EDUCACIÓN

SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de Orizaba

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

OPCIÓN I.- TESIS

TRABAJO PROFESIONAL

"Desarrollo de un módulo de reconocimiento óptico de caracteres para la interpretación de documentos clínicos a formato HL7 utilizando técnicas de procesamiento del lenguaje natural"

QUE PARA OBTENER EL GRADO DE:

MAESTRO EN SISTEMAS COMPUTACIONALES

PRESENTA:

I.T.I. Irving Jesús Ramírez Alacio

DIRECTOR DE TESIS:

Dr. José Luis Sánchez Cervantes

CODIRECTOR DE TESIS:

Dr. Isaac Machorro Cano

ORIZABA, VERACRUZ, MÉXICO. **MARZO 2024**

Orizaba, Veracruz, 12/marzo/2024
Dependencia: **División de Estudios de Posgrado e Investigación**
Asunto: **Autorización de Impresión**
OPCION: I

C. IRVING JESÚS RAMÍREZ ALACIO
CANDIDATO A GRADO DE MAESTRO EN:
SISTEMAS COMPUTACIONALES
P R E S E N T E.-

De acuerdo con el Reglamento de Titulación vigente de los Centros de Enseñanza Técnica Superior, dependiente de la Dirección General de Institutos Tecnológicos de la Secretaría de Educación Pública y habiendo cumplido con todas las indicaciones que la Comisión Revisora le hizo respecto a su Trabajo Profesional titulado:

" Desarrollo de un módulo de reconocimiento óptico de caracteres para la interpretación de documentos clínicos a formato HL7 utilizando técnicas de procesamiento del lenguaje natural"

comunico a Usted que este Departamento concede su autorización para que proceda a la impresión del mismo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
CIENCIA – TÉCNICA - CULTURA®

DRA. OFELIA LANDETA ESCAMILLA
JEFA DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

EDUCACIÓN SECRETARÍA DE EDUCACIÓN PÚBLICA | TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLOGICO DE ORIZABA
DIV. DE EST. DE POSGRADO E INVEST.

OG-13-F06

Orizaba, Veracruz, 29/enero/2024
Asunto: **Revisión de trabajo escrito**

C. CUAUHTÉMOC SÁNCHEZ RAMÍREZ
JEFE DE LA DIVISIÓN DE ESTUDIOS
DE POSGRADO E INVESTIGACIÓN
P R E S E N T E.-

Los que suscriben, miembros del jurado, han realizado la revisión de la Tesis del (la) C.

IRVING JESÚS RAMÍREZ ALACIO

La cual lleva el título de:

Desarrollo de un módulo de reconocimiento óptico de caracteres para la interpretación de documentos clínicos a formato HL7 utilizando técnicas de procesamiento del lenguaje natural

Y concluyen que se acepta.

ATENTAMENTE
Excelencia en Educación Tecnológica®
CIENCIA – TÉCNICA - CULTURA®

PRESIDENTE:	DR. JOSÉ LUIS SÁNCHEZ CERVANTES	FIRMA
SECRETARIO:	M.C.E BEATRIZ ALEJANDRA OLIVARES ZEPAHUA	FIRMA
VOCAL:	DR. GINER ALOR HERNÁNDEZ	FIRMA
VOCAL SUP.:	DR. ISAAC MACHORRO CANO	FIRMA

TA-09-18

CARTA DE ORIGINALIDAD

En la ciudad de Orizaba, Veracruz, el día 19 del mes de marzo del año 2024, el (la) que suscribe Ramírez Alacio Irving Jesús, alumno (a) del programa de Maestría en Sistemas Computacionales con número de control M21011180, manifiesta que es autor (a) del trabajo de tesis titulado "Desarrollo de un módulo de reconocimiento óptico de caracteres para la interpretación de documentos clínicos a formato HL7 utilizando técnicas de procesamiento del lenguaje natural" y declara que el trabajo es original, ya que su contenido es producto de su directa contribución intelectual. Todos los datos y las referencias a materiales ya publicados están debidamente identificados con su respectivo crédito e incluidos en las notas bibliográficas y en las citas que se destacan como tal y, en los casos que así lo requieran, se tienen las debidas autorizaciones de quienes poseen los derechos patrimoniales. Por lo tanto, se hace responsable de cualquier litigio o reclamación relacionada con derechos de propiedad intelectual, exonerando de toda responsabilidad al Tecnológico Nacional de México / Instituto Tecnológico de Orizaba.

Irving J. Ramírez Alacio

Nombre y firma autógrafa

CARTA DE CESIÓN DE DERECHOS

En la ciudad de Orizaba, Veracruz, el día 19 del mes de marzo del año 2024, el (la) que suscribe Ramírez Alacio Irving Jesús alumno (a) del programa de Maestría en Sistemas Computacionales con número de control M21011180, manifiesta que es autor (a) intelectual del trabajo de tesis bajo la dirección del Dr. José Luis Sánchez Cervantes y Dr. Isaac Machorro Cano ceden los derechos del trabajo intitulado "Desarrollo de un módulo de reconocimiento óptico de caracteres para la interpretación de documentos clínicos a formato HL7 utilizando técnicas de procesamiento del lenguaje natural" al TecNM/Instituto Tecnológico de Orizaba para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y del director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección: msc@orizaba.tecnm.mx . Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Irving Jesús Ramírez Alacio

Nombre y firma

Agradecimientos

En primer lugar, deseo expresar mi agradecimiento a Dios y a la Santísima Virgen María de Juquila, quienes me permitieron conocer y vivir la experiencia de los estudios de posgrado, ya que, sin mi fe, nada hubiera sido posible.

De igual forma agradezco al Tecnológico Nacional de México – Instituto Tecnológico de Orizaba por darme la oportunidad de ingresar y formar parte de la Maestría en Sistemas Computacionales. Asimismo, agradezco al Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT), por el apoyo económico recibido para la realización de mis estudios.

Un trabajo de investigación siempre es el resultado de un esfuerzo constante. En este caso, deseo expresar mi más sincero agradecimiento al Dr. José Luis Sánchez Cervantes. Reconozco su amabilidad y confianza al permitirme abordar su tema de tesis, así como su tiempo y paciencia. Sé que mi desempeño no fue el mejor y que en muchas ocasiones me equivoqué, cometí errores, pero él siempre estuvo dispuesto a apoyarme con la mejor actitud, lo cual, para mí, fue una de las mayores lecciones que aprendí: mantener una buena actitud ante los desafíos que presenta la vida. De igual forma, agradezco a mi codirector de tesis, el Dr. Isaac Machorro Cano, por su paciencia, apoyo y guía en la realización de este trabajo profesional.

De manera especial, agradezco a cada uno de los profesores que conforman el Consejo de la Maestría en Sistemas Computacionales por todas sus enseñanzas a lo largo de este camino. Puedo decir con toda seguridad que me llevo mucho de ellos.

A mis padres, Lorenzo Ramírez Juárez y Lucía Alacio López, extiendo mi más profundo agradecimiento. Aunque el camino hacia una maestría les era desconocido, no dudaron ni un instante en brindarme su apoyo incondicional para alcanzar mis sueños y metas. Con todo mi corazón, valoro y reconozco el inmenso esfuerzo y sacrificio que han hecho para permitirme llegar a este importante punto de mi vida. De igual forma, mi gratitud hacia Roberto Contreras Heredia es inmensurable. Su compañía, soporte inquebrantable en los momentos más desafiantes, y sus ánimos han sido un pilar fundamental en mi vida. Su presencia ha sido un regalo que trasciende cualquier agradecimiento posible.

Índice general

Índice de figuras

Índice de tablas

Índice de listas

Resumen

En el ámbito de la salud, la necesidad de intercambiar información se ha vuelto cada vez más crítica e imprescindible. El intercambio de datos y la utilización de la información intercambiada son fundamentales, no solo para el personal de salud que busca acceder al historial clínico del paciente de manera oportuna y eficiente, sino también para los propios pacientes. Esta necesidad se vuelve aún más apremiante a medida que los sistemas de salud de todo el mundo se acercan a una representación de gestión de salud basada en el valor. En este contexto, la interoperabilidad se convierte en un desafío clave y requiere expandir sus límites para garantizar un intercambio efectivo de información.

Como respuesta a esta creciente necesidad, surge la demanda de contar con una herramienta que facilite el proceso de interoperabilidad entre los sistemas de información clínica. En este proyecto, se presenta una propuesta de solución que consiste en el desarrollo e implementación de un módulo de reconocimiento óptico caracteres. Este módulo tiene la capacidad de extraer información de documentos clínicos y, posteriormente, interpretarla al estándar HL7. De esta manera, se logra una estandarización de la información clínica, unificando los datos bajo un mismo formato. Este enfoque busca abordar los desafíos que plantea la interoperabilidad en el sector de la salud, promoviendo una gestión más eficiente y efectiva de la información clínica.

Abstract

In healthcare, the need to exchange information has become increasingly critical and essential. The exchange of data and the utilization of the information exchanged are critical, not only for healthcare providers seeking to access patient records in a timely and efficient manner, but also for patients themselves. This need becomes even more pressing as health systems around the world move toward a value-based representation of health management. In this context, interoperability becomes a key challenge and requires expanding its boundaries to ensure effective information exchange.

In response to this growing need, there is a demand for a tool to facilitate the process of interoperability between clinical information systems. In this project, a proposed solution is presented that consists of the development and implementation of an optical character recognition module. This module has the capacity to extract information from clinical documents and, subsequently, to interpret it into the HL7 standard. In this way, a standardization of clinical information is achieved, unifying the data under the same format. This approach seeks to address the challenges posed by interoperability in the health sector, promoting a more efficient and effective management of clinical information.

Introducción

En los últimos años, las Tecnologías de la Información y Comunicación (TIC) han demostrado su versatilidad y utilidad en diversos sectores, abarcando desde empresas hasta instituciones educativas y, de manera significativa, el ámbito de la salud. Las TIC engloban un conjunto de herramientas esenciales que facilitan el procesamiento y la transferencia de información, desempeñando un papel transformador en la sociedad.

Dentro del vasto campo de las TIC, las aplicaciones Web han emergido como un elemento crucial en el nuevo paradigma de gestión de salud basada en estándares de *E-health* que se adopta en todo el mundo. En este contexto, la interoperabilidad se posiciona como un elemento esencial para lograr una atención médica eficiente y efectiva. La capacidad de intercambiar información de manera fluida y segura se convierte en un requisito crítico para todos los actores involucrados en la gestión y el cuidado de la salud, incluyendo hospitales, pacientes, laboratorios clínicos y médicos.

El presente proyecto de tesis se enfoca en abordar el desafío de interoperabilidad en el contexto de la gestión de la salud. A lo largo de este documento, se desarrollan cinco capítulos que comprenden: el primero aborda el marco teórico, el planteamiento del problema, los objetivos (tanto generales como específicos) y la justificación de la investigación. El segundo se centra en el estado de la práctica, donde se presentan trabajos relacionados con el proyecto de tesis, se describen las tecnologías seleccionadas y se explica la metodología a seguir en el desarrollo del proyecto. El tercero detalla las actividades realizadas durante cada una de las fases de la metodología especificada. En el cuarto, se presentan los resultados obtenidos al implementar este proyecto. Finalmente, en el quinto se ofrecen las conclusiones y recomendaciones derivadas de esta investigación, con el objetivo de fortalecer la interoperabilidad en el sector de la salud y mejorar la atención médica en general.

Capítulo 1 Antecedentes

En este primer capítulo, se presentan los conceptos principales de esta tesis con el objetivo de facilitar su comprensión, de igual manera se aborda una descripción del problema detectado, así como los objetivos (general y específicos) que dan respuesta a la problemática del presente trabajo.

1.1 Marco teórico

A continuación, se realiza una descripción de los conceptos relacionados con el tema de investigación.

1.1.1 Inteligencia Artificial

Como presenta [1] el concepto de IA (Inteligencia Artificial) se comenzó a emplear tres décadas atrás. Donde se explicó que la inteligencia artificial se basa en la imitación de la inteligencia humana en una máquina, con esto la máquina tendrá la capacidad de identificar y usar el conocimiento necesario para la resolución de problemas. Otra definición de la IA hace referencia a la capacidad de los programas informáticos para operar en la forma que lo hace el pensamiento humano, otorgando al programa la capacidad de reconocer y aprender.

Entre las características más importantes de la IA se encuentran los múltiples campos de aplicación tales como la robótica, la traducción de textos, el reconocimiento y aprendizaje de palabras y los sistemas expertos, los cuales imitan el comportamiento humano [2].

A continuación, se describen algunas áreas de investigación de mayor importancia en la IA:

- La representación del conocimiento que tiene por objetivo el encontrar técnicas expresivas para representar información acerca del mundo real.
- El desarrollo de algoritmos que se construyen y ejecutan sin ningún tipo de intervención humana.
- El desarrollo de ontologías que buscan la construcción de catálogos de

conocimiento.

- El análisis e identificación de imágenes y textos que permite el desarrollo de algoritmos para estructurar y presentar información.

1.1.2 Procesamiento del Lenguaje Natural (NLP)

El NLP (*Natural Language Processing*) es una rama de la IA, la cual permite a las Tecnologías de la Información (TI) y a la lingüística computacional tener la capacidad de procesar el lenguaje natural humano. Por otro lado, el NL (*Natural Language*) es toda forma que poseen los seres humanos de expresar sentimientos, pensamientos o emociones a través de un código estructurado conocido como lenguaje. Este generalmente se forma de lo que se lee, escribe, escucha o se habla, dando como resultado una fuente de lenguaje natural. Con apoyo en lo anterior se define el NLP como el tratamiento automatizado y/o semiautomatizado del lenguaje natural humano, de acuerdo a [3] el NLP se conforma de los siguientes elementos:

- NLU (*Natural Language Understanding*) se refiere al proceso de transformar datos de entrada NL a una representación electrónica a través del uso de alguna herramienta informática.
- NLG (*Natural Language Generation*) se refiere al proceso de transformar datos de salida de una máquina a NL.

Diferencias entre el NLU y NLG

La tabla 1.1 muestra las principales diferencias entre NLU y NLG.

Tabla 1.1 Diferencias entre NLU y NLG.

NLU	NLG
El NLU permite la comprensión de NL a partir de texto escrito o sonidos de voz.	El NLG permite obtener NL a partir de información de salida de máquinas.
Es el proceso de lectura e interpretación del lenguaje.	Es el proceso de escribir o generar lenguaje.

En [3] se mencionan algunos ejemplos de empresas que utilizan el NLP en sus productos son Google® con Google Asistant y Apple® con Siri y Amazon® con Alexa.

1.1.2.1 Categorías de procesamiento de lenguaje natural

A continuación, se explorarán las categorías de procesamiento de lenguaje natural para obtener una visión más completa de este campo de estudio.

- Reconocimiento óptico de caracteres: es el proceso de identificar y extraer información de documentos escaneados, imágenes de la cámara y PDF(s) (*Portable Document Format*) lo que permite reutilizar la información [4] .
- Reconocimiento de voz: es la tarea que lleva a cabo un programa computacional para identificar la voz humana, procesarla y convertirla en datos [5].
- Traducción automática: es un proceso que se apoya en el aprendizaje automático, consiste en convertir determinada cantidad de texto de un idioma de origen a otro idioma de destino de manera automática [6].
- Generación de lenguaje natural: es el proceso que realizan las máquinas para generar una respuesta textual en lenguaje natural a partir de datos de entrada [7].
- Análisis de sentimientos: consiste en analizar y procesar el NL a partir de textos de diferentes orígenes para identificar y obtener información representativa [8].
- Búsqueda semántica: es un grupo de particularidades que optimizan la calidad de los resultados de búsqueda. Como primer paso, agrega una clasificación secundaria en un grupo de resultados inicial, generando los resultados de mayor importancia desde un enfoque semántico. En segundo paso se extraen y se devuelven subtítulos y respuestas en la respuesta principal [9].
- Aprendizaje automático: es una la de las ramas de la IA, consiste en enseñar a un sistema a partir de los datos con los que interactúa, generalmente se trata de un algoritmo que consume datos de entrenamiento [10].
- Programación en lenguaje natural: consiste en interpretar automáticamente código basado en datos de entrada expresados en NL en código ejecutable para una máquina [11].

1.1.2.2 Reconocimiento Óptico de Caracteres (ROC)

El OCR (por sus siglas del inglés *Optical Character Recognition*) o Reconocimiento Óptico de Caracteres, de acuerdo a lo descrito por [12] forma parte del grupo de técnicas de identificación automática, lo cual consiste en el proceso de reconocimiento de objetos de manera automatizada en donde se acumulan datos pertenecientes a los objetos identificados y posteriormente se envían a un sistema informático de manera automática. El OCR es el proceso por el cual se clasifican los patrones ópticos presentes en imágenes digitales que pertenecen a caracteres alfanuméricos o de otro tipo. Para que el OCR se lleve a cabo se requiere el realizar algunas fases o pasos de gran importancia como son la segmentación, extracción de características y la clasificación. A continuación, se describen algunos de los elementos que conforman el proceso de OCR para la identificación, extracción y clasificación de caracteres.

Escaneo óptico (*Optical Scanning*)

Después de la entrada de texto o datos es el primer elemento del OCR, consiste en la captura de la imagen del documento original, la cual se lleva a cabo con ayuda de escáneres ópticos, los cuales están conformados por el mecanismo de transporte y el dispositivo de detección el cual transforma la intensidad de luz en diferentes niveles de color gris.

Un ejemplo de esto se encuentra en los documentos impresos, donde la mayoría del contenido consiste en texto en color negro sobre un fondo blanco, lo que origina una imagen en niveles de gris. Al aplicar el proceso de OCR, esta imagen se transforma en una imagen binaria, un proceso conocido como umbralización (thresholding). Este proceso se realiza en el escáner con el objetivo de optimizar el uso de recursos computacionales y evitar su desperdicio.

Por otro lado, en la actualidad existen documentos con una gama bastante amplia de características, es decir, no se limitan únicamente a texto negro sobre fondo blanco. Para abordar estos casos, se requieren métodos de umbralización más sofisticados a fin de obtener el resultado deseado [12].

Segmentación de la ubicación (*Location Segmentation*)

Una vez terminado el proceso de escaneo óptico, lo siguiente es realizar la

segmentación, la cual identifica los elementos presentes en la imagen, ya que es preciso localizar las áreas del documento que realmente tienen datos impresos y que se diferencian de figuras, gráficos, entre otras.

Al aplicar segmentación al texto este separa los caracteres o las palabras. Una gran parte de los algoritmos de OCR separa las palabras en caracteres individuales porque por lo general la segmentación se lleva a cabo separando cada elemento conectado [12].

Preprocesamiento (*Pre-processing*)

Después de concluir la etapa de segmentación de la ubicación, se continúa con el siguiente paso del proceso de OCR, denominado preprocesamiento. En esta fase, los datos se someten a una serie de pasos destinados a mejorar su calidad y facilitar su uso en el análisis de caracteres. Esto es esencial debido a que los caracteres a menudo se encuentran sucios o dañados, lo que puede generar ruido en la imagen resultante del proceso de escaneo.

Es por lo anterior que se obtiene un índice de reconocimiento bajo, por lo cual durante el preprocesamiento se elimina el ruido presente en la imagen por medio del suavizado de caracteres digitalizados, el suavizado consiste en el relleno y el adelgazamiento de caracteres. El relleno consiste en la eliminación de los huecos y cortes en los caracteres, mientras que el adelgazamiento minimiza la anchura de la línea que forma los caracteres.

El preprocesamiento también se conforma del subproceso de normalización, el cual se observa en la figura 1.1 donde se transforman los caracteres para que tengan un tamaño, inclinación y rotación uniforme [12].

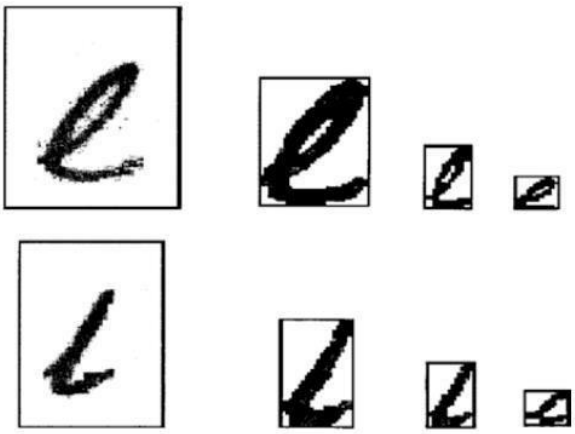


Figura 1.1 Subproceso de normalización de caracteres.

Fuente: A. Chaudhuri, K. Mandaviya, P. Badelia, and S. K. Ghosh, “Studies in Fuzziness and Soft Computing Optical Character Recognition Systems for Different Languages with Soft Computing.” [Online]. Available: http://www.springer.com/series/2941.

Segmentación (*Segmentation*)

Con la etapa de preprocesamiento se obtiene una imagen digital de los caracteres de alta calidad, es decir, contiene una cantidad apropiada de información y un porcentaje bajo de ruido. Con lo cual se da paso a la etapa de segmentación donde la imagen se divide en subcomponentes.

La segmentación se considerada una de las etapas de mayor importancia dentro del proceso de OCR porque el porcentaje de división de las diferentes líneas que forman los caracteres afecta al reconocimiento. Por esta razón se utiliza segmentación interna, la cual es el aislamiento de las líneas y curvas de los caracteres escritos de forma cursiva.

La segmentación de caracteres se divide en tres categorías:

- Segmentación explícita: reconoce los segmentos en función de las cualidades del carácter. El proceso de fragmentar la imagen en componentes significativos es por medio de la disección. La disección analiza la imagen del carácter sin hacer uso de una clase propia de información de forma.
- Segmentación implícita: busca en la imagen o imágenes los componentes que concuerden con las clases predefinidas. La segmentación se realiza a través de

la confianza del reconocimiento, lo que contiene la corrección sintáctica o semántica del resultado completo.

- Estrategias mixtas: se trata de la combinación de segmentación explícita e implícita. Para ello, utiliza un algoritmo de disección en la imagen del carácter con el propósito de dividir la imagen en un número determinado, con el objetivo de incluir los límites de segmentación correctos dentro de los cortes realizados. Una vez concluido esto, se busca la segmentación más óptima por medio de la evaluación de subconjuntos de los cortes realizados, donde cada subconjunto involucra una hipótesis de segmentación y clasificación [12].

Representación (*Representation*)

La representación junto a la segmentación forma parte de las fases de mayor importancia en el proceso de OCR. Por lo cual busca una representación más compacta. Para lograr esto, se extraen un conjunto de características para cada clase, las cuales ayudan a distinguirlas de otras clases y, al mismo tiempo, se mantienen constantes con respecto a las diferencias entre las clases.

Los métodos que permiten la representación de imágenes de caracteres se agrupan en tres conjuntos, los cuales se describen a continuación.

- Transformación global y expansión de series: se representa una señal a través de una combinación lineal de una serie de funciones simples y precisas. Los coeficientes de la combinación lineal proveen una codificación compacta. Los métodos más utilizados de la transformación global y expansión de series en él proceso de OCR son: la transformada de Fourier, la transformada de Gabor, transformada wavelet, los momentos y la expansión de karhunen loeve.
- Representación estadística: es la representación de una imagen de carácter a través de una distribución estadística de puntos, se utilizan las variaciones de estilo. Por otro lado, este tipo de representación no admite la reconstrucción de la imagen original, sino que se utiliza para minimizar la dimensión del conjunto de características, proporcionando una alta velocidad y una baja complejidad. Las características estadísticas de mayor uso en la representación de caracteres son: zonificación, cruces y distancias y proyecciones.

- Representación geométrica: es tipo de representación utiliza características geométricas y topológicas que contengan tolerancia a las distorsiones y variaciones de estilo. De igual forma, esta representación codifica conocimiento sobre la estructura del objeto o proporciona conocimiento sobre qué tipo de elementos forman ese objeto. Las representaciones de este tipo se clasifican en: extracción y recuento de estructuras topológicas, medición y aproximación de las propiedades geométricas, grafos y árboles [12].

Extracción de características (*Feature Extraction*)

Como siguiente elemento se tiene a la extracción de características cuyo objetivo es capturar los rasgos principales de los símbolos identificados.

La manera más concreta de describir un carácter es mediante una imagen de trama real, por otro lado, existe otro enfoque que extrae los rasgos que describen a los símbolos, descartando a los atributos no relevantes.

Las técnicas de extracción de características se clasifican en los siguientes tres grupos: distribución de puntos, transformaciones y expansiones, así como análisis estructural. Estos se evalúan tomando en cuenta la sensibilidad al ruido, la deformación, y la facilidad de aplicación y uso.

Entre algunas de las técnicas de extracción de características más utilizadas se encuentran la coincidencia y correlación de plantillas, las transformaciones, la distribución de puntos y el análisis estructural. Estas técnicas se evalúan en función de su sensibilidad al ruido, su capacidad para manejar deformaciones, así como su facilidad de aplicación y uso [12].

Entrenamiento y reconocimiento (*Training and Recognition*)

En esta fase, el OCR hace uso de metodologías de reconocimiento de patrones que establecen una muestra desconocida a una clase predefinida.

El OCR se apoya en cuatro enfoques de reconocimiento de patrones, los cuales son: comparación de plantillas, técnicas estadísticas, técnicas estructurales y ANNs (*Artificial Neural Networks*). Estos enfoques no son independientes entre sí, ya que una técnica de OCR se considera miembro de otros enfoques.

Los enfoques anteriormente descritos hacen uso de estrategias holísticas o analíticas

para la etapa de entrenamiento y reconocimiento, puesto que la estrategia holística utiliza enfoques descendentes para reconocer los caracteres de forma completa, aunque esto genera que OCR posea un vocabulario limitado, además de que se agrega una mayor complejidad a la representación de un solo carácter.

Por otra parte, las estrategias analíticas utilizan un enfoque ascendente, comenzando por trazos y caracteres hasta llegar a la producción de textos significativos. Para ello, este tipo de estrategias necesita de algoritmos de segmentación explícitos o implícitos que, además de agregar complejidad, añaden errores de segmentación en el sistema de OCR.

No obstante, con el apoyo de la etapa de segmentación, la problemática descrita en el párrafo anterior se reduce al reconocimiento de caracteres simples o trazos aislados, los cuales se utilizan para un vocabulario ilimitado con un alto porcentaje de reconocimiento [12].

Posprocesamiento *(Posprocessing*)

La etapa de posprocesamiento se conforma por las actividades de agrupación y la detección y corrección de errores.

Por un lado, la agrupación consiste en la asociación de símbolos individuales entre sí, formando palabras y números. La agrupación de símbolos en cadenas se apoya en la posición de los símbolos dentro del documento, los símbolos que se encuentran los suficientemente cerca se agrupan, tal es el caso de los tipos de letra con tono fijo, ya que en ellos el proceso de agrupación es sencillo, puesto que la posición de cada carácter se conoce [12].

La figura 1.2 muestra el orden que lleva cada uno de los elementos dentro del proceso de reconocimiento óptico de caracteres.

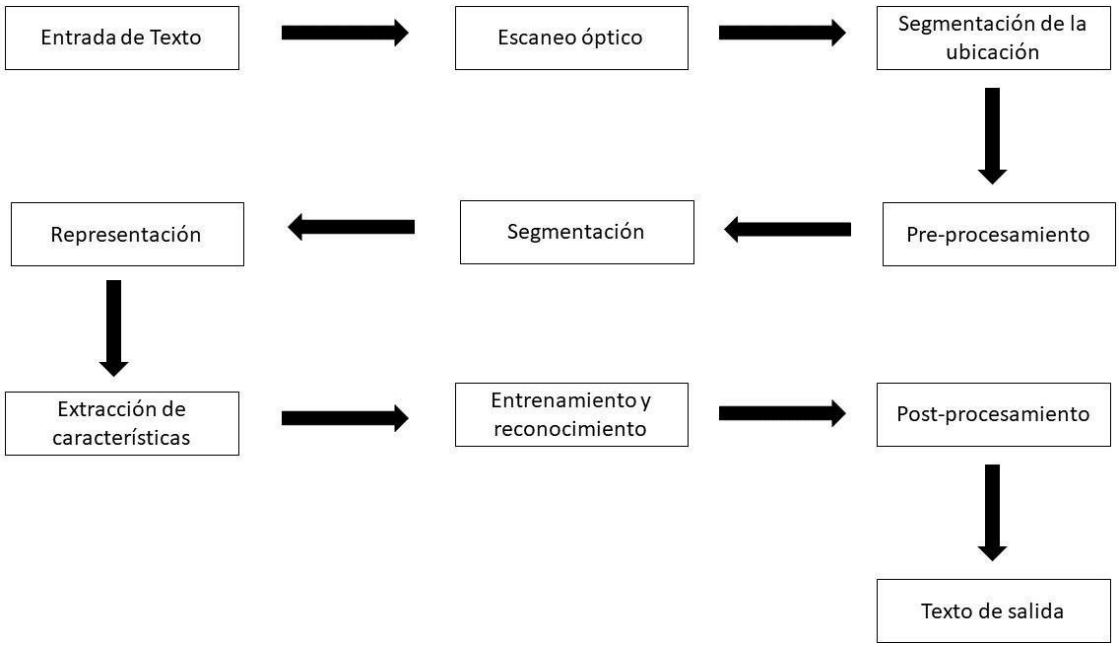


Figura 1.2 Etapas del proceso de reconocimiento óptico de caracteres.

Fuente: A. Chaudhuri, K. Mandaviya, P. Badelia, and S. K. Ghosh, "Studies in Fuzziness and Soft Computing Optical Character Recognition Systems for Different Languages with Soft Computing." [Online]. Available: http://www.springer.com/series/2941.

1.1.2.3 Bolsa de palabras (BoW)

El modelo de bolsa de palabras (BoW por sus siglas del inglés *Bag of Words*) como lo presenta [13] es un método popular de presentación de documentos basado en la premisa de que "a frecuencia de las palabras en un párrafo es una medida útil de la importancia de las palabras".

El modelo bolsa de palabras no tiene en cuenta la gramática ni el orden de las palabras: por ejemplo, dos oraciones son semánticamente diferentes. Por ejemplo "Es una chica bonita, ¿Verdad?", y "No es bonita, ¿Es una chica?". En el ejemplo anterior se considera el mismo texto. Entonces, en esto modelo, cada documento parece una "bolsa" que contiene algunas palabras del diccionario. En el caso más simple, solo se tiene en cuenta la presencia/ausencia de ciertas palabras en el documento, y el modelo resultante se denomina "binario". De manera que el modelo bolsa de palabras generalmente se utiliza en las palabras representativas, donde están

presentes las características más destacadas del documento y las agrupa en un conjunto denominado “bolsa”.

1.1.3 Expediente clínico electrónico

De acuerdo la norma mexicana NOM-024-SSA3-2012 [14] el expediente clínico electrónico, es el conjunto de información almacenada en medios electrónicos, centrada en el paciente, que documenta la atención médica prestada por profesionales de la salud con arreglo a las disposiciones sanitarias dentro de un establecimiento de salud.

Por otro lado, la revista [15] describe al expediente clínico (EC) por las siglas del español como el conjunto de datos médicos, clínicos, sistemáticos y detallados en forma ordenada, que permiten al profesional de la salud emitir un diagnóstico sindrómico y nosológico, con su posterior pronóstico, para más adelante llevar un registro del desarrollo de un tratamiento.

De igual forma, se mencionan ventajas del uso del expediente clínico electrónico, tales como: la disponibilidad de los datos de manera rápida, independiente de la ubicación espacial de la información, el utilizar un lenguaje estandarizado para mejorar comunicación entre profesionales de la salud, facilitando toma de decisiones de forma rápida, y la atención médica del paciente, ya que facilita el seguimiento de patrones de diagnóstico y tratamiento de enfermedades.

A continuación, en la tabla 1.2, se describen los elementos mínimos que contiene el expediente clínico electrónico de acuerdo a lo investigado por [16].

Tabla 1.2 Elementos básicos del expediente clínico electrónico.

A.	**Logística de atención a pacientes** • Registro. • Ingreso de pacientes ambulatorios. • Ingreso, egreso y transferencia de pacientes hospitalizados. • Programación de servicios y	**B.**	**Operación de servicios de apoyo técnico de diagnóstico y tratamiento** • Laboratorio clínico. • Imagenología médica de diagnóstico y de intervención. • Radioterapia.

gestión de citas. • Prescripción y solicitudes. **C. Gestión de datos clínicos** • Registros médicos. • Atención de enfermería. • Auditoría clínica.	• Farmacia. • **Transfusiones y banco de sangre.** • Servicio de Nutrición.

Fuente: R. Peña and S. López-Silva, "ANÁLISIS ESTRUCTURAL Y FUNCIONAL DEL EXPEDIENTE CLÍNICO ELECTRONICO," 2003, Accessed: Apr. 05, 2022. [Online]. Available: https://www.researchgate.net/publication/278678814.

De igual forma R. Peña y S. López Silva en [16] realizaron una clasificación de los datos utilizados en registros de salud generando los siguientes siete grupos como se presenta en la tabla 1.3.

Tabla 1.3 Ejemplo de grupos de datos utilizados en registros de salud.

Tipo de dato	**Ejemplo**
Datos codificados.	Diagnóstico, procedimientos, resultados de
Texto.	laboratorio.
Imágenes de documentos.	Radiología, informes de anatomía patológica,
Registros de señal biológica.	notas.
Objetos de voz.	Expedientes médicos escaneados
Imágenes sin movimiento.	Electrocardiogramas (ECG),
Video con movimiento total.	Informes dictados.

Fuente: R. Peña and S. López-Silva, "ANÁLISIS ESTRUCTURAL Y FUNCIONAL DEL EXPEDIENTE CLÍNICO ELECTRONICO," 2003, Accessed: Apr. 05, 2022. [Online]. Available: https://www.researchgate.net/publication/278678814.

1.1.3.1 Norma Oficial Mexicana NOM 168-SSA1-1998

La norma oficial mexicana define las pautas científicas, tecnológicas y administrativas obligatorias para la gestación, integración, uso y archivo de los documentos clínicos. La presente norma [17] establece que los proveedores de servicios de salud públicos

y privados deberán integrar y mantener registros clínicos conforme a lo dispuesto en esta norma. La organización es solidariamente responsable del cumplimiento de esta obligación en la medida en que sus empleados de servicio participen en ello.

De acuerdo con la norma mexicana se establece que todo expediente clínico, deberá tener los siguientes datos generales:

- Tipo de establecimiento.
- Nombre del establecimiento.
- Domicilio del establecimiento.
- Nombre de la institución a la que pertenece.
- La razón y denominación social del propietario o concesionario.

1.1.3 2 Norma Oficial Mexicana NOM-024-SSA3-2010

Esta norma oficial mexicana tiene por objetivo establecer las funcionales y las funciones que deben cumplir los productos de expediente clínico electrónico para garantizar la compatibilidad, el procesamiento, la interpretación, la confidencialidad, la seguridad y el uso de estándares y categorías de la información de la historia clínica electrónica.

La norma mexicana presentada en [18] dispone para las características de recolección de datos demográficos y de identidad del paciente, que estos deben registrarse utilizando un código o nomenclatura estandarizados, o registrarse como datos no estructurados, según su naturaleza. El personal autorizado ingresará los datos, dependiendo del servicio o campo médico que atienda al paciente. Esta norma instituye que los expedientes clínicos electrónicos están destinados a cumplir los siguientes propósitos en los servicios de atención médica:

- Consulta Externa.
- Hospitalización
- Urgencias
- Farmacia
- Laboratorio
- Imagenología

- Quirófano

La norma también sugiere que el usuario de sistemas de expediente clínico electrónico debe estar autenticado en el sistema mediante una firma digital simple, es decir, un nombre de usuario de más de 6 caracteres, una contraseña alfanumérica compuesta por números, letras minúsculas y mayúsculas.

Asimismo, en el apéndice normativo A, clave 3.2.1, referente a la funcionalidad de "Interoperabilidad basada en estándares", esta norma mexicana establece que se debe utilizar el estándar internacional HL7 *(Health Level Seven*) en su versión 3 o HL7 FHIR (*Fast Healthcare Interoperability Resources*) para fines de interoperabilidad.

