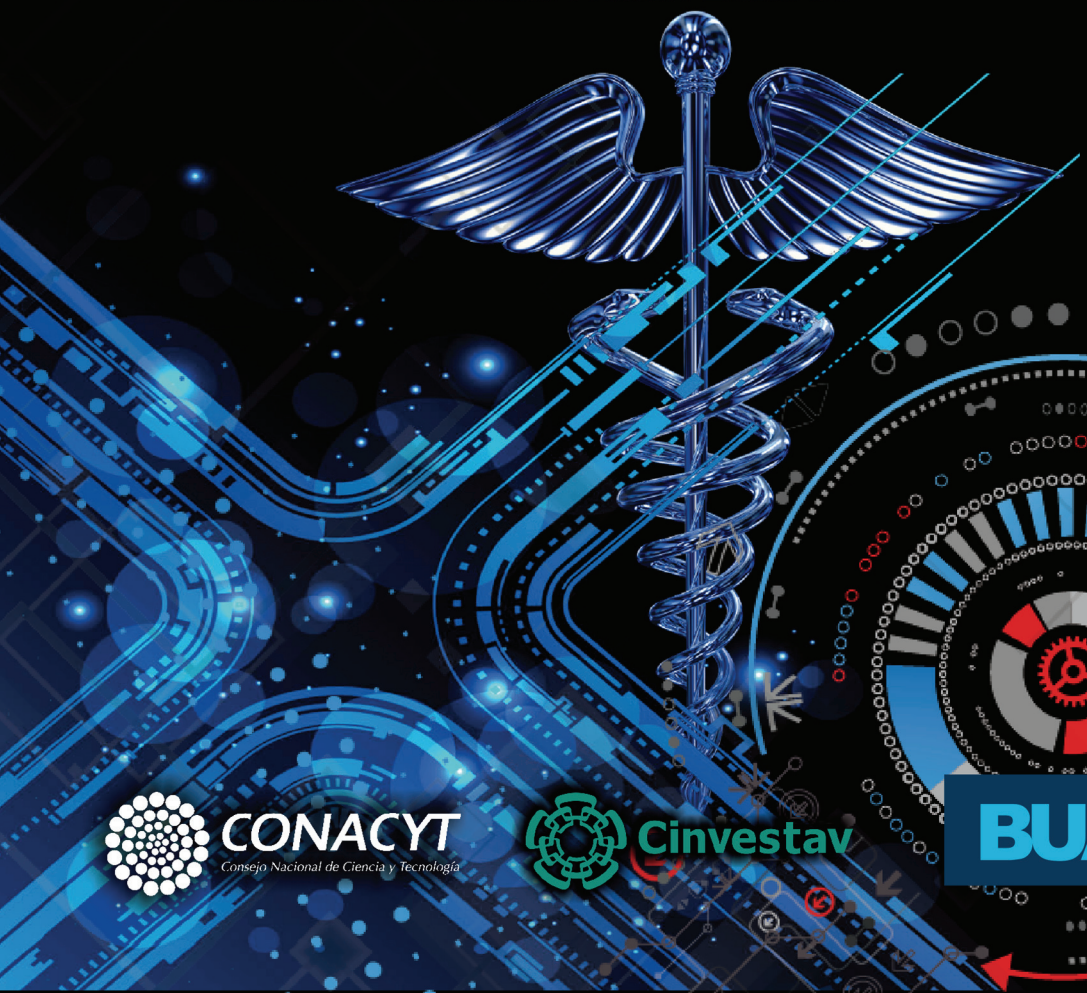
Judith Pérez Marcial
coordinadora

BIGDATA EN SALUD

TECNOLOGÍAS EMERGENTES Y APLICACIONES

EDITORES:
José Luis González Compeán
María del Carmen Santiago Díaz
Miguel Morales Sandoval
Gustavo Trinidad Rubín Linares



CONACYT
Consejo Nacional de Ciencia y Tecnología
Cinvestav
BUAP
Montiel & Soriano
EDITORES

Bigdata en salud: tecnologías emergentes y aplicaciones

Bigdata en salud: tecnologías emergentes y aplicaciones

José Luis González Compeán
María del Carmen Santiago Díaz
Miguel Morales Sandoval
Gustavo Trinidad Rubín Linares
Editores

Judith Pérez Marcial
Coordinadora

Bigdata en salud: tecnologías emergentes y aplicaciones
José Luis González Compeán
María del Carmen Santiago Díaz
Miguel Morales Sandoval
Gustavo Trinidad Rubín Linares
Editores

Judith Pérez Marcial
Coordinadora

Primera Edición: Noviembre 2022
ISBN versión impresa: 978-607-8857-23-4
ISBN versión digital: 978-607-8857-25-8

Montiel & Soriano Editores S.A. de C.V.
15 sur 1103-6 Col. Santiago Puebla, Pue.

Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional - CINVESTAV
Avenida Instituto Politécnico Nacional #2508 Col. San Pedro Zacatenco
C.P. 07360, Delegación Gustavo A. Madero, Ciudad de México.

CINVESTAV Unidad Tamaulipas
Dirección: Carretera Victoria- Soto la Marina Kilómetro 5.5,
Ciudad Victoria - Soto la Marina, 87130 Cd Victoria, Tamps.

Este trabajo ha sido financiado por el proyecto No. 41756 titulado "Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud" del fondo PRONACES-CONACYT.

ESTA OBRA, PARA SER PUBLICADA, FUE ARBITRADA A DOBLE CIEGO Y AVALADA POR EL SISTEMA DE PARES ACADÉMICOS.

Impreso y Hecho en México / Printed and bound in México

Índice

Introducción

Los datos, ahora mayormente disponibles en formato digital, se han convertido en el principal activo de personas, organizaciones, gobiernos y cualquier entidad que, con las herramientas adecuadas, pueden procesarlos y obtener información útil para tomar decisiones. El sector salud, incluido el de nuestro país, está produciendo cúmulos de información sobre la práctica médica y las interacciones entre los profesionales de la salud y los pacientes. Los datos producidos en el dominio de la salud, con el fin de ponerlos a disponibilidad de la comunidad científica, se están recolectando de manera acelerada y sostenida; se capturan, se condensan y almacenan en repositorios públicos o privados que concentran bases de datos producidas por diferentes instituciones públicas del sector salud. Un modelo computacional llamado Big Data se ha popularizado como una solución para analizar, manejar, almacenar cúmulos de datos de una gran variedad a una gran velocidad, y en un constante crecimiento. Este modelo se aplica para desarrollar Ciencia de Datos en donde se desarrolla investigación basada en los datos disponibles para producir información que permita tomar decisiones tales como diagnósticos, pronósticos o predicciones.

Actualmente, en el contexto de Ciencia de Datos y de Big Data, se han comenzado a procesar estos grandes repositorios de datos sobre salud para producir información que permita obtener el conocimiento que se requiere para que los servicios de salud mejoren sus procesos de toma de decisiones con el fin de mejorar la administración de los recursos del sistema de salud, mejorar los procesos de atención a los pacientes y extraer conocimiento para las enfermedades que los aquejan.

Para crear sistemas de ciencia de datos y aplicaciones que aprovechen la alta disponibilidad de datos en salud, se requiere la integración de múltiples tecnologías, métodos, algoritmos y aplicaciones en sistemas eficientes seguros y tolerantes a fallos.

En este libro de Big Data en Salud: Tecnologías Emergentes y Aplicaciones, se presentan al lector las tecnologías emergentes, métodos, algoritmos, aplicaciones y contribuciones de la comunidad científico-académica que trabaja en el desarrollo de Big Data en Salud, abarcando tecnologías para la adquisición de datos en salud, su procesamiento, almacenamiento, distribución, manejo, acceso y uso. El libro brinda una visión de la complejidad detrás de los sistemas de ciencia de datos en salud y una ruta que permita a científicos, tecnólogos, estudiantes y público en general, conocer y adentrarse en el área de Big Data en Salud, la cual es crítica para avanzar de manera significativa hacia una inteligencia de datos en el sector salud, que se traduzca en una mejor atención sanitaria y permita diseñar políticas de salud preventivas efectivas.

El libro recopila 14 capítulos sobre el trabajo desarrollado por miembros de la comunidad académica, científico y tecnológica en materia de Big Data en

Salud. Estas contribuciones se han agrupado en técnicas, servicios y aplicaciones.

Por el lado de las técnicas, en el Capítulo 1 se aborda el tema de la seguridad de los datos en salud, un tema crucial, dado que los datos médicos son sensibles por naturaleza, y su adquisición, almacenamiento, acceso y uso requiere de mecanismos robustos que garanticen la seguridad y privacidad de estos. En el Capítulo 2, se continúa el tema de la privacidad de datos desde el punto de vista de su uso, en el caso de la minería de procesos. En este tipo de minería, el objetivo es mejorar los procesos de salud (como los servicios que brinda un hospital a las personas) haciendo uso de los datos en bitácoras de los sistemas de información de las instituciones de salud, pero sin comprometer la privacidad de dicha información. En el Capítulo 3 se describen técnicas de vanguardia relacionadas con el almacenamiento, distribución y acceso a datos médicos, de forma segura, confiable y con capacidades de trazabilidad. La confiabilidad es necesaria para evitar pérdida de datos y para habilitar la tolerancia a fallas en el acceso a datos médicos. La trazabilidad es un servicio que permite auditar un sistema de ciencia de datos. Estos servicios son generalmente requeridos para hacer cumplir normas sobre manejo y uso de datos médicos, preservando el derecho a la privacidad de pacientes y profesionales de la salud. En el Capítulo 4 se describen técnicas que permiten desplegar sistemas de ciencia de datos de manera transparente para el usuario en distintas plataformas, ya sean equipos personales, organizacionales o en una infraestructura de un tercero. Esta capacidad es requerida, ya que uno de los principales inconvenientes que se han observado en el desarrollo de sistemas en distintos ámbitos, incluidos los de salud, es la dependencia de dichos sistemas con las infraestructuras donde se ejecutan, afectando la portabilidad y reutilización de dichos sistemas.

Por el lado de los servicios, en el Capítulo 5 se presenta una plataforma para crear sistemas de ciencia de datos de manera simple e intuitiva, como una conexión de cajas negras, los sistemas de análisis de datos se crean bajo demanda y es posible, después, desplegar dichos sistemas en prácticamente cualquier infraestructura de cómputo. En el Capítulo 6 se presenta una evaluación de algunos retos tecnológicos y de mayor involucramiento y concientización de las personas que permitan un cambio de paradigma de la medicina que, a la vez, habilite nuevos y más servicios para contar con un servicio médico más personalizado.

