Ontology-based clinical information extraction from physician’s free-text notes,” *J Biomed Inform*, vol. 98, p. 103276, 2019, doi: https://doi.org/10.1016/j.jbi.2019.103276.

[6] S. N. Laique *et al.*, “Application of optical character recognition with natural language processing for large-scale quality metric data extraction in colonoscopy reports,” *Gastrointest Endosc*, vol. 93, no. 3, pp. 750–757, 2021, doi: https://doi.org/10.1016/j.gie.2020.08.038.

[7] H. Goodrum, K. Roberts, and E. v Bernstam, “Automatic classification of scanned electronic health record documents,” *Int J Med Inform*, vol. 144, p. 104302, 2020, doi: https://doi.org/10.1016/j.ijmedinf.2020.104302.

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