De igual manera, en el apéndice normativo A, clave 3.2.2 funcionalidad "Estándares de intercambio de información" establece en los puntos a y b que se debe usar el estándar HL7 en su versión 3 y/o HL7 FHIR para el intercambio de información y que se debe intercambiar toda aquella información disponible definida en cada tipo de mensaje del estándar HL7.

1.1.3.3 Norma Oficial Mexicana NOM-024-SSA3-2012

Esta norma tiene como objetivo regular los sistemas de información de historias clínicas electrónicas y establecer mecanismos para el registro, intercambio y consolidación de información para los proveedores de servicios de salud en el sistema nacional de salud.

El Diario de la Federación Mexicana, publicado en [19] menciona que la mejor estrategia es establecer reglas y estándares que se apliquen a todas las soluciones tecnológicas que proporcionen "comunicación" o interacción entre diferentes sistemas; por lo cual, aunque el sistema de cada proveedor de servicios es diferente, todos utilizan el mismo lenguaje, garantizando en todo momento la confidencialidad y seguridad de la información de la historia clínica electrónica, de acuerdo con la normativa vigente.

Dentro de las especificaciones de esta norma se menciona que los lineamientos y formatos los cuales se apoyan en la arquitectura de referencia y los procedimientos proporcionados por la Secretaría de salud a través de la DGIS (Dirección General de Información en Salud). Esta estructura de referencia toma en cuenta estándares y

recomendaciones publicados internacionalmente por IHE (*Integrating the Healthcare Enterpris*), HL7, o estándares determinados por la propia Secretaría a través de la DGIS de acuerdo con los avances regulatorios y tecnológicos de la SNS (Servicio Nacional de Salud).

Otra especificación de la norma índica los estándares a considerar para la comunicación entre sistemas de salud son: HL7 CDA (*Health Level Seven Clinical Document Architecture*), HL7 en la versión 3.0, XML (*eXtensible Markup*) y/o un estándar aprobado según lo determine la Secretaría a través de la DGIS de conformidad con el reglamento de avances tecnológicos y SNS. Los proveedores de atención médica eligen los estándares que mejor se adapten a las necesidades de intercambio de su organización.

1.1.4 Análisis clínicos

De acuerdo a [20] un análisis clínico es un tipo de prueba exploratoria que consiste en la toma de muestras biológicas de un paciente y su examen en laboratorio para confirmar o descartar un diagnóstico realizado por un médico, detectar anomalías u obtener la información necesaria para aplicar un determinado tratamiento o cualquier otro procedimiento terapéutico.

Otra forma de referirse a un análisis clínico es como una prueba de laboratorio.

A continuación, se mencionan algunas de las finalidades que tienen los exámenes clínicos tales como:

- En la prevención de enfermedades, se recurre al empleo de análisis clínicos con el fin de detectar tempranamente posibles anomalías. Un ejemplo de ello es la realización de exámenes clínicos de rutina, que se utilizan para evaluar el estado de los biomarcadores y verificar si estos se encuentran dentro de los valores normales. En caso contrario, se toman medidas preventivas y correctivas con el propósito de evitar el desarrollo de enfermedades.
- La validación de diagnósticos médicos para corroborar la exactitud de los mismos. Dependiendo del tipo de prueba utilizada, los resultados se expresan de manera cualitativa, es decir, como positivos o negativos, o de manera cuantitativa, apoyándose en una cifra absoluta o en el nivel alcanzado en una

escala preestablecida.

1.1.4.1 Tipos de análisis clínicos

A continuación, se presentan los principales tipos de análisis clínicos de acuerdo a una clasificación realizada por el INER (Instituto Nacional de Enfermedades Respiratorias) como lo muestran en su página Web [21].

La tabla 1.4 muestra los análisis clínicos descritos y clasificados por el INER.

Tabla 1.4 Tipos de análisis clínicos clasificación INER.

Tipo de análisis clínico	
Química Clínica	Química de 5 elementos.
	Perfil hepático.
	Perfil de lípidos.
	Perfil pancreático.
	Hemoglobina glucosilada.
	Fructosamina.
	Curva de tolerancia a la glucosa depuración de creatinina.
	Electrólitos.
Hematología	Biometría hemática.
	Velocidad de sedimentación.
	Reticulocitos.
	Tiempos de Coagulación, Antitrombina III.
	Proteína C.
	Proteína S.
	Dímero D.
Líquidos orgánicos	ADA (Adenosín deaminasa).
	Citoquímicos.
Uroanálisis	EGO (Examen general de orina).
	Proteína en orina.
	Electrólitos.
Parasitología	Coproparasitoscópico.
	Sangre oculta en heces.

Tipo de análisis clínico	
Inmunología	Perfil reumático.
	Perfil inmunológico.
	Perfil tiroideo.
	Perfil de Torch.
	Marcadores Tumorales.
	ANA (Anticuerpos Antinucleares).
	ANCA (Anticuerpos anticitoplasmáticos).
	Reacciones febriles.
	VDRL (Venereal Disease Research Laboratory) Examen para detectar sífilis.
	Antígeno Aviario.
	Procalcitonina.

1.1.5 Estándar HL7 (*Health Level Seven*)

En [22] se presenta al estándar HL7® como un protocolo de la capa de aplicación del modelo de referencia OSI (*Open Systems Interconnection*) el cual se encuentra acreditado por la ANSI (*American National Standards Institute*), su principal uso es en el intercambio de datos clínicos y administrativos entre sistemas de salud.

En otras palabras, se define al estándar HL7 como un lenguaje que permite el intercambio de información entre sistemas de salud.

Dentro de los sistemas de salud existen tres estándares generales para cubrir la necesidad de integración, los cuales se mencionan en seguida: un estándar de transferencia de mensajes (¿Cómo se envía un mensaje?), un estándar de formato de mensaje (¿Cómo se verá un mensaje?) y un estándar de datos (¿Cómo se organiza la información clínica en un mensaje?). Siendo estos dos últimos estándares los que abarca HL7, ya que define estándares para el formato del mensaje y estándares para los datos del contenido del mensaje.

1.1.5.1 Estándar de formato de mensaje

El estándar de formato de mensaje especifica la estructura de un mensaje, lo que

permite al sistema receptor dar sentido a su contenido. Sin embargo, si los datos no siguen una estructura estándar, el sistema tendrá dificultades para localizar la información en el mensaje. En esta cuestión, en conjunto con el mensaje, el sistema emisor envía información sobre dónde buscar los datos y el significado de cada valor, gracias al uso del estándar de formato de mensaje se tiene un mensaje corto y directo, comprensible para los sistemas de salud como lo presenta [22] en su libro.

1.1.5.2 Estándar de datos

El estándar de datos descrito en [22], se utiliza para representar la información del mensaje, HL7 define un formato de datos estándar teniendo como ejemplo el siguiente caso donde dos sistemas requieren compartir información, por lo tanto, es necesario que ambos sistemas utilicen el estándar de datos HL7 para la mensajería. De esta forma se evitan problemas de ambigüedad con el contenido de los campos y evitarse el recurrir al tortuoso proceso de conciliación de campos cada vez que se integren dos sistemas de salud.

1.1.5.3 Versiones de HL7

En [22] se comenta que a partir de la publicación de la primera versión de HL7 en 1987, la organización HL7 International® busca verificaciones constantes del estándar.

Cada vez que surge una nueva versión de HL7, mejora mediante la adición de reglas, la actualización de las reglas existentes y la búsqueda de soluciones a los problemas planteados por las personas que han implementado el estándar. También se consideró el garantizar que cada nueva versión sea compatible con las versiones inferiores. Es decir, una nueva versión es capaz de comunicarse con sistemas que utilizan versiones anteriores.

Asimismo, HL7 busca garantizar que, con cada nueva versión, la estructura y el formato previos no se vean afectados. La versión 2.1 de HL7 fue la primera versión implementable y posterior a su lanzamiento tuvo ocho revisiones del estándar: v2.2, v2.3, v2.3.1, v2.4, v.2.5, v2.5.1, v2.6 y v2.7. De ellas, las versiones HL7 v2.3 y v2.3.1 fueron las más implementadas. Llegando a representar más de tres cuartas partes de toda la mensajería HL7.

1.1.5.4 HL7 versión 3

En [22] se menciona que HL7 v3.0, vio la luz por primera vez en 2005 y es aquí donde tuvo un gran cambio, comenzando con el número de versión, el cual ya no cambia, pero se valida de forma anual. HL7 v3.0 concurre a la par HL7 v2.x. Sin embargo, no son compatibles. La diferencia entre v2.x y v3.0 quiere decir que los sistemas que lleguen a implementar la versión v3.0 no tendrá la posibilidad comunicarse con los sistemas que implementaron los HL7 v2. x.

1.1.5.5 HL7 FHIR (*Fast Health Interoperability Resources*)

La página oficial del estándar HL7 FHIR® en [23] especifica cómo la información de salud se intercambia entre sistemas informáticos diferentes, independientemente de su formato de almacenamiento. Esto posibilita que la información de salud, que incluye datos clínicos y administrativos, esté accesible de forma segura para quienes necesitan acceder a ella, en beneficio de los pacientes que reciben atención médica. La organización de desarrollo de estándares HL7utiliza un enfoque colaborativo para mejorar y desarrollar FHIR.

El inicio de la creación de FHIR se remonta a 2012, cuando se detectó la necesidad en el mercado de agilizar y simplificar el intercambio de la creciente cantidad de datos de salud de manera más eficaz. El aumento en la disponibilidad de nuevos datos de salud, impulsado por el auge de las aplicaciones “apps”, generó la demanda de una forma ágil y en tiempo real para que tanto profesionales de la salud como consumidores compartieran información utilizando tecnologías y estándares de internet más modernos.

FHIR toma como apoyo estándares de internet que son ampliamente empleados en diversas industrias, además de la salud. En particular, integra la filosofía REST (*Representational State Transfer*), que detalla cómo se pueden compartir con facilidad unidades individuales de información denominadas recursos. Al hacer uso de estándares y tecnologías ya establecidas y familiares para los desarrolladores de software, FHIR disminuye de manera considerable los obstáculos para que nuevos desarrolladores de software puedan abordar las demandas de la atención médica.

1.2 Planteamiento del problema

Por mucho tiempo, los sistemas de salud presentaron problemas de comunicación e interacción entre ellos, por tal motivo para intercambiar información fluidamente es indispensable la adopción de estándares sobre los que coincidan todos los sistemas de salud. Por ello, surgen organizaciones que tienen como propósito unificar criterios en pro de la interoperabilidad en el sector salud, tales como HL7 International, HIMSS (*Healthcare Information and Management Systems Society*) o NEMA (*National Electrical Manufacturers Association*). Por consiguiente, tanto para la interoperabilidad sintáctica (referida a la estructura de la comunicación) como para la interoperabilidad semántica (que hace referencia al significado de la comunicación), el sector salud propone estándares para varios propósitos relacionados con mensajería, terminología, documentos, esquemas conceptuales, aplicación y arquitecturas. Entre los estándares de documentos, que indican el tipo de información que se incluye en un documento y cómo este se estructura en secciones de contenido, se encuentran los estándares HL7 CDA (*Clinical Document Architecture*), CCDA (*Consolidated CDA*), y CCR (*Continuity of Care Record*) que definen una vista consolidada o resumen de información de salud de un paciente, incluyendo alergias, tratamientos, plan de cuidados y lista de problemas activos, para compartir información entre los profesionales de salud. Además, la interoperabilidad es un medio y no un fin en sí mismo, que se potencia utilizando estándares. No obstante, en el campo de la salud, la presencia de estándares divergentes y sistemas propietarios a menudo complica la toma de decisiones en relación a la interoperabilidad entre los sistemas con los que un hospital interactúa. Estos estándares contradictorios y la falta de uniformidad en los formatos de datos plantean desafíos significativos para lograr una comunicación efectiva y una gestión de datos eficiente. Por lo tanto, resulta de suma importancia definir y establecer políticas, estándares y guías de implementación de manera adecuada para superar estas barreras y facilitar la interoperabilidad.

Sin embargo, ante el nuevo paradigma 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. Por tal motivo, es

necesario un sistema de intercambio de información entre los actores que colaboran en la gestión del valor en salud, tales como hospitales, pacientes, laboratorios clínicos, doctores, entre otros.

1.3 Objetivo general y Objetivos específicos

En seguida, se presentan a detalle el objetivo general y los objetivos especifico planteados para este trabajo.

1.3.1 Objetivo general

Desarrollar un módulo de reconocimiento óptico de caracteres para la interpretación de documentos clínicos a formato HL7 utilizando técnicas de procesamiento del lenguaje natural.

1.3.2 Objetivos específicos

- Investigar el estado del arte para identificar las propuestas similares al proyecto.
- Investigar las diferentes versiones del estándar HL7 para determinar cuál es la versión y/o versiones utilizadas por los sistemas de salud.
- Analizar la estructura del expediente clínico electrónico en México para identificar las secciones más importantes que lo conforman.
- Comparar las tecnologías y APIs disponibles para realizar el reconocimiento óptico de caracteres.
- Diseñar las interfaces gráficas del módulo a desarrollar siguiendo un estilo amigable con el usuario para facilitar el uso del software.
- Construir un componente de software que permita aplicar el proceso de reconocimiento óptico de caracteres en documentos clínicos para su interpretación a estándar HL7.
- Construir un componente de software para almacenar los datos transformados a estándar HL7.
- Probar el módulo construido mediante una prueba de concepto para validar su funcionalidad.

1.4 Justificación

La interoperabilidad se refiere a la capacidad de diversos sistemas de información y aplicaciones de software para comunicarse, intercambiar datos y utilizar la información compartida. En el contexto del sector salud, la interoperabilidad desempeña un papel fundamental al aumentar la seguridad de la información personal del paciente, ya que permite el acceso y la disponibilidad de los datos clínicos de manera eficiente y segura. Además, acceder a los datos clínicos en tiempo real permite atender pacientes desde cualquier punto, mejorando la calidad y continuidad asistencial. Por ello, es necesario que los diferentes sistemas de salud intercambien información y la transfieran de un sistema a otro a través de interfaces específicas, adaptadas o personalizadas, que estructuren la información de manera similar. Además, en materia tecnológica se garantiza el intercambio y portabilidad de los datos para lograr sistemas de salud conectados.

Por otro lado, el OCR es un proceso dirigido a la digitalización de textos, que utilizan un software para identificar automáticamente a partir de una imagen, símbolos o caracteres que pertenecen a un determinado alfabeto, para luego almacenarlos en forma de datos. El OCR es una tecnología transversal, aplicable en distintos ámbitos como el área de la salud. Asimismo, se tiene que el HL7 se encarga de definir un marco de trabajo y estándares para el intercambio, la integración y el acceso a la información electrónica de salud. Adicionalmente, la tecnología semántica permite identificar datos con un mayor significado, a inferir relaciones y a extraer información de enormes conjuntos de datos primarios almacenados en diferentes formatos y en diferentes fuentes.

Es por lo anterior que el presente proyecto busca promover la interoperabilidad entre los sistemas de información clínica, unificando los diferentes formatos de análisis clínicos bajo el estándar HL7 con el objetivo de facilitar la obtención de información del paciente. Para ello se propone desarrollar un módulo que aplique el proceso de reconocimiento óptico de caracteres a múltiples documentos clínicos y posteriormente unifique sus datos bajo un solo formato.

Capítulo 2 Estado de la práctica

El presente capítulo aborda el estado de la práctica, el cual se clasifica en tres grupos principales: 1) Procesamiento de información en el estándar HL7, 2) Reconocimiento Óptico de Caracteres en múltiples dominios de aplicación y, 3) Reconocimiento Óptico de Caracteres en el área de la salud. Asimismo, se describen las tecnologías de la información seleccionadas para llevar a cabo el presente trabajo y la metodología para su realización.

2.1 Trabajos relacionados

El aumento en la expansión de los servicios de internet y el crecimiento del uso de dispositivos móviles han llevado a un incremento significativo en el número de personas que utilizan las redes sociales. Esta tendencia se traduce en un uso cada vez más común de estos dispositivos para mejorar los servicios de salud, a pesar de que subsisten desafíos en materia de privacidad y seguridad de la información.. A partir de lo anterior, en [24] Trigo y sus colaboradores 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 incluyó un escenario a pequeña escala y otro un escenario de límites. En el primero, probaron dos tamaños de imágenes y en el segundo, dos versiones del algoritmo de incrustación. Los resultados mostraron que el sistema fue lo suficientemente rápido (menor a 1 s). La arquitectura ofreció a los usuarios servicios de salud móvil, amigables y seguros a través imágenes compartidas con ayuda de las redes sociales.

En [25] se habló de la necesidad de interoperabilidad y de cómo es fundamental para los sistemas de información de salud. Por esta razón, el utilizar estándares

especializados en el manejo de datos relacionados con la salud, como HL7 FHIR, es de suma importancia, ya que estos estándares ayudan a definir el significado sintáctico y semántico de la información que se comparte e intercambia. Es uno de esos estándares que se desarrolló para el intercambio de información médica HIE (*Health Information Exchange*). Si bien el estándar HL7 FHIR permite el uso de arquitectura REST (*Representational State Transfer*) y SOA (*Service Oriented Architectures*) para el intercambio de información de forma segura, también trae consigo la inflexibilidad y la complejidad propias de arquitecturas como REST. Por otro lado, se mostró las ventajas que tiene GraphQL para evitar los problemas mencionados. En el presente trabajo, se utilizó GraphQL y HL7 FHIR para HIE y se presentó un algoritmo para mapear recursos HL7 FHIR a un esquema GraphQL, y se desarrolló un prototipo de implementación, el cual se compara con un enfoque RESTful. El resultado obtenido muestra que la combinación de GraphQL y las API (*Application Programming Interface*) Webs basadas en HL7 FHIR para HIE es capaz de satisfacer los requisitos de los clientes Web y móviles.

Actualmente, la necesidad de datos clínicos requiere una orientación basada en los datos, lo que ofrece la oportunidad de mecanizar las actividades relacionadas con la atención médica, proporcionando una mejor forma para la detección de enfermedades, un pronóstico preciso, un progreso rápido de la investigación clínica y una mejor forma de administrar la información de los pacientes. Compartir la información con las personas involucradas en el proceso de atención médica y mantener un expediente clínico requiere interoperabilidad, ya que es la única forma sostenible de permitir que los sistemas se comuniquen entre sí y obtengan la imagen completa de un paciente a partir de los datos recabados. En [26] 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 planteó el proyecto fue el desarrollar ontologías de datos, las cuales se almacenan 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 [27] V. Kilintzis y sus colaboradores presentaron un marco de gestión de datos de telemedicina, que tenía 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*. Con lo mencionado anteriormente, aún lado con el almacenamiento persistente y la construcción de servicios Web de comunicación que permitan la gestión de los datos de la HCE, que ayuden asegurar integridad y veracidad de la información de los pacientes compartidos como instancias ontológicas autodescriptivas. El marco de trabajo presentado se basa en la flexibilidad y la reutilización, tomando en cuenta la ontología como un único punto de cambio. La solución presentada se implementó en la gestión de datos en un sistema de telemonitorización para pacientes con EPOC (Enfermedad Pulmonar Obstructiva Crónica) y comorbilidades. Gracias a los resultados obtenidos se comprobó que el marco de trabajo logro adoptarse satisfactoriamente en diferentes escenarios de atención médica integrada.

HL7 FHIR es uno de los estándares de comunicación datos más importantes en el área de la salud. Investigaciones realizadas confirman que HL7 FHIR es útil para modelar datos estructurados y no estructurados de la HCE (Historia Clínica Electrónica). Pero a pesar de la capacidad de HL7 FHIR para admitir análisis de datos clínicos, no se han realizado las investigaciones suficientes. En el caso de estudio presentado por [28] se observó el siguiente resultado representación apoyada en HL7 FHIR de los datos no estructurados de la HCE, los cuales se introducen en modelos de aprendizaje profundo para la categorización de textos en el fenotipo clínico. Usando las ventajas que ofrece la línea de normalización de datos clínicos NLP2FHIR se realizó un estudio con dos grupos de datos de obesidad. Se evaluó la funcionalidad de varios clasificadores de

texto basados en el aprendizaje profundo, tales como las redes neuronales convolucionales, la unidad recurrente cerrada y las redes convolucionales de grafos de texto, tanto en texto crudo como en entradas NLP2FHIR. Donde se observó que la combinación de la entrada NLP2FHIR y las redes convolucionales de grafos de texto tiene la calificación F1 más alta. Por esta razón, se llegó a la conclusión que los métodos de aprendizaje profundo apoyados en HL7 FHIR tienen el potencial y son utilizados para apoyar el fenotipado de la HCE, haciendo que los algoritmos de fenotipado sean manejables entre los sistemas de HCE y las instituciones dedicadas al ámbito de la salud.

El análisis es una de las etapas más importantes en la obtención de información de imágenes escaneadas, por esta razón se presenta el trabajo desarrollado por [29] S. Tomovic, K. Pavlovic, y M. Bajceta. El cual consistió en un algoritmo para organizar los modelos creados con diferentes motores de OCR. El requisito principal del proyecto fue obtener el mismo diseño para la imagen del documento escaneado, independiente al OCR utilizado para el procesamiento de imágenes. El algoritmo que se presentó se desarrolló para procesar documentos administrativos con diseños complejos. Gracias a la orientación presentada, fue posible deducir que los sistemas de comprensión de documentos son independientes de los pasos de preprocesamiento que dependen del algoritmo de segmentación de páginas y el OCR implementado. Dado lo anterior se determinó que la opción para este tipo de casos es utilizar diferentes extractores para cada OCR. Lo cual lleva a que solo sean utilizados si se conoce qué motor de OCR es utilizado para procesar la imagen del documento y generar un documento PDF con capacidad de búsqueda correspondiente.

En [30] se habló acerca de la gran cantidad de personas que posee una discapacidad visual y que padecen diferentes limitaciones con respecto a la comunicación, la interacción y la independencia personal. Una de estas limitaciones es la poca cantidad de literatura en braille que hay disponible, otra limitación son situaciones económicas. Este proyecto propuso un sistema de lectura para personas con discapacidad visual

para dispositivos móviles. El presente trabajo mostró la mezcla de técnicas de segmentación, extracción de rasgos y aprendizaje automático para convertir texto a braille de una manera rápida y precisa. El OCR presentado en este proyecto se dividió en dos fases. La primera utilizó técnicas de filtrado, y algoritmos de localización, segmentación y ordenamiento de los caracteres y símbolos gráficos. En la segunda fase, se trabajó con técnicas de reconocimiento de patrones para identificar cada uno de los caracteres y símbolos segmentados obtenidos a partir la fase anterior. Las pruebas realizadas indicaron que el extractor que utiliza Perceptrón Multicapa fue la mejor opción para el sistema OCR construido con un 99,86% de precisión y un 99,93% de especificidad. De igual forma, se evaluó usabilidad del dispositivo portátil con personal docente de nivel primaria y alumnado de una asociación de personas con discapacidad visual.

El reconocimiento correcto de caracteres manuscritos combinados es un gran desafío para los sistemas de OCR en bangla. En [31] se propuso una técnica de segmentación apoyada en la descomposición de formas para los caracteres combinados. Esta desintegración de la forma reduce la complejidad de la clasificación, minimizando el número de clases a reconocer, y simultáneamente mejorando la eficacia del reconocimiento. La descomposición se llevó a cabo en el área de segmentación, donde las dos formas básicas se acoplan para formar un carácter combinado. Se utilizó un grupo de características de histograma de código de cadena con un clasificador basado en un MLP (Perceptrón Multicapa) con aprendizaje por retropropagación para la clasificación. Por otra parte, la metodología propuesta se divide en cuatro partes: preprocesamiento y detección del área de segmentación, formación de grupos y descomposición de formas, extracción de características y clasificación. De igual forma se llevaron a cabo diferentes pasos de forma secuencial para identificar el área de segmentación en la que las dos formas básicas se mezclan para formar un carácter combinado. Una vez localizada el área, se revisó la presencia de una línea de referencia, nombrada RL. Con ayuda del conocimiento de la escritura bangla se elaboraron las reglas base para dividir los caracteres combinados en cinco

grupos. Usando estas reglas, el algoritmo presentado coloca el carácter combinado a su grupo correspondiente y descompone el carácter en formas básicas prominentes. En las pruebas realizadas, se observó que el método presentado provee una buena precisión de reconocimiento en comparación con otros métodos existentes.

En [32] se presentó un sistema de reconocimiento óptico de caracteres que reconoce el contenido de la escritura hecha de forma manual en los folletos. Teniendo en cuenta que los folletos que ocupan contienen una mezcla de fórmulas japonesas y matemáticas, se infiere que solo un tipo de sistema de reconocimiento óptico de caracteres no proveería una precisión suficiente. Por esta razón, el sistema propuesto, toma en cuenta la mezcla de dos tipos de sistemas OCR. Los OCR que se eligieron son Tesseract y Mathpix, donde se comprobó su eficiencia para identificar fórmulas japonesas y matemáticas. Si se utilizaron dos OCR, lo más apropiado es seleccionar el resultado final admisible. El resultado del OCR es un valor de autoevaluación que se obtiene aún lado al reconocimiento. Sin embargo, carece de seguridad de que la calificación OCR y el resultado de reconocimiento real muestren la misma tendencia. Por lo que, en el presente proyecto, después a clasificar los contenidos escritos en tres categorías, palabras, frases y fórmulas matemáticas, se comprobó que los resultados de reconocimiento de cada OCR se eligieron apropiadamente en función de la puntuación OCR. Ya en esta parte resultó necesaria una mejora adicional para realizar la selección, basándose únicamente en la calificación del OCR. De igual manera se consiguió clasificar los contenidos escritos manualmente con el sistema presentado mediante un aprendizaje adicional.

Desde la aparición de la visión por computadora, el OCR tomó un rol de suma importancia. Con ayuda del reconcomiendo óptico de caracteres, las imágenes de documentos escritos de forma manual, se convierten en texto codificado por las máquinas. El reconocimiento de caracteres escritos a mano, es un campo de investigación esencial del reconocimiento de patrones, lo cual llevó a realizar amplias investigaciones, ejemplo de ello es CNN (*Convolucional Neural Network* por sus siglas

en inglés) o Red Neuronal Convolucional la cual es una de las opciones más utilizadas para realizar aprendizaje profundo, en el reconocimiento de caracteres escritos de forma manual. Actualmente, se han realizado diversos trabajos de investigación en este campo, ejemplo de ello es [33] donde describió como mejorar la exactitud del reconocimiento de los caracteres escritos en idioma inglés a través del uso de una CNN para aprendizaje profundo, produciendo nuevas instancias de entrenamiento a partir de las instancias existentes utilizando el aumento y GAN (*Generative Adversarial Networks*). En este trabajo se utilizó un modelo de aprendizaje profundo, GAN para obtener el reconocimiento de caracteres escritos en inglés tanto letras mayúsculas como minúsculas. Además, se ilustró cómo los caracteres generados se utilizaron para aumentar el rendimiento de la clasificación de caracteres escritos a mano en idioma inglés.

La extracción de texto o caracteres a partir de imágenes es un objeto de investigación en el campo de procesamiento y análisis de imágenes, las redes neuronales, la inteligencia artificial y la visión artificial. La idea principal del reconocimiento de caracteres es obtener y reconocer el texto de las imágenes y, usar esto para analizar contenido de documentos de diferente tipo o recuperar información. Para ello se tiene que el análisis de imágenes se compone variados procesos y algoritmos para la obtención de textos, tales como la detección de bordes, la detección de puntos, la detección de esquinas, entre otros. Para extraer los textos presentes en imágenes. De igual manera, el reconocimiento de textos conlleva varias etapas, las cuales son: el preprocesamiento, la segmentación, la extracción de características y la clasificación. Es por lo anterior que en [34] se mostró el uso del algoritmo FAST (*Features from accelerated segment test*) que se utiliza para la detección de esquinas aplicado a las imágenes para el reconocimiento de texto. El algoritmo FAST detecta y determina la presencia de una esquina probando un área pequeña alrededor del centro potencial de la esquina con la ayuda del valor de intensidad. Este enfoque de reconocimiento de texto se utilizó eficazmente incluso con imágenes con contenido difícil de trabajar, como es el contenido borroso.

En [35] se expuso que el reconocimiento de figuras y caracteres es especialmente utilizado para obtener texto de imágenes. No obstante, una de las tareas más difíciles es mantener una buena tasa de reconocimiento, sin que las interferencias de la luz afecten a esta en un entorno poco favorable. Con el objetivo de minimizar las interferencias de luz para optimizar la tasa de reconocimiento OCR. Por esta razón, en el presente trabajo se diseñó un método que utiliza imágenes compuestas de un solo píxel para conseguir imágenes sin obstrucciones de luz para el reconocimiento de caracteres. En contraste a las imágenes tradicionales basadas en CCD/CMOS que se ven atadas al posprocesamiento de la imagen para descartar las interferencias, el método planteado no requiere de un posprocesamiento de la imagen, que es un proceso complicado que termina reduciendo la calidad de la imagen obtenida. El análisis teórico dice que las imágenes de un solo píxel se deshacen de las interferencias de luz bajo determinadas condiciones. Los experimentos realizados para comprobar exponen que la imagen de un solo píxel obtiene la imagen del objetivo e ignorar una franja resaltada introducida, mientras que la imagen tradicional basada en CCD/CMOS no eliminan los efectos de la franja sin el posprocesamiento de la imagen. En las pruebas de comparación realizadas en contraste con la tasa de reconocimiento de caracteres de la imagen conseguida a partir de imágenes tradicionales CCD/CMOS expuso que la contrapartida obtenida a partir de imágenes de un solo píxel aumentó del 88,64% al 97,73%.

En la actualidad el porcentaje de documentos digitalizados aumenta velozmente durante los últimos años, tal es el caso de los documentos históricos. Es por esta razón se requiere generar métodos eficaces para la recuperación de datos y extracción de conocimiento para acceder a esta información. Este tipo métodos requiere del reconocimiento óptico de caracteres OCR, el cual transforma las imágenes de documentos en representaciones textuales. Hoy en día los métodos de OCR no están optimizados para tratar con documentos históricos, además, requieren un gran número de documentos escritos. Por esta razón, el trabajo presentado en [36] muestra un

conjunto de métodos que permiten realizar un OCR en imágenes de documentos históricos, requiriendo una cantidad pequeña de datos de entrenamiento reales y escritos a mano. El sistema OCR que se presentó se conforma de dos tareas de suma importancia como son: el análisis del diseño de la página, que incluye la segmentación de bloques de texto y líneas, y el OCR empleado. Los métodos de segmentación ocupados utilizan redes convolucionales, y el enfoque de OCR usa redes neuronales recurrentes. De igual manera, con ayuda del portal *Porta fontium*® se obtuvo un nuevo conjunto de datos para el OCR, cabe destacar que este corpus se encuentra disponible gratuitamente para la investigación, y los métodos presentados en este trabajo se evalúan con estos datos.

En [37] se habló de la visión artificial y como esta es un gran éxito en el campo del OCR. Esto también abarca tareas de detección y reconocimiento de textos. No obstante, la extracción de información clave denominada KIE (*Knowledge Information Extraction*) de los documentos como tarea del OCR, tiene múltiples posibilidades de uso. Sin embargo, esto sigue siendo un desafío, ya que los documentos no solo tienen características textuales que se extraen con los sistemas de OCR, sino que también presentan características visuales semánticas que no se abordan correctamente y tienen un rol crítico en KIE. En la actualidad existe poco trabajo a realizar un uso eficiente de las características textuales y visuales de los documentos. En el presente trabajo, mostró “PICK”, que es un marco eficaz y robusto en el trato de la disposición de documentos complejos para KIE, esto es gracias a que combina el aprendizaje de grafos con la operación de convolución de grafos, generando una representación semántica más completa que cuenta con características textuales, visuales y la disposición global sin que haya ambigüedad. De esta forma se llevaron a cabo experimentos muy completos en diversos conjuntos de datos del ámbito real que muestran que el método presentado supera a los métodos de referencia con márgenes significativos.

La incorrecta interpretación y documentación de notas clínicas, llega a afectar las

historias clínicas electrónicas, lo cual dificulta el trabajo de los médicos. En [38] se presentó la construcción 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). El “OB-CIE” apoya, a los médicos, a documentar las notas que generan sin afectar su flujo de trabajo. Para identificar las entidades con nombre de los datos clínicos, se emplearon los conceptos de la ontología para elaborar un diccionario de categorías semánticas, de igual manera, se utilizó el método de correspondencia exacta del diccionario para hacer coincidir las frases sustantivas con sus categorías semánticas. Con apoyo de una orientación basada en reglas para clasificar las frases clínicas en categorías predefinidas. Los resultados que se obtuvieron expusieron que el “OB-CIE” se desempeñó correctamente durante la extracción de conceptos clínicos en una comparación con las técnicas de minería de datos. Gracias a esto, se concluye que el sistema se emplea en otros campos, adaptando su ontología y su conjunto de reglas de extracción.

Los textos presentes dentro de imágenes médicas en su mayoría contienen una gran cantidad de información valiosa a cerca de la condición clínica de los pacientes. El objetivo del presente trabajo fue obtener información textual estructurada de imágenes médicas semiestructuradas. Para lo cual, en [39] se presentó un lenguaje específico del dominio, denominado ODL, que permite a los usuarios describir el valor y la disposición de los datos de texto contenidos en las imágenes médicas. Apoyándose en las restricciones de valor y espaciales descritas en el ODL, el analizador sintáctico del ODL asocia los valores encontrados en la imagen con la estructura de datos de la descripción del ODL, ajustándose a las restricciones. La sintaxis de ODL utiliza información de valor y diseño para describir el formato de datos de las imágenes. El analizador sintáctico ODL presentado genera las mejores alineaciones entre los datos estructurados en ODL y los textos reconocidos por el motor. De igual forma, a partir de los múltiples textos candidatos del OCR y las correcciones anotadas de forma manual,

el modelo de corrección se adaptó con el analizador sintáctico ODL, y realizó correcciones sobre la marcha de los errores de reconocimiento más frecuentes. Al observar las pruebas realizadas, el analizador en sintáctico ODL supera metódicamente a los enfoques existentes en términos de precisión de la extracción.

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 agregados en los registros clínicos electrónicos. Por esta razón, la información difícilmente se encuentra disponible para acelerar la gestión de la calidad o notificar de los factores de riesgo concretos del paciente. Por esta razón, en [40] 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. Tomando lo anterior en cuenta, se realizó un estudio en la Cleveland Clinic®, Cleveland, Ohio, y en la Universidad de Minnesota®. Donde se seleccionó una lista de muestras aleatorias de procedimientos de colonoscopia e informes de patología. Posteriormente, se eligieron las variables deseadas. Después de esto se utilizó OCR/NLP para obtener las mismas variables de tres HCE (Historia Clínica Electrónica) utilizadas en la institución, las cuales fueron las siguientes: *Epic, ProVation* utilizadas para los informes de endoscopia, y *Sunquest PowerPath* utilizada para los informes de patología.

El COVID-19 provocó la innovación digital en el sector salud, para mejorar la eficiencia operativa de las instituciones, trayendo como resultado la cantidad de datos de pacientes almacenados de forma digital, tales como cartas de alta, imágenes de escáner, resultados de pruebas o notas de texto libre por parte del personal de salud como son médicos o enfermeros lo cual aumento notablemente. En el año 2020, se llegó a 2314 *exabytes* de datos médicos en todo el mundo. Estos datos médicos no se estructuran de una forma genérica, se encuentran en gran parte en forma de documentos no estructurados, generados digitalmente o escaneado y almacenados

como parte de los informes clínicos del paciente. Los datos no estructurados se digitalizan mediante un proceso de reconocimiento óptico de caracteres. Sin embargo, una de las principales dificultades es la precisión del proceso de OCR esta suele variar por causa de la incapacidad de los motores de OCR actuales para transcribir correctamente los documentos escaneados o escritos a mano o en los que el texto muestra sesgo, ensombrecido o ser poco legible. A esto se agrega el hecho de que el texto procesado está formado por terminologías médicas específicas que no forman parte del léxico lingüístico común. En [41] se utilizó una técnica de preentrenamiento autosupervisado basada en una red neuronal RoBERTa (*Robustly Optimized Bidirectional Encoder Representations from Transformers*) que tiene la capacidad de aprender a predecir secciones ocultas (disfrazadas) de los textos para rellenar los huecos de las partes no transcribibles de los documentos procesados. La estimación del método presentado en junto a los datos de dominios específicos que forman parte de documentos médicos reales, presenta una tasa de error de palabras significativamente reducida que comprueba la eficacia del enfoque desarrollado.