En lo que respecta a las aplicaciones, en el Capítulo 7 se presentan los detalles de una plataforma tecnológica integral que permite la gestión, intercambio y preservación de grandes volúmenes de datos en salud, desde la perspectiva de la ciencia de datos. La plataforma integra técnicas y servicios descritos en los capítulos previos, lo que la hace ideal para la implantación de sistemas de ciencia de datos en instituciones del sector salud, ya que cubre prácticamente todo el ciclo de vida en Big data: recopilación, limpieza de datos, análisis y visualización de resultados. Como un estudio de caso de aplicación de esta plataforma, en el Capítulo 8 se presenta una herramienta para el diagnóstico de cáncer de hueso largo. En dicha herramienta, se integran técnicas de aprendizaje máquina, así como su implementación y despliegue para operar sobre datos médicos reales (imágenes DICOM). En el Capítulo 9 se presenta un análisis y reflexión sobre los sistemas de almacenamiento digital, lo cual tiene un impacto importante en el sector salud. En el Capítulo 10 se describe un

modelo de interoperabilidad de sistemas de expediente clínico electrónico, así como el detalle de una plataforma que implementa dicho modelo. Esto permite integrar prácticamente cualquier sistema de expediente clínico electrónico con el fin de consultar la información de pacientes que se encuentra distribuida en los diversos sistemas. El modelo permite, además, la recolección de información relevante para conformar un repositorio público de datos en salud. En el Capítulo 11 se presenta el detalle de un servicio para el transporte y almacenamiento de datos médicos. Dicho servicio permite crear catálogos para que instituciones de salud coloquen sus sistemas, servicios o aplicaciones, de manera que otras instituciones puedan accederlos y utilizarlos. La solución incluye de manera implícita requerimientos de eficiencia, seguridad y fiabilidad. En el Capítulo 12 se describe una metodología para la creación de un chatbot para detección de depresión. El chatbot identifica y califica el nivel de depresión en el usuario, incorporando reconocimiento de voz con un modelo de aprendizaje mixto, un modelo de procesamiento de lenguaje natural y un diagrama de árbol de decisión para orientar la interacción con el usuario. La interfaz gráfica cuenta con un diseño amigable que favorece la usabilidad y la interacción voz a voz con los usuarios. En el Capítulo 13 se discuten diferentes tipos de datos no numéricos en el contexto clínico y las alternativas para lograr una transformación adecuada que permita extraer su valor informativo y poder usarlos en algoritmos de análisis numérico para el apoyo a la toma de decisiones. Se presentan varios casos de estudio en los que se aplican diferentes técnicas de transformación y extracción de características para texto, imágenes y señales fisiológicas. Finalmente, en el Capítulo 14 se discuten diversos retos científicos y tecnológicos para desplegar sistemas inteligentes de e-Salud, entre los más importantes los relacionados con las redes de datos, seguridad e inteligencia artificial. Como caso de aplicación, la discusión se orienta a sistemas y redes de telemedicina.

Este libro es producido en el marco de los Programas Nacionales Estratégicos - PRONACES - en Salud, en el Proyecto Nacional de Investigación e Incidencia – PRONAII – Ciencia de datos en Salud, del proyecto específico No. 41756 “Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud”.

Dr. José Luis González Compeán
Dr. Miguel Morales Sandoval

Seguridad y Privacidad de Datos en Sistemas de Ciencia de Datos en Salud

Melissa Brigitthe Hinojosa-Cabello[0000−0002−0404−0398], Miguel Morales-Sandoval[0000−0003−1702−8467], Diana Elizabeth Carrizales-Espinoza[0000−0002−3925−031X], y José Luis González-Compeán[0000−0002−2160−4407]

Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Ciudad Victoria, Tamaulipas, 87130, México
{melissa.hinojosa,miguel.morales,diana.carrizales, joseluis.gonzalez}@cinvestav.mx

Resumen El término *Big Data* se refiere a la producción de datos a gran velocidad, de una gran variedad y con un alto volumen. En el ámbito de la salud, *Big Data* se refiere a los procesos involucrados en la gestión, almacenamiento, tratamiento y uso de datos médicos que pueden provenir de distintas fuentes. En este contexto, paradigmas como el Internet de las Cosas Médicas (IoMT) o el cómputo en la nube han acelerado la producción masiva de datos en el área de la salud. Los datos originados en nodos IoMT (e.g., monitoreo de frecuencia cardiaca, niveles de glucosa, etc.) se almacenan en la nube y después son consumidos o accedidos desde la misma por los usuarios finales (pacientes, profesionales de la salud, personal médico, entre otros) a partir de diversas aplicaciones. Dada la naturaleza sensible de los datos médicos, que incluyen información tanto personal como médica de los pacientes, resulta necesario preservar el derecho a la privacidad. Si hay datos sensibles, éstos deben estar protegidos en todo momento, durante el ciclo de vida de los datos, ante cualquier divulgación o modificación no autorizada. Es decir, los propietarios de los datos esperan que éstos únicamente estén disponibles y puedan ser accedidos por usuarios autorizados, sin que el proveedor del servicio u otras entidades no autorizadas sean capaces de obtener y procesar dichos datos. En este sentido, los requerimientos de seguridad que se deben cubrir son confidencialidad y control de acceso, principalmente.
En este capítulo presentamos una descripción y detalles de construcción del concepto de *sobres digitales con capacidades de búsqueda*, los cuales son objetos criptográficos que permiten garantizar la privacidad de datos sensibles, como los de salud. Así, los datos únicamente serán accedidos por entidades autorizadas, descritas por un conjunto de atributos que los caracterizan e identifican. De igual forma, el almacenamiento y recuperación segura de datos médicos es indispensable en el desarrollo de sistemas de ciencia de datos. Por ello, al final de este capítulo se describe un caso de uso de los sobres digitales en este tipo de aplicaciones.

Palabras Clave: Confidencialidad · Control de Acceso · Sobres Digitales · eSalud · Big Data.

1. Introducción

La seguridad informática se refiere a todos aquellos mecanismos y recursos utilizados para prevenir accesos no autorizados a los sistemas de información, que incluyen recursos o infraestructura de cómputo, sistemas y datos. De entre éstos, la seguridad de datos es la última línea de defensa, ya que cuando un atacante logra romper la seguridad de la red y del dispositivo, éste tiene la posibilidad de acceder a los datos y comprometer su confidencialidad. En este capítulo, a menos que se indique lo contrario, nos enfocamos únicamente en la seguridad de los datos. Desde esta perspectiva, la seguridad de datos la abordamos desde dos requerimientos principales: confidencialidad de datos y control de acceso hacia éstos.

Definición 1. *Confidencialidad* [6], [7]: La confidencialidad garantiza la privacidad de datos sensibles al impedir su divulgación mediante la restricción del acceso a éstos a personas, recursos o procesos no autorizados, permitiendo que únicamente aquellos con autorización legítima puedan acceder a los datos, consumirlos o procesarlos. Éste es el requerimiento más antiguo y también el más demandado cuando se habla de seguridad de datos.

Definición 2. *Control de acceso* [7]: El objetivo del control de acceso lógico es la protección de cualquier tipo de recurso (datos, aplicaciones, servicios, entre otros) de operaciones inadecuadas llevadas a cabo por usuarios malintencionados. Éste consiste en la definición de una serie de restricciones, normalmente basadas en políticas, que describen quién puede acceder a los recursos y las operaciones permitidas sobre éstos, e impiden el acceso no autorizado mediante soluciones tecnológicas. El control de acceso involucra herramientas y protocolos para gestionar el acceso a sistemas y recursos mediante la identificación, autenticación y autorización de los usuarios.

La confidencialidad de datos puede alcanzarse mediante el cifrado de los mismos. Cifrar significa, a grandes rasgos, una transformación de los datos (D), de un formato legible a uno ilegible (CT), mediante un procedimiento (P) bien definido y conocido. Para realizar dicha transformación se usa una llave criptográfica k_c que, en términos simples, corresponde a una secuencia de bits de longitud n con suficiente aleatoriedad. Este proceso de cifrado se representa por la Ecuación 1. Una vez cifrados, los datos D ya no son accesibles por nadie, salvo por aquellos que posean una llave para descrifrar k_d, y mediante un proceso inverso al cifrado (i.e., descifrado, P^{-1}) puedan transformar nuevamente los datos cifrados CT en D, como se muestra en la Ecuación 2.

$$CT = P(D, k_c) \tag{1}$$

$$D = P^{-1}(CT, k_d) \tag{2}$$

En este sentido, la premisa del cifrado es que, para cualquier entidad que desconozca k_d resulta prácticamente imposible obtener los datos legibles a partir del

texto transformado CT tras ejecutar P^{-1}, incluso siendo éste un procedimiento bien conocido. Todos los cifradores, tanto antiguos como modernos, basan su funcionamiento en los preceptos previamente descritos y, en principio, garantizan el servicio de confidencialidad. Una vez cifrados, los datos pueden almacenarse en un medio inseguro (e.g., en una unidad de disco), transmitirse también por un medio inseguro (como internet), o enviarse a la nube. Sin embargo, es necesario que el propietario de datos imponga y maneje las restricciones de acceso a través de un control sobre las llaves de descifrado. Ante un escenario donde existen grandes colecciones de datos que deben cifrarse para garantizar su confidencialidad, como en el caso de *big data* en salud, se deben resolver al menos tres problemáticas principales: eficiencia, compartición y recuperación.

- La *eficiencia*: El cifrado, finalmente, es un requerimiento no funcional que convella una sobrecarga, tanto en procesamiento como en almacenamiento, para ejecutar los procedimientos P y P^{-1}. El nivel de seguridad está correlacionado con la longitud de la llave k_c, por lo que entre más grande k_c, mayor es el nivel de seguridad, pero también más lentos los procedimientos P y P^{-1}.
- La *compartición*: Los datos generalmente no son consultados solo por el propietario de los mismos. En el caso de eSalud, los datos deben ser accedidos por distintos actores, como los mismos pacientes, médicos, enfermeros, especialistas y, en general, profesionales de la salud. Ante un creciente volumen de datos se hace evidente la necesidad de contar con mecanismos efectivos y eficientes de control de acceso hacia dichos datos, para una compartición no solamente segura, sino también eficiente.
- La *recuperación*: Dado un gran volumen de datos, la recuperación de información es necesaria para localizar rápidamente datos de interés y recuperarlos para su acceso y consumo. Pero, ¿qué pasa si los datos están cifrados y, por lo tanto, se encuentran en formato ilegible? ¿Cómo un motor de búsqueda puede localizar y recuperar datos ilegibles de interés? Éste es uno de los desafíos más relevantes en el contexto del cifrado de grandes colecciones de datos. Afortunadamente, existen mecanismos que permiten abordar este problema, en lo que se conoce como *Searchable Encryption (SE)* [21]. Bajo este enfoque, los usuarios de los datos pueden hacer búsquedas cifradas, esto es, enviar al proveedor del servicio de almacenamiento de los datos cifrados un *'token'* que indica palabras clave en formato cifrado, de tal forma que éste pueda usarlo para buscar en los datos cifrados y localizar aquellos que empaten con los criterios de búsqueda. Al estar cifrados el *token* y los datos, el servidor realiza algo parecido a una búsqueda a ciegas pero efectiva, recuperando los datos de interés sin aprender acerca de los criterios de búsqueda o de los datos localizados. Sin embargo, no cualquier esquema SE podría ser adecuado para un entorno particular.