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, por mencionar algunos tal como menciona [42] donde 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. De igual forma, 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. Se puso a prueba el sistema en tres niveles diferentes de clasificación, usando documentos completos como páginas individuales de documentos y para finalizar se comparó el efecto de diferentes métodos de procesamiento de texto.

La aparición de la HCE (Historia Clínica Electrónica) es un avance significativo en el creciente mundo de la medicina moderna y telemedicina. A pesar de ello, normalmente 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. Tal es el caso que en [43] 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. El trabajo se conforma de dos módulos, los cuales se mencionan a continuación: detección y reconocimiento de textos. Para el caso de la detección de textos, se emplea una estrategia de entrenamiento basada en "*parch"*, la cual alcanzado un nivel recuerdo del 99,5% en las pruebas que se realizaron. 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.

2.2 Análisis Comparativo

La tabla 2.1 muestra una comparación cada uno de los artículos con mayor relacionados con el proyecto de tesis con el objetivo de identificar los elementos más importantes de cada uno de estos.

Tabla 2.1 Análisis comparativo de trabajos.

Artículo	Problema	Contribución	Tecnología	Resultados	Estado
Trigo et al. [24]	Falta de seguridad y privacidad de la información que se comparte.	Arquitectura genérica para el desarrollo de servicios de salud móvil estandarizados y seguros.	Estándar HL7, estándar DICOM, protocolo de comunicación SCP-ECG y	Servicios de salud móvil rápidos, amigables y seguros a través imágenes compartidas	Terminado

Artículo	Problema	Contribución	Tecnología	Resultados	Estado
			protocolo HTTP	con ayuda de las redes sociales.	
Mukhiya et al. [25]	Necesidad de interoperabilidad entre sistemas de salud.	Algoritmo para Mapear recursos HL7 FHIR a un esquema GraphQL.	Estándar HL7 FHIR, protocolo HTTP y GraphQL.	Prototipo de implementación de la combinación de GraphQL y Web API Basadas en HL7 FHIR.	Terminado
Kiourtis et al. [26]	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	Estándar HL7 FHIR, almacén RDF (*Resource Description Framework*) y XML.	Ontologías de datos con almacenamiento RDF y traducción al estándar HL7 FHIR.	Terminado
Kilintzis et al. [27]	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.	Estándar HL7 FHIR, XML, protocolo HTTP, SPARQL (*Protocol and RDF Query*	Sistema de Telemonitorización para pacientes con EPOC y comorbilidades	Terminado

Artículo	Problema	Contribución	Tecnología	Resultados	Estado
			Language).		
Tomovic et al. [29]	Necesidad de obtención de información de imágenes escaneadas.	Algoritmo para organizar los modelos creados con diferentes motores de OCR.	PdfMiner, XML, Python 3 y Tesseract OCR.	Algoritmo para procesar documentos administrativos con diseños complejos.	Terminado
Holanda et al. [30]	Personas que posee una discapacidad visual y que padecen diferentes limitaciones con respecto a la comunicación, la interacción y la independencia personal.	Mezcla de técnicas de segmentación, extracción de rasgos y aprendizaje automático para convertir texto a braille de una manera rápida y precisa.	Módulo AUBTM- 2, microcontrolador PIC18F25K22 y protocolo PS2.	Sistema de lectura para personas con discapacidad visual para dispositivos móviles.	Terminado
Kobayashi et al. [32]	Folletos que contienen una mezcla de fórmulas japonesas y	Sistema de reconocimiento óptico de caracteres que reconoce el	Tesseract OCR y Mathpix.	Clasificador de contenido escrito en tres categorías,	Terminado

Artículo	Problema	Contribución	Tecnología	Resultados	Estado
	matemáticas, por lo cual solo un tipo de sistema de reconocimiento óptico de caracteres no provee una precisión suficiente.	contenido de la escritura hecha de forma manual		palabras, frases y fórmulas matemáticas.	
Sasipriyaa et al. [33]	Reconocimiento de caracteres escritos a mano.	Modelo de aprendizaje profundo, GAN para obtener el reconocimiento de caracteres escritos en inglés tantas letras mayúsculas como minúsculas.	Red neuronal convolucional y Red generativa Adversa	Aumento en el rendimiento de la clasificación de caracteres escritos a mano en idioma inglés.	Terminado
Martínek et al. [36]	Los métodos de OCR no suelen estar optimizados o pensados para tratar con documentos	Conjunto de métodos que permite realizar un OCR en imágenes de documentos históricos	Red LSTM, OCRopus, Tesseract, ABBYY Finereader Engine, XML y Transcribus.	OCR conformado de tareas las cuales son: el análisis del diseño de la página, que incluye	Terminado

Artículo	Problema	Contribución	Tecnología	Resultados	Estado
	históricos, y se requiere un gran número de caracteres escritos.	requiriendo una cantidad pequeña de datos de entrenamiento reales y escritos a mano.		la segmentación de bloques de texto y líneas.	
Luo et al. [39]	Necesidad de extraer los textos presentes dentro de imágenes médicas que contienen información acerca de la condición clínica de los	Lenguaje específico del dominio, denominado ODL, que permite a los usuarios describir el valor y la disposición de los datos de textos contenidos en las imágenes médicas.	Tesserac y XML	Modelo de corrección adaptado con el analizador sintáctico ODL, capaz de realizar correcciones sobre la marcha de los errores de reconocimiento más frecuentes.	Terminado
Karthikeyan et al. [41]	Los datos médicos no se estructuran de una forma genérica se encuentran	Técnica de preentrenamiento autosupervisada basada en una red neuronal	Tesseract	EL método presentado muestra tasa una de error de palabras significativa	Terminado

Artículo	Problema	Contribución	Tecnología	Resultados	Estado
	en forma de documentos no estructurados generados digitalmente o escaneado y almacenados como parte de los informes clínicos del paciente.	con capacidad de aprender a predecir secciones ocultas de los textos para rellenar los huecos de las partes no transcribibles.		mente reducida.	
H. Goodrum et al. [42]	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,	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.	Reconocimiento óptico de caracteres. Bolsa de palabras (BoW).	Sistema para clasificar los documentos clínicos en importantes y no importantes.	Terminado

Artículo	Problema	Contribución	Tecnología	Resultados	Estado
	informes de radiología, correspondencia clínica.				

Después de analizar los trabajos relacionados con el tema de tesis, se determinó que aunque comparte similitud con algunos trabajos presentados, como es el caso de [43], este trabajo se distingue de otros, ya que se centra en la interpretación de formatos de análisis clínicos y dado 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, 4) análisis clínicos generales. Además, el presente trabajo 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.

2.3 Propuesta de solución

En esta sección se muestra la propuesta de solución desarrollada para dar respuesta a la problemática presentada que consiste módulo de reconocimiento óptico de caracteres para interpretar documentos clínicos a estándar HL7 utilizando técnicas de procesamiento del lenguaje natural.

Para lo cual en la tabla 2.2 se presenta la combinación de las tecnologías de la información necesarias para construir la solución propuesta.

Tabla 2.2 Combinación de tecnologías de la información.

Lenguaje	Editor de código	Sistema Gestor de Base de Datos	Marco de trabajo	OCR API	Metodología	Vocabulario terminológico
Python +	Visual	MongoDB	Flask +	Pytess	XP (eXtreme	MeSH

Lenguaje	Editor de código	Sistema Gestor de Base de Datos	Marco de trabajo	OCR API	Metodología	Vocabulario terminológico
HTML5, CSS3, JavaScript y TypeScript	Studio Code		Angular	eract	Programming)	+ DeCS

A continuación, se describe cada una de las tecnologías de la información presentadas en la tabla anterior.

2.3.1 Python

En [44] se presenta a Python como un lenguaje de programación interpretado, multipropósito y multiparadigma débilmente tipado, pero aun pese a esto este lenguaje de programación tiene una sintaxis simple y concisa cercana al lenguaje natural que permite desarrollar algoritmos comprensibles desde la primera lectura.

2.3.2 HTML 5

HTML (*HyperText Markup Language*) es el elemento de mayor importancia y sobre el cual se construyó la Web, ya que define una estructura para los documentos que conforman la Web. El cinco hace referencia a la versión más estable de HTML, en este punto se incluyó una estructura más completa donde se agregaron nuevos elementos tales como: header, nav, section, article, video y audio por mencionar algunos [45].

2.3.3 CSS 3

De acuerdo a [46] CSS (*Cascading Style Sheets*) es un lenguaje de estilo utilizado para describir cómo se deben presentar los documentos HTML o XML incluidos varios lenguajes basados en XML, como SVG (*Scalable Vector Graphics*), MathML (*Mathematical Markup Language*) o XHTML (*eXtensible HyperText Markup Language*). Por otro lado, tres hace referencias a la versión de CSS.

CSS se utiliza para diseñar páginas Web, por ejemplo, cambiar la fuente, el color, el tamaño y el espaciado del contenido, dividir el contenido en varias columnas o

agregando animaciones y elementos decorativos, entre otros.

2.3.4 JavaScript

Es un lenguaje de programación liviano, interpretado con funcionalidad de primera clase. Aunque se conoce como lenguaje de secuencias de comandos para páginas Web y se usa en muchos entornos externos al navegador Web, como Node.js, Apache CouchDB y Adobe Acrobat, JavaScript es un lenguaje de programación, con soporte para programación orientada a objetos, imperativa y declarativa como es el caso de la programación funcional como se presenta en [47].

2.3.5 TypreScript

Es un lenguaje de programación moderno que permite crear aplicaciones Web complejas en JavaScript. Es un “transpilador”, es decir, un compilador que se encarga de traducir las instrucciones de un lenguaje a otro, TypeScript es un lenguaje pre-compilado, es decir, un lenguaje el cual es compilado finalmente a JavaScript como se menciona en [48].

2.3.6 MongoDB

MongoDB es una base de datos denominada como NoSQL de acuerdo a [49] es decir, no utiliza SQL (*Structured Query Language*) como lenguaje de consulta, ya que MongoDB no se estructura en tablas, se basa en documentos ofreciendo una mejor escalabilidad, flexibilidad y un modelo de consultas superior.

2.3.7 Angular

Angular es una plataforma y un marco de trabajo para desarrollar aplicaciones de una sola página de lado del cliente, utiliza HTML y TypeScript. Angular está escrito en TypeScript. Implemente funciones básicas y opcionales como un conjunto de bibliotecas de TypeScript que importa a sus aplicaciones.

La arquitectura de una aplicación Angular se basa en los siguientes conceptos: Los bloques de construcción del marco Angular son los componentes estructurados *NgModules* lo cuales ordenan el código relacionado en grupos de funciones. Una aplicación Angular siempre tiene al menos un módulo raíz y generalmente tiene más módulos funcionales como se menciona en [50].

2.3.8 Flask

La página Web oficial [51] presenta a Flask como un marco de trabajo para aplicación Web WSGI (*Web Server Gateway Interface*) ligero. Está diseñado para ser rápido y fácil de usar con la capacidad de escalar para construir aplicaciones más complejas. Comenzó como un simple envoltorio para Werkzeug y Jinja y se ha convertido en uno de los marcos de trabajo para aplicaciones Web en Python más populares, llegando a competir con el marco de trabajo Django.

El marco de trabajo Flask hace recomendaciones, pero no aplica ninguna de las dependencias o diseños al proyecto. Es el desarrollador el que tiene que elegir qué herramientas y bibliotecas quiere usar. Hay muchas extensiones proporcionadas por la comunidad de Flask que facilitan la adición de nuevas funciones.

2.3.9 Pytesseract

Pytesseract como se menciona en [52] es una herramienta que permite realizar el proceso de reconocimiento óptico de caracteres utilizando el lenguaje de programación Python. Pytesseract funciona como una envoltura para el motor Tesseract, pero tiene la cualidad de funcionar como un script Python de invocación independiente a Tesseract, esto le da la ventaja de leer casi todos los formatos de imágenes soportados por el lenguaje Python.

2.3.10 Visual Studio Code

Es un editor de código que ha llegado a posicionarse junto a poderosos IDEs tales como Visual Studio, Netbeans, Eclipse e IntelliJ IDEA, pero sin perder lo ligero de un editor de código, se encuentra disponible para la plataforma de Windows, macOS y Linux. Visual Studio Code viene por defecto con la integración de soporte para los lenguajes JavaScript, TypeScript y Node.js

Además, cuenta con un gran número de extensiones para aumentar su funcionalidad y extender sus capacidades como editor de código [53].

2.3.11 XP (*eXtreme Programming*)

"*eXtreme Programming*" o "Programación Extrema" es una metodología ágil de desarrollo de software muy exitosa en la actualidad, es un enfoque de ingeniería de

software formulado por Kent Beck.

XP surge como una nueva manera de encarar proyectos de software, proponiendo una metodología apoyada esencialmente en la simplicidad y agilidad. Es una metodología ágil centrada en potenciar las relaciones interpersonales como clave para el éxito en desarrollo de software, promoviendo el trabajo en equipo, preocupándose por el aprendizaje de los desarrolladores, y propiciando un buen marco de trabajo.

La metodología XP se caracteriza por tener cuatro fases de gran importancia las cuales se presentan en la figura 2.1.

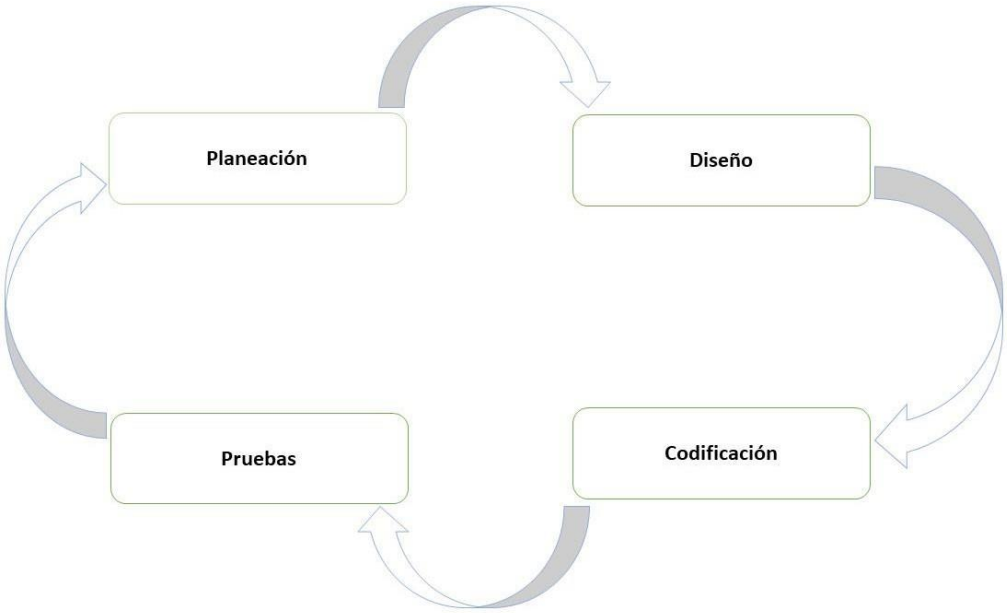


Figura 2.1 Fases de la metodología XP.

Inmediatamente, se describen cada una de las fases de la metodología de desarrollo de software XP.

2.3.11.1 Planificación

La planificación se plantea como un diálogo entre las personas involucradas en el proyecto, incluido al cliente y los programadores. El proyecto comienza recopilando “Historias de usuario”, teniendo como apoyo los requerimientos del proyecto. Una vez obtenidas las historias de usuario, los programadores evalúan el tiempo de desarrollo de cada una, utilizando el gráfico de Gantt en el cual se designa un determinado lapso de tiempo a cada actividad.

2.3.11.2 Diseño

La metodología XP hace especial énfasis en los diseños simples y claros. Uno de los conceptos más importantes de diseño en esta metodología es el siguiente:

Simplicidad

Un diseño simple se implementa más rápidamente que uno complejo. Por ello, XP propone implementar el diseño más simple posible que funcione. El principal objetivo de la fase de diseño es crear una visión global de lo que se quiere desarrollar.

2.3.11.3 Codificación

Para la fase de codificación XP propone los siguientes puntos a considerar:

Disponibilidad del cliente

Uno de los requerimientos de XP es tener al cliente disponible durante todo el proyecto. No solamente como apoyo a los desarrolladores, sino formando parte del grupo de trabajo.

Uso de estándares

XP promueve la programación apoyada en estándares, de manera que sea fácilmente entendible por todo el equipo, y que facilite la recodificación.

2.3.11.4 Pruebas

Las pruebas XP sugiere que todos los módulos o componentes de software deben de pasar las pruebas antes de ser aceptados. Es por ello, la metodología sugiere el uso de pruebas unitarias, pruebas funcionales y pruebas de aceptación con el fin de validar el correcto funcionamiento.

2.3.12 MeSH® (*Medical Subject Headings*)

En [54] se explica que MeSH es un léxico regulado desarrollado por la Biblioteca Nacional de Medicina, el cual se emplea para indexar, clasificar y buscar información y documentos relacionados con la salud y la biomedicina.

Es por ello que expertos en distintas áreas se encargan de mantener actualizado el vocabulario MeSH, realizando cada año la incorporación de cientos de nuevos conceptos y llevando a cabo miles de modificaciones.

2.3.13 DeCS (Descriptores de Ciencias de la Salud)

La página oficial DeCS [55], menciona que la Organización Panamericana de la Salud (OPS) y la Biblioteca Regional de Medicina (BIREME) desarrollaron el sistema, es una herramienta que consiste en un vocabulario estructurado y controlado. Su propósito principal es facilitar la búsqueda y clasificación de información científica en el campo de las ciencias de la salud.

El objetivo fundamental de DeCS es mejorar el intercambio de información en salud al proporcionar un lenguaje común y estandarizado. Este vocabulario abarca diversas áreas temáticas relacionadas con la salud, como medicina, epidemiología, enfermería, odontología, entre otras, permitiendo una organización más precisa y una recuperación eficiente de la información.

Capítulo 3 Aplicación de la metodología

En este capítulo, se describe el proceso de desarrollo llevado a cabo para construir el módulo de software que abordará el problema mencionado en el capítulo uno. Cabe destacar que el proceso de desarrollo sigue las fases de la metodología ágil de desarrollo XP.

3.1 Metodología de desarrollo

La programación extrema (XP) es una metodología ágil de desarrollo de software que se enfoca en el fortalecimiento de las relaciones interpersonales como clave para el éxito en el desarrollo de software. XP promueve el trabajo en equipo, se preocupa por el aprendizaje de los desarrolladores y proporciona un sólido marco de trabajo para equipos pequeños y proyectos con requisitos cambiantes.

3.1.1 Fase de planeación

La planificación de este proyecto de investigación se desarrolló siguiendo la metodología XP. Durante esta fase, se llevó a cabo la recopilación de información mediante diversos instrumentos de investigación, como el análisis documental y la observación, con el propósito de identificar los requisitos del sistema los cuales se describen a continuación.

3.1.1.1 Requisitos del sistema

Esta etapa es uno de los aspectos más críticos en la fase de planificación, ya que los requisitos del sistema desempeñan un papel fundamental al describir las actividades, servicios y características que se requieren en un sistema. Esto asegura que el sistema sea una herramienta capaz de satisfacer de manera efectiva las necesidades del usuario.

Requisitos funcionales

A continuación, se enumeran los requisitos funcionales críticos para garantizar el funcionamiento óptimo del módulo.

En la tabla 3.1, se detallan los requisitos específicos de la página de acceso encargada

de llevar a cabo el proceso de autenticación. Estos requisitos son esenciales para asegurar la seguridad y eficiencia sistema.

Tabla 3.1 Requisito funcional de autenticación.

Identificador del requisito:	RF001
Nombre del requisito:	Autenticar usuario
Prioridad del requisito: **(Alta, Media, baja)**	Alta
Descripción del requisito:	El módulo incluye una caja de texto destinada a la inserción del correo electrónico y otra caja para la contraseña. Ambas cajas cuentan con validación para recibir y enviar los datos ingresados por el usuario, facilitando así el proceso de autenticación.

En la tabla 3.2, se presentan los requisitos correspondientes a la página de registro, donde los usuarios que deseen registrarse deberán completar un formulario con su información personal.

Tabla 3.2 Requisito funcional de registro de usuarios.

Identificador del requisito:	RF002
Nombre del requisito:	Registro de usuarios
Prioridad del requisito: **(Alta, Media, baja)**	Alta
Descripción del requisito:	El módulo incluye un formulario que permita a los usuarios ingresar su nombre, apellido paterno, apellido materno, número de teléfono, correo electrónico y contraseña. Todos estos campos son validados antes de proceder con el registro exitoso.

En la tabla 3.3, se presentan los requisitos correspondientes a la página de

recuperación de contraseña de acceso. Para recuperar su contraseña, los usuarios deben proporcionar su correo electrónico y solicitar una nueva contraseña de acceso.

Tabla 3.3 Requisito funcional recuperar contraseña de acceso.

Identificador del requisito:	RF003
Nombre del requisito:	Recuperar contraseña de acceso
Prioridad del requisito: **(Alta, Media, baja)**	Alta
Descripción del requisito:	El módulo incluye un formulario que consta de una caja de texto diseñada para recibir la dirección de correo electrónico del usuario que desea recuperar su contraseña. El sistema envía y valida el correo electrónico proporcionado.

En la tabla 3.4, se presentan los requisitos correspondientes a la página de cambio de contraseña de acceso. Los usuarios que deseen modificar su contraseña ingresan la nueva contraseña, junto con una confirmación de la misma.

Tabla 3.4 Requisito funcional modificar contraseña de acceso.

Identificador del requisito:	RF004
Nombre del requisito:	Modificar contraseña de acceso
Prioridad del requisito: **(Alta, Media, baja)**	Alta
Descripción del requisito:	El módulo incluye una caja de texto destinada a la inserción de la nueva contraseña, así como otra caja de texto para confirmar la contraseña. Ambos campos se validan para recibir y procesar los datos ingresados por el usuario con el fin de efectuar la modificación de la contraseña.

En la tabla 3.5, se detallan los requisitos correspondientes a la página del panel de

administración. Una vez que el usuario se autentique de manera exitosa, visualiza un listado de los documentos clínicos que necesiten su atención.

Tabla 3.5 Requisito funcional consultar información relevante.

Identificador del requisito:	RF005
Nombre del requisito:	Consultar información relevante.
Prioridad del requisito: **(Alta, Media, baja)**	Alta
Descripción del requisito:	El módulo incluye una tabla que lista los documentos clínicos que requieran la atención del usuario.

En la tabla 3.6, se presentan los requisitos de la ventana Herramienta OCR, donde el usuario procesa los archivos de análisis clínicos (imágenes o documentos PDF).

Tabla 3.6 Requisito funcional procesar archivos con OCR.

Identificador del requisito:	RF006
Nombre del requisito:	Procesar archivos con OCR
Prioridad del requisito: **(Alta, Media, baja)**	Alta
Descripción del requisito:	El módulo incluye un control tipo "*File Chooser*" que permite al usuario seleccionar al menos un archivo de su dispositivo. Además, en la página de la herramienta OCR, se encuentra un campo para seleccionar el idioma del archivo a procesar. También, contiene un botón que inicia el proceso de OCR y presenta la información extraída en una tabla con filas editables.

Para generar archivos en formato HL7 v3 o HL7 FHIR, es necesario hacer clic en un botón ubicado en la parte inferior de la tabla que contiene los datos extraídos del

análisis clínico. El proceso se describe en la tabla 3.7.

Tabla 3.7 Requisito funcional generación de documentos HL7.

Identificador del requisito:	RF007
Nombre del requisito:	Generar documentos HL7
Prioridad del requisito: **(Alta, Media, baja)**	Alta
Descripción del requisito:	El módulo incluye una tabla que lista los caracteres extraídos de los documentos clínicos. Asimismo, dispone de un botón para iniciar el proceso de transformación de los documentos al formato HL7 v3 o HL7 FHIR para su posterior descarga o validación.

La tabla 3.8, presenta el requisito de guardar documentos de caracteres extraídos mediante OCR en la capa de datos para su posterior consulta y/o transformación al estándar HL7 v3 o HL7 FHIR.

Tabla 3.8 Requisito funcional guardar documentos OCR.

Identificador del requisito:	RF008
Nombre del requisito:	Guardar documentos OCR
Prioridad del requisito: **(Alta, Media, baja)**	Alta
Descripción del requisito:	El módulo incluye la opción de guardar los documentos OCR con los caracteres extraídos para su posterior transformación a HL7.

La tabla 3.9, detalla el requisito de validar documentos HL7 (v3 o FHIR). Para determinar que el proceso de transformación al estándar HL7 se ha llevado a cabo correctamente, es necesario realizar la validación de los documentos HL7 generados a través del módulo utilizando las herramientas de HAPI FHIR® o Gazelle hl7®.

Tabla 3.9 Requisito funcional validar documentos HL7.

Identificador del requisito:	RF009
Nombre del requisito:	Validar documentos HL7
Prioridad del requisito: **(Alta, Media, baja)**	Alta
Descripción del requisito:	El módulo incluye la opción de validar los documentos HL7 previamente almacenados, permitiendo al usuario seleccionar la herramienta de validación de su elección.

La tabla 3.10, describe el requisito de descargar documentos clínicos generados a partir de la extracción de caracteres y su transformación en formato HL7.

Tabla 3.10 Requisito funcional descargar documentos procesados.

Identificador del requisito:	RF010
Nombre del requisito:	Descargar documentos procesados
Prioridad del requisito: **(Alta, Media, baja)**	Alta
Descripción del requisito:	El módulo incluye la opción de descargar los documentos procesados y almacenados. Una vez que el documento HL7 ha sido validado, se lista en el registro histórico de documentos, donde el usuario tiene la opción de descargarlo en formato PDF, XML o HL7.

Requisitos no funcionales

La siguiente sección presenta los requisitos no funcionales que se tendrán en cuenta en el desarrollo del módulo de software.

La tabla 3.11, enumera los requisitos no funcionales que el diseño de la interfaz de la página de autenticación de usuarios debe satisfacer.

Tabla 3.11 Requisito no funcional de diseño de interfaz autenticación de usuarios.

Identificador del requisito:	RFN001
Nombre del requisito:	Diseño de interfaz autenticación de usuarios
Prioridad del requisito: **(Alta, Media, baja)**	Media
Descripción del requisito:	• El módulo incluye un nombre para su identificación. • El módulo proporciona una caja de texto para que el usuario ingrese su nombre de usuario. • El módulo contiene una caja de contraseña para que el usuario introduzca su contraseña. • El módulo dispone en la página de autenticación de un botón de "Ingresar" para permitir al usuario acceder. • En la página de registro el módulo dispone de un formulario tipo horizontal para el registro del usuario. • En la página de autenticación el módulo contiene un enlace para acceder a la página de recuperar contraseña.

Los requisitos no funcionales para el diseño de la interfaz de la página de administración del módulo se detallan en la tabla 3.12.

Tabla 3.12 Diseño de la interfaz página de administración.

Identificador del requisito:	RFN002
Nombre del requisito:	Diseño de interfaz página de administración
Prioridad del requisito:	Media

(Alta, Media, baja)	
Descripción del requisito:	• Al ingresar al módulo a través del proceso de autenticación, la primera página inicial muestra una tabla que lista los registros de análisis clínicos que requieren atención por parte del usuario. • El módulo cuenta con una barra de navegación en el lado derecho que contiene enlaces para acceder a las diferentes secciones del módulo.

La tabla 3.13 presenta los requisitos no funcionales relacionados con el diseño de la interfaz de la página de la “Herramienta OCR”.

Tabla 3.13 Requisito no funciona diseño de interfaz de la página herramienta OCR.

Identificador del requisito:	RFN003
Nombre del requisito:	Diseño de interfaz página herramienta OCR
Prioridad del requisito: **(Alta, Media, baja)**	Media
Descripción del requisito:	• La página de herramienta OCR incluye un formulario en la parte superior que permite la selección de los documentos a procesar, así como la elección del idioma de origen de los documentos. • Después de completar el proceso de OCR en los documentos especificados, la información extraída se muestra en una tabla con filas editables para permitir ajustes si es necesario.

	• En la parte inferior de la tabla que contiene información del documento, se encuentran los botones "Generar HL7 y Guardar", "Guardar" y "Eliminar".

En la tabla 3.14, se detallan los requisitos no funcionales relacionados con el diseño de la interfaz de la página de "Mis Documentos y Pendientes".

Tabla 3.14 Requisito no funcional diseño de interfaz página mis documentos y pendientes.

Identificador del requisito:	RFN004
Nombre del requisito:	Diseño de interfaz página mis documentos y pendiente
Prioridad del requisito: **(Alta, Media, baja)**	Media
Descripción del requisito:	• En la sección inicial de la página "Mis Documentos y Pendientes", se muestran documentos cuyos caracteres se extrajeron utilizando OCR y se eligió la opción de guardado. • En la segunda sección, la página presenta una lista de los documentos que se han transformado al formato HL7. Cada registro incluye la opción de validar el documento a través de un servicio externo.

Asimismo, en la tabla 3.15, se detallan los requisitos no funcionales relacionados con respecto al diseño de la interfaz de la página "Histórico de Documentos".

Tabla 3.15 Requisito no funcional diseño de interfaz página documentos históricos.

Identificador del requisito:	RFN005

Nombre del requisito:	Diseño de interfaz página documentos históricos
Prioridad del requisito: **(Alta, Media, baja)**	Media
Descripción del requisito:	• La página incluye una tabla en la que se enumeran los documentos clínicos que han sido procesados con la herramienta OCR y convertidos al formato HL7. • La página proporciona opciones para cada registro que permiten la descarga del archivo en formatos como XML y PDF.

La tabla 3.16 contiene los requisitos no funcionales relacionados con los mecanismos de autenticación.

Tabla 3.16 Requisito no funcional mecanismos de autenticación.

Identificador del requisito:	RFN006
Nombre del requisito:	Mecanismos de autenticación
Prioridad del requisito: **(Alta, Media, baja)**	Media
Descripción del requisito:	Definir los mecanismos de autenticación JWT (*JSON Web Token*) que se utilizarán para proteger los servicios REST.

La tabla 3.17 contiene los requisitos no funcionales relacionados con la autorización de recursos.

Tabla 3.17 Requisito no funcional de autorización sobre recursos.

Identificador del requisito:	RFN007
Nombre del requisito:	Autorización sobre recursos

Prioridad del requisito: **(Alta, Media, baja)**	Media
Descripción del requisito:	Especificar quién tiene acceso a qué recursos y qué acciones realizan sobre los recursos.

3.1.1.2 Cronograma de actividades

Para la planificación de actividades se utilizó el gráfico de Gantt, ya que este permite programar tareas en un período de tiempo determinado, lo que facilita el seguimiento y control detallado del avance de las actividades descritas.

En la figura 3.1 se presenta la planificación elaborada para llevar a cabo este proyecto con el propósito de administrar y distribuir eficazmente el tiempo dedicado a su realización.

Id	Nombre de tarea	Duración	Comienzo
1	**Fase 1 Planeación**	**25 días**	**lun 04/07/22**
2	1.1 Analísis del planteamiento del problema	10 días	lun 04/07/22
3	1.2 Elaboración de las historias de usuario	15 días	lun 18/07/22
4	**Fase 2 Diseño**	**45 días**	**lun 15/08/22**
5	2.1 Elaboración de Diagramas UML	5 días	lun 15/08/22
6	2.2 Evaluación de bibliotecas OCR	15 días	lun 22/08/22
7	2.3 Clasificación de los analísis clínicos	20 días	lun 19/09/22
8	2.4 Diseño de mockups	5 días	lun 10/10/22
9	**Fase 3 Codificación**	**145 días**	**lun 17/10/22**
10	3.1 Codificación de interfaces	13 días	lun 17/10/22
11	3.2 Codifiacación de servicio web basado en rest e integración con interfaces	130 días	lun 07/11/22
12	3.3 Codificación de la base de datos e integración con servicio web	90 días	lun 02/01/23
13	**Fase 4 Pruebas**	**40 días**	**lun 03/04/23**
14	4.1 Pruebas de funcionales	40 días	lun 03/04/23

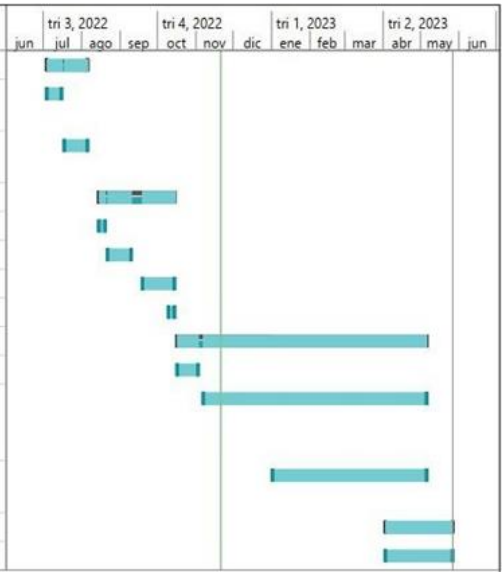


Figura 3.1 Cronograma de actividades utilizando gráfico de Gantt.

3.1.2 Fase de diseño

La fase de diseño reviste una importancia fundamental en el modelo de XP, ya que es en este periodo donde se lleva a cabo la identificación de las funciones y el rendimiento del software, así como la comunicación que mantendrá con otros componentes del proyecto. Además, en esta etapa se realiza el diseño del proyecto, el cual abarca

diversos subprocesos, como la creación de diagramas UML (*Unified Modeling Language*), el establecimiento de la arquitectura del software y la elaboración de prototipos de interfaz de usuario para la maquetación del software.

3.1.2.1 Diagramas UML

Los diagramas de casos de uso representan de manera gráfica la interacción entre los diversos usuarios y la aplicación, así como los procesos que la aplicación llevará a cabo.

Por otro lado, los diagramas de secuencia proporcionan una visualización de cómo las diferentes partes de un sistema interactúan entre sí a lo largo del tiempo, mostrando el orden en que se realizan las acciones y las interacciones entre los componentes del sistema.

Asimismo, los diagramas de clases ofrecen una representación visual de las clases en un sistema, mostrando sus atributos y relaciones. Estos diagramas son esenciales para comprender la estructura del sistema y las relaciones entre las diferentes partes del software, lo que facilita su diseño, implementación y mantenimiento.

Diagramas de casos de uso

La figura 3.2 presenta el diagrama de casos de uso relacionado con los procesos de autenticación de usuarios. En este diagrama, el actor (Usuario) cuenta con varias opciones, incluyendo "Iniciar sesión", "Cerrar sesión", "Registrar nuevo usuario", "Recuperar contraseña" y "Modificar contraseña". Este diagrama ilustra cómo los usuarios interactúan con el sistema en estas diferentes situaciones.

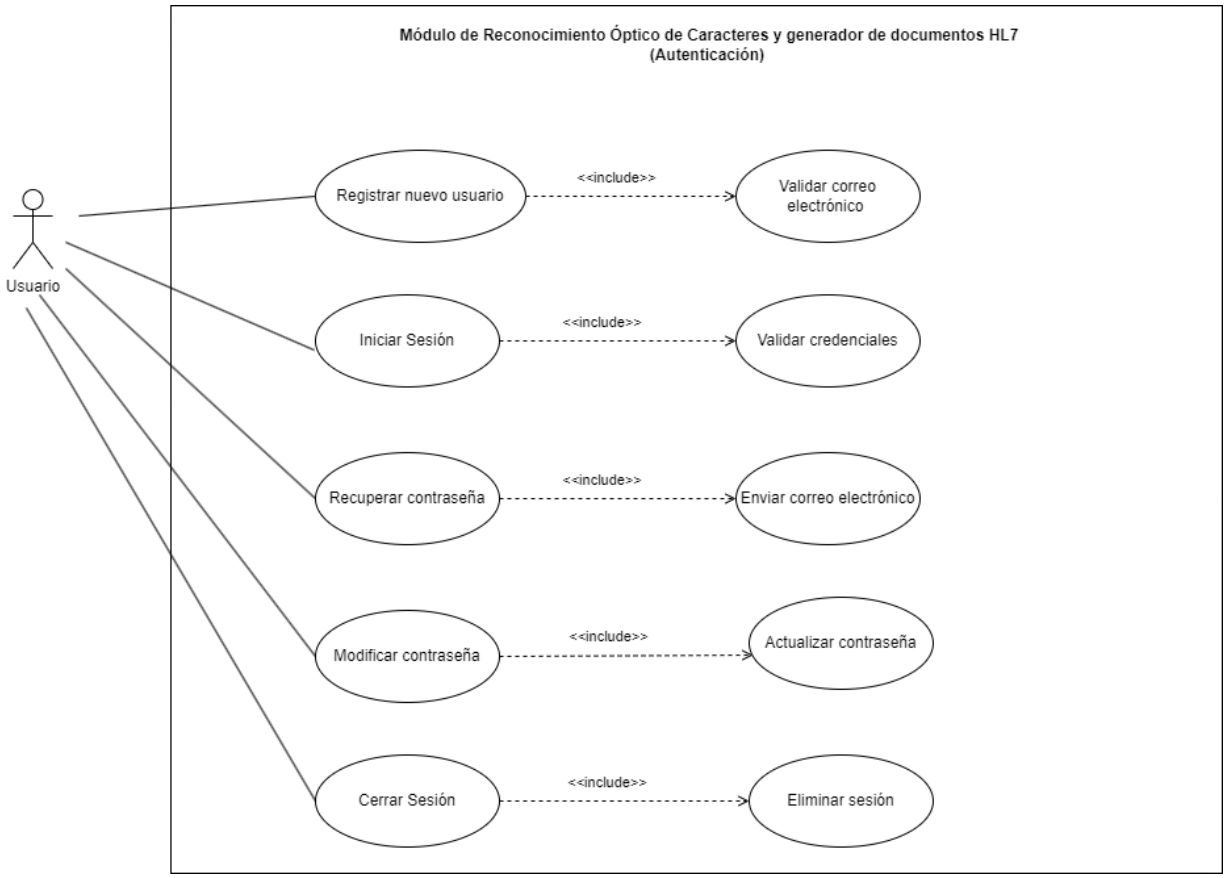


Figura 3.2 Diagrama de casos de uso autenticación.

La figura 3.3 muestra el diagrama de casos de uso para el proceso de consulta de registros de documentos que requieren atención por parte del usuario. Este diagrama ilustra cómo los usuarios interactúan con el sistema para acceder y revisar los documentos clínicos que necesitan su atención.

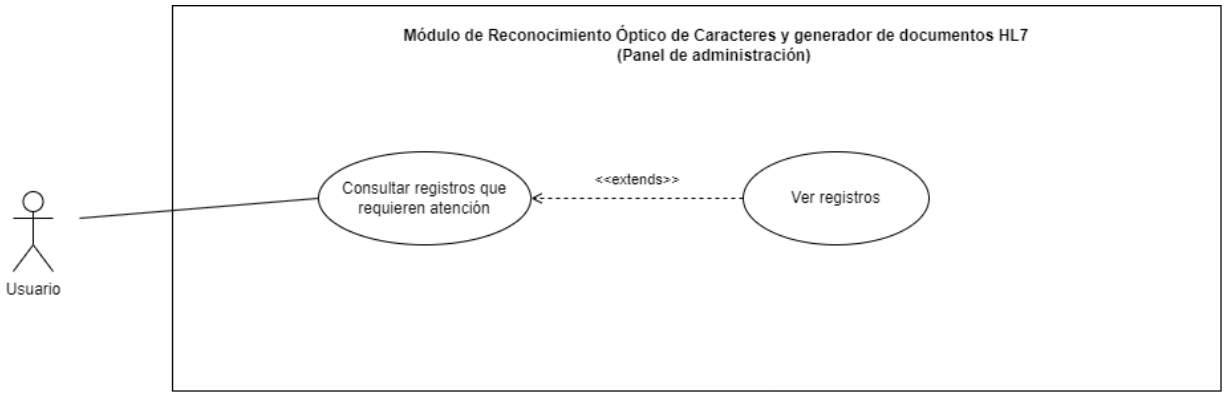


Figura 3.3 Diagrama de casos de uso panel de administración.

En la figura 3.4 se presenta el diagrama de casos de uso que detalla el procesamiento de documentos mediante OCR. Este diagrama incluye el caso de uso "Ver caracteres extraídos" y se extiende con los casos de uso "Guardar documentos" y "Generar HL7 y guardar". Estos casos de uso describen cómo se gestionan los documentos después de haber sido procesados por OCR, ya sea para guardarlos en el sistema o para generar archivos en formato HL7.

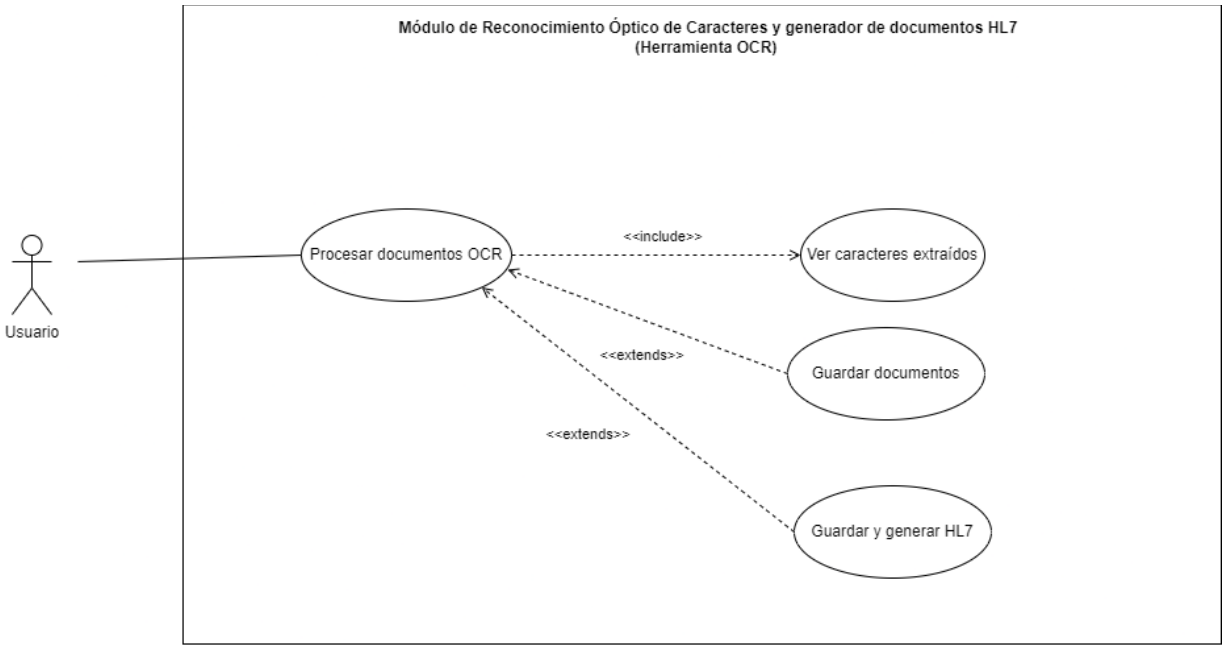


Figura 3.4 Diagrama de casos de uso procesar documentos OCR.

En la figura 3.5, se exhibe el diagrama de casos de uso que ilustra el proceso de consulta de documentos generados y almacenados, los cuales han sido procesados mediante OCR y transformados a formato HL7. Esto se refleja en los casos de uso “Consultar documentos procesados OCR” y “Consultar documentos HL7”.

Del caso de uso “Consultar documentos procesados OCR” se extiende con el caso de “Generar documento HL7”. Por otro lado, el caso de uso “Consultar documentos HL7” se extiende con el caso de uso “Validar documento HL7”. Estas extensiones indican cómo se gestionan los documentos en el sistema y cómo se les da seguimiento después de haber sido consultados.

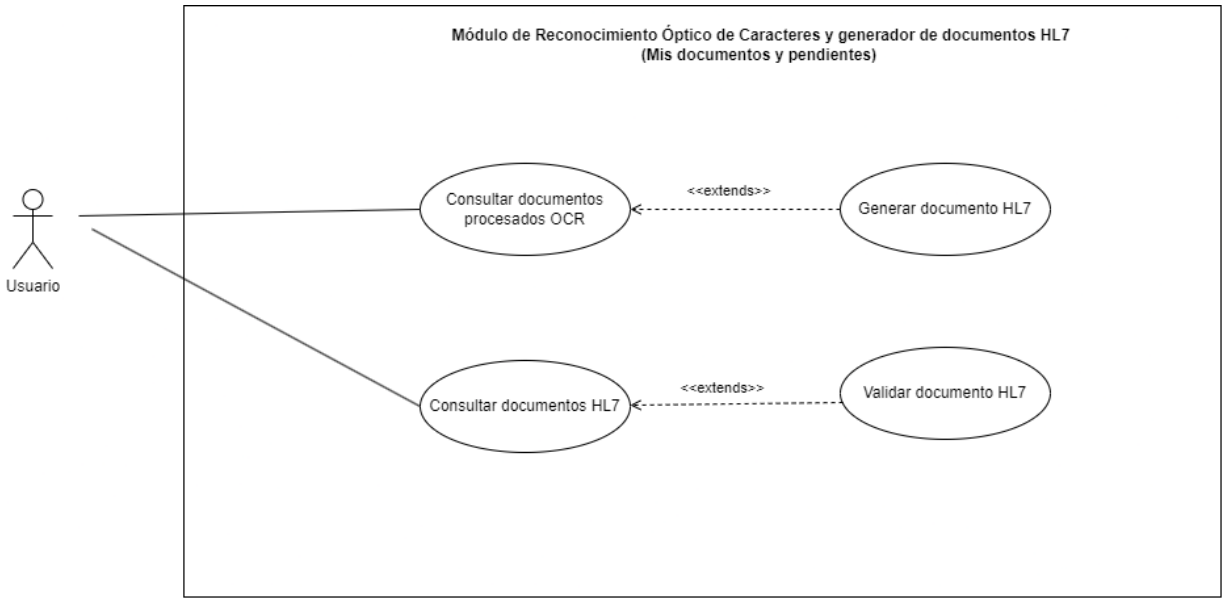


Figura 3.5 Diagrama de casos de uso consultar documentos OCR y HL7.

La figura 3.6 muestra el diagrama de casos de uso que describe el proceso de consulta de documentos históricos. De este proceso se derivan los casos de uso “Descargar PDF” y “Descargar XML”. Estos casos de uso extienden la funcionalidad del sistema al permitir a los usuarios descargar documentos históricos en diferentes formatos.

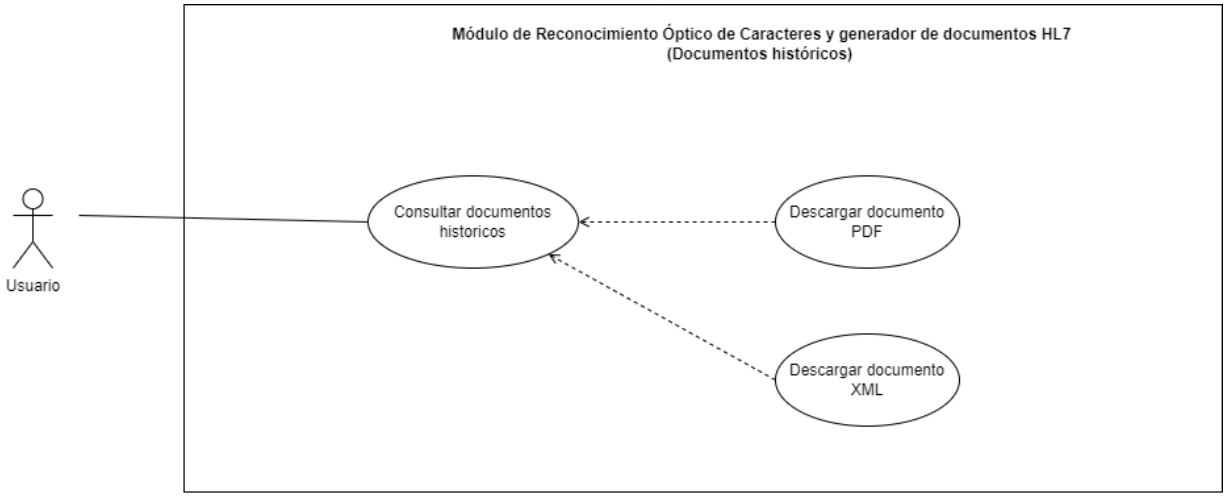


Figura 3.6 Diagrama de casos de uso consular y descargar histórico.

Diagramas de secuencias

La figura 3.7 muestra el diagrama de secuencias que describe el proceso e interacción de los componentes involucrados en el registro de un nuevo usuario en el módulo. Este

diagrama representa visualmente cómo se lleva a cabo el proceso de registro, mostrando las interacciones entre los diferentes componentes y cómo fluye la información durante este procedimiento.

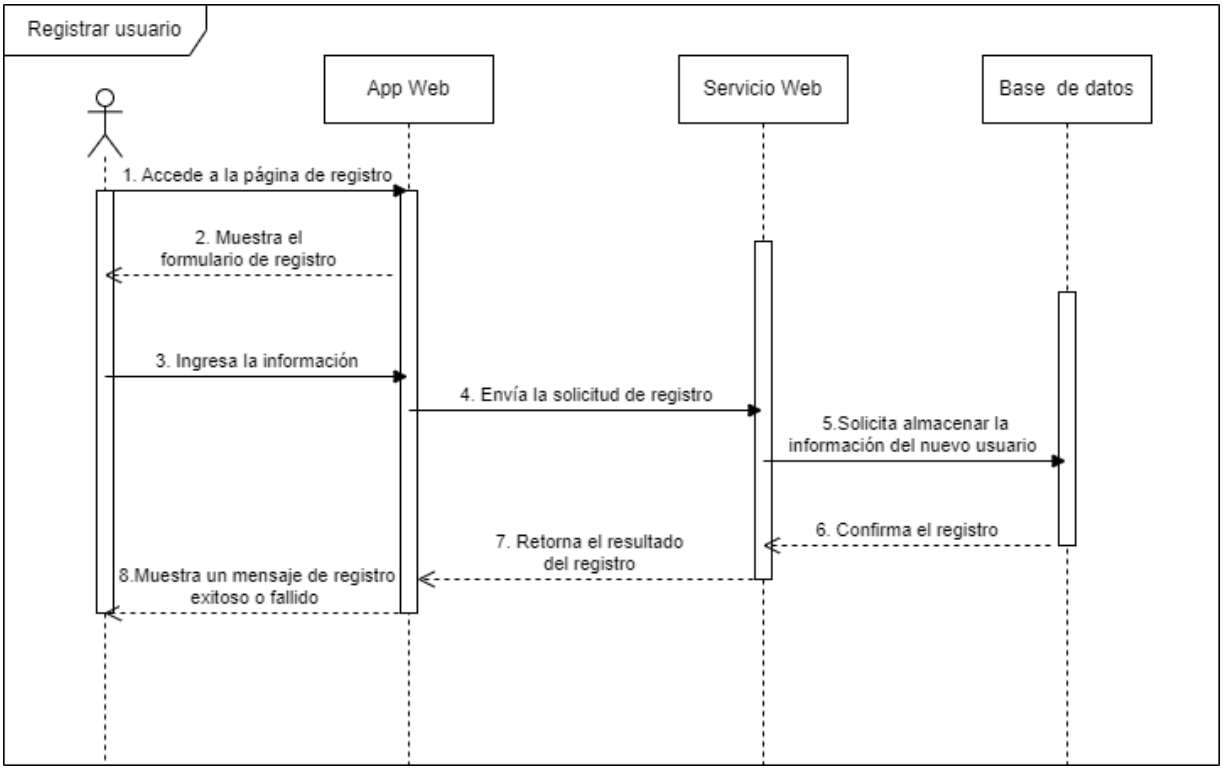


Figura 3.7 Diagrama de secuencia registro de usuario nuevo.

En la figura 3.8, se detalla el diagrama de secuencia que describe el proceso de autenticación que lleva a cabo un usuario previamente registrado al iniciar sesión en el sistema. Este diagrama proporciona una representación visual de las interacciones entre los componentes y las etapas involucradas en el proceso de autenticación, mostrando cómo se verifica la identidad del usuario y se le permite acceder al sistema.

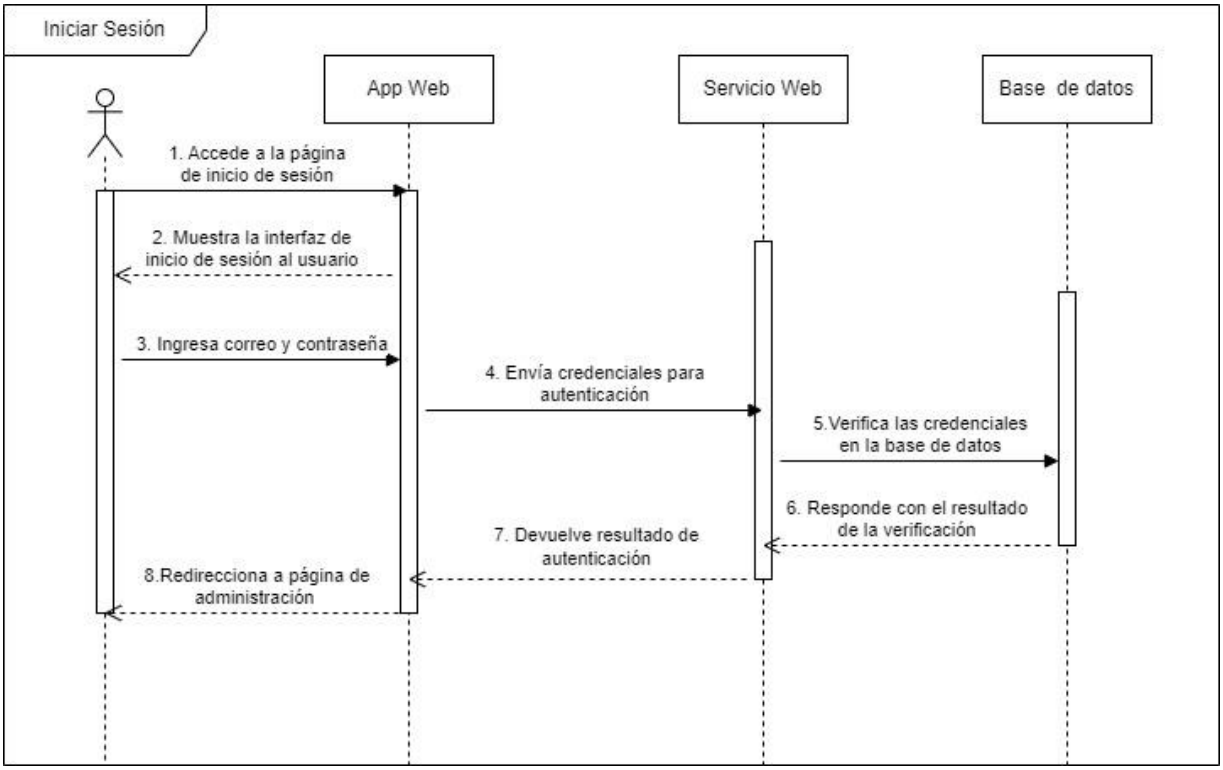


Figura 3.8 Diagrama de secuencia iniciar sesión.

En la figura 3.9, se presenta el diagrama de secuencia que detalla el proceso de recuperación de la contraseña de acceso. Este diagrama proporciona una representación visual de las interacciones y pasos involucrados en la solicitud y generación de una nueva contraseña por parte de un usuario que ha olvidado su contraseña anterior.

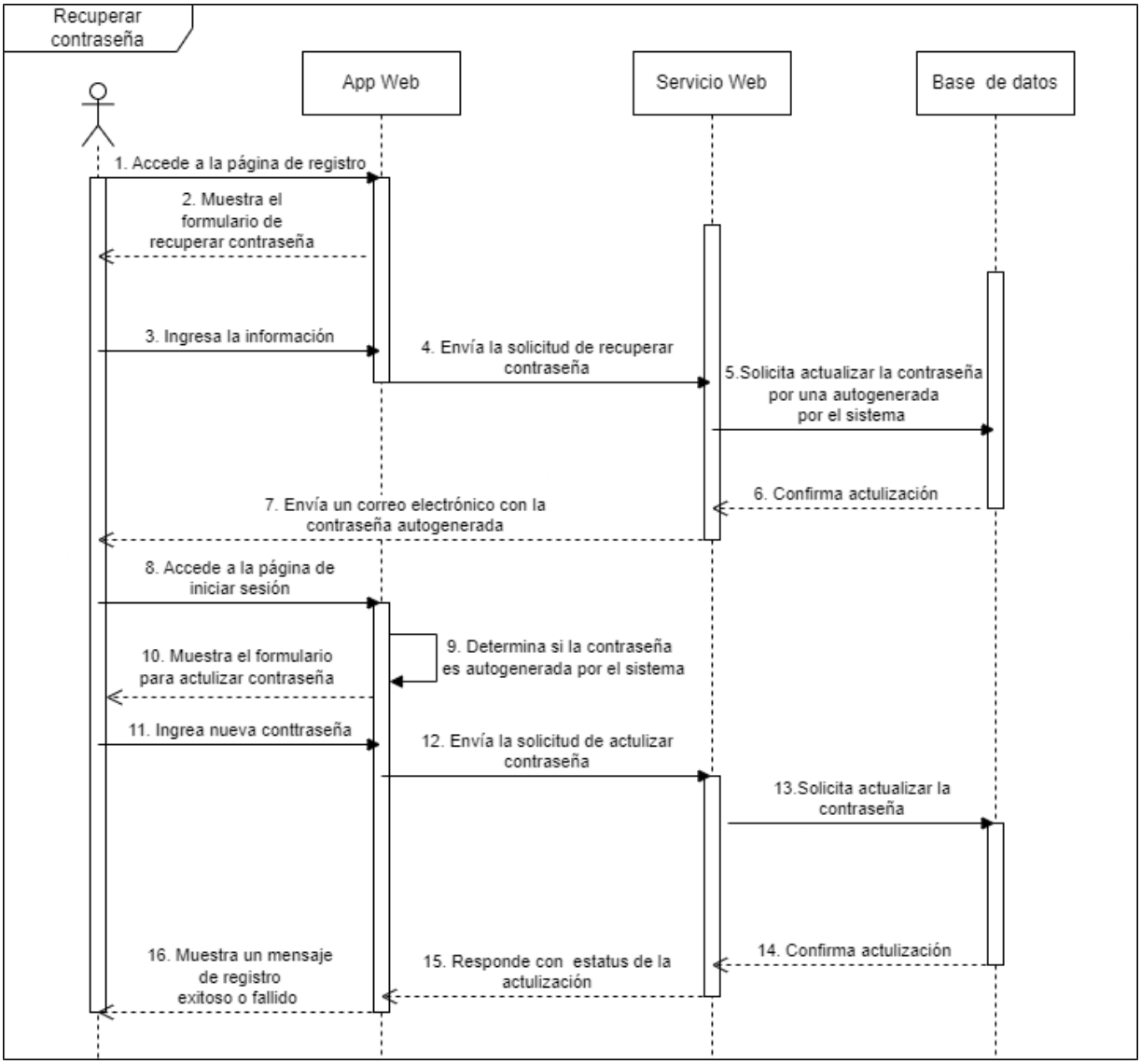


Figura 3.9 Diagrama de secuencia recuperar contraseña.

El diagrama de secuencia presentado en la figura 3.10 describe el proceso de consulta de documentos pendientes que requieren atención por parte del usuario. Este diagrama proporciona una representación visual de las interacciones y pasos involucrados en la revisión de documentos que están pendientes de ser atendidos por el usuario.

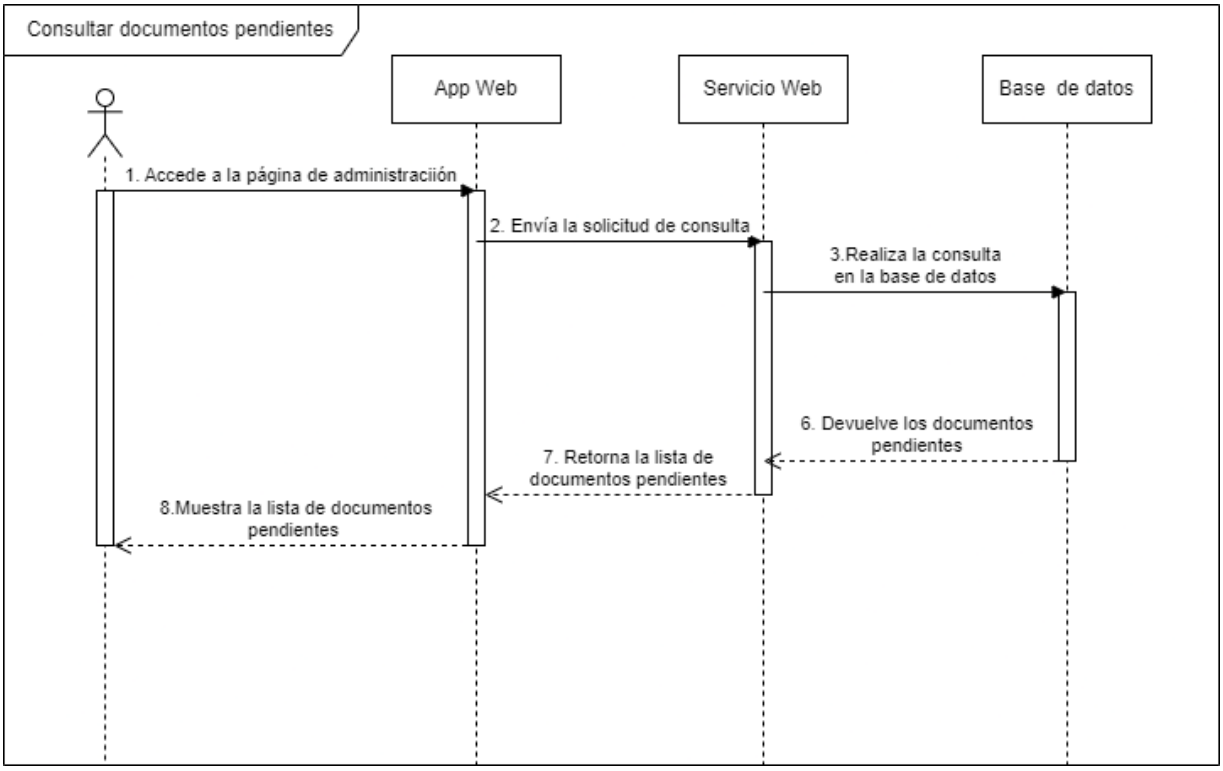


Figura 3.10 Diagrama de secuencia consultar documentos pendientes.

La figura 3.11 presenta el diagrama de secuencia que describe el procesamiento de documentos clínicos con OCR para la extracción de caracteres. Este diagrama ilustra de manera visual las interacciones y pasos involucrados en el uso del OCR para obtener información textual a partir de documentos clínicos.

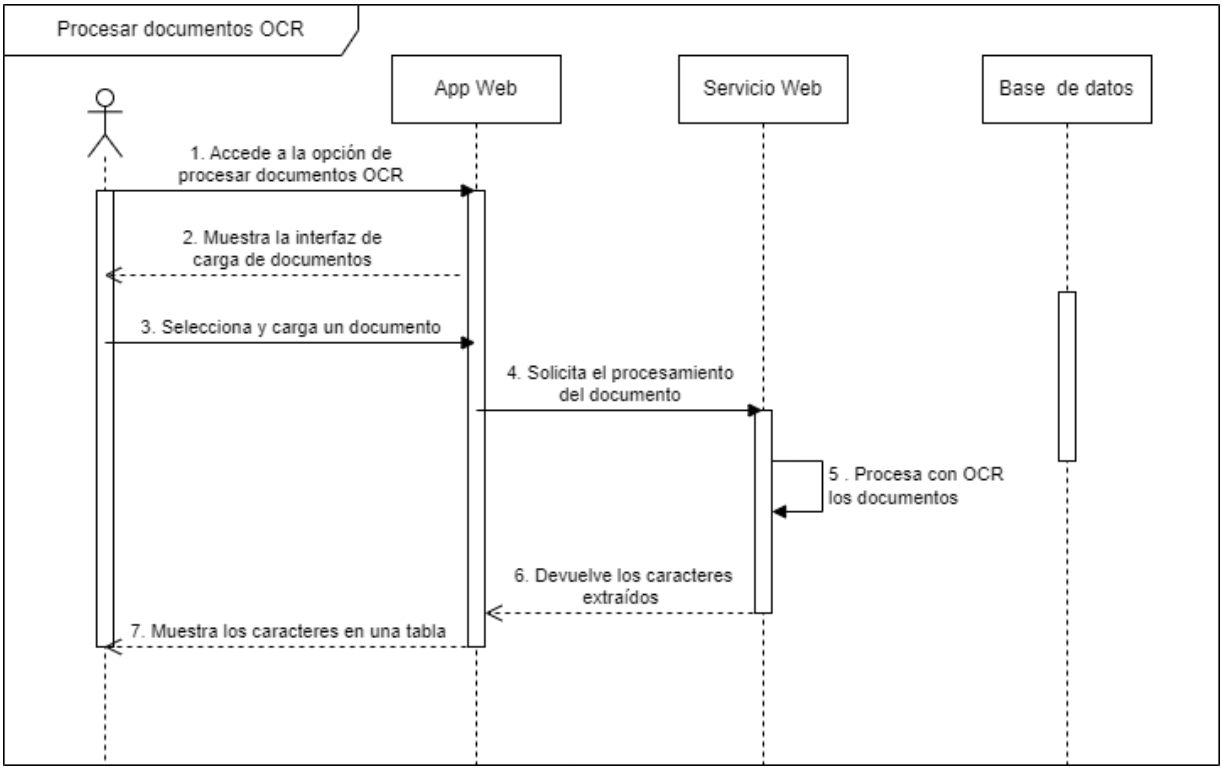


Figura 3.11 Diagrama de secuencia procesar documentos con OCR.

En la figura 3.12, se presenta el diagrama de secuencia que detalla el procedimiento de guardado de los caracteres extraídos de un documento. Este diagrama visualiza las interacciones y pasos involucrados en el proceso de almacenamiento de la información extraída de un documento.

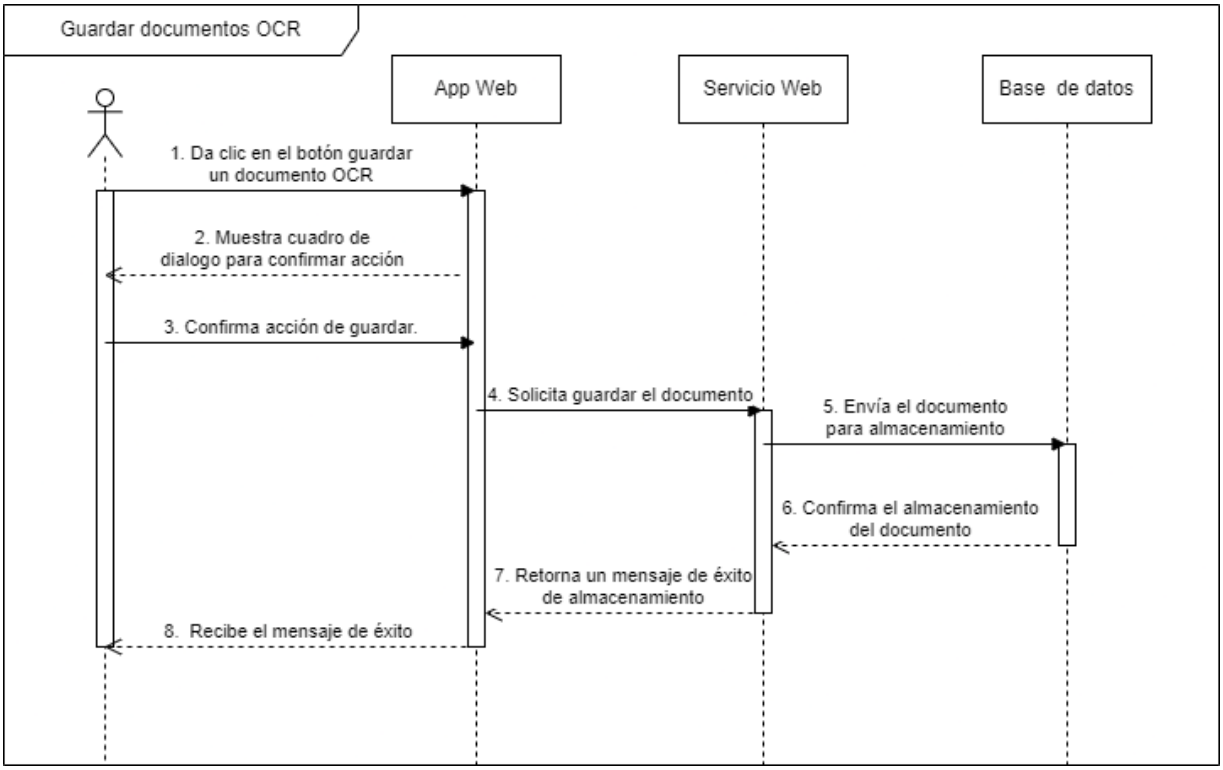


Figura 3.12 Diagrama de secuencia guardar documento.

Por otro lado, en la figura 3.13, se detalla el proceso de generación y almacenamiento de documentos en formato HL7. Este diagrama de secuencia representa las acciones e interacciones involucradas en la creación y almacenado de documentos en formato HL7.

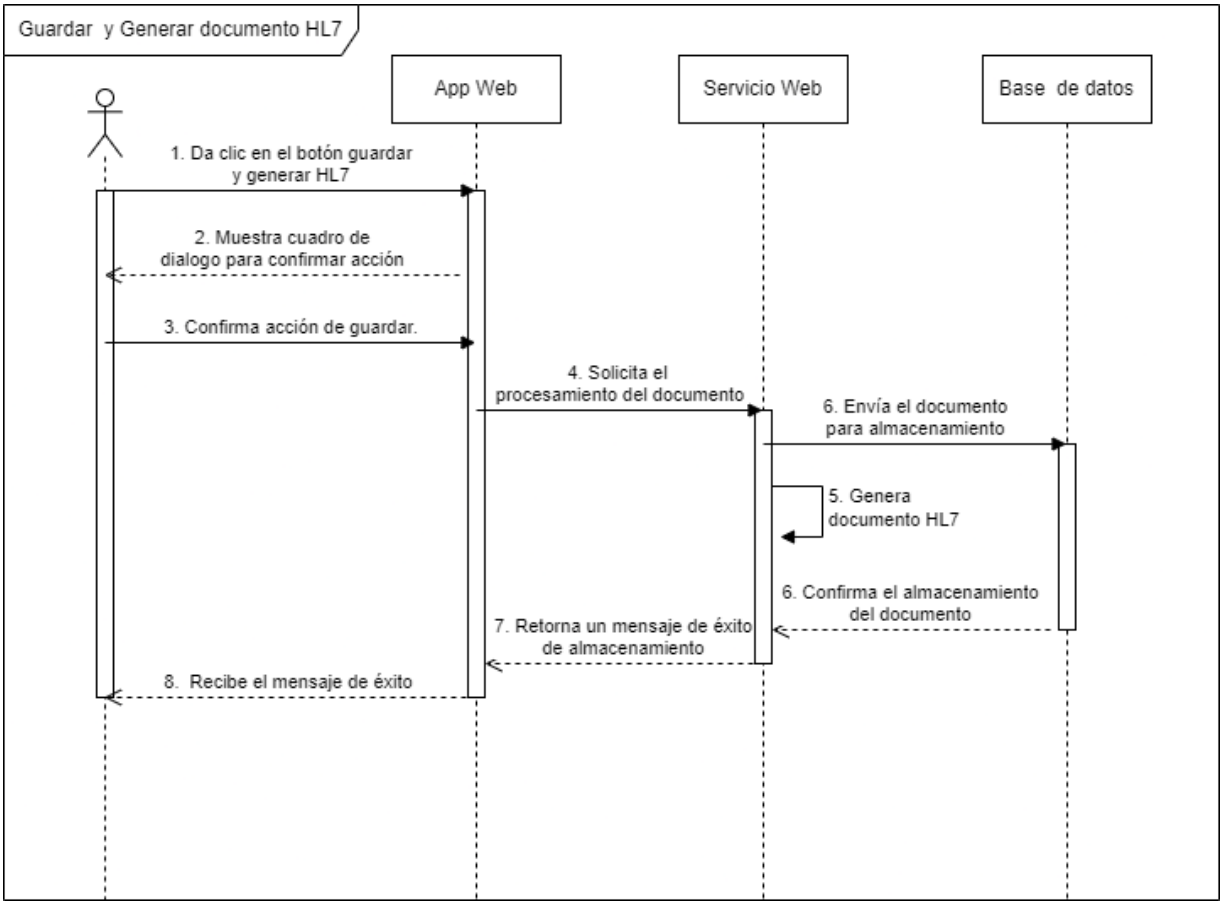


Figura 3.13 Diagrama de secuencia guardar y generar HL7.