En este capítulo presentamos el concepto, diseño, implementación y evaluación en un caso de uso de *sobres digitales con capacidades de búsqueda (SDB)*,

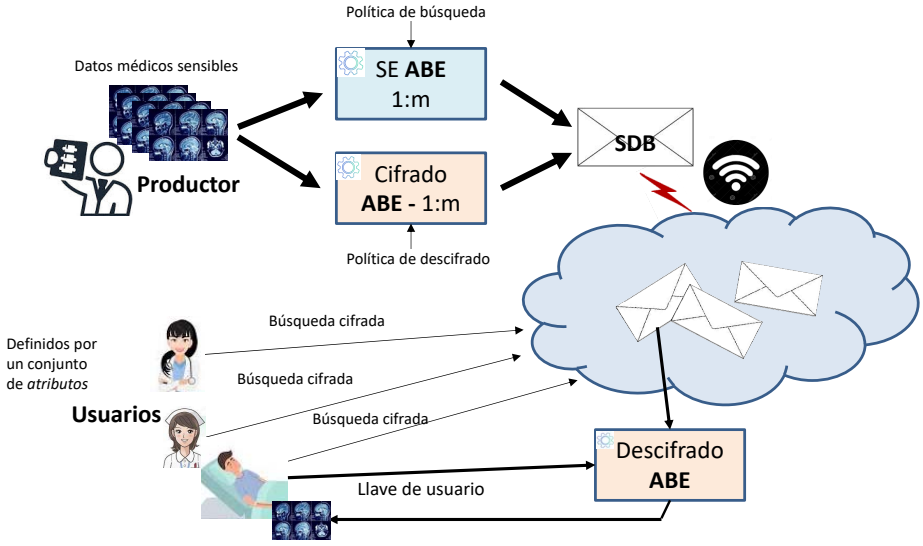


Figura 1: Vista general del concepto de sobres digitales con capacidades de búsqueda y su aplicación en el dominio de *big data* en salud.

los cuales son abstracciones fundamentadas en el cifrado de datos no convencional, llamado cifrado basado en atributos (ABE) [17]. Mediante ABE y bajo el concepto de SDB abordamos los tres problemas previamente descritos. En las siguientes secciones se darán detalles tanto de ABE como de su implementación con capacidades de búsqueda (SE-ABE). Por ahora, el enfoque de solución y concepto de SDB se muestra en la Figura 1. Con base en ello, el proceso de aseguramiento de datos sensibles en salud mediante SDB se lleva a cabo de la siguiente forma:

1. El productor de datos, que generalmente es un médico, especialista en salud, o incluso algún dispositivo médico, genera datos sensibles, como radiografías;
2. Estos datos en algún momento serán requeridos por otro médico o profesional de la salud, incluso por el mismo paciente; al ser las radiografías en nuestro ejemplo datos sensibles, éstas no pueden almacenarse en un disco o memoria, o enviarse mediante correo o algún servicio de mensajería a través de un teléfono inteligente; los datos deben asegurarse y almacenarse en un medio desde el cual después dichos datos puedan recuperarse; para ello, el productor de los datos debe ejecutar dos macro-procesos: *i*) el algoritmo de cifrado ABE, el cual requiere de una política de control de acceso que, en esencia, determina qué usuarios podrán descifrar esos datos más adelante; *ii*) el algoritmo de búsquedas cifradas SE-ABE, el cual también requiere de una política de control de acceso, pero que determina qué usuarios podrán consultar los datos cifrados mediante *tokens* igualmente cifrados;

3. Tanto el cifrado ABE como parte de la búsqueda cifrada SE-ABE se ejecutan del lado del productor de datos; el resultado de cada proceso es, por un lado, los datos cifrados y, por otro, un índice seguro de búsqueda sobre los datos cifrados; todo esto es lo que conforma el SDB;
4. Una vez creado, el SDB es enviado a un repositorio (la nube), de donde más adelante los usuarios autorizados podrán hacer consultas (si sus atributos satisfacen la política usada en SE-ABE), recuperar la información de interés (cifrada) y acceder a ella descifrándola usando su llave de usuario estrechamente relacionada con los atributos que le describen (médico, enfermera, paciente y demás datos relacionados con dicho rol);
5. Un usuario de los datos está descrito por los atributos que lo caracterizan; por ejemplo, un profesional de la salud puede tener atributos como su especialidad, nivel jerárquico en la organización donde labora, datos del lugar donde labora que describen su pertenencia a dicha organización, datos personales y cualquier otro que sea relevante para el propósito; con base en estos atributos, cada usuario cuenta con una llave, la cual es intransferible y necesaria para crear los *tokens* de búsqueda y para descifrar los datos recuperados desde el repositorio de datos cifrados.

Nuestra construcción de SDB se realiza sobre emparejamientos bilineales asimétricos, los cuales son estructuras matemáticas en el dominio de la teoría de grupos y campos finitos. Con ello, ABE no solo es lo suficientemente seguro para preservar la confidencialidad de los datos, sino también para garantizar el control de acceso a éstos únicamente a entidades específicas mediante el cifrado de uno a muchos. Asimismo, se habilitan las búsquedas cifradas también bajo el concepto de cifrado basado en atributos, reutilizando las estructuras algebraicas de esta técnica criptográfica. El cifrado de datos en SDB consta de dos capas de cifrado a partir de las cuales se toma ventaja a la par de la eficacia y del elevado nivel de seguridad provisto por ABE y de la eficiencia de los cifradores simétricos [18]. Los datos sensibles son cifrados mediante llaves de sesión de un cifrador simétrico, mientras que las llaves de sesión se cifran a partir de ABE. De esta forma, es posible preservar la confidencialidad de los datos y decidir selectivamente los usuarios autorizados para acceder y consumir determinados conjuntos de datos.

Este capítulo está organizado de la siguiente manera: en la Sección 2 presentamos los conceptos más relevantes para la definición de los SDB; en la Sección 3 se presentan los detalles de diseño de sobres digitales, mientras que en la Sección 4 se presenta el diseño de los SDB; en la Sección 5 se describen las estrategias de eficiencia y paralelismo de los SDB, mientras que en la Sección 6 se detalla la validación de la construcción de SDB en el dominio de la salud, siendo un componente principal en el despliegue de un servicio que permite el aseguramiento de datos médicos desde su producción hasta su consumo [4]; en la Sección 7 se discuten, desde la perspectiva de este trabajo, los retos para proteger los datos en el sector salud; finalmente, en la Sección 8 se presentan las conclusiones.

2. Antecedentes

Día con día, usuarios e instituciones generan y recolectan grandes cantidades de datos derivados de actividades cotidianas, tales como compras en línea, realización de trámites, transacciones bancarias, entre muchas otras. Gran parte de estos datos se encuentra disponible de forma pública en sitios web o redes sociales, por citar algunos ejemplos. Sin embargo, existen datos que son sensibles y requieren protección ante el acceso no autorizado por parte de terceros [6]. Ejemplos de datos sensibles son los registros médicos o financieros, números de cuenta bancarios, datos de identificación personal o número de seguridad social; planes de adquisición, información personal de clientes, datos financieros o derechos de propiedad intelectual. De esta forma, a medida que se generan y distribuyen grandes cantidades de datos, la protección de éstos se vuelve indispensable para sus propietarios, ya sean individuales o grandes empresas e instituciones.

La importancia de la seguridad de la información radica en la protección de los datos y sistemas que los producen o utilizan del daño, uso, divulgación o destrucción no autorizados [11]. Con el volumen y variedad de datos, así como la velocidad con la que éstos son generados por los usuarios y las operaciones diarias del negocio, la confidencialidad, integridad y disponibilidad de los datos son esenciales para la seguridad de la información [6]. Uno de sus principios fundamentales es la confidencialidad, la cual garantiza la privacidad de los datos al restringir el acceso a éstos través del cifrado de su contenido. Para ello, se apoya, además, en mecanismos de autenticación o concesión de niveles de privilegios, permitiendo que solamente personas autorizadas puedan ver o manipular datos. De esta manera se evita que entidades no autorizadas puedan derivar u obtener información a partir de dichos datos [6], [7].

2.1. Criptografía

Uno de los métodos más utilizados para asegurar la confidencialidad de los datos es el cifrado. La criptografía se encarga de implementar el cifrado realizando transformaciones a los datos de manera que, al almacenarlos y transmitirlos, solo los destinatarios autorizados puedan accederlos y procesarlos [10]. Dichas transformaciones se llevan a cabo del lado de los propietarios de datos, convirtiendo información legible en texto incomprensible, y de los destinatarios, aplicando el proceso inverso de descifrado para obtener los datos originales [12], [10]. Los métodos criptográficos modernos emplean algoritmos seguros desde el punto de vista computacional con la finalidad de que la información protegida no pueda ser comprometida fácilmente. Dicho objetivo se logra mediante mecanismos como el cifrado de los datos o de los canales de comunicación durante su transmisión, así como la generación de códigos de autenticación de mensajes y firmas digitales [12], [7].

Los procesos de cifrado y descifrado requieren dos componentes elementales: un algoritmo criptográfico, o cifrador, y una llave. El algortimo consiste en la aplicación de funciones matemáticas que, en conjunto con una llave que se utiliza

como parámetro de entrada durante el procedimiento, realizan las transformaciones pertinentes a los datos [12]. Al llevar a cabo el cifrado de los datos, éstos se protegen al convertir su forma legible en texto cifrado que resulta incomprensible. En contraste, el descifrado constituye el proceso opuesto, a partir del cual se obtienen los datos sensibles al remover la protección suministrada a través del cifrado [7]. Por otra parte, la llave o clave criptográfica es un componente indispensable de cualquier algoritmo de cifrado. Usualmente, estas claves se generan de manera aleatoria previa al cifrado de los datos, aunque éstas también pueden ser especificadas por el usuario [12], [11].

Cabe destacar que la ventaja del uso de un algoritmo sobre otro radica en la eficacia de la generación y administración de las llaves utilizadas para los procesos de cifrado y descifrado. Cuanta mayor dificultad asociada a la clave criptográfica, mayor seguridad será capaz de brindar el algoritmo; no obstante, su ejecución se vuelve más compleja. En consecuencia, la seguridad del cifrado reside sustancialmente en el secreto de las claves, no en el algoritmo. En la actualidad, existen diversos algoritmos de cifrado y, debido a que éstos son de acceso público, las claves criptográficas que dichos algoritmos utilizan son las que garantizan la discreción de los datos [7]. Son diversos los sistemas criptográficos que han presentado fallas debido a errores en sus procedimientos de administración de llaves. En la práctica, la mayoría de los ataques implican vulnerar el sistema de gestión de claves, en lugar del algoritmo criptográfico en sí. Es por ello que la gestión de claves constituye la parte más difícil de abordar al momento de diseñar un sistema criptográfico.

Cada método de cifrado utiliza un algoritmo específico (P, como ha sido descrito en la Ecuación 1) que se compone de una serie de pasos bien definidos, generalmente estandarizados, utilizados para cifrar y descifrar los datos. Existen diversos métodos para crear texto cifrado, siendo los más antiguos la transposición o sustitución de caracteres y, el más reciente, la combinación de datos con claves secretas. Si bien los algoritmos de cifrado contemporáneos emplean técnicas más robustas basadas en problemas matemáticos, no hay un algoritmo que resulte idóneo para cualquier caso de aplicación. La adopción de éste dependerá del nivel de sensibilidad y cantidad de información que se requiera proteger, así como de los inconvenientes que se pretendan mitigar. Asimismo, la forma de almacenamiento o de transmisión de los datos y los recursos computacionales con los que se cuente determinarán, en gran medida, la opción que habrá de elegirse para llevar a cabo esta tarea [22], [7]. Según la manera en que se gestionen las llaves, los algoritmos criptográficos se pueden dividir en simétricos y asimétricos.

2.2. Criptografía de Clave Privada: Cifradores Simétricos

Los algoritmos simétricos se caracterizan por hacer uso de una cantidad menor de recursos computacionales que su contraparte asimétrica. Esto se debe a que en el cifrado simétrico se utiliza una misma llave para realizar el cifrado y descifrado de los datos, tal como se observa en la Figura 2. A esta llave se le denomina clave secreta o previamente compartida, ya que el emisor y recep-

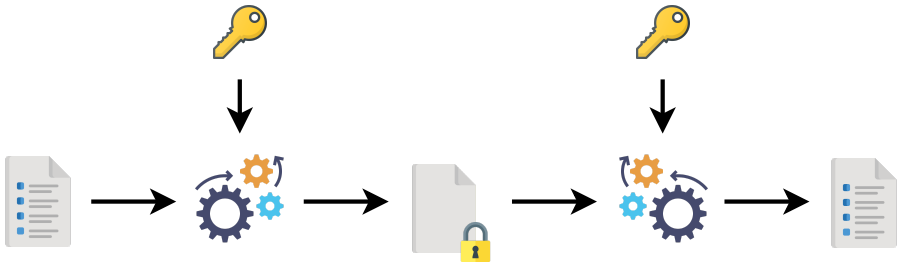


Figura 2: Flujo de operaciones en el cifrado simétrico.

tor deben conocerla antes de que inicie el proceso de cifrado [7], [12]. Dado un mensaje, la clave secreta sirve como parámetro de entrada para el algoritmo de cifrado que aplica las transformaciones necesarias para producir como salida un texto cifrado. Cabe destacar que dicho mensaje se procesa a nivel de arreglos de bytes, por lo que éste puede representar desde una cadena de caracteres hasta un archivo de cualquier extensión. Por el contrario, el algoritmo de descifrado recibe como entrada el texto cifrado, así como la misma clave previamente compartida, y produce como resultado el texto plano del mensaje original.

Los algoritmos simétricos pueden dividirse en cifradores por bloque o por flujo en función de la cantidad de datos de entrada que manejan. Es decir, la diferencia entre ambos recae en la forma de realizar el agrupamiento de bits para los procesos de cifrado y descifrado. Los algoritmos de cifrado por bloque dividen los datos de entrada en bloques de tamaño fijo, usualmente de 64 ó 128 bits, y posteriormente realizan el procesamiento de dichos bloques. En cambio, los cifradores por flujo procesan los datos de entrada conforme éstos se van recibiendo, esto es, un byte o un bit a la vez [7]. De acuerdo con Barker [2], el estándar de cifrado avanzado (AES) es el algoritmo recomendado en la actualidad por el Instituto Nacional de Estándares y Tecnología (NIST) para el cifrado-descifrado de datos. AES es un cifrador por bloque desarrollado para reemplazar al ya obsoleto estándar de cifrado de datos (DES), por lo que constituye el algoritmo más utilizado en la actualidad. Este algoritmo procesa datos en bloques de 128 bits utilizando claves de 128, 192 ó 256 bits, con lo cual se considera que es capaz de proveer niveles de seguridad válidos más allá del año 2030.

2.3. Criptografía de Clave Pública: Cifradores Asimétricos

Para el cifrado y descifrado, los algoritmos asimétricos utilizan un par de claves –una pública y otra privada– matemáticamente relacionadas entre sí, lo cual implica que estos algoritmos sean más complejos que su contraparte simétrica. Si bien existe una relación matemática entre las claves, no es posible obtener la llave privada a partir de la llave pública debido a la complejidad de las operaciones involucradas en la generación del par de claves. Es por ello que estos algoritmos requieren una mayor cantidad de recursos computacionales y su pro-

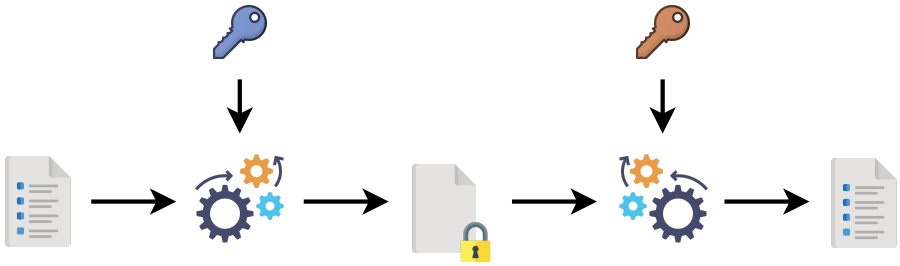


Figura 3: Flujo de operaciones en el cifrado asimétrico.

ceso de ejecución es más lento. En el ejemplo expuesto en la Figura 3, en un sistema de cifrado de clave pública el emisor cifra los mensajes utilizando la clave pública de un receptor en particular [12], [11]. De esta forma, únicamente la clave privada asociada a la clave pública usada para el cifrado puede descifrar los datos. Debido a ésto, no es necesario que ambas partes hagan uso de una llave previamente compartida para intercambiar mensajes de manera segura.

Al igual que en la criptografía de clave privada, los algoritmos asimétricos son de acceso público, por lo que es posible conocer cómo trabajan de manera general. Sin embargo, sus operaciones se basan en problemas matemáticos complejos pero bien conocidos, como la factorización de enteros y los logaritmos discretos. Dado lo anterior, la seguridad de este tipo de algoritmos recae en el par de claves de las cuales se hace uso en los procesos de cifrado y descifrado [7]. Al utilizar funciones del álgebra abstracta en lugar de números reales se vuelve poco viable la búsqueda de la clave privada asociada a una clave pública, incluso si se conoce con qué algoritmo se crearon ambas, dado el tiempo computacional requerido y el elevado costo asociado a dicho procedimiento [12]. Rivest-Shamir-Adleman (RSA), Diffie-Hellman (DH), ElGamal y la criptografía de curva elíptica (ECC) son ejemplos de algoritmos asimétricos utilizados hoy en día.

Como se puede observar, existen evidentes diferencias entre los algoritmos simétricos y asimétricos.

Los primeros son más eficientes en cuanto al tiempo de procesamiento requerido y la cantidad de datos que pueden manejar, pero la gestión de claves es mucho más difícil por el uso de claves compartidas. Otra de las grandes diferencias entre ambos es el tamaño de las llaves, lo cual determina su susceptibilidad a ataques por fuerza bruta. Al ser las claves asimétricas de mayor tamaño en comparación con las claves simétricas –1024 bits contra 128 bits, respectivamente–, el rango de valores posibles es también mucho mayor, lo cual hace inviables este tipo de ataques [11]. En la práctica, ambos esquemas se utilizan en conjunto: el cifrado simétrico se utiliza para garantizar la confidencialidad de grandes volúmenes de datos y la criptografía de clave pública para el intercambio de las claves secretas.

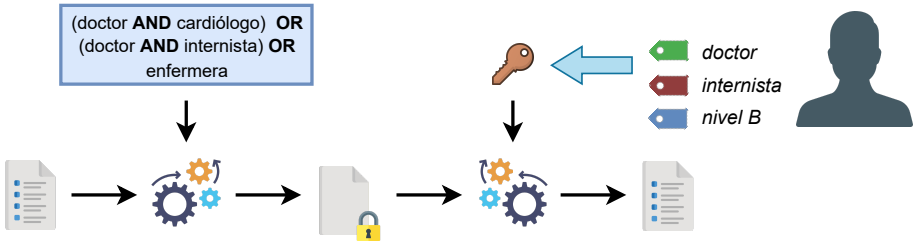


Figura 4: Flujo de operaciones en el cifrado basado en atributos.

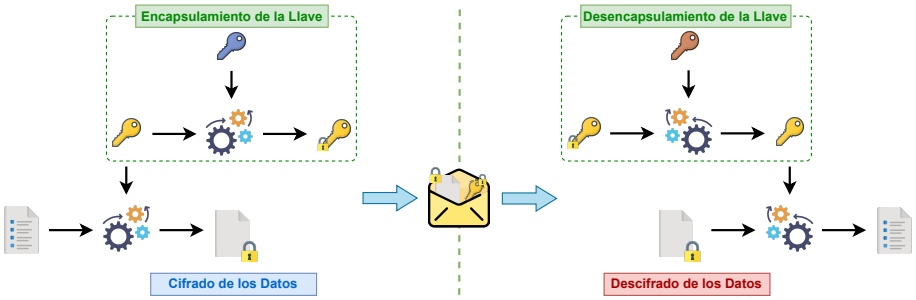
2.4. Cifrado Basado en Atributos

El cifrado basado en atributos (ABE) es una técnica de la criptografía de clave pública que tiene su fundamento en un mecanismo de control de acceso en el que una entidad se identifica a partir de un conjunto de atributos descriptivos. Éste permite compartir datos de forma segura con múltiples usuarios, a la vez que ofrece una gran flexibilidad de gestión de acceso a los datos [17], [1]. En lugar de utilizar las tradicionales claves públicas o privadas, en ABE los datos se cifran mediante la especificación de los atributos que un potencial usuario debe poseer para poder descifrar un mensaje utilizando su clave secreta, tal como se muestra en la Figura 4. Dichos atributos se especifican en estructuras denominadas políticas de control de acceso, las cuales establecen las reglas de acceso a los datos de los propietarios mediante compuertas lógicas (*AND*, *OR*) o de tipo umbral (*k-of-n*) [14].

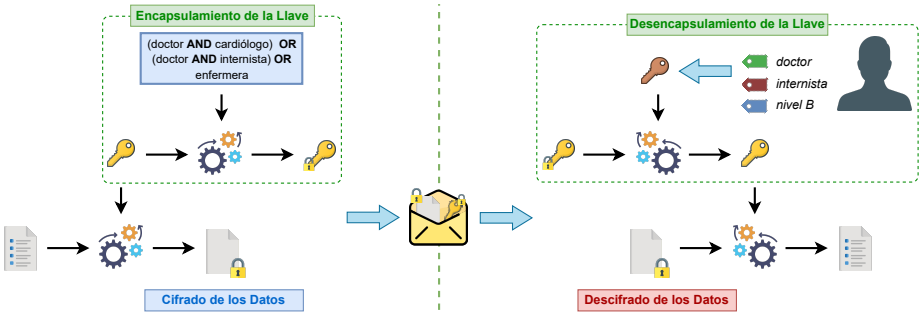
Cabe mencionar que, a diferencia de otros algoritmos de clave pública, ABE es un esquema de cifrado de muchos a muchos, por lo que los propietarios de datos no tiene que conocer de antemano a todos los posibles usuarios. Dado que los algoritmos asimétricos tradicionales utilizan un par de claves relacionadas matemáticamente, un mismo mensaje se tiene que cifrar tantas veces como destinatarios existan para éste. Por el contrario, puesto que en ABE un atributo puede ser común a múltiples usuarios, el cifrado mediante políticas de acceso permite abarcar un mayor número de destinatarios. Esta característica representa una de las principales ventajas de este esquema, ya que permite un control de acceso de grano fino, sin incurrir en sobrecargas de almacenamiento y comunicación asociadas a algoritmos como RSA [19]. De este modo, ABE resulta más adecuado para escenarios de almacenamiento y compartición de datos en la nube, ya que los datos de los propietarios permanecen confidenciales incluso en entornos poco confiables.

3. Sobres Digitales

Como se mencionó anteriormente, la confidencialidad de datos sensibles se logra a partir del cifrado, y éste puede realizarse mediante algoritmos simétricos o asimétricos. Ambos realizan transformaciones a los datos de modo que solamente aquellos destinatarios que posean la correspondiente llave de descifrado



(a) Criptografía de clave pública tradicional.



(b) Cifrado basado en atributos.

Figura 5: Flujo de operaciones en la creación y uso de sobres digitales.

puedan acceder a dichos datos. Sin embargo, estos dos tipos de cifradores conllevan desventajas que dificultan su uso de forma aislada; por ello, éstos suelen emplearse de forma conjunta en aplicaciones prácticas. La ventaja de los algoritmos simétricos sobre los asimétricos es la capacidad de cifrar una gran cantidad de datos eficientemente en términos de tiempos de respuesta. No obstante, éstos implican un problema de distribución y gestión de llaves debido a que la llave usada para cifrar es la misma requerida por el proceso de descifrado. Así, es necesario que los propietarios de datos compartan las claves de descifrado con los destinatarios de dichos datos a través de canales de comunicación seguros, algo que no es posible garantizar en todos los casos.