Asimismo, en la figura 3.14 se describe la secuencia del proceso de validación de documentos en formato HL7. Este diagrama de secuencia representa las interacciones y las acciones involucradas en la verificación de documentos HL7.

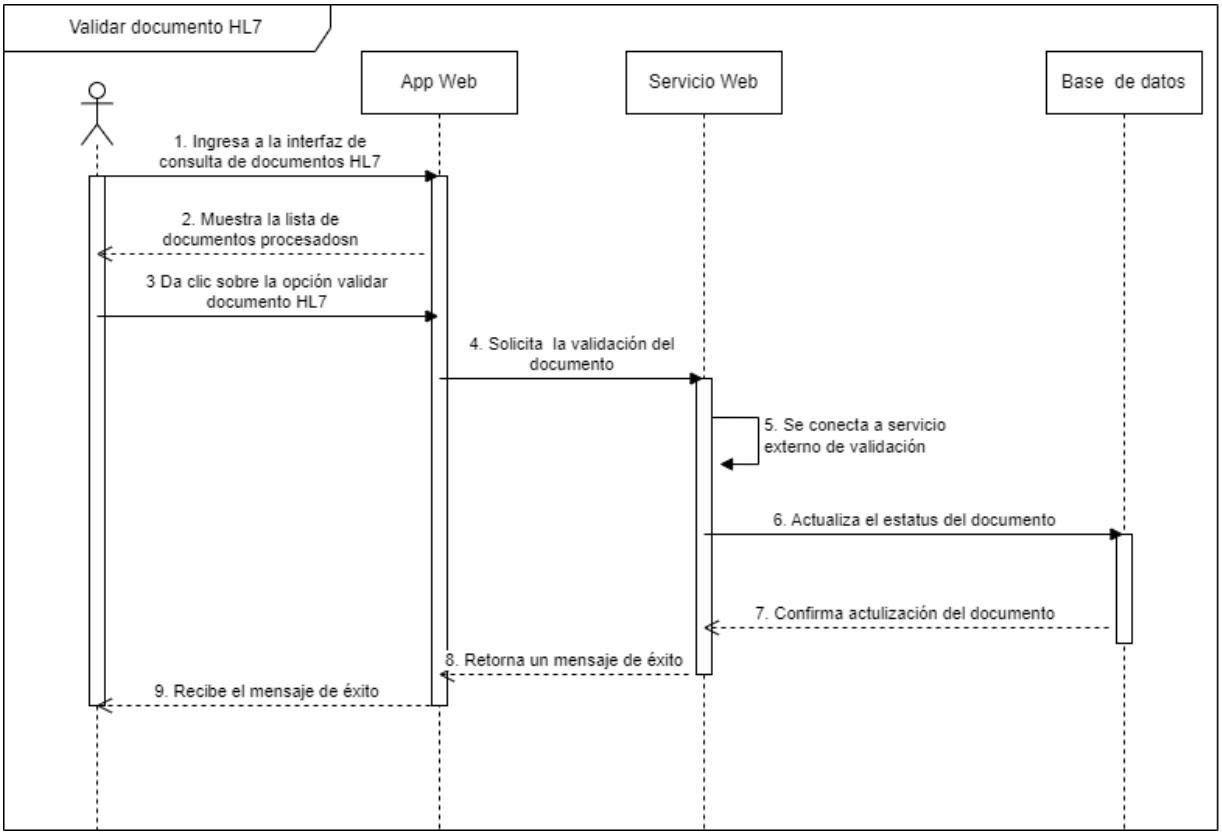


Figura 3.14 Diagrama de secuencia validar documento HL7.

La figura 3.15 representa la secuencia de interacciones y acciones involucradas en el proceso de descarga de documentos desde la aplicación. Este diagrama ilustra cómo un usuario interactúa con la interfaz de la aplicación para seleccionar un documento específico que desea descargar. Muestra la secuencia de eventos que ocurren desde el momento en que el usuario solicita la descarga hasta que el documento se descarga con éxito en el formato deseado PDF o XML.

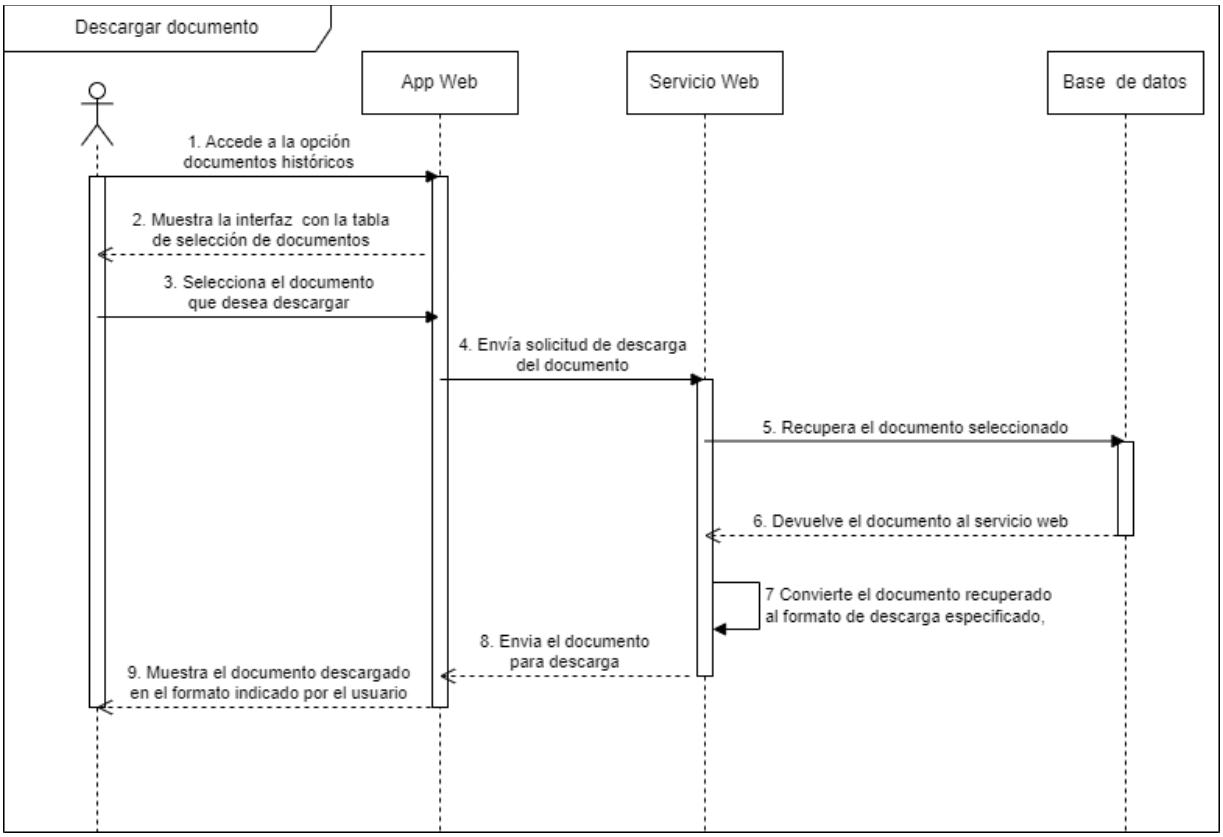


Figura 3.15 Diagrama de secuencia descargar documento.

Diagrama de clases

La figura 3.16 ilustra un diagrama de clases que representa las clases clave y sus relaciones en la capa lógica del módulo. Este diagrama proporciona una visión general de cómo estas clases interactúan en el sistema. Cada clase se presenta con sus atributos y métodos. Es importante destacar que el diagrama de clases no es particularmente complejo, ya que Python es un lenguaje multiparadigma por lo cual se hizo uso del paradigma de programación funcional la cual no se contempló en el diagrama de clases.

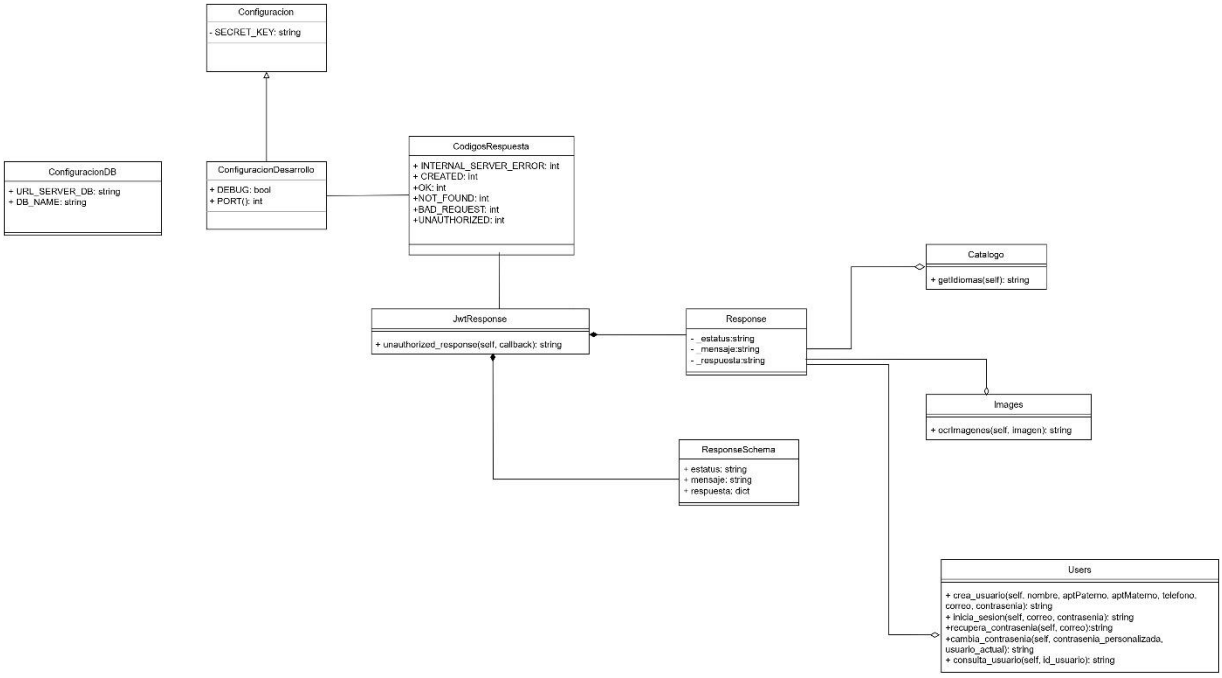


Figura 3.16 Diagrama de clases utilizadas.

3.1.2.2 Arquitectura del módulo de reconocimiento óptico de caracteres para interpretación de documentos clínicos a formato HL7

En este apartado, se presenta la arquitectura para desarrollar un módulo de reconocimiento óptico de caracteres destinado a la interpretación de documentos clínicos al estándar HL7, utilizando técnicas de procesamiento del lenguaje natural. La arquitectura, que se muestra en la figura 3.17, se basa en un diseño por capas que simplifica su mantenimiento. Cada capa engloba los componentes necesarios para su funcionamiento y se describen de manera concisa a continuación.

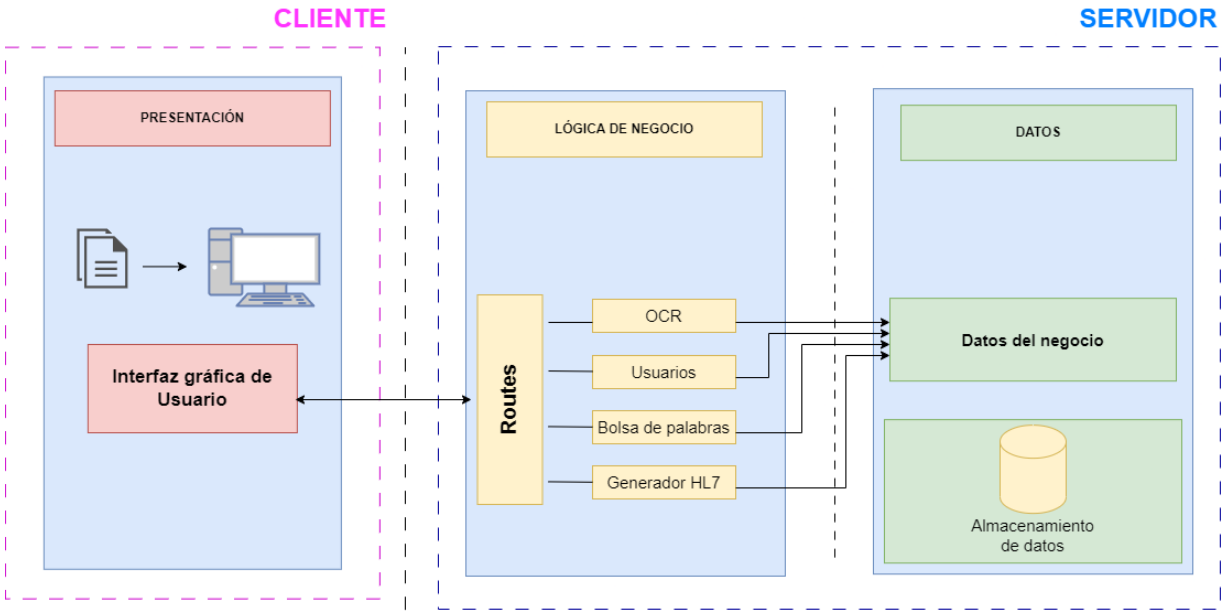


Figura 3.17 Arquitectura propuesta para el módulo.

Capa de presentación: Esta capa presenta la interfaz gráfica del módulo, que facilita la interacción del usuario con el sistema, incluyendo la manipulación de documentos digitales. En esta capa, se permite el envío de documentos para su procesamiento en uno de los siguientes formatos: JPG, PNG o PDF.

Capa lógica de negocio: En la capa, se integran funcionalidades esenciales que abarcan desde el reconocimiento óptico de caracteres hasta la traducción de términos clínicos al formato HL7. El proceso se inicia con el OCR, que comprende el escaneo óptico, la segmentación, el preprocesamiento, la representación, la extracción de

características, el entrenamiento y reconocimiento, el posprocesamiento y la generación del texto de salida. Asimismo, se utiliza una bolsa de palabras que alberga términos clínicos relevantes y se verifica su correspondencia semántica con los términos extraídos por OCR. La validación del documento se efectúa mediante técnicas de NLP. Finalmente, se traducen los términos validados al formato HL7.

Capa de datos: Tras el proceso de transformación de documentos al estándar HL7, los datos se almacenan de forma permanente en un Sistema de Gestión de Bases de Datos (SGBD).

La arquitectura propuesta en la figura 3.17 es una contribución al sector de la salud, ya que se espera que facilite la interoperabilidad entre sistemas de salud. Esto ayudará al personal especializado en el campo de la salud a intercambiar información clínica del paciente de manera más eficiente al unificar los diferentes formatos clínicos bajo un estándar común.

3.1.2.3 Modelo de datos

En la Figura 3.18 se presenta el modelo de datos que ha sido implementado para almacenar tanto la información de los usuarios como los documentos que han sido procesados mediante OCR o transformados al formato HL7.

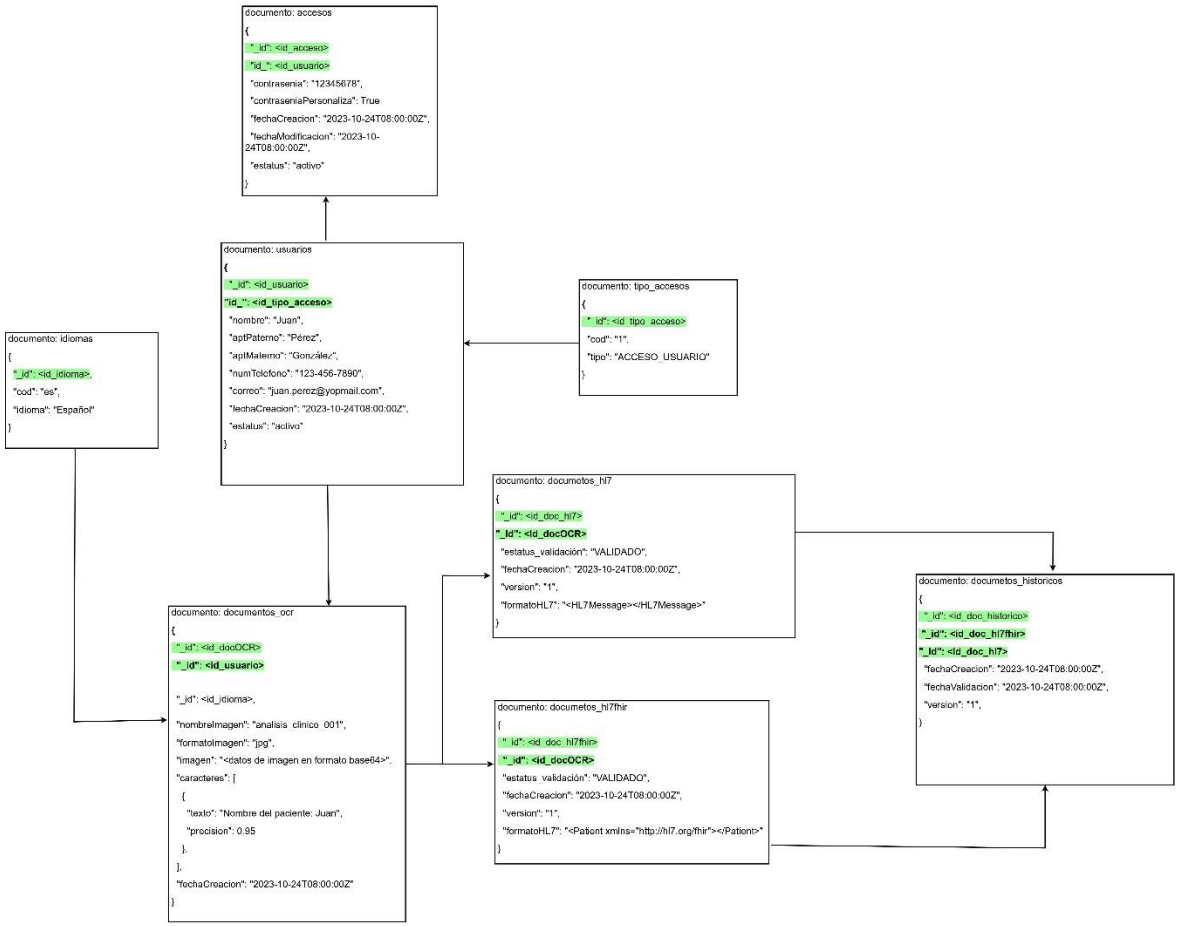


Figura 3.18 Modelo de datos.

3.1.2.4 Maquetación del módulo (*Mockups*)

La maquetación del módulo se llevó a cabo utilizando mockups, que son modelos que permiten mostrar al cliente la apariencia propuesta para la aplicación. Un mockup representa tanto la arquitectura de la aplicación como su apariencia.

La figura 3.19 ilustra la interfaz de la página de inicio, donde el usuario debe seleccionar una de las dos opciones disponibles: “Registro” o “Acceso”.

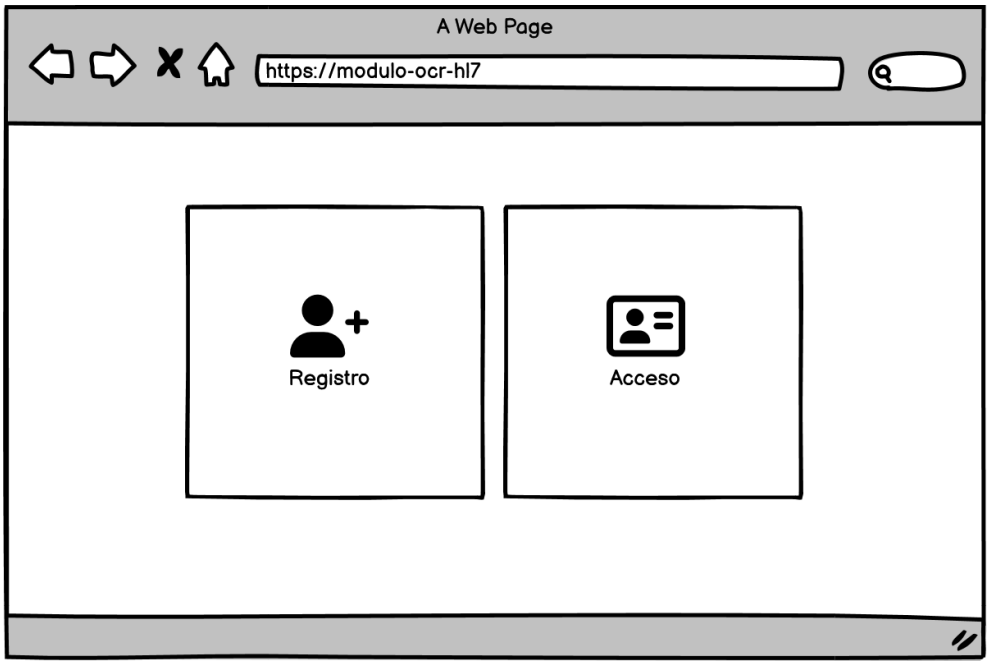


Figura 3.19 Interfaz página de inicio.

Además, la figura 3.20 muestra la interfaz de la página de registro de usuarios. En la parte inferior, se encuentra un enlace para acceder a la página de acceso, mientras que en el centro se ubica un formulario para registrarse.

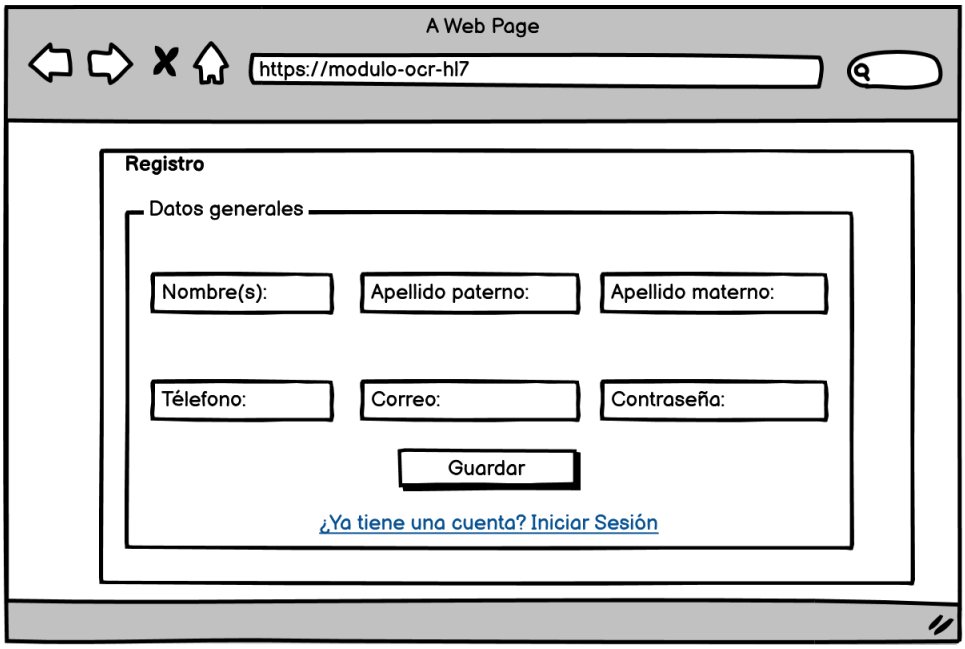


Figura 3.20 Interfaz registro de usuarios.

La Figura 3.21 ilustra la interfaz de la página de autenticación de usuarios (*login*). En la parte superior, se encuentra un icono de usuario, y en el centro, se sitúan los campos para ingresar el correo electrónico y la contraseña.

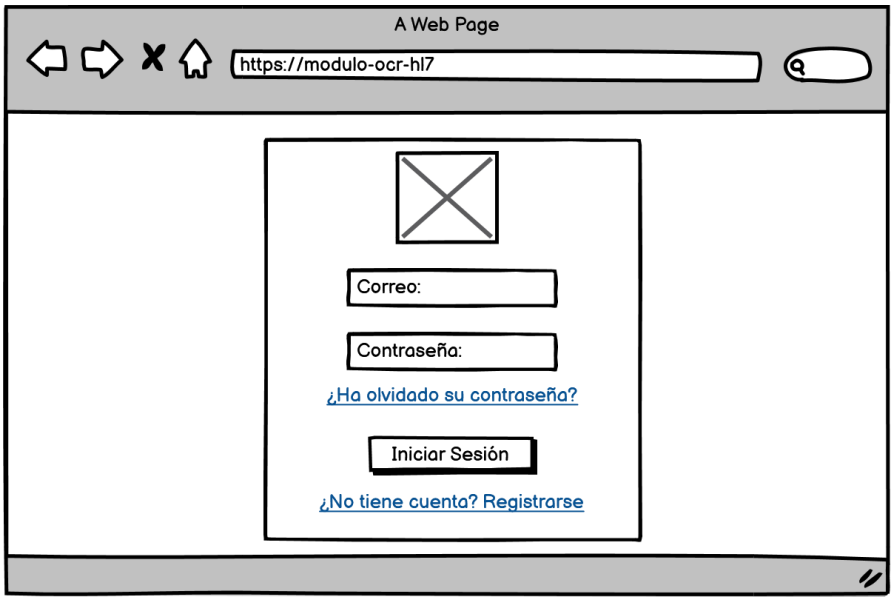


Figura 3.21 Interfaz de autenticación.

Por otro lado, la figura 3.22 muestra la página de recuperación de contraseña de

acceso. En esta página, se encuentra un campo donde el usuario ingresa su dirección de correo electrónico para solicitar la recuperación de su contraseña.

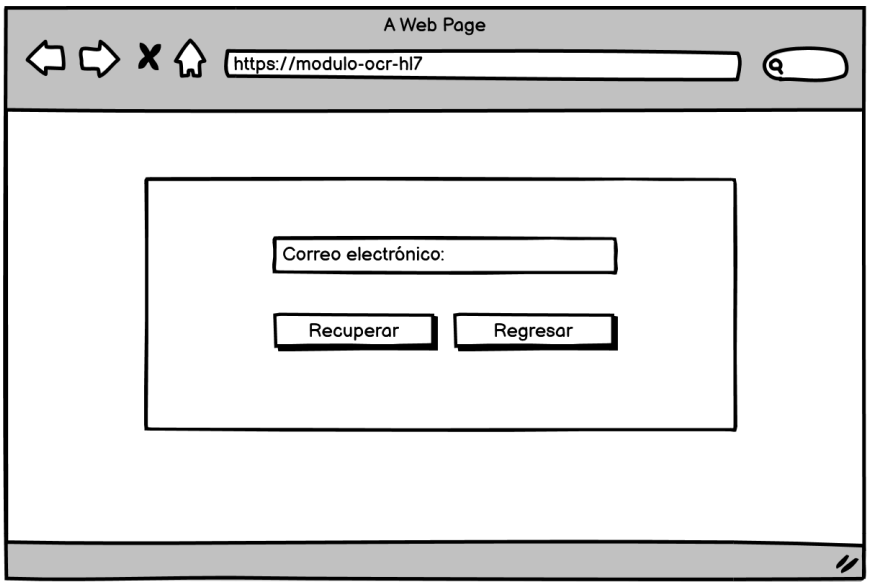


Figura 3.22 Interfaz recuperar contraseña.

Además de lo anterior, la figura 3.23 muestra la página para cambiar la contraseña, la cual incluye un campo para ingresar la nueva contraseña, así como otro campo para confirmar la contraseña.

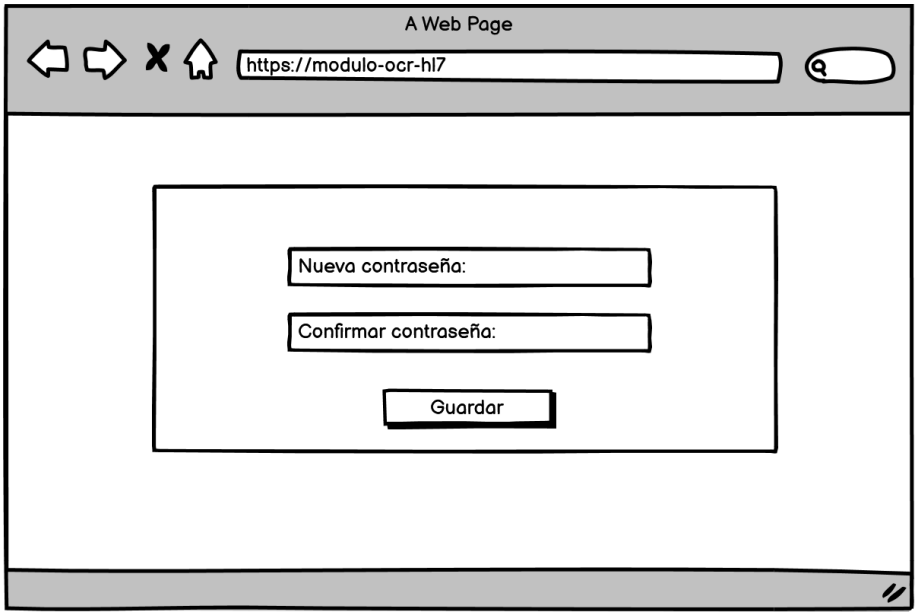


Figura 3.23 Interfaz cambiar contraseña.

Por otro lado, en la figura 3.14 se muestra el panel de administración el cual es

accesible para el usuario una vez realizada autenticación. En esta sección, el usuario visualiza, en primera instancia, los documentos generados más recientemente, así como el listado de documentos que requieren atención.

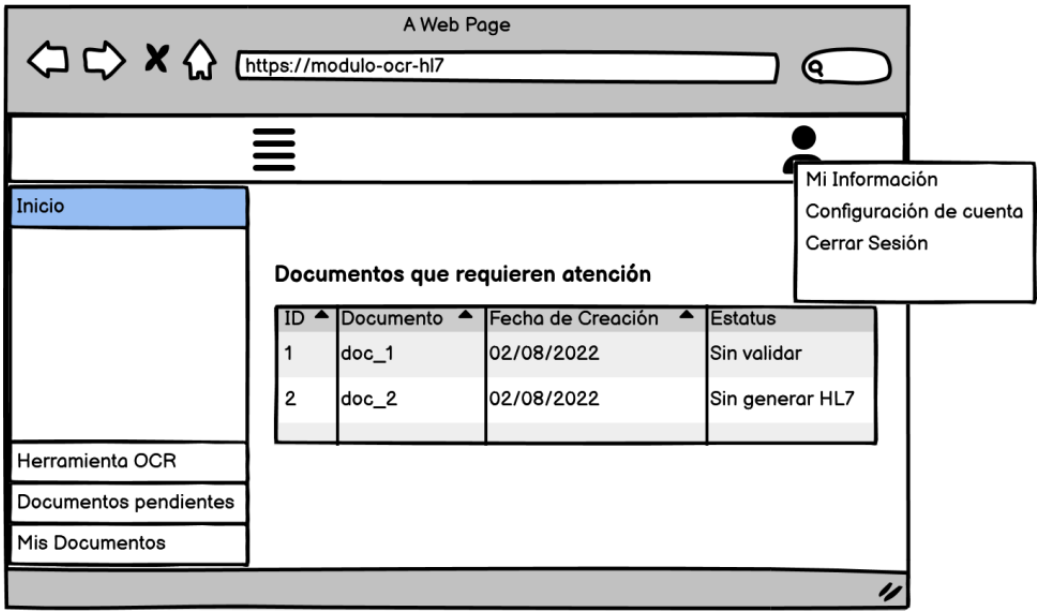


Figura 3.24 Interfaz de administración.

Otra sección importante dentro del área administrativa es la herramienta OCR, la cual se muestra en la figura 3.25. Esta página dispone de un explorador de archivos que permite seleccionar los documentos clínicos para su procesamiento mediante la biblioteca OCR. Además, en esta ventana se incluye una sección para generar documentos en formato HL7 a partir de los datos extraídos mediante el OCR.

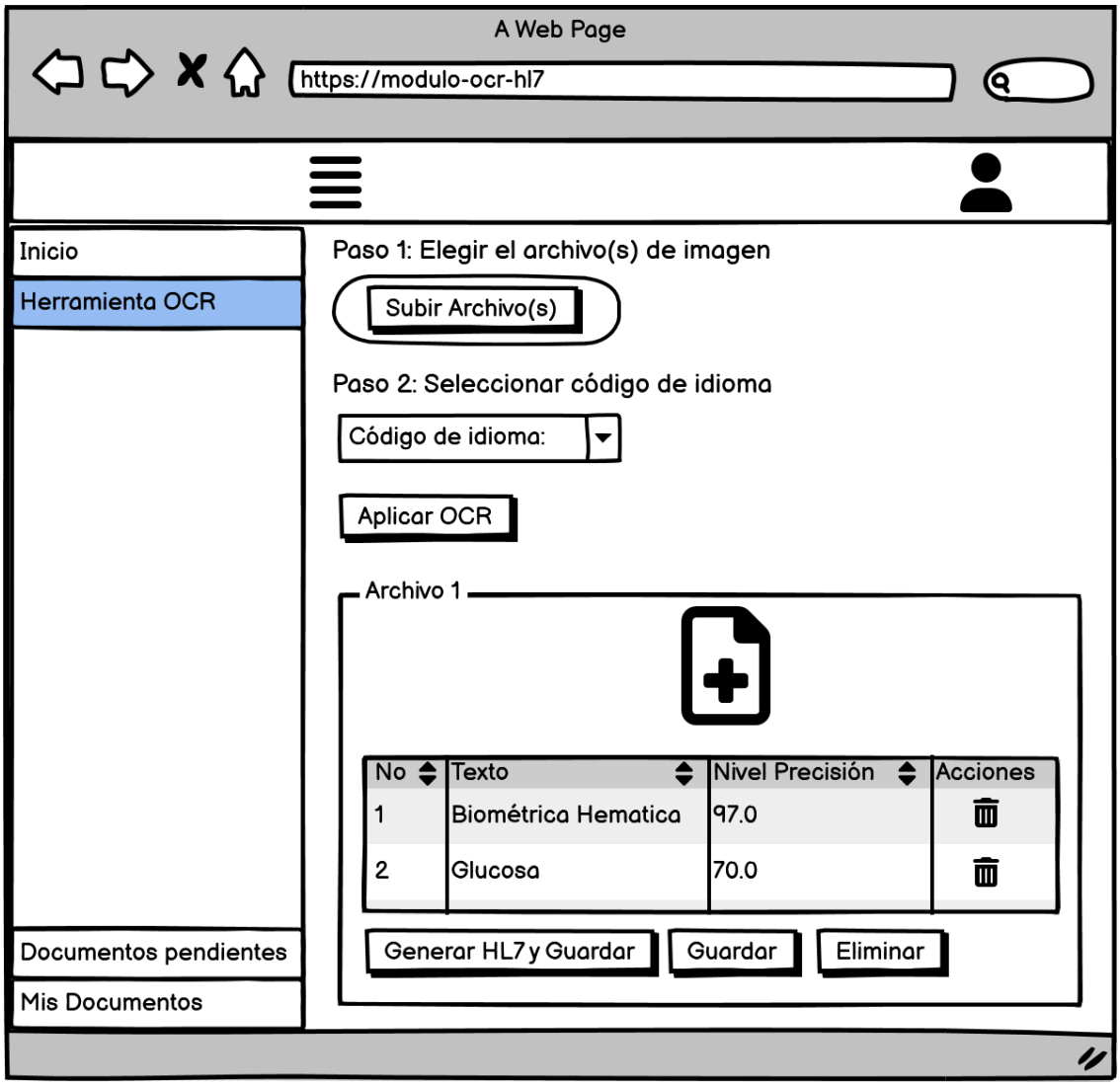


Figura 3.25 Interfaz herramienta OCR.