En este sentido, una forma de sortear el problema de compartición de claves es mediante el uso de una técnica criptográfica denominada *sobre digital*. Un sobre digital se define como un objeto criptográfico que consta de dos capas de cifrado a partir de las cuales se transporta y distribuye una llave de sesión de forma segura. Mediante éstos, es posible tomar ventaja simultáneamente tanto de la criptografía de clave pública, como de la criptografía de clave privada. Como se puede observar en la Figura 5a, los datos sensibles son cifrados mediante llaves de sesión de cifradores simétricos, mientras que dichas llaves de sesión se cifran o encapsulan a partir de criptografía de clave pública. Es decir, los datos se cifran y descifran con una misma llave simétrica, que a su vez se cifra utilizando la clave

pública del destinatario de los datos y éste los descifra usando su clave privada. De esta forma es posible preservar la confidencialidad de una gran cantidad de datos en un tiempo razonable, mientras que es posible compartir las llaves de descifrado con destinatarios específicos, aun utilizando canales de comunicación inseguros.

Además de proporcionar mayor robustez contra ataques, los algoritmos asimétricos eluden el problema de compartición de llaves al utilizar un par de claves relacionadas matemáticamente. No obstante, dada dicha relación entre llaves, es necesario conocer a priori a los potenciales usuarios de un mismo conjunto de datos. Por ello, una forma de abordar esta problemática en el contexto de los sobres digitales implica la remoción del algoritmo asimétrico empleado y la incorporación en su lugar del cifrado basado en atributos, tal como se muestra en la Figura 5b. De esta manera, ABE permite compartir datos de forma segura con múltiples usuarios, incluidos aquellos no definidos a priori. Solamente aquellos usuarios que posean un conjunto de atributos que satisfaga de forma criptográfica la política de control de acceso definida previo al cifrado podrán acceder a los datos en texto plano. Es decir, únicamente quienes cumplan con los criterios establecidos en la política de acceso podrán acceder a la llave de sesión y, con ella, a los datos sensibles.

De esta manera, al emplear sobres digitales en conjunto con ABE desaparece la necesidad de implementar mecanismos adicionales de gestión de llaves. Lo anterior, considerando que los atributos permiten describir las características de los usuarios, así como sus inherentes derechos de acceso. De esta forma, se evitan sobrecargas de cómputo, resulta poco significativo si la transmisión de datos se realiza mediante canales de comunicación seguros o no, y se impone un control de acceso de grano fino. Incluso si un sobre es filtrado pero su portador no cuenta con los atributos que satisfacen la política utilizada en la creación de dicho sobre, éste no será capaz de acceder al contenido legible del sobre digital. Cabe destacar que, a partir de los atributos que posean los usuarios, una autoridad de confianza (TA) se encarga de generarle a cada usuario su correspondiente llave secreta, la cual permite corroborar si éste satisface la política utilizada en el cifrado. Además, la TA tiene la facultad de implementar mecanismos de revocación de acceso para el caso de aquellos usuarios que dejen de pertenecer a la organización donde se gestionan los datos sensibles o aquellos que hagan uso indebido de los mismos.

4. Sobres Digitales con Capacidades de Búsqueda

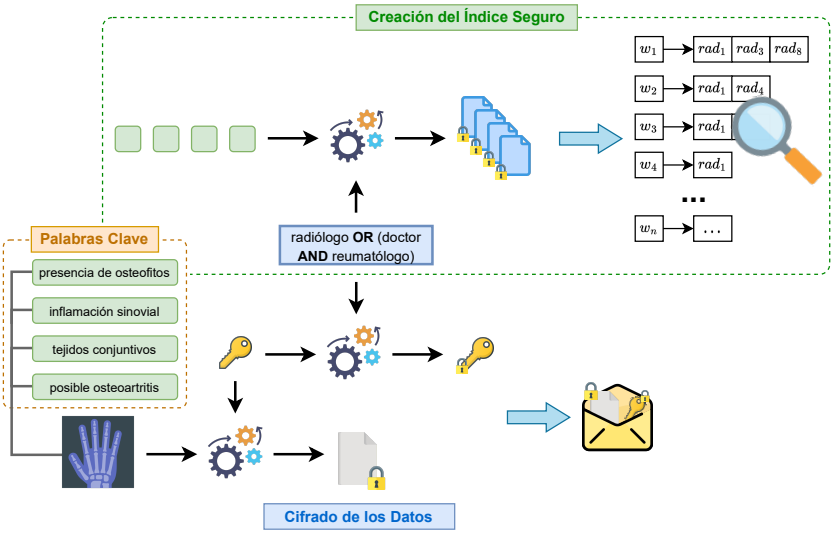
El cifrado de datos permite garantizar la confidencialidad de datos sensibles en entornos poco confiables, como en el caso de escenarios de almacenamiento en la nube. Asimismo, el almacenamiento en la nube facilita el acceso conveniente a los datos y su compartición con múltiples usuarios, proveyendo, además, capacidades de búsqueda y recuperación de información. Sin embargo, el hecho de que los propietarios cifren sus datos previo a externalizarlos a la nube introduce dos grandes problemas. En primer lugar, la compartición se vuelve una tarea

compleja que implica que los propietarios gestionen los mecanismos de control de acceso hacia sus datos. Dicha problemática se puede abordar a través del uso de ABE, el cual ofrece una gestión flexible mediante controles de acceso de grano fino a la vez que garantiza la confidencialidad de los datos. En segundo lugar, las capacidades de búsqueda del proveedor del servicio de almacenamiento no se pueden aprovechar debido a que los datos se encuentran en formato ininteligible.

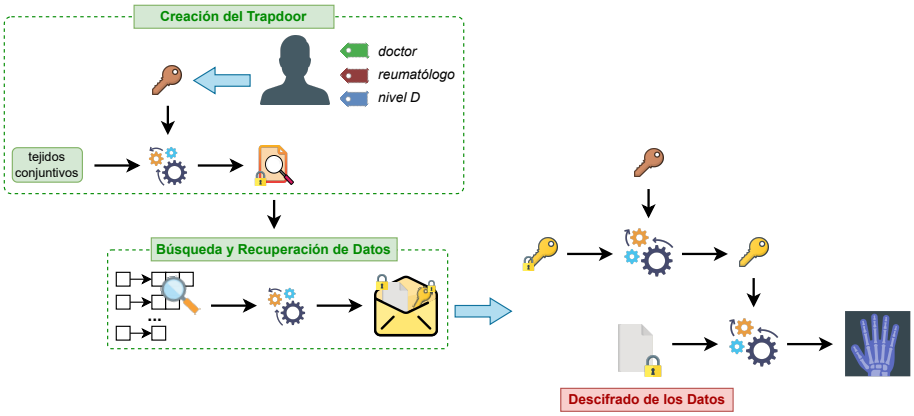
Si bien los usuarios podrían descargar todos los datos (cifrados), descifrarlos y aplicar localmente algoritmos de búsqueda y recuperación tradicionales, este enfoque es totalmente inviable en la práctica por varias razones. Por ejemplo, se introducen sobrecargas de comunicación innecesarias al descargar todo un conjunto de datos que, en el peor de los casos, pudiera no contener información relevante. Además, se generan sobrecargas de procesamiento donde, al poseer recursos heterogéneos, no todos los dispositivos pueden ejecutar procedimientos exhaustivos de búsqueda. En este contexto, surge *Searchable Encryption* (SE), una técnica criptográfica que permite realizar búsquedas sobre datos cifrados. Su objetivo es mantener la confidencialidad de los datos mientras el proveedor del servicio de almacenamiento es capaz de preservar sus capacidades de búsqueda [5]. SE ha sido implementado mediante tres enfoques principales, siendo el cifrado basado en atributos con capacidades de búsqueda (ABSE) el más adecuado para escenarios de almacenamiento y compartición de datos cifrados.

ABSE se apoya en la creación de un índice seguro que contiene palabras clave representativas del contenido o características de los datos sensibles y, a partir del cual, posteriormente se realizan las búsquedas [1]. Al ser un enfoque basado en atributos, ABSE opera de forma muy similar a ABE: se emplean políticas para establecer reglas de acceso y atributos para describir a los usuarios y, por ende, sus restricciones de acceso. Tras identificar las palabras clave que describen el contenido de los datos, éstas se cifran una sola vez mediante una política de acceso, definida sobre un conjunto de atributos y, a partir de ellas, se construye el índice seguro. Una vez creado éste, tanto los datos sensibles como su correspondiente llave de sesión son cifrados, produciendo el sobre digital que habrá de enviarse en conjunto con el índice seguro para su almacenamiento en la nube [9]. Cabe resaltar que la política de acceso utilizada para cifrar las palabras clave puede ser la misma o una política diferente a la usada para cifrar la llave de sesión, dependiendo de las necesidades de acceso que caractericen a los datos. Este proceso se ilustra en la Figura 6a.

Para realizar las búsquedas, el índice seguro es consultado por el proveedor del servicio de almacenamiento dado un *token* cifrado, denominado trampilla de búsqueda o *trapdoor*, creado por el usuario que solicita una búsqueda. Al igual que en ABE, cada usuario posee una llave secreta que se genera con base en el conjunto de atributos que lo caracterizan. De este modo, solo los usuarios que poseen el conjunto de atributos adecuados (dada una política) pueden buscar y recuperar datos de interés [1], [13]. Derivado de una necesidad de información, a partir de la clave secreta de usuario se genera una representación cifrada de la consulta del usuario, la cual permite realizar la búsqueda en el índice seguro [20]. Por ello, es importante señalar que el proveedor de servicio no es capaz de



(a) Creación del índice seguro y cifrado.



(b) Creación del *trapdoor*, búsqueda y descifrado.

Figura 6: Flujo de operaciones en la creación y uso de sobres digitales con capacidades de búsqueda.

derivar información, saber qué está buscando o el contenido de los resultados que encuentra dado un trapdoor en particular.

Si los atributos del potencial usuario de los datos satisfacen la política de cifrado y si existen resultados para su consulta, se retornan los sobres digitales correspondientes. Finalmente, el usuario podrá descifrar la llave de sesión mediante su clave secreta y los datos por medio de dicha llave de sesión, tal como se muestra en la Figura 6b [9], [20]. De esta forma se garantiza confidencialidad y control de acceso, así como la capacidad de compartir datos de forma segura con múltiples usuarios al incorporar ABE en el contexto de los sobres digitales.

Además, se preservan las capacidades de búsqueda y recuperación de información del proveedor del servicio de almacenamiento al hacer uso de esta técnica en conjunto con el cifrado con capacidades de búsqueda.

5. Eficiencia y Seguridad de Sobres Digitales

Crear y abrir SDBs requiere suficiente poder de cómputo. La complejidad en tiempo de ejecución y demanda de recursos de cómputo está asociada a los algoritmos criptográficos para cifrar los datos con el cifrador simétrico y para proteger la llave de sesión de dicho cifrador simétrico mediante el cifrado basado en atributos. Esta complejidad queda determinada por:

- El tamaño de los datos a cifrar, que impacta directamente en la complejidad en tiempo para el cifrador simétrico;
- El nivel de seguridad, que impacta directamente en el número de operaciones y la longitud de los operandos del cifrado basado en atributos.

El problema de eficiencia en SDBs se aborda a través del uso de patrones de paralelismo. El problema de la seguridad en SDBs se aborda mediante construcciones basadas en emparejamientos asimétricos. Se consideran dos esquemas de paralelismo: *i*) *pipeline*; y *ii*) *overlapped*. En el esquema llamado *pipeline* se despliegan dos patrones de paralelismo diferentes: el patrón *pipe & filters* en combinación con el patrón conocido como *manejador/trabajador*. De esta manera, el patrón *pipe & filters* se encarga de organizar el sistema en tuberías y el patrón manejador/trabajador se encarga de desplegar dichas tuberías como trabajadores ejecutados en paralelo. Este esquema se encarga de cifrar y descifrar

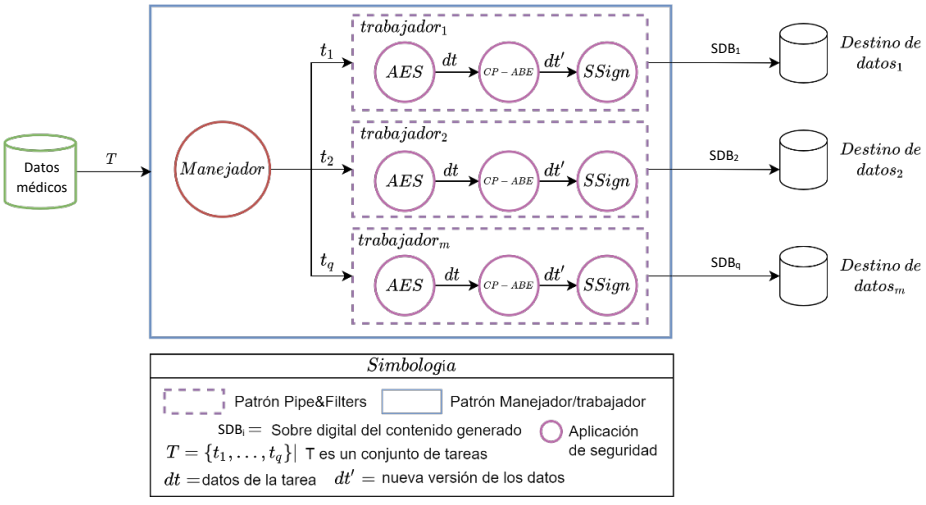


Figura 7: Representación conceptual de un esquema *pipeline*.

conjuntos pequeños de datos dividiendo las tareas entre el número de trabajadores disponibles desplegados en el sistema, lo cual permite reducir el tiempo de ejecución de los procesos de cifrado, los cuales dependen del tamaño de los datos, más que del nivel de seguridad.

La Figura 7 muestra la representación conceptual del ejemplo de un esquema *pipeline*, el cual cuenta con un patrón *pipe & filters* que incluye las aplicaciones de AES (como cifrador simétrico), CP-ABE (como cifrador basado en atributos) y SSign (para firma digital, basado en identidad). En este esquema, cada tubería es clonada en q trabajadores para eficientizar el procesamiento de los datos mediante la distribución de tareas a través de un trabajador dedicado. Una vez que los datos, expresados como archivos t_i en un repositorio de entrada, son procesado a través de toda la tubería por un trabajador dedicado, éstos se encapsulan para dar origen al SDB_i correspondiente, que puede ser enviado a su destino.

Por otro lado, el esquema *overlapped* permite el acoplamiento de sistemas independientes para que se ejecuten de forma suprapuesta (mediante el patrón *fork/join*), y el acoplamiento en forma de tubería para aquellos sistemas que cuenten con algún tipo de dependencia. Este esquema permite que los procesos asociados con la creación de SDBs se ejecuten en forma de una tubería y se gestionen como si fueran trabajadores en un patrón manejador/trabajador, permitiendo que la ejecución de tareas se realice de forma paralela. El esquema *overlapped* fue diseñado para cifrar y descifrar grandes conjuntos de datos de forma paralela.

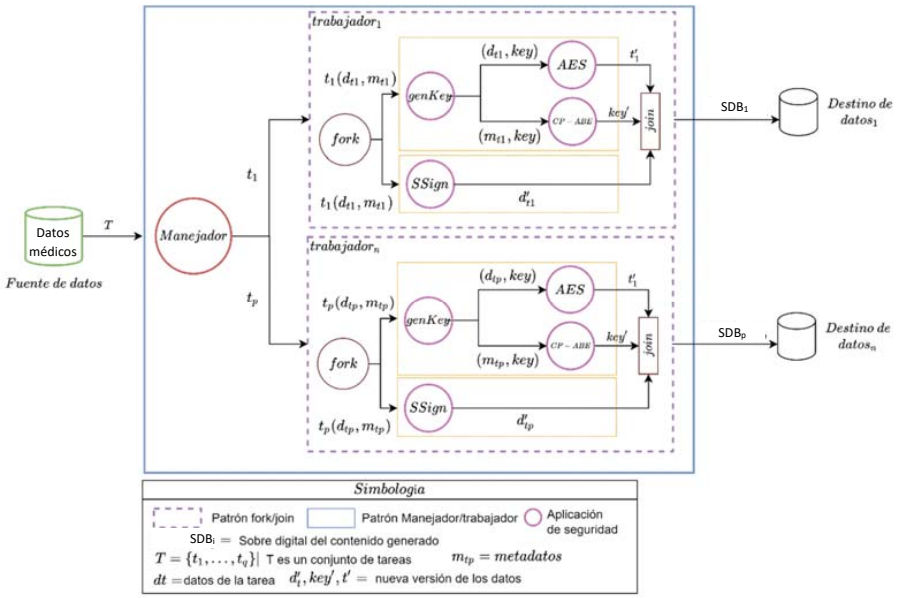


Figura 8: Representación conceptual de un esquema *overlapped*.

La Figura 8 muestra la representación conceptual de un ejemplo de un esquema *overlapped*, en donde un conjunto de tareas es extraído desde una fuente de datos y, posteriormente, éstas son distribuidas a los trabajadores a través de un manejador. Cada trabajador contiene un patrón *fork/join*, el cual permite la ejecución de tareas de forma suprapuesta (en este caso, la ejecución de la generación de llaves al mismo tiempo que la ejecución de la firma digital de los contenidos). Esto permite ejecutar dos tuberías al mismo tiempo y, una vez que la ejecución de ambas tuberías ha terminado, el contenido procesado es integrado en un sobre digital y, posteriormente, enviado a un destino (ya sea para su consumo, compartición o procesamiento en un entorno distinto).

6. Sobres Digitales para Ciencia de Datos

Hasta nuestro conocimiento, los SDBs han sido explorados y propuestos como tal, por primera vez, en el grupo de investigación del Cinvestav Unidad Tamaulipas [9]. De igual forma, y en el marco del programa de apoyo a la investigación en salud PRONACES - Salud, PRONAII Ciencia de Datos en Salud, en el proyecto *"Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud"* [15], se han implementado los sobres digitales para garantizar la seguridad, de extremo a extremo, de los datos médicos durante su ciclo de vida, que incluye:

1. **Creación**: Los datos se originan en un dispositivo médico (como radiografías, por ejemplo) o sistema de información (expediente clínico electrónico).
2. **Almacenamiento**: Una vez creados, los datos se encapsulan en SDBs y se almacenan en repositorios, locales o externos. En el caso de usar medios de almacenamiento externo, la comunicación desde el origen al destino se realiza comúnmente por un medio público o inseguro. Las propiedades de seguridad inherentes de los SDBs permiten emplear, incluso, canales de comunicación inseguros.
3. **Uso**: Los datos que ya se encuentran en un repositorio pueden ser consumidos o consultados por usuarios autorizados mediante técnicas de búsqueda, recuperación y acceso a SDB. Las características inherentes a los SDBs permiten realizar estas operaciones que habilitan a los usuarios autorizados (médicos, especialistas, profesionales de la salud, u otros dispositivos o sistemas) acceder a los datos de manera segura.

La Figura 9 describe de manera gráfica el ciclo de vida descrito previamente. Los datos se crean en el ámbito de una organización A (hospital, unidad médica familiar, consultorio, laboratorio) y es ahí donde se aseguran mediante la creación del SDB. Más adelante, desde el repositorio donde se encuentre dicho SDB, los usuarios (especialistas de la salud, sistemas, dispositivos) pueden acceder a él y únicamente aquellos con los atributos necesarios podrán abrir el SDB y acceder a los datos en claro. Durante la creación y apertura de SDB,

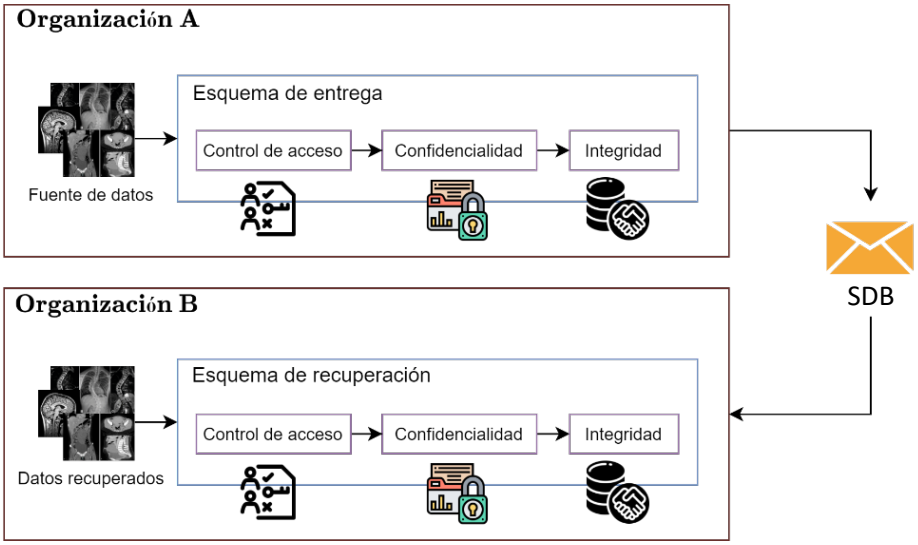


Figura 9: Creación y acceso (apertura) de sobres digitales buscables (SDB) en salud.

los esquemas de paralelismo permiten la viabilidad de implantar este concepto, puesto que los datos médicos, como las tomografías, generalmente ocupan una cantidad considerable de almacenamiento y, bajo un escenario de *big data*, la complejidad en tiempo incrementa considerablemente. El transporte seguro de los datos sensibles médicos, posible mediante los SDB, permite la distribución de datos sensibles de forma segura, lo cual es un requerimiento en las normas oficiales para tratamiento de datos médicos. Los usuarios finales de los datos, que los recuperan mediante operaciones de descifrado, podrán usarlos en los procesos correspondientes para su análisis y obtención de conocimiento útil mediante técnicas de ciencia de datos.

7. Retos y Perspectivas para Proteger Datos en el Sector Salud

Como es sabido, los datos sensibles generalmente demandan servicios de seguridad y de privacidad. Éste es un requerimiento impuesto, incluso, por regulaciones y leyes que varían en cada país. En México, la norma oficial NOM-024-SSA3-2012 establece los objetivos funcionales y funcionalidades que deberían observar los productos de Sistemas de Expediente Clínico Electrónico para garantizar la interoperabilidad, procesamiento, interpretación, confidencialidad, seguridad, así como uso de estándares y catálogos de la información de los registros electrónicos en salud. En este sentido, los SDBs son mecanismos que permiten lograr objetivos en materia de seguridad de datos y que permiten dar cumplimiento a los requerimientos de seguridad de éstos. Específicamente, la confidencialidad y

control de acceso son esenciales en el manejo, acceso e intercambio de datos en salud, que se consiguen con el uso de SDB mediante las técnicas criptográficas que incorporan.