Asimismo, en el área administrativa, como se observa en la figura 3.26, se encuentra la sección de “Documentos y Pendientes” que presenta un listado de los documentos clínicos procesados con OCR, así como los documentos clínicos transformados a HL7 v3 y HL7 FHIR.

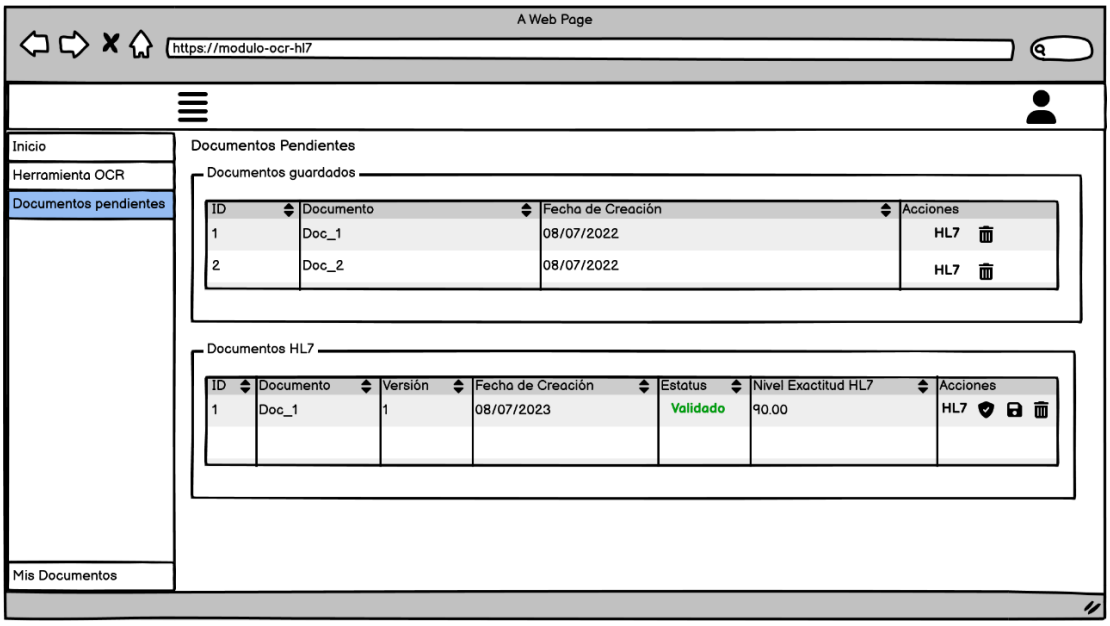


Figura 3.26 Interfaz documentos y pendientes.

Además, la sección “Mis Documentos” muestra un histórico de todos los documentos clínicos en formato HL7 v3 y HL7 FHIR generados, como se ilustra en la Figura 3.27.

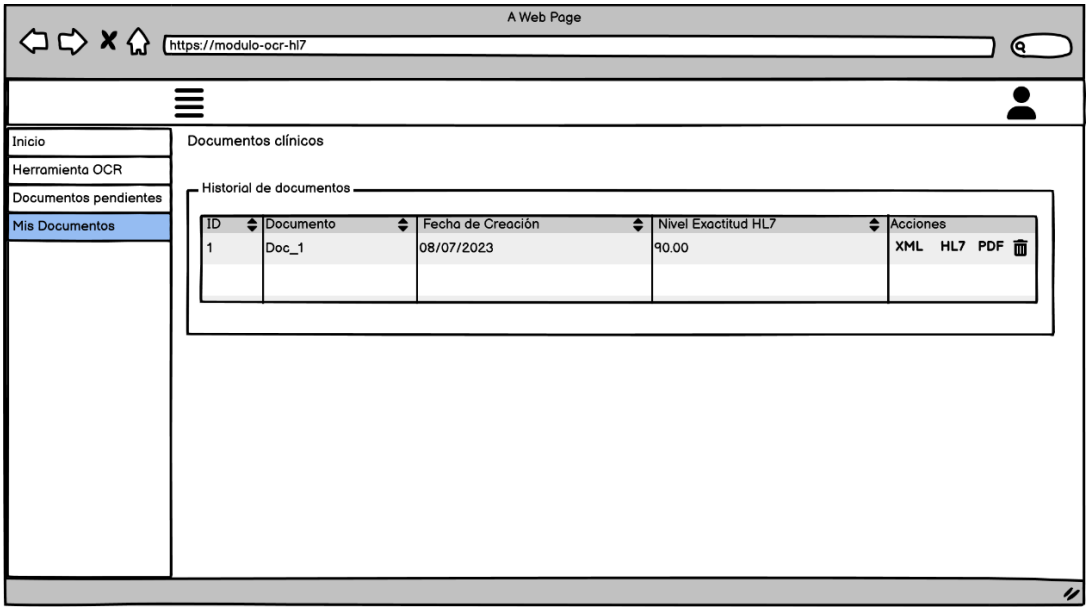


Figura 3.27 Interfaz mis documentos históricos.

3.1.3 Fase de codificación

Durante la fase de codificación, siguiendo la metodología XP, se procede a desarrollar el módulo Web. En esta etapa, los requisitos y los diagramas se traducen en código fuente, y se emplean las tecnologías mencionadas en el capítulo anterior. A continuación, se proporciona una descripción detallada del proceso de desarrollo de cada uno de los componentes del módulo, junto con una explicación de su funcionamiento.

3.1.3.1 Codificación del módulo (lado del servidor)

Como se mencionó previamente, el diseño del módulo sigue una arquitectura en capas con el propósito de separar la lógica del negocio y facilitar su mantenimiento en el futuro. En este apartado, se detalla el funcionamiento del módulo en el lado del servidor, que se compone de un proveedor de servicios Web basado en REST. Este proveedor ofrece una serie de puntos de acceso (*endpoints*) que la aplicación cliente consume.

A continuación, se presenta la estructura del proveedor de servicios Web, desarrollado utilizando el marco de trabajo Flask en su versión 2.0.3. El proyecto se organiza en tres carpetas principales: "*app*", "*routes*" y "*models*", las cuales se describen brevemente a continuación.

- **app**: Esta carpeta alberga el módulo principal de la aplicación, que se encarga de incorporar las dependencias y configuraciones necesarias para el funcionamiento adecuado de la aplicación.
- **routes**: En esta carpeta se encuentran los archivos donde se registran todas las rutas a las que se puede acceder dentro de la aplicación. Es importante mencionar que en esta carpeta se especifican los prefijos de las URL (*Uniform Resource Locator*) para acceder a los recursos.
- **models**: En esta carpeta se encuentran las clases que se llaman en diversas partes de la aplicación. Esto se hace con el objetivo de facilitar el acceso a los métodos y datos necesarios para el funcionamiento de la aplicación.

A continuación, se muestran fragmentos de código correspondientes a los procesos

más significativos que realiza el proveedor de servicios Web.

En la lista 3.1 se muestra un fragmento de la clase "*Images*" ubicada dentro de la carpeta "*models*", que proporciona la funcionalidad para realizar OCR en una imagen y entregar los resultados correspondientes. Estos resultados incluyen la imagen original con rectángulos que indican el texto reconocido, el texto extraído y otros detalles relacionados con la imagen.

```
1. def ocrImagenes(self, imagen):
2.         arrTextos = []
3.
4.         img = cv2.imdecode(numpy.fromstring(imagen.read(), numpy.uint8),
   cv2.IMREAD_COLOR)
5.
6.         texto_reader = pytesseract.image_to_data(img,
   output_type='data.frame')
7.         texto_img = pytesseract.image_to_data(img,
   output_type=Output.DICT)
8.
9.         texto_reader = texto_reader[texto_reader.conf != -1]
10.        texto_reader.head()
11.        print('Texto extraido ', texto_reader)
12.        n_textos = texto_reader['text']
13.        n_precision = texto_reader['conf']
14.        for t, p in zip(n_textos, n_precision):
15.            arrTextos.append({'texto': t, 'precision': p})
16.
17.        n_box = len(texto_img['level'])
18.        for index in range(n_box):
19.            (x, y, w, h) = (texto_img['left'][index],
   texto_img['top'][index],
20.
   texto_img['width'][index],texto_img['height'][index])
21.            cv2.rectangle(img, (x, y), (x + w, y + h), (0, 255, 0), 2)
22.
23.        file_name, file_extension = os.path.splitext(imagen.filename)
24.
25.        _, img_encoded = cv2.imencode(file_extension, img)
26.        img_b46 = base64.b64encode(img_encoded).decode()
27.
28.        data = {
```

```
            'nombreImagen': file_name,
            'formatoImagen': file_extension,
            'imagen': str(img_b46),
            'caracteres': arrTextos
        }
        return data
```

Lista 3.1 Fragmento de código para procesar archivos con OCR.

En la lista 3.2, se presenta un fragmento de código localizado en la carpeta "models." Este método se encarga de validar la solicitud (*request*) que contiene los caracteres extraídos de los documentos clínicos. La validación se realiza utilizando un modelo de bolsa de palabras (BoW) y los vocabularios clínicos proporcionados por DeCS en formato XML que se depositan en la carpeta *assets* del proyecto.

```
def buscar_palabras_clinicas(corpus):
    directorio_vocabulario = 'assets/decs/'

    fecha_actual = datetime.now()

    fecha_mas_cercana = None
    archivo_mas_cercano = None

    response = Response()
    responseSchema = ResponseSchema()

    for archivo in os.listdir(directorio_vocabulario):
        if archivo.endswith(".xml"):
            nombre_archivo, extension = os.path.splitext(archivo)
            fecha_archivo = datetime.strptime(nombre_archivo.split('_')[-1], '%d-%m-%Y')

            diferencia_dias = abs((fecha_actual - fecha_archivo).days)

            if fecha_mas_cercana is None or diferencia_dias < fecha_mas_cercana:
                fecha_mas_cercana = diferencia_dias
                archivo_mas_cercano = archivo

    if archivo_mas_cercano:
```

```
        ruta_archivo = os.path.join(directorio_vocabulario, archivo_mas_cercano)

        # Cargar el archivo XML del vocabulario DeCS
        tree = ET.parse(ruta_archivo)
        root = tree.getroot()

        palabras_clinicas = []
        for term in root.iter('term'):
            palabra_clinica = term.text.strip()
            palabras_clinicas.append(palabra_clinica)

        # Crear un objeto CountVectorizer
        vectorizer = CountVectorizer(vocabulary=palabras_clinicas)

        # Inicializar el contador de palabras
        palabras_clinicas_presentes = 0

        for entrada in corpus:
            texto = entrada.get("texto", "")

            # Transformar el texto en una matriz BoW
            X = vectorizer.transform([texto])

            # Obtener una representación en formato de matriz
            bow_matrix = X.toarray()

            # Verificar si las palabras de DeCS están presentes en el texto
            palabras_presentes = [palabras_clinicas[j] for j in range(len(palabras_clinicas)) if bow_matrix[0][j] > 0]

            # Incrementar el contador de palabras
            palabras_clinicas_presentes += len(palabras_presentes)

        # Devolver el numero total de palabras presentes
        return palabras_clinicas_presentes
    else:
      response.estatus = 'ERROR'
        response.mensaje = 'No se encontraron archivos XML en el directorio.'
        response.respuesta = {}
        result = responseSchema.dump(response)
```

```
63.         return jsonify(result), CodigosRespuesta.INTERNAL_SERVER_ERROR
```

Lista 3.2 Fragmento de código para validar términos clínicos.

Por otro lado, en la lista 3.3 se muestra un fragmento de código que se encarga de transformar los datos de análisis clínicos al formato HL7. Es importante destacar que este método tiene la capacidad de interpretar documentos clínicos y convertirlos al formato HL7 v3.

```
1. from lxml import etree
2. 
3. def resultados_a_hl7(resultados):
4.     hl7_message = etree.Element("HL7Message")
5. 
6.     # Segmento PID (Informacion del Paciente)
7.     pid = etree.SubElement(hl7_message, "PID")
8.     pid.text =
   f"1|123456|{resultados['nombre']}|{resultados['fecha_nacimiento']}|M"
9. 
10.    # Segmento OBR (Orden de Laboratorio) desde los argumentos
11.    obr = etree.SubElement(hl7_message, "OBR")
12.    obr.text =
   f"1|{resultados['codigo_orden']}|{resultados['nombre_orden']}|{resultados
   ['fecha_orden']}|{resultados['fecha_resultado']}||||{resultados['id_solici
   tante']}^{resultados['nombre_solicitante']}||||{resultados['fecha_resultad
   o']}"
13. 
14.    # Segmento OBX (Resultados de Pruebas)
15.    for i, resultado in enumerate(resultados['pruebas'], start=1):
16.        obx = etree.SubElement(hl7_message, "OBX")
17.        obx.text =
   f"{i}|NM|{resultado['codigo']}^{resultado['nombre']}^{resultado['ubicacio
   n']}^L|1|{resultado['valor']}|{resultado['unidad']}|||F"
18. 
19.    return etree.tostring(hl7_message, pretty_print=True,
   encoding="unicode")
```

Lista 3.3 Fragmento de código transformar datos a HL7.

Asimismo, en la lista 3.4 se muestra un fragmento de código que permite transformar los datos clínicos al formato HL7 FHIR. Esto se realizó con la finalidad de contar con otro formato HL7 que sea compatible con un mayor número de sistemas de salud en México.

```
1. def resultados_a_fhir(paciente_id, resultados):
2.     root = ET.Element("Bundle")
3.     root.set("xmlns", "http://hl7.org/fhir")
4.
5.     # Agregar informacion del paciente
6.     patient_entry = ET.SubElement(root, "entry")
7.     patient_resource = ET.SubElement(patient_entry, "Patient")
8.     ET.SubElement(patient_resource, "id").text = paciente_id
9.     # Agregar resultados de laboratorio
10.    for result in resultados:
11.        result_entry = ET.SubElement(root, "entry")
12.        result_resource = ET.SubElement(result_entry, "Observation")
13.
14.        ET.SubElement(result_resource, "status").text = "final"
15.        ET.SubElement(result_resource, "category", {"value":
   "laboratory"})
16.        ET.SubElement(result_resource, "code", {"valueSet":
   "http://loinc.org", "value": result["loinc_code"]})
17.        ET.SubElement(result_resource, "valueQuantity", {"value":
   str(result["value"]), "unit": result["unit"]})
18.
19.    tree = ET.ElementTree(root)
20.    return ET.tostring(root, encoding="utf-8", method="xml")
```

Lista 3.4 Fragmento de código transformar datos a HL7 FHIR.

3.1.3.2 Codificación del módulo (lado del cliente)

En el contexto anterior y siguiendo el enfoque de arquitectura en capas, se desarrolla una aplicación Web cliente utilizando el marco de trabajo Angular en su versión 14.0.7, que actualmente cuenta con un amplio soporte.

Una vez completada la configuración de las dependencias necesarias, se procede a la codificación de la aplicación cliente y a la integración del proveedor de servicios Web previamente desarrollado. Estos pasos se aprecian en las siguientes figuras.

Página de inicio

Cuando un usuario accede a la URL a través de un navegador Web, se encontra con la página de inicio que se muestra en la figura 3.28. En esta página, se presentan las opciones disponibles para el usuario.

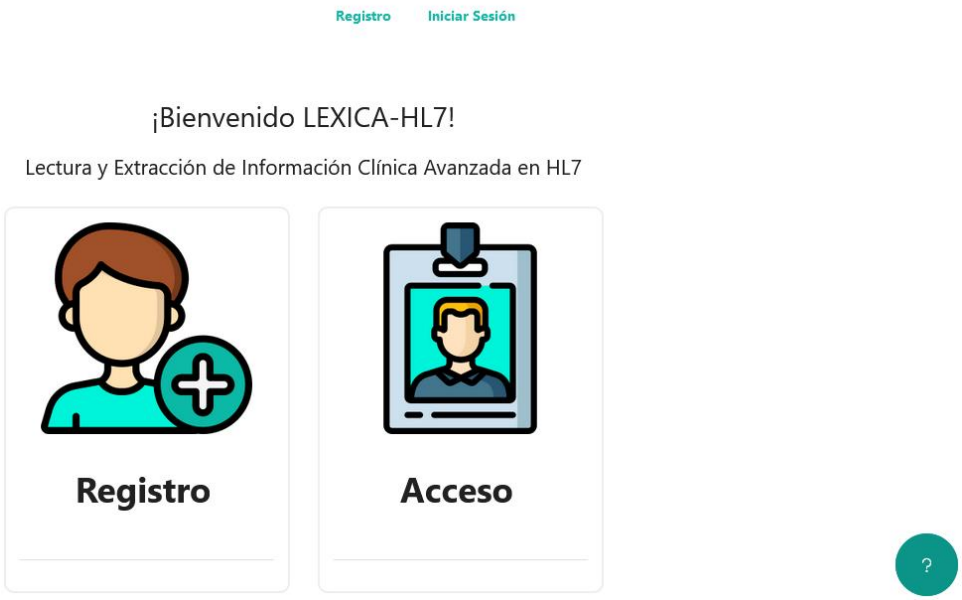


Figura 3.28 Página de inicio.

Página de registro

La página de registro, muestra en la figura 3.29, se diseñó con un formulario para completar por los usuarios que acceden a la URL y no cuenten con un perfil. Este formulario valida cada dato proporcionado por el usuario para asegurar que se complete correctamente.

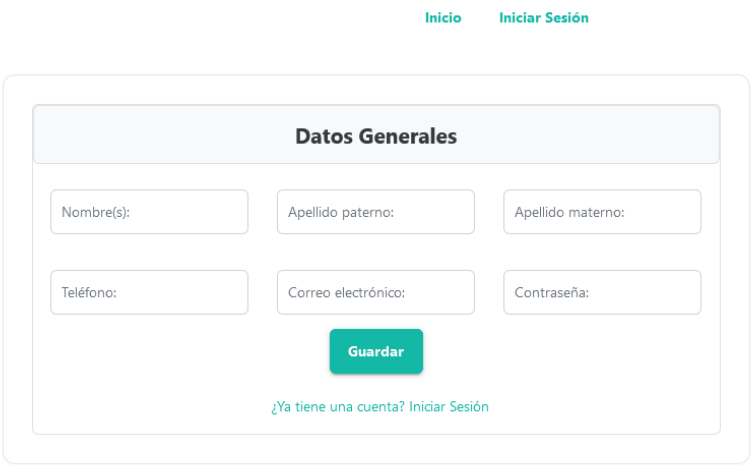


Figura 3.29 Página de registro.

Página de inicio de sesión

La página de inicio de sesión, como se muestra en la figura 3.30, se desarrolló para permitir que los usuarios se autentiquen mediante la validación de su dirección de correo electrónico y contraseña, lo que les permite acceso a sus perfiles.

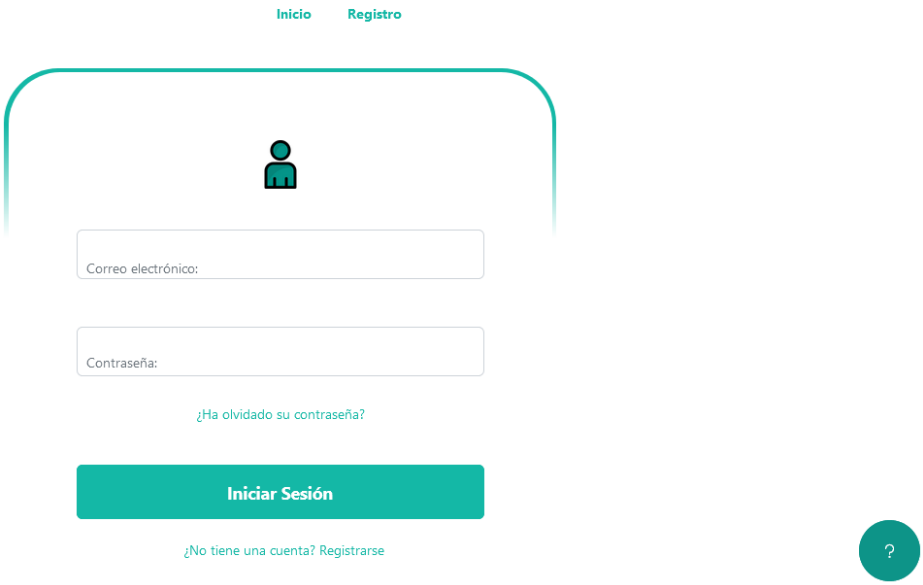


Figura 3.30 Página de inicio de sesión.

Página de panel de administración

Cuando un usuario inicia sesión, accede al panel administrativo como se muestra en la figura 3.31. En este panel, se presenta un listado de documentos clínicos que requieren atención, junto con varias opciones de navegación.

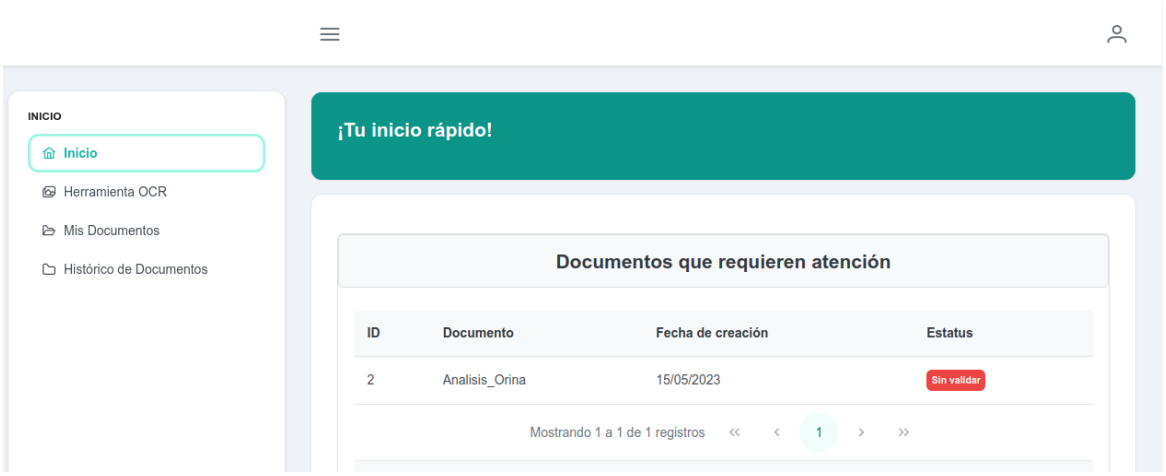


Figura 3.31 Página panel de administración.

Página de herramienta OCR

La figura 3.32 presenta la página de la herramienta OCR, desarrollada con un campo de exploración de archivos que permite a los usuarios cargar los documentos clínicos, ya sean imágenes o archivos PDF. Esto marca el inicio del proceso de detección de caracteres en los documentos clínicos.

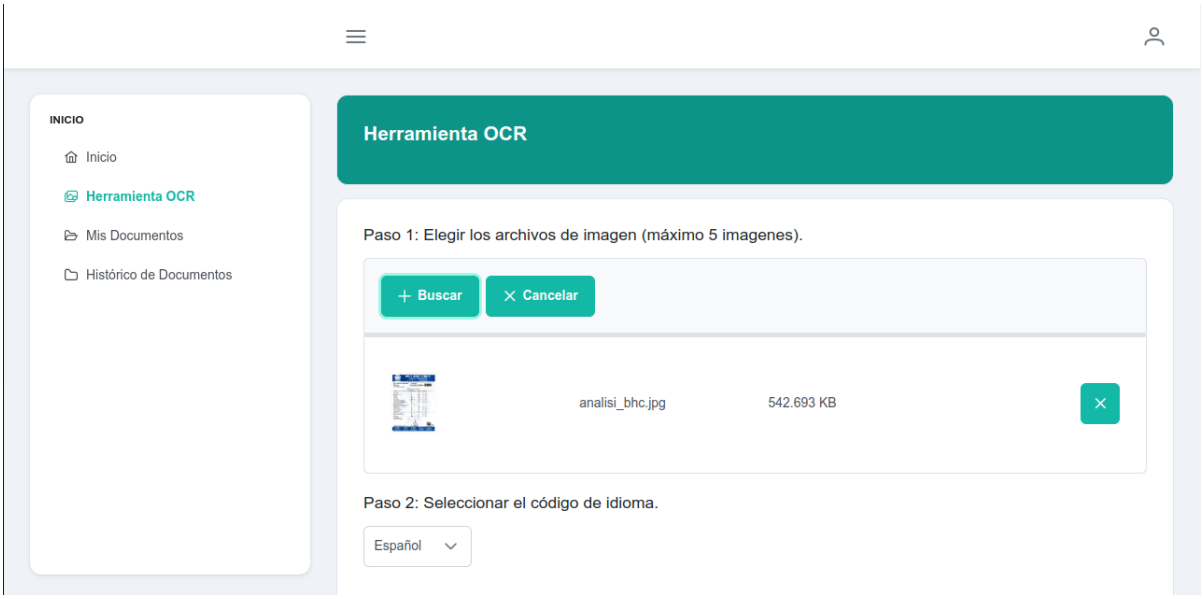


Figura 3.32 Página herramienta OCR.

Posteriormente, se le otorga al usuario la opción editar y, si es necesario, eliminar los caracteres extraídos. Una vez completada esta acción, se tiene la opción de guardar el documento digitalizado o de guardar y generar el documento HL7, como se muestra en la figura 3.33. Se difuminaron datos del documento clínico utilizado en la prueba, por razones de privacidad y confidencialidad del paciente y la institución de salud emisora del análisis.

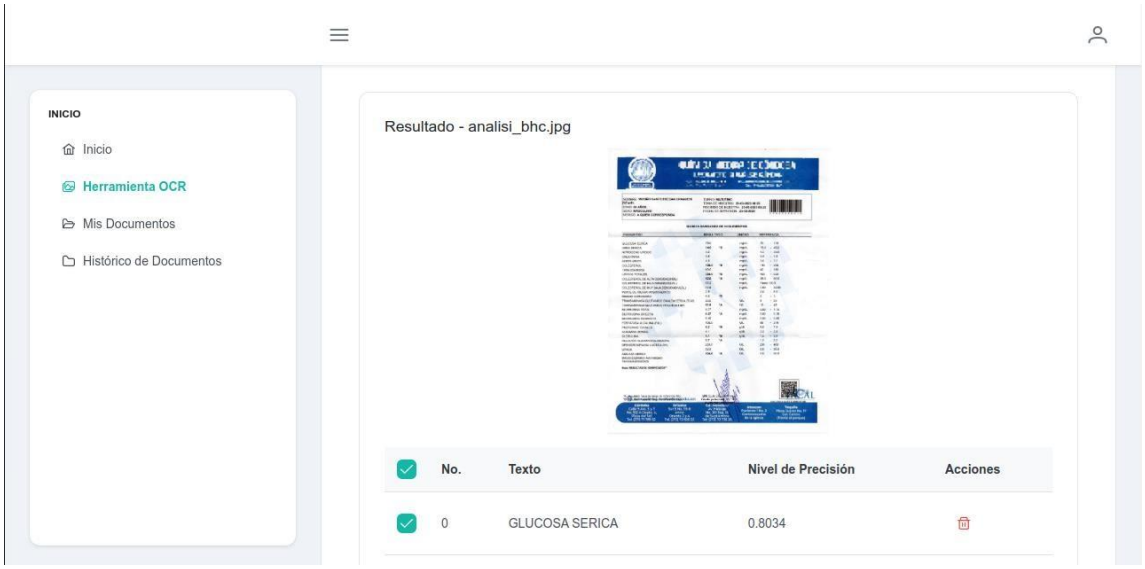


Figura 3.33 Página herramienta OCR (resultados).

Además, en la página de la herramienta OCR, se tiene la capacidad de generar un documento personalizado a partir de los documentos clínicos procesados. Para realizar esto, el usuario marcar las casillas junto a los textos extraídos que desee agregar al documento personalizado, como se ilustra en la figura 3.34.

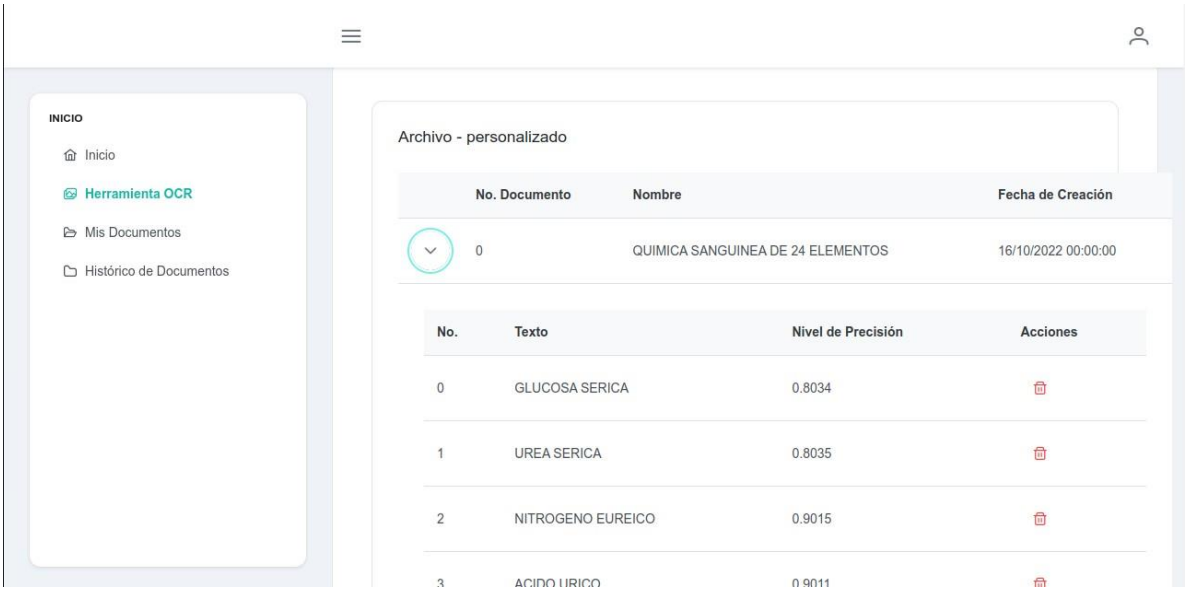


Figura 3.34 Página herramienta OCR (documentos personalizados).

Página mis documentos y pendientes

En esta página, se incorpora la opción de generar documentos clínicos en formato HL7 a partir de los caracteres extraídos, como se muestra en la figura 3.35. En esta sección, se listan las versiones del documento HL7 y se proporcionan opciones para guardar, validar, descargar o eliminar el archivo. Además, en la misma página, se listan los documentos digitalizados aún no transformados en formato HL7.

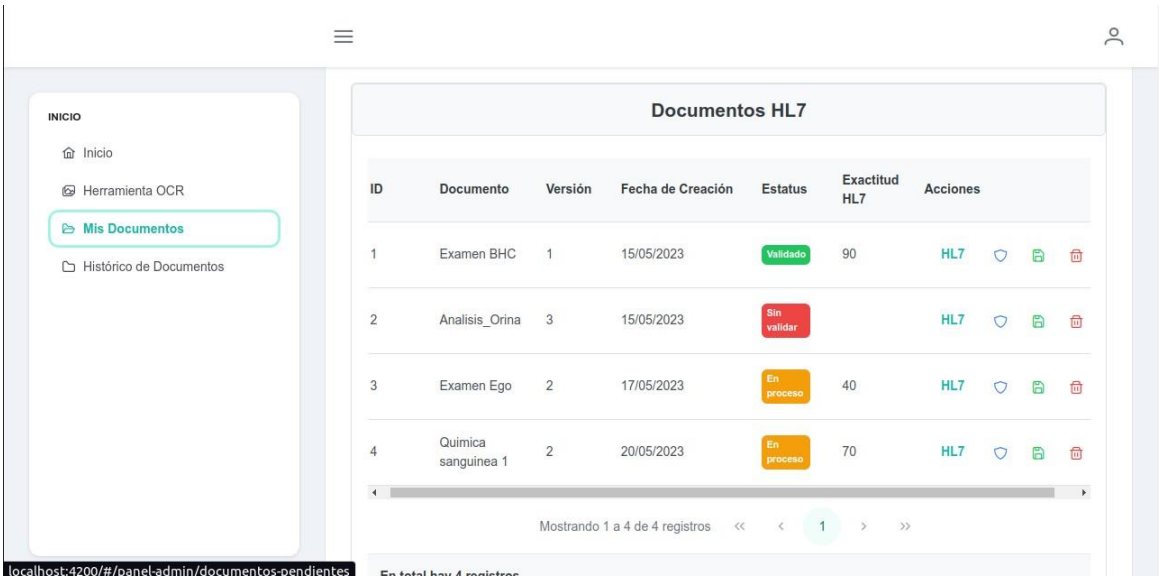


Figura 3.35 Página mis documentos y pendientes.

El estatus de los documentos se rige por una clasificación de alertas tipo semáforo. Esto es a causa de que la validación de documentos HL7 se realiza a través de servicios Web externos proporcionados por las herramientas HAPI FHIR® o Gazelle hl7®., las cuales no siempre se encuentra disponibles.

Página de histórico de documentos clínicos

La página muestra un listado de los documentos HL7 generados hasta el momento. Estos documentos son trasladados al historial cuando su fecha de creación supera los treinta días, conforme a una solicitud del laboratorio clínico que proporcionó los análisis clínicos, como se observa en la Figura 3.36.

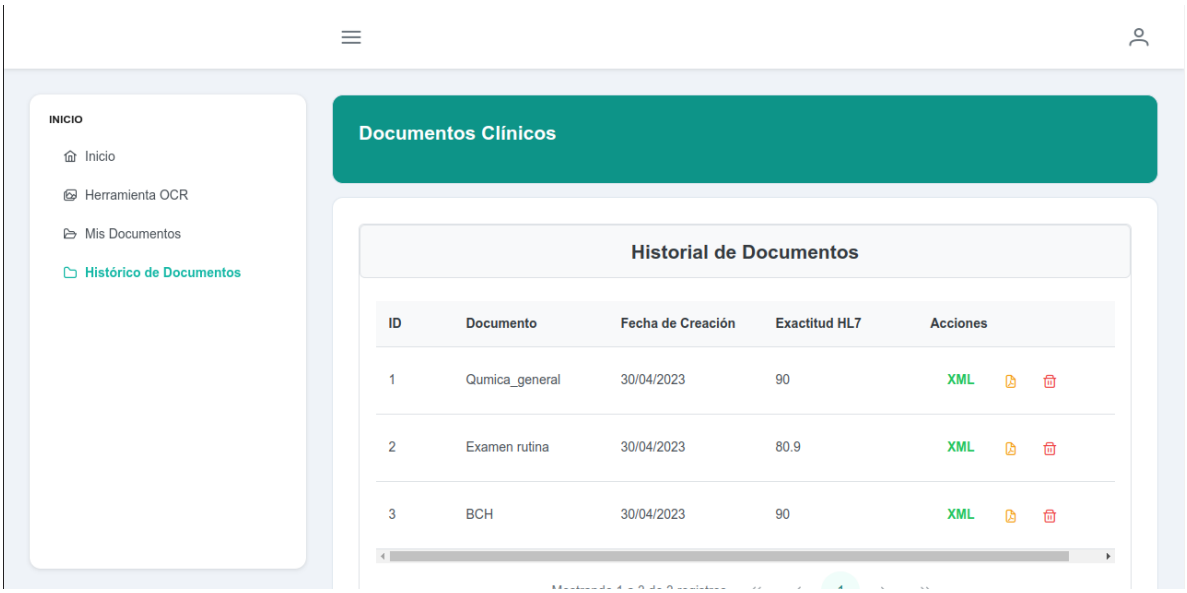


Figura 3.36 Página histórico de documentos.