El cifrado usado en los SDBs es de dos tipos: simétrico y basado en atributos. Ambos enfoques actualmente son suficientemente seguros. El cifrado simétrico sustenta su seguridad en que el trabajo de un atacante para vulnerarlo es exponencial respecto al tamaño de las llaves usadas. En la actualidad, se considera que una llave de 128 bits es imposible de atacar (en la práctica), dado que el costo computacional para lograrlo es de 2^{128} operaciones. Por lo anterior, a una computadora convencional le tomaría centenas de años realizar este trabajo. En el caso del cifrado basado en atributos, las llaves utilizadas son de un tamaño de al menos de 3×256 bits. En este caso, el costo para un atacante depende de la dificultad para resolver un problema matemático con un costo asociado de 2^{256} operaciones en una computadora convencional, pero la llave involucra al menos tres componentes de 256 bits.

Sin embargo, desde hace algunos años se vienen teniendo avances significativos en cómputo cuántico. Incluso, algunas empresas como Google han declarado haber alcanzado ya la supremacía cuántica, esto es, haber logrado diseñar una computadora que resuelve problemas que una computadora convencional no habría podido resolver. La computación cuántica es ahora la mayor amenaza a las soluciones de seguridad de datos basadas en algoritmos criptográficos, como lo son los SDBs. Aunque por ahora no existe una computadora cuántica con el suficiente poder de cómputo para atacar a los sistemas de cifrado como los que se usan en un SDB, se estima que en un futuro cercano, en 2030 según expertos [8], se cuente con dicha capacidad. El impacto en la seguridad del cifrado simétrico no será tan alto como lo será para el cifrado basado en atributos y para otros tipos de cifrado que basan su seguridad en la dificultad de resolver problemas matemáticos, ya que el poder de cómputo cuántico podrá resolver dichos problemas. Por ello, los mecanismos de seguridad que se apoyan en el cifrado, como los SDB, deben fundamentar su seguridad en una criptografía postcuántica [3], misma que ya se viene desarrollando desde 2014.

Existe un ataque llamado *harvest now, decrypt later*, que en español literalmente se traduce como *colecta datos cifrados ahora, descífralos después*. Esto es, los datos cifrados, al ser ilegibles, aunque estén disponibles para un atacante, no tienen ninguna utilidad para dicho atacante. Pero el atacante podría recolectarlos ahora y, cuando se tenga ya una computadora cuántica con suficiente poder computacional, descifrar esos datos recolectados. Por ejemplo, en la Figura 9, si un atacante (como el proveedor del servicio de almacenamiento) hace un respaldo de todos los SDBs, más adelante, con la ayuda de una computadora cuántica, éste podría abrir todos esos SDBs y, por lo tanto, tener acceso a los datos que, para ese entonces, pudieran resultar de alguna utilidad. Por ejemplo, los SDBs asociados a altos mandos militares o políticos de hoy podrían revelarse en un futuro cercano (8 años aproximadamente). Cabe destacar que en el ámbito de la salud se recomienda conservar los datos médicos de una persona (i.e., historia clínica) por al menos 5 años y hasta 15 años, incluso, después de su muerte.

Por tanto, en este capítulo los autores afirmamos que uno de los principales retos para la seguridad de los datos en el sector salud es contar con mecanismos de seguridad robustos, no solo bajo los modelos de ataque actuales, sino también para aquellos modelos de ataque que se vislumbran en un futuro no muy lejano. Por otro lado, si bien los SDBs son ahora eficientes y seguros, solamente cubren los servicios de confidencialidad, integridad y control de acceso. Sin embargo, es necesario tener en cuenta que existen otros requerimientos de seguridad en salud, tales como la trazabilidad. Es muy deseable explorar el desarrollo de métodos efectivos que pudieran garantizar estos servicios, como puede ser la incorporación adecuada de tecnologías disruptivas como *Blockchain* [16].

8. Conclusiones

En este capítulo hemos introducido el concepto de sobres digitales con capacidades de búsqueda (SDB). Se trata de una abstracción que permite garantizar dos servicios de seguridad principales: la confidencialidad y el control de acceso a datos sensibles. Por ello, los SDBs son idóneos para proteger la seguridad y privacidad de datos médicos. Al estar basados en dos capas criptográficas, una fundamentada en el cifrado simétrico (rápido para cifrado) y otra en el cifrado basado en atributos (efectivo para la distribución de llaves y el control de acceso criptográfico), los SDBs integran controles de acceso de grano fino aplicables, incluso, a grandes colecciones de datos, como ocurre en el ámbito del *big data* en salud.

La fortaleza de los SDBs está probada y recae en la seguridad de ambas capas de cifrado. Su eficiencia recae en la efectividad de los patrones de paralelismo que se usan en el despliegue de los SDB, bajo la premisa de que existen recursos de cómputo disponibles (para explotar el paralelismo de datos y de tareas). La alta seguridad y eficiencia de los SDBs los hacen viables para proveer los requerimientos de confidencialidad y de control de acceso que demanda el tratamiento de datos sensibles médicos, tal como lo exige la norma mexicana NOM-024-SSA3-2012. Ante un desarrollo continuo de capacidades de un computador cuántico, los esquemas de seguridad de datos basados en cifrado se ven amenazados en el corto plazo. El trabajo futuro se está enfocando en analizar y desarrollar metodologías eficientes para incorporar cifrado postcuántico en el diseño de los SDBs.

Agradecimientos

Este trabajo forma parte del Proyecto No. 41756 CONACYT - PRONAII Ciencia de Datos en Salud *"Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud"*, financiado por FORDECYT-PRONACES.

Referencias

[1] Aubrey Alston. *Attribute-Based Encryption for Attribute-based Authentication, Authorization, Storage, and Transmission in Distributed Storage Systems*. Inf. téc. arXiv:1705.06002v1. Cornell University, 2017. DOI: 10.48550/arXiv.1705.06002.

[2] Elaine Barker. *Recommendation for Key Management. Part 1: General*. Inf. téc. National Institute of Standards and Technology, 2020. DOI: 10.6028/NIST.SP.800-57pt1r5.

[3] Johannes Buchmann, Kristin Lauter y Michele Mosca. "Postquantum cryptography—state of the art". En: *IEEE Security & Privacy* 15.4 (2017), págs. 12-13. DOI: 10.1109/MSP.2017.3151326.

[4] Diana Elizabeth Carrizales-Espinoza, José Luis González-Compeán y Miguel Morales-Sandoval. "Zamna: a tool for the secure and reliable storage, sharing, and usage of large data sets in data science applications". En: *2022 IEEE Mexican International Conference on Computer Science (ENC)*. IEEE, 2022. ISBN: 978-1-6654-7347-7. DOI: 10.1109/ENC56672.2022.9882938.

[5] Yunling Wang, Jianfeng Wang, Xiaofeng Chen. "Secure Searchable Encryption: A Survey". En: *Communications and Information Networks* Vol. 1. No. 4 (2016), págs. 52-65. DOI: 10.11959/j.issn.2096-1081.2016.043.

[6] Cisco Networking Academy. *Introduction to Cybersecurity*. Inf. téc. Cisco Systems, Inc., 2016.

[7] Cisco Networking Academy. *Cybersecurity Essentials*. Inf. téc. Cisco Systems, Inc., 2017.

[8] Vikas Hassija et al. "Present landscape of quantum computing". En: *IET Quantum Communication* 1.2 (2020), págs. 42-48. DOI: 10.1049/iet-qtc.2020.0027.

[9] Melissa Brigitthe Hinojosa-Cabello. "An Attribute-Based Encryption Scheme for Storage, Sharing and Retrieval of Digital Documents in the Cloud". Tesis de mtría. Cinvestav, 2020.

[10] Richard Kuhn et al. *Introduction to Public Key Technology and the Federal PKI Infrastructure*. Inf. téc. National Institute of Standards and Technology, 2001.

[11] Badrinarayanan Lakshmiraghavan. *Pro ASP.NET Web API Security. Securing ASP.NET Web API*. Ed. por Apress Media, LLC. 1.ª ed. Springer, 2013. 416 págs. ISBN: 978-1-4302-5782-0. DOI: 10.1007/978-1-4302-5783-7.

[12] Elaine Barker, William Barker, Annabelle Lee. *Guideline for Implementing Cryptography in the Federal Government*. Inf. téc. National Institute of Standards and Technology, 2005.

[13] Antonis Michalas. "The Lord of the Shares: Combining Attribute-Based Encryption and Searchable Encryption for Flexible Data Sharing". En: *SAC '19: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. Association for Computing Machinery, 2018. ISBN: 978-1-4503-5933-7. DOI: 10.1145/3297280.3297297.

[14] Praveen Kumar Premkamal, Syam Kumar Pasupuleti y Pja Alphonse. "Attribute Based Encryption in Cloud Computing: A Survey, Gap Analysis, and Future Directions". En: *Network and Computer Applications* 108 (2018), págs. 37-52. DOI: 10.1016/j.jnca.2018.02.009.

[15] Conacyt PRONACES. *Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud. PRONACES Salud, FORDECYT 2019-06 CONACyT, proyecto número 41756.* http://adaptivez.org.mx/e-SaludData/. 2022.

[16] Nabil Rifi et al. "Towards using blockchain technology for eHealth data access management". En: *2017 fourth international conference on advances in biomedical engineering (ICABME)*. IEEE, 2017. ISBN: 978-1-5386-1642-0. DOI: 10.1109/ICABME.2017.8167555.

[17] Amit Sahai y Brent Waters. "Fuzzy Identity-Based Encryption". En: *Advances in Cryptology – EUROCRYPT 2005*. Springer Berlin Heidelberg, 2005. ISBN: 978-3-540-32055-5. DOI: 10.1007/11426639_27.

[18] Douglas Selent. "Advanced encryption standard". En: *Rivier Academic Journal* 6.2 (2010), págs. 1-14.

[19] Víctor Jesús Sosa-Sosa et al. "Protecting Data in the Cloud: An Assessment of Practical Digital Envelopes from Attribute based Encryption". En: *KDCloudApps 2017*. SciTePress, 2017. ISBN: 978-989-758-255-4. DOI: 10.5220/0006484603820390.

[20] Hui Bin Yin et al. "CP-ABSE: A Ciphertext-Policy Attribute-Based Searchable Encryption Scheme". En: *IEEE Access* 7 (2019), págs. 5682-5694. DOI: 10.1109/ACCESS.2018.2889754.

[21] Rui Zhang, Rui Xue y Ling Liu. "Searchable encryption for healthcare clouds: a survey". En: *IEEE Transactions on Services Computing* 11.6 (2017), págs. 978-996. DOI: 10.1109/TSC.2017.2762296.

[22] Eduardo Palma Ávila. "Criptografía Basada en Hardware". En: *Revista Seguridad. Cultura de prevención para TI*. No. 21. Universidad Nacional Autónoma de México, jun. de 2014.