Además, por cada documento registrado se cuenta la opción de descargar su versión en formato XML o PDF, y se proporciona la opción de eliminar el registro.

3.1.4 Fase de pruebas

La fase de pruebas desempeña un papel fundamental en la identificación y corrección de posibles errores en el software. En la mayoría de los proyectos, la ausencia de errores es crucial, ya que las pruebas son necesarias para asegurar que el producto final sea de alta calidad y cumpla con su propósito.

En la siguiente sección, se detallan las pruebas que se llevaron a cabo en el módulo de reconocimiento óptico de caracteres para la interpretación de documentos clínicos en el formato HL7. Este módulo fue bautizado como “LEXICA-HL7”, un nombre que refleja su función de “Lectura y Extracción de Información Clínica Avanzada en el estándar HL7”.

A continuación, se describen los criterios que se utilizaron al realizar las pruebas.

- Caso de uso
- Páginas involucradas
- Datos de entrada
- Acción a probar

- Resultados esperados
- Resultados obtenidos

Estos criterios se evaluaron durante la realización de las pruebas unitarias en las distintas unidades de código empleadas en el desarrollo del módulo LEXICA-HL7.

Capítulo 4 Resultados

El propósito de este capítulo es proporcionar los resultados del proyecto de tesis y el caso de estudio relacionado con la problemática investigada. El módulo “LEXICA-HL7” se enfoca en interpretar análisis clínicos según el estándar internacional HL7 en su versión 3 y HL7 FHIR, con el objetivo de cumplir con los criterios de interoperabilidad establecidos por la Norma Oficial Mexicana NOM-024-SSA3-2010.

4.1 Caso de estudios

A continuación, se proporcionan detalles sobre el caso de estudio aplicado en el presente proyecto de tesis.

En este proyecto, se obtuvieron los datos requeridos y se realizaron varias pruebas en el marco de la investigación en el Laboratorio de Análisis Clínicos “PASTEUR” ubicado en Cuautlapan, Veracruz. El laboratorio clínico ofrece los siguientes servicios:

- Examen general de orina (EGO)
- Biometría Hemática Completa (BHC)
- Perfil de lípidos
- Perfil tiroideo
- Prueba de embarazo
- Antígeno prostático
- Cortisol en sangre (solo a mujeres)
- Química sanguínea
- Triglicéridos
- Examen Colesterol total
- Cultivos (Urocultivo y Procultivo)
- Prueba de Virus de la Inmunodeficiencia Humana (VIH)

Se recopilaron documentos de análisis clínicos de los clientes del laboratorio a partir del mes de abril de 2023. Durante un período de un mes y medio, se obtuvieron 70 documentos clínicos que incluían varios tipos de análisis clínicos. El responsable del

área de pruebas clínicas proporcionó los documentos necesarios.

En la tabla 4.1 exhibe algunas muestras de análisis proporcionadas por el laboratorio. Se ha excluido información sensible y privada, sustituyéndola por nombres ficticios de forma artificial.

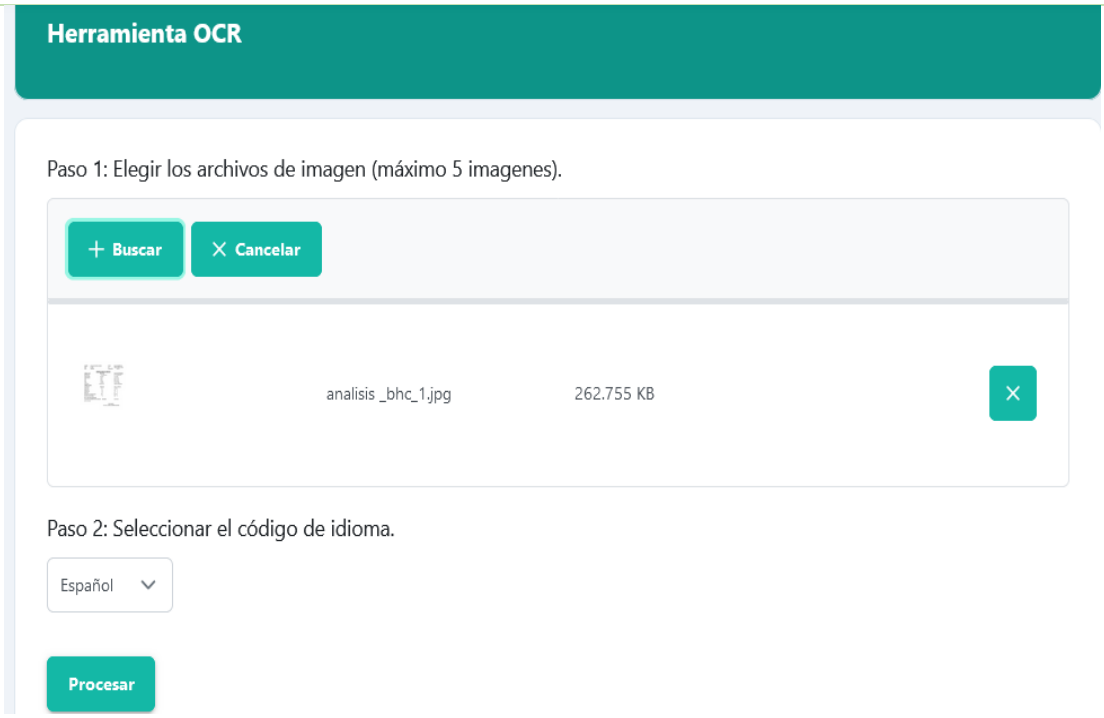
Tabla 4.1 Resumen de análisis clínicos proporcionados por el laboratorio.

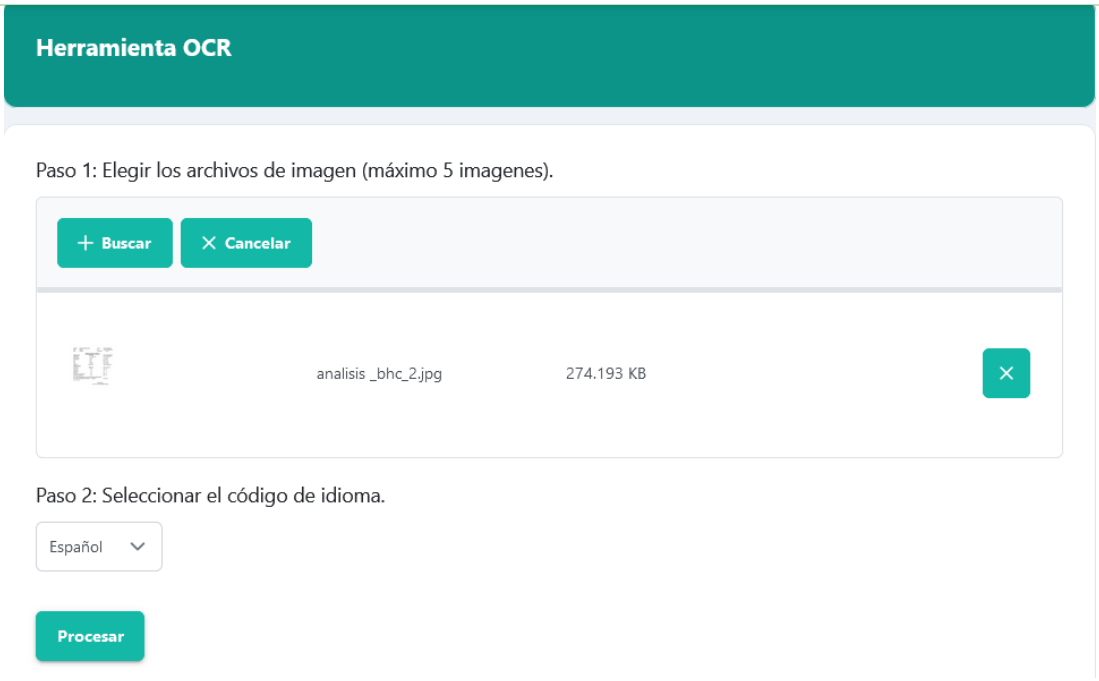
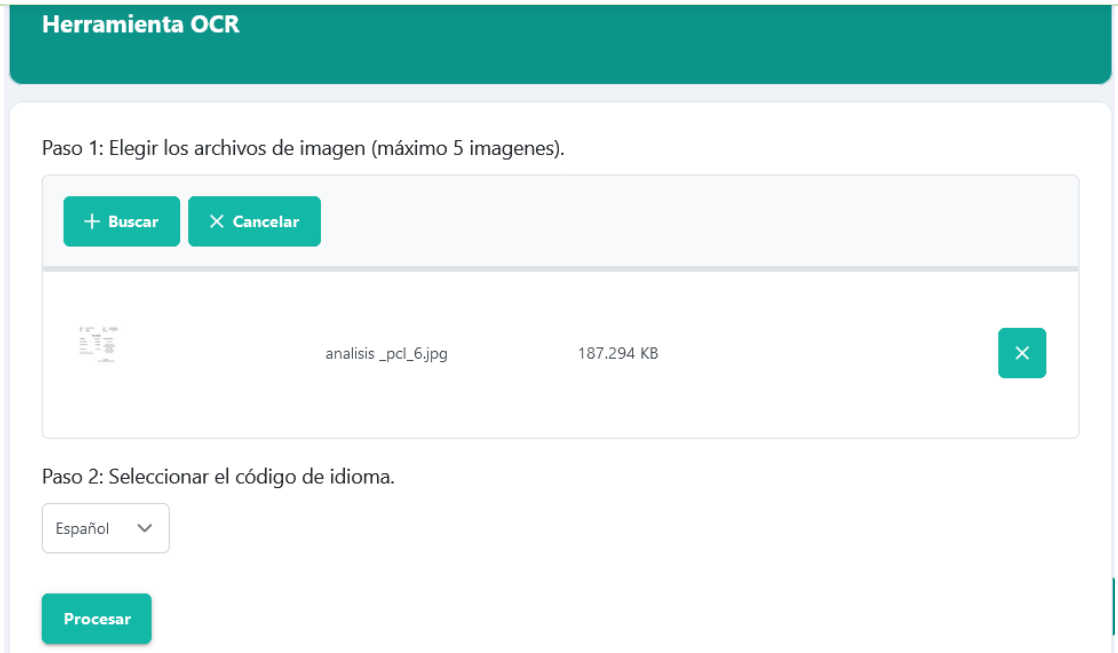
Nombre del documento	Paciente	Ejemplo de contenido del documento	Formato de documento
Biometría hemática completa			
analisis _bhc_1	Paciente 1	ERITROCITOS 4.71 millones/mmc	JPG
analisis _bhc_2	Paciente 2	HEMOGLOBINA 14.60 g/dL	JPG
analisis _bhc_3	Paciente 3	HEMATOCRITO 44.0 %	JPG
analisis _bhc_4	Paciente 4	VCM 91.7 fL	JPG
analisis _bhc_5	Paciente 5	HCM 30.4 pg	JPG
Perfil coronario I			
analisis _pcI_6	Paciente 6	GLUCOSA SÉRICA 103.00 mg/dL	JPG
analisis _pcI_7	Paciente 7	COLESTEROL 187.0 mg/dL	JPG
analisis _pcI_8	Paciente 8	TRIGLICÉRIDOS 97.0 mg/dL	JPG
analisis _pcI_9	Paciente 9	Aconsejable < 200 mg/dL	JPG
analisis _pcI_10	Paciente 10	MÉTODO: ESPECTROFOTOMETRÍA.	JPG
Determinación de glucosa sérica			
analisis _dgs_11	Paciente 11	GLUCOSA SÉRICA 99.00 mg/dL	PNG
analisis _ dgs _12	Paciente 12	GLUCOSA SÉRICA 100.00 mg/dL	PNG
analisis _ dgs _13	Paciente 13	GLUCOSA SÉRICA 144.00	PNG

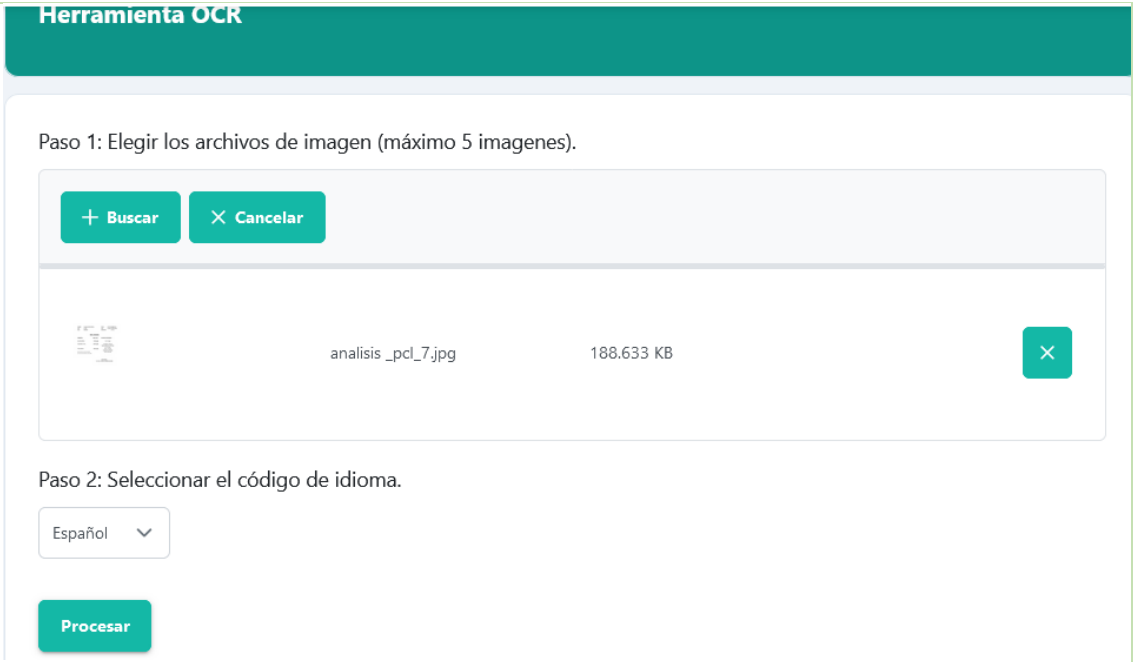
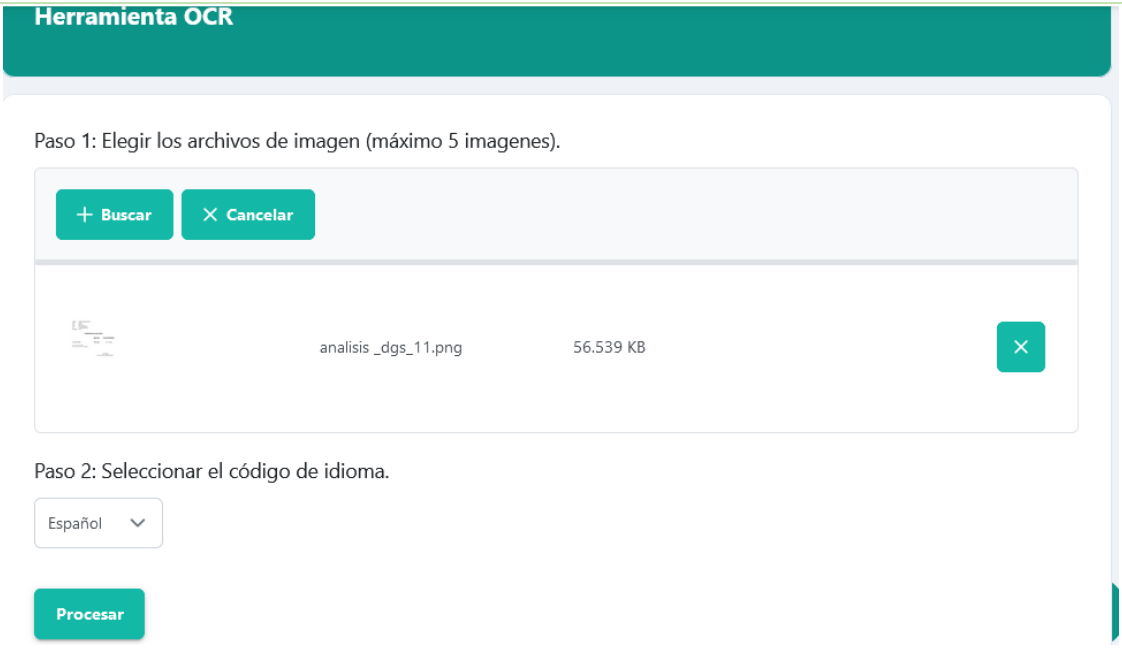
Nombre del documento	Paciente	Ejemplo de contenido del documento	Formato de documento
		mg/dL	
analisis _ dgs _14	Paciente 14	GLUCOSA SÉRICA 79.00 mg/dL	PNG
analisis _ dgs _15	Paciente 15	GLUCOSA SÉRICA 94.00 mg/dL	PNG
Urocultivo			
analisis _u _16	Paciente 16	Células epiteliales Escasas (+)	PDF
analisis _u_17	Paciente 17	Bacterias Abundantes (+++)	PDF
analisis _u_18	Paciente 18	Filamento de mucina No se observa	PDF
analisis _u_19	Paciente 19	NORFLOXACINA Resistente	PDF
analisis _u_20	Paciente 20	CIPROFLOXACINO Resistente	PDF

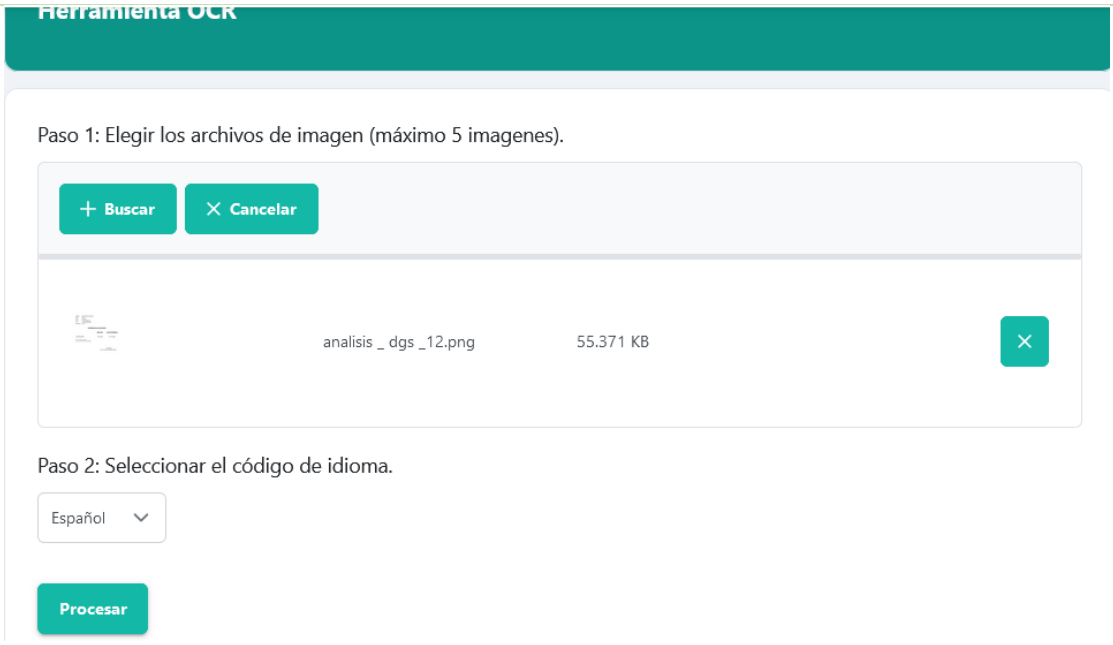
De muestra, se procesaron dos documentos de análisis clínicos a través del módulo LEXICA-HL7 para cada categoría mencionada anteriormente. La tabla 4.1 muestra la carga de cada uno de los documentos utilizados como muestra de este caso de estudio.

Tabla 4.2 Completado de formulario de carga de análisis clínicos para procesamiento OCR.

Paciente 1	**analisis _bhc_1**	**Biometría hemática completa**	**JPG**
	Carga de análisis		
			
Paciente 2	**analisis _bhc_2**	**Biometría hemática completa**	**JPG**
	Carga de análisis		

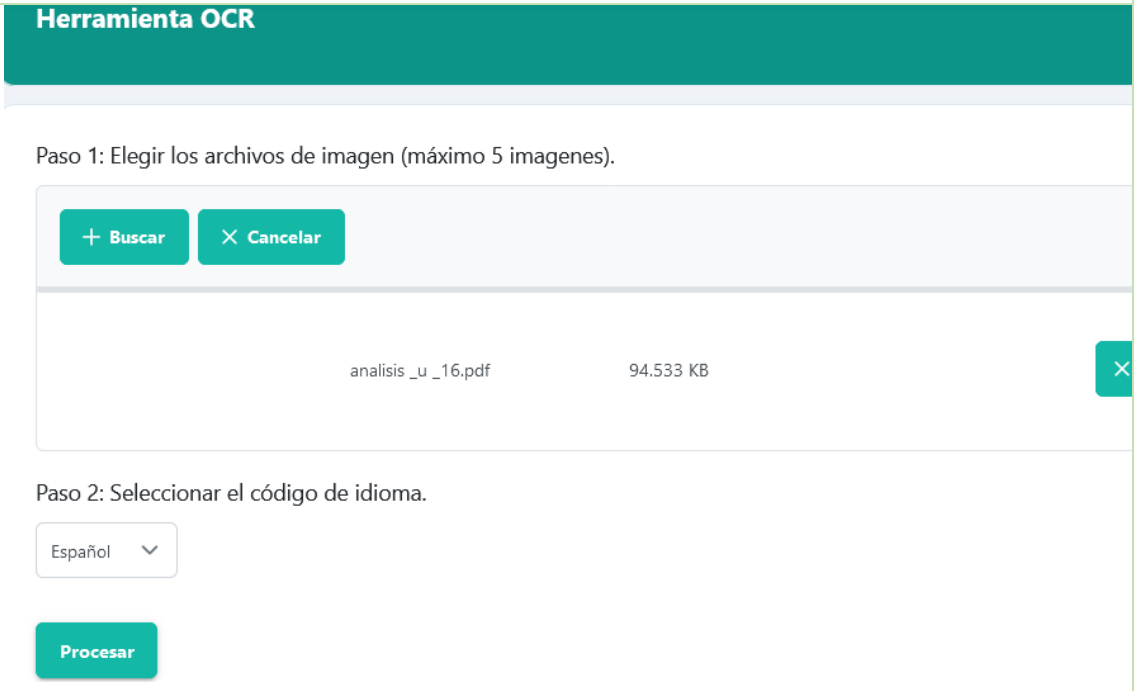
			
Paciente 6	**analisis _pcl_6**	**Perfil coronario I**	**JPG**
	Carga de análisis		
			
Paciente 7	**analisis _pcl_7**	**Perfil coronario I**	**JPG**
	Carga de análisis		

			
Paciente 11	**analisis _dgs_11**	**Determinación de glucosa sérica**	**PNG**
	Carga de análisis		
			
Paciente 12	**analisis _dgs_12**	**Determinación de glucosa sérica**	**PNG**
	Carga de análisis		

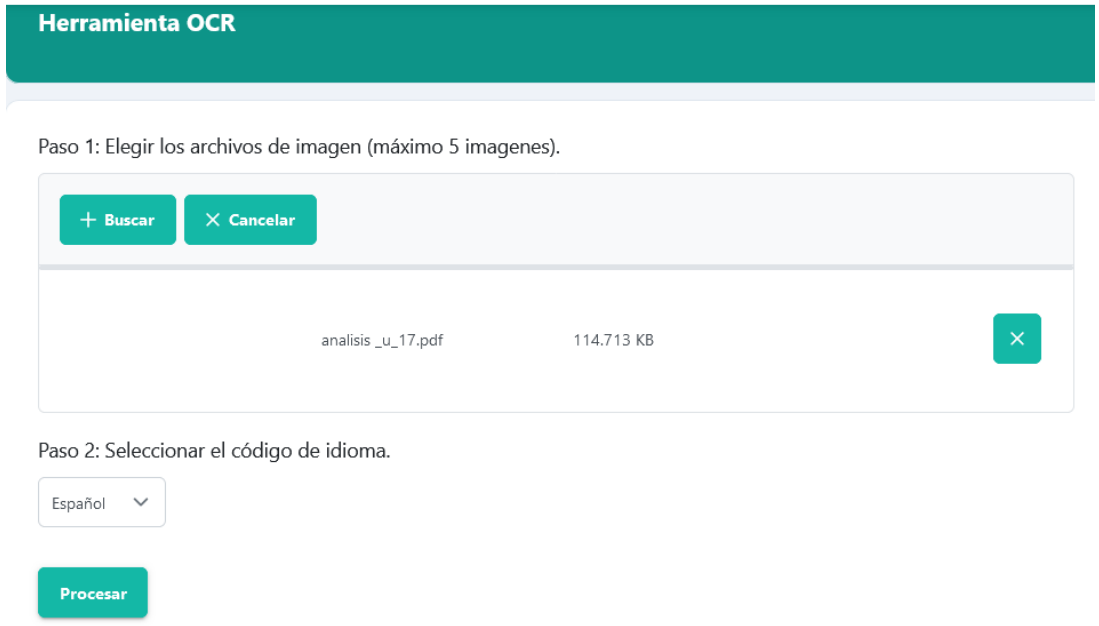


	analisis _u _16	**Urocultivo**	**PDF**
	Carga de análisis		

Paciente 16



Paciente 17	**analisis _u _16**	**Urocultivo**	**PDF**
	Carga de análisis		



Después de completar el formulario para procesar análisis, el usuario hace clic en el botón “Procesar”. En la siguiente tabla 4.3, se presentan los resultados obtenidos de la aplicación Web.

Tabla 4.3 Resultados de procesamiento de OCR por cada documento de muestra para el caso de estudio.

Paciente 1	**analisis _bhc_1**	**Biometría hemática completa**	**JPG**
	Resultado de procesamiento OCR		

Resultado - analisis _bhc_1.jpg

	No.	Texto	Nivel de Precisión	Acciones
	0	NOMBRE:	0.8034	
	1		0.8035	
	2	FECHA:	0.9015	

Paciente 2	**analisis _bhc_2**	**Biometría hemática completa**	**JPG**

Resultado de procesamiento OCR

Resultado - analisis _bhc_2.jpg

	No.	Texto	Nivel de Precisión	Acciones
	0	NOMBRE:	0.9034	
	1		0.9012	
	2	FECHA:	0.9015	

Paciente 6	**analisis _pcl_6**	**Perfil coronario I**	**JPG**
	Resultado de procesamiento OCR		
	Ana Laura Ramírez Ruiz Resultado - analisis _pcl_6.jpg 		

	No.	Texto	Nivel de Precisión	Acciones
☐	0	NOMBRE:	0.9004	🗑
☐	1		0.8011	🗑
☐	2	FECHA:	0.9015	🗑

Paciente 7	**analisis _pcl_7**	**Perfil coronario I**	**JPG**
	Resultado de procesamiento OCR		

Resultado - analisis _pcl_7.jpg

	No.	Texto	Nivel de Precisión	Acciones
☐	0	NOMBRE:	0.8994	🗑
☐	1		0.7141	🗑
☐	2	FECHA:	0.9015	🗑

	analisis _dgs_11	**Determinación de glucosa sérica**	**PNG**

Resultado de procesamiento OCR

Paciente 11

Resultado - analisis _dgs_11.png

	No.	Texto	Nivel de Precisión	Acciones
☐	0	NOMBRE:	0.9001	🗑
☐	1		0.8341	🗑
☐	2	EDAD:	0.8805	🗑

<table>
<tr><td rowspan="3">Paciente 12</td><td>analisis _dgs_12</td><td>Determinación de glucosa sérica</td><td>PNG</td></tr>
<tr><td colspan="3">Resultado de procesamiento OCR</td></tr>
<tr><td colspan="3">Resultado - analisis _ dgs _12.png
 
<table><tr><th></th><th>No.</th><th>Texto</th><th>Nivel de Precisión</th><th>Acciones</th></tr><tr><td></td><td>0</td><td>NOMBRE:</td><td>0.9561</td><td></td></tr><tr><td></td><td>1</td><td></td><td>0.808</td><td></td></tr><tr><td></td><td>2</td><td>EDAD:</td><td>0.7712</td><td></td></tr></table></td></tr>
<tr><td rowspan="3">Paciente 16</td><td>analisis _u _16</td><td>Urocultivo</td><td>JPG</td></tr>
<tr><td colspan="3">Resultado de procesamiento OCR</td></tr>
<tr><td colspan="3"></td></tr>
</table>

Resultado - analisis _u _16.jpg

	No.	Texto	Nivel de Precisión	Acciones
☐	0	NOMBRE:	0.9108	🗑
☐	1		0.8471	🗑

Paciente 17

analisis _u _17	Urocultivo	JPG

Resultado de procesamiento OCR

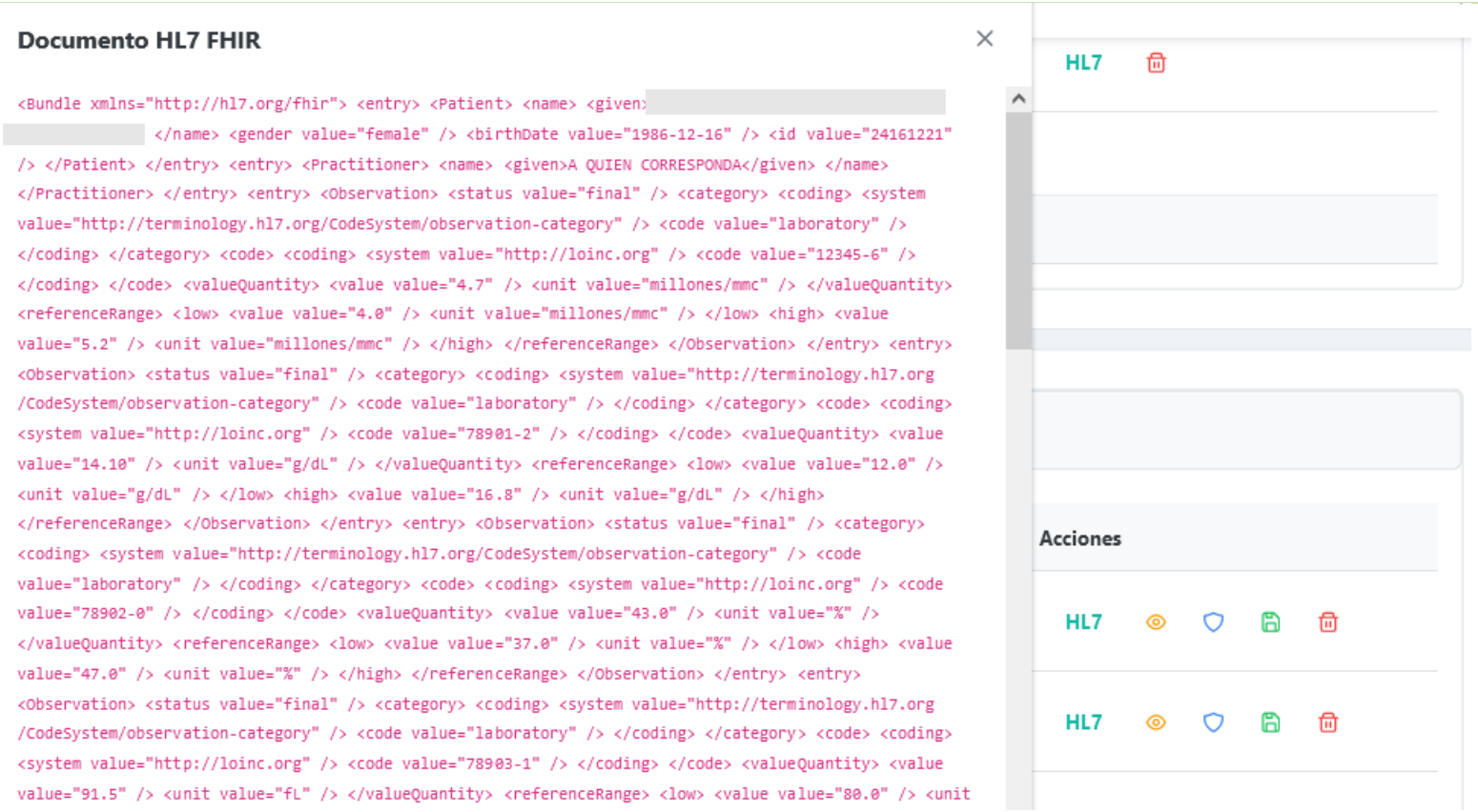
Resultado - analisis _u_17.jpg

	No.	Texto	Nivel de Precisión	Acciones
☐	0	NOMBRE:	0.9037	🗑
☐	1		0.7018	🗑
☐	2	FECHA:	0.9172	🗑

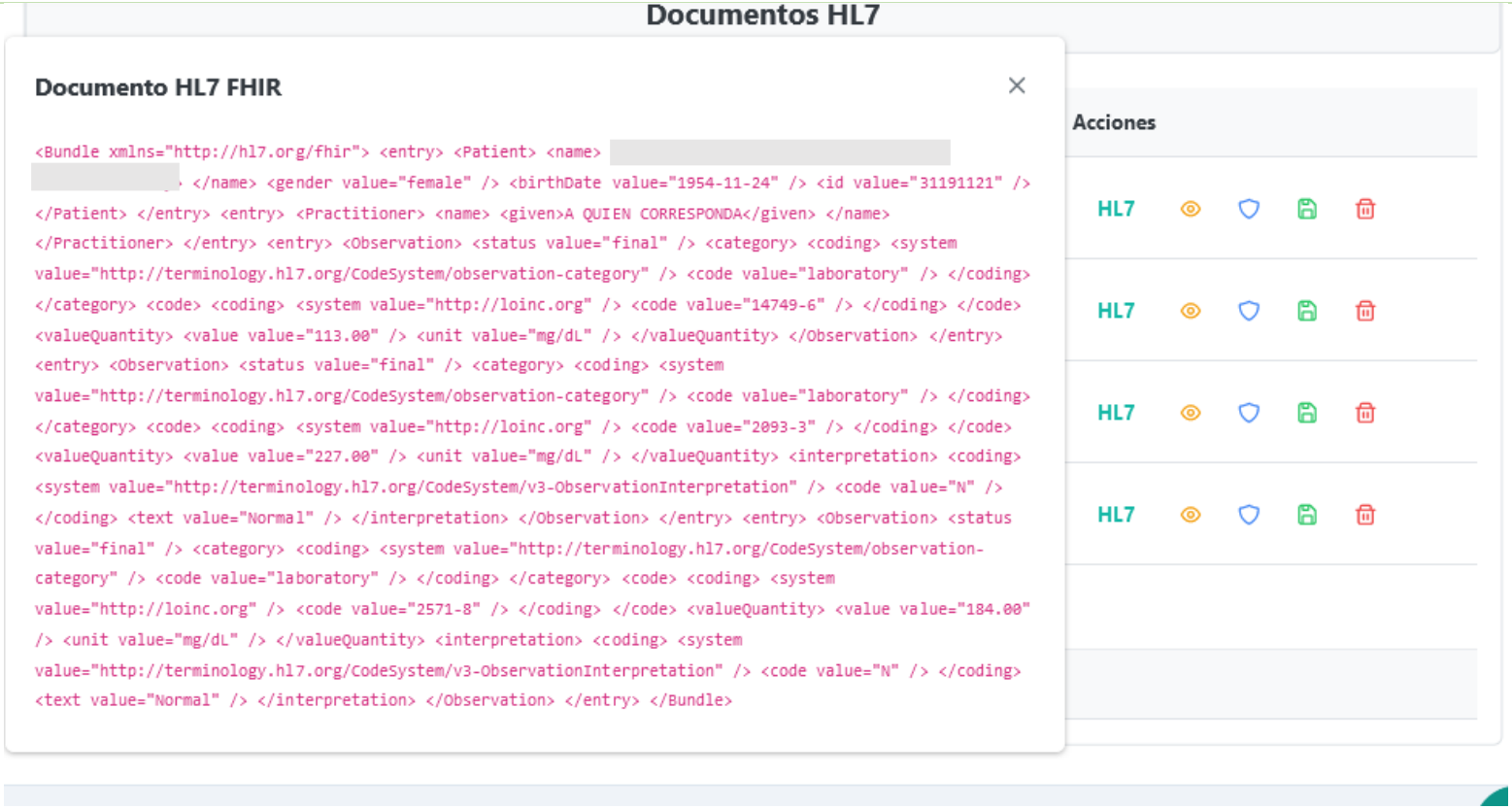
La tabla 4.4 presenta los documentos generados en formato HL7 v3 y HL7 FHIR para para los análisis clínicos utilizados en esta muestra del caso de estudio.

Tabla 4.4 Resultado de transformación de análisis clínicos a formato HL7.

<table>
<tr><td rowspan="3">Paciente 1</td><td>analisis _bhc_1</td><td>Biometría hemática completa</td><td>JPG</td></tr>
<tr><td colspan="3">Resultado de transformación a formato HL7 v3</td></tr>
<tr><td colspan="3">Documento HL7 v3
<HL7Message> <PID>1||[redacted]|09 DE DICIEMBRE 2021|61 AÑOS|FEMENINO||||12091221</PID>
<OBR>1|BIOMETRIA HEMATICA COMPLETA|||||||FREDDY BERNAL PALMA||||09 DE DICIEMBRE 2021</OBR>
<OBX>1|ST|ERITROCITOS||5.0 millones/mmc|4.0 - 5.2 millones/mmc</OBX> <OBX>2|ST|HEMOGLOBINA||15.45 g/dL|12 a 16 g/dL</OBX> <OBX>3|ST|HEMATOCRITO||46.0 %|37.0 a 47.0 %</OBX> <OBX>4|ST|VCM||92.0 fL|80.0 a 110.0 fL</OBX> <OBX>5|ST|HCM||30.9 pg|26.0 a 34.0 pg</OBX> <OBX>6|ST|CHCM||33.6 g/dL|31.0 a 37.0 g/dL</OBX>
<OBX>7|ST|PLAQUETAS||181,000 /mmc|150,000 a 450,000/mmc</OBX> <OBX>8|ST|LEUCOCITOS||4,700 /mmc|5,000 a 10,000/mmc</OBX> <OBX>9|ST|LINFOCITOS||27 %|17 a 45 %</OBX> <OBX>10|ST|MONOCITOS||5 %|2 a 8 %</OBX>
<OBX>11|ST|EOSINOFILOS||2 %|1 a 4 %</OBX> <OBX>12|ST|BASOFILOS||0 %|0 a 1 %</OBX> <OBX>13|ST|NEUTROFILOS SEGMENTADOS||65 %|55 a 75 %</OBX> <OBX>14|ST|NEUTROFILOS EN BANDA||1 %|0 a 3 %</OBX> <OBX>15|ST|VELOCIDAD DE SEDIMENTACION GLOBULAR||26.0 mm/hr|0 a 20 mm/hr</OBX> <OBX>16|ST|MÉTODO||WINTROBE</OBX>
<OBX>17|ST|NOTA||RESULTADOS VERIFICADOS.</OBX> </HL7Message>
Acciones
HL7
En total hay 1 registros.</td></tr>
<tr><td rowspan="3">Paciente 2</td><td>analisis _bhc_2</td><td>Biometría hemática completa</td><td>JPG</td></tr>
<tr><td colspan="3">Resultado de transformación a formato HL7 FHIR</td></tr>
</table>

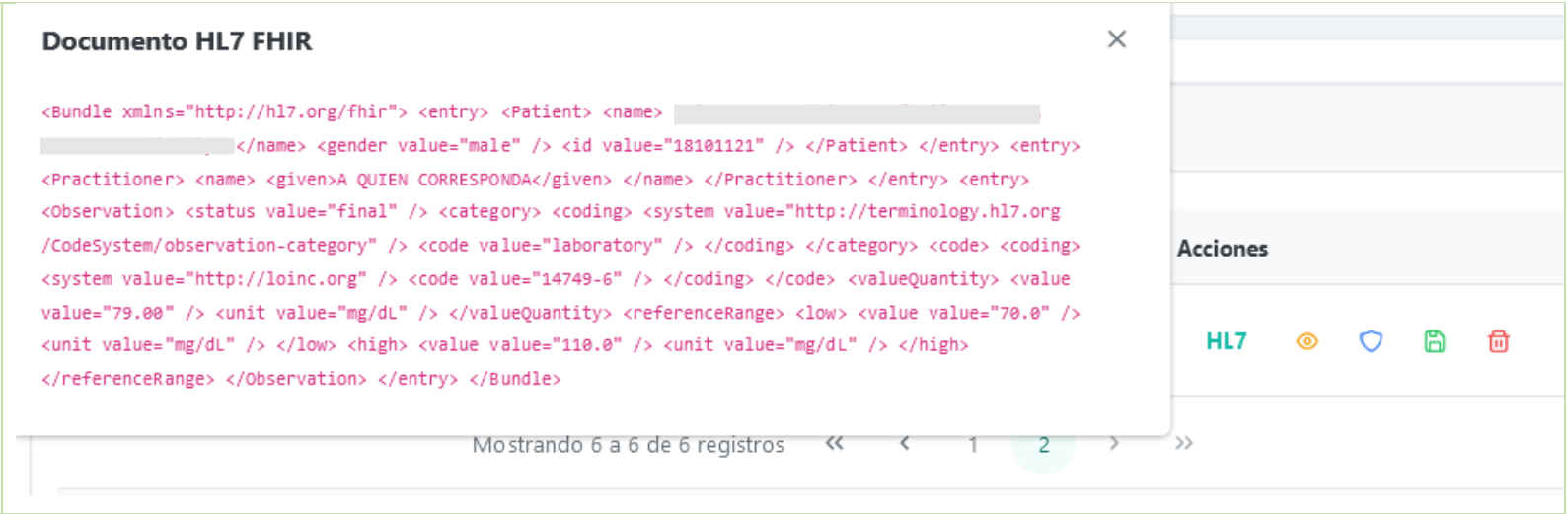
<table>
<tr><td rowspan="3">Paciente 6</td><td colspan="3">Documento HL7 FHIR
 
<Bundle xmlns="http://hl7.org/fhir"> <entry> <Patient> <name> <given> </name> <gender value="female" /> <birthDate value="1986-12-16" /> <id value="24161221" /> </Patient> </entry> <entry> <Practitioner> <name> <given>A QUIEN CORRESPONDA</given> </name> </Practitioner> </entry> <entry> <Observation> <status value="final" /> <category> <coding> <system value="http://terminology.hl7.org/CodeSystem/observation-category" /> <code value="laboratory" /> </coding> </category> <code> <coding> <system value="http://loinc.org" /> <code value="12345-6" /> </coding> </code> <valueQuantity> <value value="4.7" /> <unit value="millones/mmc" /> </valueQuantity> <referenceRange> <low> <value value="4.0" /> <unit value="millones/mmc" /> </low> <high> <value value="5.2" /> <unit value="millones/mmc" /> </high> </referenceRange> </Observation> </entry> <entry> <Observation> <status value="final" /> <category> <coding> <system value="http://terminology.hl7.org/CodeSystem/observation-category" /> <code value="laboratory" /> </coding> </category> <code> <coding> <system value="http://loinc.org" /> <code value="78901-2" /> </coding> </code> <valueQuantity> <value value="14.10" /> <unit value="g/dL" /> </valueQuantity> <referenceRange> <low> <value value="12.0" /> <unit value="g/dL" /> </low> <high> <value value="16.8" /> <unit value="g/dL" /> </high> </referenceRange> </Observation> </entry> <entry> <Observation> <status value="final" /> <category> <coding> <system value="http://terminology.hl7.org/CodeSystem/observation-category" /> <code value="laboratory" /> </coding> </category> <code> <coding> <system value="http://loinc.org" /> <code value="78902-0" /> </coding> </code> <valueQuantity> <value value="43.0" /> <unit value="%" /> </valueQuantity> <referenceRange> <low> <value value="37.0" /> <unit value="%" /> </low> <high> <value value="47.0" /> <unit value="%" /> </high> </referenceRange> </Observation> </entry> <entry> <Observation> <status value="final" /> <category> <coding> <system value="http://terminology.hl7.org/CodeSystem/observation-category" /> <code value="laboratory" /> </coding> </category> <code> <coding> <system value="http://loinc.org" /> <code value="78903-1" /> </coding> </code> <valueQuantity> <value value="91.5" /> <unit value="fL" /> </valueQuantity> <referenceRange> <low> <value value="80.0" /> <unit</td></tr>
<tr><td>analisis _pcl_6</td><td>Perfil coronario I</td><td>JPG</td></tr>
<tr><td colspan="3">Resultado de transformación a formato HL7 v3</td></tr>
</table>

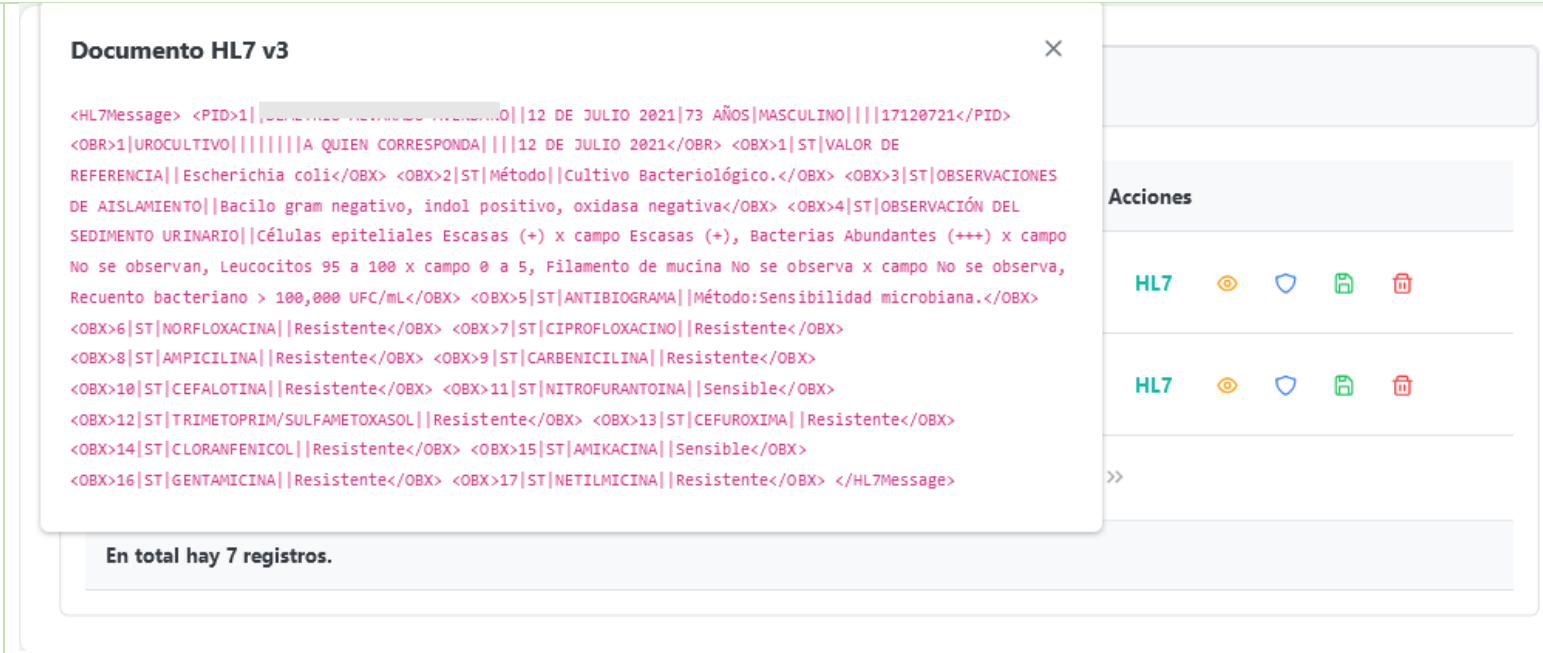
<table>
<tr><td rowspan="4">Paciente 7</td><td colspan="3"> 
</td></tr>
<tr><td>analisis _pcl_7</td><td>Perfil coronario I</td><td>JPG</td></tr>
<tr><td colspan="3">Resultado de transformación a formato HL7 FHIR</td></tr>
</table>

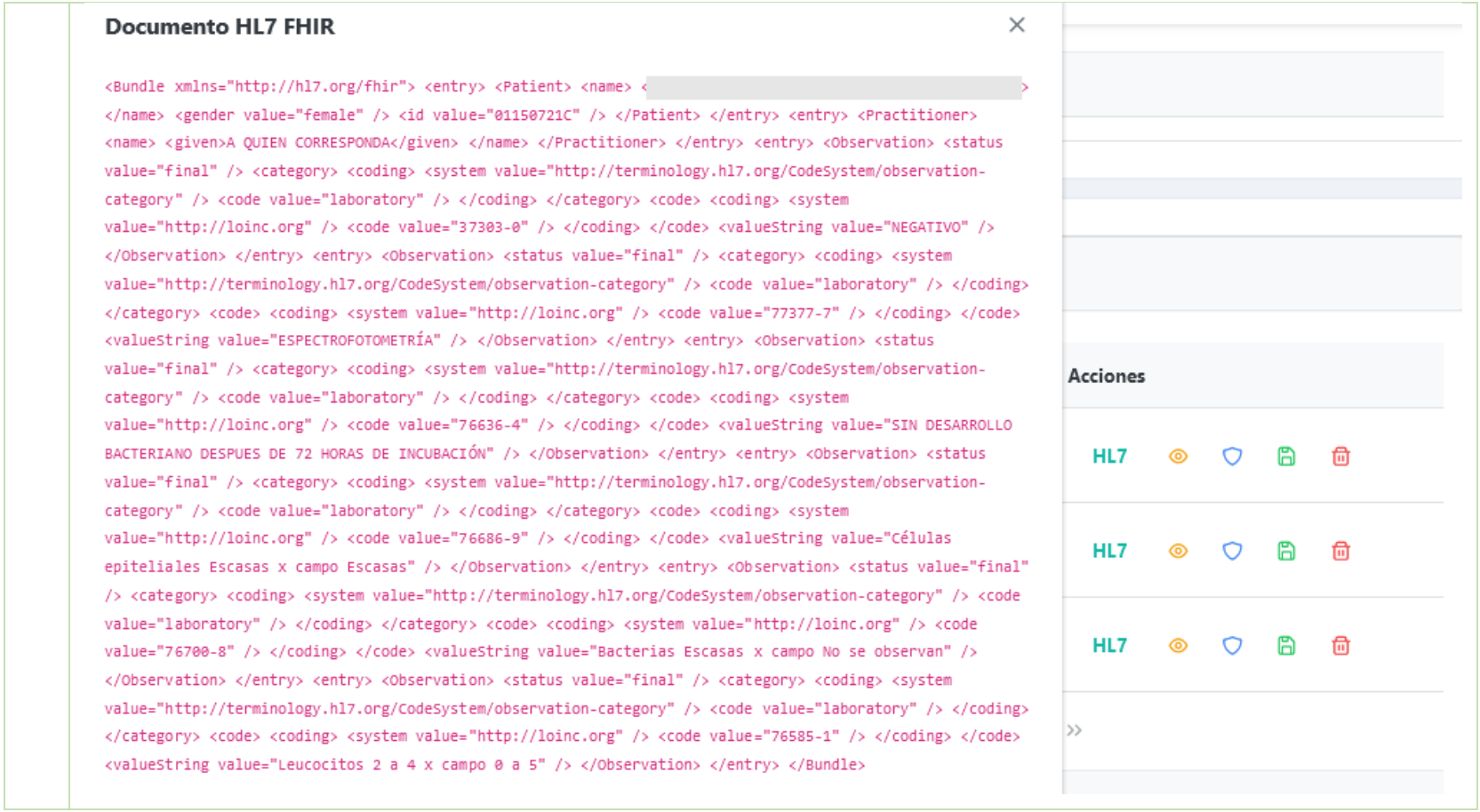


Paciente 11	**analisis _dgs_11**	**Determinación de glucosa sérica**	**PNG**
	Resultado de transformación a formato HL7 v3		

Paciente 12			
	analisis _dgs_12	**Determinación de glucosa sérica**	**PNG**
	Resultado de transformación a formato HL7 FHIR		

<table>
<tr><td rowspan="3">Paciente 16</td><td colspan="3"> 
</td></tr>
<tr><td>analisis _u _16</td><td>Urocultivo</td><td>JPG</td></tr>
<tr><td colspan="3">Resultado de transformación a formato HL7 v3</td></tr>
</table>

<table>
<tr><td rowspan="4">Paciente 17</td><td colspan="3"> 
</td></tr>
<tr><td>analisis _u _17</td><td>Urocultivo</td><td>JPG</td></tr>
<tr><td colspan="3">Resultado de transformación a formato HL7 FHIR</td></tr>
</table>


Documento HL7 FHIR
<Bundle xmlns="http://hl7.org/fhir"> <entry> <Patient> <name> <
</name> <gender value="female" /> <id value="01150721C" /> </Patient> </entry> <entry> <Practitioner>
<name> <given>A QUIEN CORRESPONDA</given> </name> </Practitioner> </entry> <entry> <Observation> <status
value="final" /> <category> <coding> <system value="http://terminology.hl7.org/CodeSystem/observation-
category" /> <code value="laboratory" /> </coding> </category> <code> <coding> <system
value="http://loinc.org" /> <code value="37303-0" /> </coding> </code> <valueString value="NEGATIVO" />
</Observation> </entry> <entry> <Observation> <status value="final" /> <category> <coding> <system
value="http://terminology.hl7.org/CodeSystem/observation-category" /> <code value="laboratory" /> </coding>
</category> <code> <coding> <system value="http://loinc.org" /> <code value="77377-7" /> </coding> </code>
<valueString value="ESPECTROFOTOMETRÍA" /> </Observation> </entry> <entry> <Observation> <status
value="final" /> <category> <coding> <system value="http://terminology.hl7.org/CodeSystem/observation-
category" /> <code value="laboratory" /> </coding> </category> <code> <coding> <system
value="http://loinc.org" /> <code value="76636-4" /> </coding> </code> <valueString value="SIN DESARROLLO
BACTERIANO DESPUES DE 72 HORAS DE INCUBACIÓN" /> </Observation> </entry> <entry> <Observation> <status
value="final" /> <category> <coding> <system value="http://terminology.hl7.org/CodeSystem/observation-
category" /> <code value="laboratory" /> </coding> </category> <code> <coding> <system
value="http://loinc.org" /> <code value="76686-9" /> </coding> </code> <valueString value="Células
epiteliales Escasas x campo Escasas" /> </Observation> </entry> <entry> <Observation> <status value="final"
/> <category> <coding> <system value="http://terminology.hl7.org/CodeSystem/observation-category" /> <code
value="laboratory" /> </coding> </category> <code> <coding> <system value="http://loinc.org" /> <code
value="76700-8" /> </coding> </code> <valueString value="Bacterias Escasas x campo No se observan" />
</Observation> </entry> <entry> <Observation> <status value="final" /> <category> <coding> <system
value="http://terminology.hl7.org/CodeSystem/observation-category" /> <code value="laboratory" /> </coding>
</category> <code> <coding> <system value="http://loinc.org" /> <code value="76585-1" /> </coding> </code>
<valueString value="Leucocitos 2 a 4 x campo 0 a 5" /> </Observation> </entry> </Bundle>
Acciones
HL7
HL7
HL7

En la tabla 4.5, se presenta el promedio de precisión obtenido para cada documento clínico procesado con OCR, además de indicar si el formato HL7 al que se interpretó el documento clínico es correcto. Esta validación se llevó a cabo utilizando las herramientas HAPI FHIR para HL7 FHIR y Gazelle HL7 Validator para HL7 v3.

Tabla 4.5 Promedio precisión y validación de formato HL7.

Nombre del documento	**Promedio de precisión alcanzada por OCR.**	**Estatus de interpretación y validación de formato HL7**
analisis _bhc_1	88 %	Exitosa
analisis _bhc_2	72 %	Exitosa
analisis _pcI_6	79 %	Exitosa
analisis _pcI_7	91 %	Exitosa
analisis _dgs_11	69 %	Exitosa
analisis _dgs_12	79 %	Exitosa
analisis _u _16	90 %	Exitosa
analisis _u _16	89 %	Exitosa

Después de probar el módulo LEXICA-HL7 en los análisis clínicos proporcionados, se determinó que, si el promedio del nivel de precisión del OCR es inferior al 50%, la conversión a formato HL7 no es exitosa. Esto se tiene porque un porcentaje por debajo del 50% indica que la mayoría de los caracteres no se identificaron correctamente y, en caso de los identificados, no lo son del todo correcto. Esto añade un nivel adicional de desafío a la conversión, ya que cada estándar HL7 v3 y HL7 FHIR requiere datos específicos para completarse correctamente.

Capítulo 5 Conclusiones y recomendaciones

En esta sección se exponen las conclusiones y las sugerencias derivadas de la implementación de la solución propuesta en el proyecto denominado “LEXICA-HL7: Lectura y Extracción de Información Clínica Avanzada en HL7”.

5.1 Conclusiones

En conclusión, el desarrollo del módulo de reconocimiento óptico de caracteres (OCR) para la interpretación de documentos clínicos en formato HL7 permitió abordar la problemática de interoperabilidad en el sector de salud. La necesidad de intercambiar información de manera fluida y precisa entre los diversos actores involucrados en la atención médica es cada vez más importante.

El objetivo general de este proyecto fue lograr la interpretación de documentos clínicos mediante técnicas de procesamiento del lenguaje natural y su conversión al estándar HL7. Los resultados obtenidos tras probar el módulo LEXICA-HL7 en los análisis clínicos demostraron que la precisión del OCR juega un papel crucial en el éxito de la transformación a formato HL7. Se determinó que cuando el promedio de precisión es inferior al 50%, la interpretación resulta fallida debido a la incorrecta identificación de caracteres.

Es evidente que la calidad y precisión del reconocimiento óptico de caracteres es fundamental para garantizar la correcta interpretación y representación de la información clínica. Por lo tanto, es necesario seguir trabajando en la mejora de las técnicas de OCR y asegurar niveles de precisión superiores al umbral establecido. Esto permitirá lograr una mayor interoperabilidad en el intercambio de información clínica y contribuirá a una atención médica más eficiente y oportuna.

Por otro lado, como parte del trabajo de investigación de este proyecto de tesis, se redactó y presentó el artículo titulado “Reconocimiento Óptico de Caracteres en el Dominio Clínico: Análisis Comparativo y Arquitectura para su Transformación al Estándar HL7” en el IX Congreso Internacional de Investigación Tijuana, y se adjunta una constancia de presentación en el anexo.

5.2 Recomendaciones

Una recomendación de suma importancia es mejorar la precisión del reconocimiento óptico de caracteres (OCR) ya que es fundamental invertir en técnicas y algoritmos avanzados de OCR para aumentar la precisión en la identificación de caracteres en los documentos clínicos. Esto ayudará a reducir el porcentaje de fallas en la conversión a formato HL7 y mejorar la calidad de la interpretación de la información.

Como segunda recomendación, es importante considerar la estructura y los datos presentes en los análisis clínicos a procesar por el módulo LEXICA-HL7, ya que para la correcta interpretación en el estándar HL7, es necesario que el documento clínico contenga las secciones de datos mínimos requeridas por el estándar para representar un documento clínico. Si estos datos mínimos no están presentes en el documento procesado, esto afectará la generación del documento HL7 resultante del proceso.

Finalmente, para garantizar que el proyecto se mantenga operativo a largo plazo y reciba el soporte necesario, es importante que realizar revisiones regulares de la documentación oficial de los estándares HL7 (FHIR y v3). Estos estándares se actualizan con regularidad con el propósito de simplificar su uso. Por lo tanto, es fundamental mantenerse al tanto de las modificaciones o actualizaciones que reciban.

Anexo

IX Congreso Internacional de Investigación Tijuana

Otorga la presente:

CONSTANCIA

a: *ITI. Irving Jesús Ramírez Alacio, Dr. José Luis Sánchez Cervantes, Dr. Isaac Machorro Cano, MCE. Beatriz Alejandra Olivares Zepahua, Dr. Giner Alor Hernández*

Por su participación con el trabajo titulado: **"Reconocimiento Óptico de Caracteres de Dominio Clínico: Análisis comparativo y Arquitectura para su transformación al estándar HL7"** mediante la presentación oral en el 6th Conference on Computer Science and Engineering en el marco del Congreso Internacional de Investigación Tijuana que se llevó a cabo del 26 al 28 de abril del 2023, en la ciudad de Tijuana, B. C. México.

Mauricio Alonso Sánchez Herrera
Track Chair de CoCSCE

Dra. Alejandra Serrano Trujillo
Coordinador General del CI2T 2023

Bibliografía

[1] M. D. C. Sosa Sierra, “Inteligencia artificial en la gestión financiera empresarial,” Pensamiento & Gestión, no. 23, pp. 153–186, Jan. 2007, [Online]. Available: https://www.redalyc.org/articulo.oa?id=64602307

[2] “La ciencia y el hombre.” Accessed: Jan. 23, 2022. [Online]. Available: https://www.uv.mx/cienciahombre/revistae/vol17num3/articulos/inteligencia/index.htm

[3] J. Thanaki, Python natural language processing : explore NLP with machine learning and deep learning techniques.

[4] “What Is Optical Character Recognition (OCR)? | IBM.” Accessed: Jan. 30, 2022. [Online]. Available: https://www.ibm.com/cloud/blog/optical-character-recognition

[5] “What is Speech Recognition? | IBM.” Accessed: Jan. 30, 2022. [Online]. Available: https://www.ibm.com/cloud/learn/speech-recognition

[6] “Machine Translation - Microsoft Translator for Business.” Accessed: Jan. 30, 2022. [Online]. Available: https://www.microsoft.com/en-us/translator/business/machine-translation/

[7] “NLP vs. NLU vs. NLG: the differences between three natural language processing concepts - Watson Blog.” Accessed: Jan. 30, 2022. [Online]. Available: https://www.ibm.com/blogs/watson/2020/11/nlp-vs-nlu-vs-nlg-the-differences-between-three-natural-language-processing-concepts/

[8] L. Banica and A. Hagiu, “Using big data analytics to improve decision-making in apparel supply chains,” Information Systems for the Fashion and Apparel Industry, pp. 63–95, Apr. 2016, doi: 10.1016/B978-0-08-100571-2.00004-X.

[9] “Semantic search - Azure Cognitive Search | Microsoft Docs.” Accessed: Jan. 30, 2022. [Online]. Available: https://docs.microsoft.com/en-us/azure/search/semantic-search-overview

[10] “¿Qué es Machine Learning? - México | IBM.” Accessed: Jan. 31, 2022. [Online].

Available: https://www.ibm.com/mx-es/analytics/machine-learning

[11] Z. Nan, H. Guan, and X. Shen, “HISyn: Human Learning-Inspired Natural Language Programming,” in Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, New York, NY, USA: Association for Computing Machinery, 2020, pp. 75–86. [Online]. Available: https://doi.org/10.1145/3368089.3409673

[12] A. Chaudhuri, K. Mandaviya, P. Badelia, and S. K. Ghosh, “Studies in Fuzziness and Soft Computing Optical Character Recognition Systems for Different Languages with Soft Computing.” [Online]. Available: http://www.springer.com/series/2941

[13] A. Bolshoy, Z. (Vladimir) Volkovich, V. Kirzhner, and Z. Barzily, “Mathematical Models for the Analysis of Natural-Language Documents,” pp. 23–42, 2010, doi: 10.1007/978-3-642-12952-0_3.

[14] “SIRES Certificados en la NOM-024-SSA3-2012.” Accessed: Apr. 05, 2022. [Online]. Available: http://www.dgis.salud.gob.mx/contenidos/intercambio/sires_certificacion_gobmx.html

[15] J. L. Tapia Vázquez, “El expediente cl\’\inico electrónico,” Revista odontológica mexicana, vol. 14, no. 2, pp. 76–77, 2010.

[16] R. Peña and S. López-Silva, “ANÁLISIS ESTRUCTURAL Y FUNCIONAL DEL EXPEDIENTE CLÍNICO ELECTRONICO,” 2003, Accessed: Apr. 05, 2022. [Online]. Available: https://www.researchgate.net/publication/278678814

[17] “NORMA OFICIAL MEXICANA NOM-168-SSA1-1998, DEL EXPEDIENTE CLINICO.” Accessed: Nov. 17, 2022. [Online]. Available: http://www.salud.gob.mx/unidades/cdi/nom/168ssa18.html

[18] “NORMA Oficial Mexicana NOM-024-SSA3-2010, Que establece los objetivos funcionales y funcionalidades que deberán observar los productos de Sistemas de Expediente Clínico Electrónico para garantizar la interoperabilidad, procesamiento, interpretación, conf.” Accessed: Nov. 18, 2022. [Online].

Available: https://www.dof.gob.mx/normasOficiales/4151/salud/salud.htm

[19] “DOF - Diario Oficial de la Federación.” Accessed: Nov. 18, 2022. [Online]. Available: https://dof.gob.mx/nota_detalle.php?codigo=5280847&fecha=30/11/2012#gsc.tab=0

[20] “¿Qué Es Análisis Clínico? Conoce La Definición | CEMP.” Accessed: Apr. 05, 2022. [Online]. Available: https://cemp.es/noticias/que-es-analisis-clinico-conoce-la-definicion/

[21] “Estudios que realiza el laboratorio.” Accessed: Nov. 18, 2022. [Online]. Available: http://www.iner.salud.gob.mx/interna/labclinico_estudios.html

[22] Rahul. Bhagat and Calvin. Hui, HL7 for busy professionals : your no sweat guide to understanding HL7. Anchiove, 2015.

[23] “Welcome to the HL7 FHIR Foundation.” Accessed: Oct. 14, 2023. [Online]. Available: https://www.fhir.org/

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] S. Liu et al., “Integration of NLP2FHIR Representation with Deep Learning Models for EHR Phenotyping: A Pilot Study on Obesity Datasets,” AMIA Jt Summits Transl Sci Proc, vol. 2021, pp. 410–419, May 2021, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34457156

[29] S. Tomovic, K. Pavlovic, and M. Bajceta, “Aligning document layouts extracted with different OCR engines with clustering approach,” Egyptian Informatics Journal, vol. 22, no. 3, pp. 329–338, 2021, doi: https://doi.org/10.1016/j.eij.2020.12.004.

[30] G. B. Holanda et al., “Development of OCR system on android platforms to aid reading with a refreshable braille display in real time,” Measurement, vol. 120, pp. 150–168, 2018, doi: https://doi.org/10.1016/j.measurement.2018.02.021.

[31] R. Pramanik and S. Bag, “Shape decomposition-based handwritten compound character recognition for Bangla OCR,” J Vis Commun Image Represent, vol. 50, pp. 123–134, 2018, doi: https://doi.org/10.1016/j.jvcir.2017.11.016.

[32] Y. Kobayashi, S. Mimuro, S. Suzuki, Y. Iijima, and A. Okada, “Basic research on a handwritten note image recognition system that combines two OCRs,” Procedia Comput Sci, vol. 192, pp. 2596–2605, 2021, doi: https://doi.org/10.1016/j.procs.2021.09.029.

[33] N. Sasipriyaa, P. Natesan, R. S. Mohana, E. Gothai, K. Venu, and S. Mohanapriya, “Design and simulation of handwritten detection via generative adversarial networks and convolutional neural network,” Mater Today Proc, vol. 47, pp. 6097–6100, 2021, doi: https://doi.org/10.1016/j.matpr.2021.05.024.

[34] P. K. Nanda and L. Goswami, “Image processing application in character recognition,” Mater Today Proc, 2021, doi: https://doi.org/10.1016/j.matpr.2021.03.697.

[35] X. Gao et al., “Removing light interference to improve character recognition rate by using single-pixel imaging,” Opt Lasers Eng, vol. 140, p. 106517, 2021, doi: https://doi.org/10.1016/j.optlaseng.2020.106517.

[36] J. Martínek, L. Lenc, and P. Král, “Building an efficient OCR system for historical documents with little training data,” Neural Comput Appl, vol. 32, no. 23, pp.

17209–17227, 2020, doi: 10.1007/s00521-020-04910-x.

[37] W. Yu, N. Lu, X. Qi, P. Gong, and R. Xiao, “PICK: Processing Key Information Extraction from Documents using Improved Graph Learning-Convolutional Networks,” in 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 4363–4370. doi: 10.1109/ICPR48806.2021.9412927.

[38] 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.

[39] K. Luo, J. Lu, K. Q. Zhu, W. Gao, J. Wei, and M. Zhang, “Layout-aware information extraction from semi-structured medical images,” Comput Biol Med, vol. 107, pp. 235–247, 2019, doi: https://doi.org/10.1016/j.compbiomed.2019.02.016.

[40] 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.

[41] S. Karthikeyan, A. G. S. de Herrera, F. Doctor, and A. Mirza, “An OCR Post-correction Approach using Deep Learning for Processing Medical Reports,” IEEE Transactions on Circuits and Systems for Video Technology, p. 1, 2021, doi: 10.1109/TCSVT.2021.3087641.

[42] 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.

[43] 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.

[44] “Python 3: los fundamentos del lenguaje - Sébastien Chazallet - Google Libros.” Accessed: Feb. 26, 2022. [Online]. Available: https://books.google.es/books?hl=es&lr=&id=KRYyvKmZvpwC&oi=fnd&pg=PA5&dq=python&ots=UG7aAtgbQ-

&sig=wzm6U1k5G1wS6LbO_gBQcTzCnnU#v=onepage&q&f=false

[45] “HTML: Lenguaje de etiquetas de hipertexto | MDN.” Accessed: Feb. 27, 2022. [Online]. Available: https://developer.mozilla.org/es/docs/Web/HTML

[46] “CSS | MDN.” Accessed: Nov. 18, 2022. [Online]. Available: https://developer.mozilla.org/es/docs/Web/CSS

[47] “JavaScript | MDN.” Accessed: May 15, 2022. [Online]. Available: https://developer.mozilla.org/es/docs/Web/JavaScript

[48] “(Manual-TypeScript.pdf) TypeScript - Tutoriales en PDF.” Accessed: May 15, 2022. [Online]. Available: https://tutorialesenpdf.com/typescript/previsualizacion/Manual-TypeScript.pdf

[49] “¿Qué Es MongoDB? | MongoDB.” Accessed: Feb. 27, 2022. [Online]. Available: https://www.mongodb.com/es/what-is-mongodb

[50] “Angular - Introduction to Angular concepts.” Accessed: May 15, 2022. [Online]. Available: https://angular.io/guide/architecture

[51] “Flask | The Pallets Projects.” Accessed: Nov. 18, 2022. [Online]. Available: https://palletsprojects.com/p/flask/

[52] “pytesseract · PyPI.” Accessed: Feb. 27, 2022. [Online]. Available: https://pypi.org/project/pytesseract/

[53] “Documentation for Visual Studio Code.” Accessed: Feb. 27, 2022. [Online]. Available: https://code.visualstudio.com/docs

[54] “Medical Subject Headings - Home Page.” Accessed: May 24, 2023. [Online]. Available: https://www.nlm.nih.gov/mesh/meshhome.html

[55] “DeCS – Descritores em Ciências da Saúde.” Accessed: May 24, 2023. [Online]. Available: https://decs.bvsalud.org/es/