Ciencia de Datos en Salud: Minería de Procesos con Preservación de Privacidad de Datos Médicos

Heidy M. Marin-Castro[1], Héctor A. De la Fuente-Anaya[2], Miguel Morales-Sandoval[2], Ana B. Ríos-Alvarado[3], and Tania Y.Guerrero-Meléndez[3]

[1] Conacyt - Facultad de Ingeniería y Ciencias, Universidad Autónoma de Tamaulipas, México
hmarisol@docentes.uat.edu.mx
[2] Cinvestav Unidad Tamaulipas, Cd. Victoria, Tamps, México
{hector.delafuente,miguel.morales}@cinvestav.mx
[3] Facultad de Ingeniería y Ciencias, Universidad Autónoma de Tamaulipas, México
{arios,tyguerre}@docentes.uat.edu.mx

Resumen Diariamente, los sistemas de información generan una gran cantidad de registros de eventos a partir de la ejecución de procesos. Esta situación ha incentivado gran interés por parte de las organizaciones por realizar análisis de estos datos a fin de generar conocimiento que apoye en la toma de decisiones. La Minería de Procesos, en combinación con la Preservación de Privacidad de Datos, es una disciplina orientada a descubrir, monitorear y mejorar el desempeño de los procesos de negocio, garantizando preservar la confidencialidad de los datos sensibles o personales de eventos producidos por la ejecución de un proceso a partir de un sistema de información. La gran mayoria de los procesos clínicos relacionados con el diagnóstico, tratamiento y organización de personal de la salud y de los pacientes requieren de algoritmos que puedan realizar análisis y estudio de sus datos y procesos asegurando la privacidad y confidencialidad de estos mientras son usados. En este capítulo se describen algunas de las características más relevantes de los procesos en el dominio de salud, así como la importancia de la Minería de Procesos como un servicio, y se presenta una estrategia de confidencialidad para proveer privacidad en los datos de las bitácoras de eventos relacionados con un proceso clínico. Una de las ventajas de la estrategia propuesta es que permite mantener la utilidad de los datos para las tareas de Minería de Procesos sin que exista alguna perdida de información o la posibilidad de revelar a terceros información que se considera confidencial.

Palabras clave: Minería de Proceso · Bitácora de Eventos · Modelo de Proceso · Confidencialidad · Cifrado · Utilidad de Datos

1. Introducción

En nuestra sociedad actual, el tema de protección y privacidad de datos ha ganado mucha atención en los últimos años debido a los frecuentes ataques

cibernéticos o filtraciones de datos que comúnmente se presentan contra los sistemas de información y las regulaciones relacionadas con Reglamento General de Protección de Datos (GDPR) de Europa [7]. En general, la privacidad puede ser descrita como el derecho de las personas a controlar cómo se recopilan, utilizan y/o divulgan sus datos personales a otros individuos, organizaciones o gobiernos [21]. La privacidad de los datos se ha vuelto mucho más crítica, especialmente para aquellas empresas que trabajan con datos sensibles, como las organizaciones en el sector salud. Los proveedores de atención médica deben asegurarse de administrar adecuadamente los datos de los pacientes para crear una cultura de confianza y transparencia mientras cumplen con las estrictas normas legales y de privacidad de datos. La privacidad de los datos en el cuidado de la salud está en constante evolución, con leyes y regulaciones continuamente actualizadas. De esta forma, los pacientes obtienen la privacidad de datos que esperan. Mientras que ha habido mucha investigación sobre lo que constituye la privacidad de datos y su importancia en los sistemas de información en el sector salud [14], [3], [13], [18], existe una clara brecha en la investigación sobre privacidad en el campo de la Minería de Procesos (MP), la cual se centra en el estudio de los procesos de negocio descritos como un conjunto de actividades interrelacionadas y desempeñadas por un grupo de participantes para lograr un objetivo de negocio. Esta disciplina conecta técnicas de la Ciencia de Datos y de la Ciencia de Procesos para llevar a cabo tareas de descubrimiento, verificación de la conformidad y mejora de los procesos de negocio a partir de extraer conocimiento de colecciones de eventos llamadas bitácoras de eventos [1].

La MP ha evolucionado a través de los años. Al inicio, los algoritmos y herramientas de MP fueron desarrolladas por grupos de investigación [2] y paulatinamente han sido utilizados por la industria a través del análisis y estudio de casos y proyectos. Actualmente, MP se ha convertido formalmente en una disciplina ante la IEEE en el 2011 [1], cuyo objetivo principal es el de contribuir a mejorar el desempeño de los procesos de las organizaciones, a fin de descubrir su verdadero comportamiento, proporcionando información de lo que se está realizando bien, diagnosticando problemas y sugiriendo automáticamente acciones o medidas correctivas o de mejora del proceso. Para ello, se apoya en dos de los elementos medulares que sirven de entrada a la mayoría de los algoritmos de MP: las bitácoras de eventos y los modelos de proceso. Por un lado, las bitácoras de eventos son creadas a partir de la ejecución de los procesos de negocio disponibles a través de sistemas de información de las organizaciones. Cada *evento* en la bitácora corresponde a una actividad o tarea que forma parte del proceso de negocio realizada por algún participante del proceso, y un conjunto de eventos conforma un *caso* o *instancia*. Por ejemplo, en la Tabla 1 se muestran tres casos y siete eventos relacionados con las actividades de un proceso en el dominio de salud. Los eventos están conformados por diversos atributos: una estampa de tiempo (periodo de tiempo en que se ejecutó dicho evento), un nombre de la actividad ejecutada, un costo (representa el costo tomado por la actividad asociada al evento), un recurso o participante (rol o nombre de la persona que llevó a cabo dicho evento), y nombre del paciente. Cada instancia corresponde

a una *traza*, por ejemplo, la secuencia de actividades {*Registro, Triaje, Examen de Sangre*} en la Tabla 1 representa una traza sobre la atención a un paciente, en este caso el paciente de nombre Brenda.

Tabla 1: Ejemplo del segmento de una bitácora de eventos del dominio de salud.

Id Caso	**Estampa de Tiempo**	**Actividad**	**Costo**	**Recurso**	**Paciente**
1	01-01-2018 15:20:15	Registro	100	Pedro	Brenda
1	01-01-2018 15:22:02	Triaje	50	Ana	Brenda
1	01-01-2018 15:25:43	Ex. sangre	800	Julio	Brenda
2	01-01-2018 15:43:08	Registro	100	Jorge	Isidro
2	01-01-2018 15:43:50	Rayos X	500	Pedro	Isidro
3	01-01-2018 15:46:27	Registro	100	Pedro	Marta
3	01-01-2018 15:48:14	Triaje	50	Ana	Marta

Por otro lado, un modelo de proceso puede verse como la representación abstracta y gráfica de las actividades llevadas a cabo por un proceso de una organización. Existen diversos lenguajes de notación de los modelos de procesos, siendo el estándar de Modelado y Notación de Procesos de Negocio (BPMN) [8] uno de lenguajes más usados. Un ejemplo sencillo de un modelo de proceso BPMN en el área de salud sobre la atención médica de un paciente puede verse en la Figura 1. El proceso comienza con la actividad *Registrar al paciente*, seguido por las actividades paralelas *Registro de signos vitales* y *Recolectar síntomas*, seguida por *Triaje*, es decir, actividades que se ejecutan en cualquier orden. Después, se realiza la actividad de *Actualización del expediente médico* y se crean dos caminos alternativos con las actividades *Hacer prueba de sangre* y *Realizar análisis de hemoglobina*, donde alguna de ellas es ejecutada. Los dos símbolos en forma de diamante con el signo '+' adentro denotan compuertas paralelas. El primero corresponde a una compuerta paralela de división comenzando con dos ramas concurrentes y el segundo es la compuerta de la unión. Los dos símbolos en forma de diamante con el signo 'X' adentro denotan compuertas exclusivas, es decir, se tienen caminos alternativos. En el modelo de proceso, el evento inicial (mostrado como un círculo rojo) es activado por el paciente a partir de su llegada y termina con un evento final (mostrado por un círculo blanco). Durante la ejecución del proceso en salud pueden intervenir diferentes actores, como personal de salud (médicos, enfermeras, asistentes, etc.) y pacientes, realizando una o varias actividades del proceso.

1.1. Requerimiento de confidencialidad

Tanto los modelos de procesos como las bitácoras de eventos son una valiosa fuente de información que permite identificar el comportamiento real de un proceso y no únicamente una vista idealizada. Sin embargo, en un dominio de estudio como el de salud, el trabajar con bitácoras con datos de eventos sensibles,

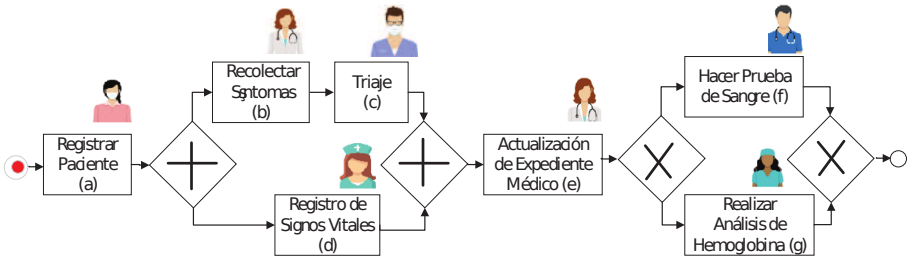


Figura 1: Modelo de proceso BPMN del seguimiento de pacientes.

así como la falta de confianza entre los participantes que ejecutan un proceso, inspira a la MP a desarrollar y utilizar métodos seguros de los datos de eventos que garanticen la confidencialidad de éstos. Sin embargo, las investigaciones que se han reportado en la literatura sobre tareas de MP pocas veces consideran cuestiones de confidencialidad de los datos, siendo éste un problema de gran relevancia, principalmente en dominios donde es necesario garantizar seguridad y privacidad de los datos.

En MP, la confidencialidad de las bitácoras de eventos no se puede lograr únicamente cifrando los datos; es necesario contar con mecanismos de seguridad y privacidad adecuados, de forma que las tareas de descubrimiento, validación de conformidad y mejora de procesos no sean afectadas y puedan seguir realizándose. Uno de los problemas comunes al intentar únicamente de anonimizar algunos atributos o registros de eventos es la posibilidad de vincular los eventos a un caso especifico e identificar atributos iguales. Esta situación de vinculación puede ocasionar que se descubra la identidad del registro. Por ejemplo, suponiendo que se tiene una bitácora de eventos de pacientes de un hospital con algunos atributos pseudo anonimizados (nombre del paciente, actividad, empleo) y un atacante desea conocer la enfermedad de cierto paciente del cual únicamente conoce su edad y algunas de las fechas en las que visitó el hospital en un período de tiempo determinado. Al identificar esta información en la bitácora de eventos puede inferir los eventos correspondientes de este paciente y, así, poder conocer su enfermedad. En este escenario se presenta un problema de ataque a la privacidad en los datos de los registros de eventos a partir de la vinculación de información.

Muchas organizaciones en salud son conscientes de la necesidad de gestionar y mejorar sus procesos, los cuales constantemente están evolucionando y cambiando dinámicamente. Algunos de los trabajos de investigación reportados en la literatura [17], [4], [11], [20], [16], [12] sobre la MP en salud, han identificado que aún no se cuenta con soluciones prácticas capaces de adaptarse a los diferentes entornos de los procesos en salud, los cuales carecen de mecanismos que garanticen la seguridad de las bitácoras de eventos construidas. En este sentido, resulta importante desarrollar y mejorar los algoritmos de MP para trabajar con datos sensibles de acuerdo con las normas internas o regulaciones externas establecidas

por instituciones médicas sin que exista algún riesgo de confidencialidad en los datos.

1.2. Objetivo y organización de este documento

En este documento se describen algunas de las características más relevantes de los procesos de negocio en el sector salud y se presenta una Estrategia de Confidencialidad para la Bitácora de Eventos (ECBE) basada en métodos criptográficos (haciendo uso de cifrado determinista, así como de técnicas de anonimización) sin perder la utilidad de los datos para su uso en algoritmos de Minería de Procesos. Esta estrategia construye una bitácora de eventos segura capaz de ser usada en tareas como el descubrimiento de procesos. La bitácora de eventos cifrada puede ser compartida por el equipo multidisciplinar que ejecuta el proceso sin que se comprometa la privacidad de los datos y el cumplimiento de las regulaciones, debido a que los participantes del equipo no podrán deducir información que no tenga que ver con la bitácora o el modelo de proceso cifrado.

El resto de este capítulo se encuentra organizado de la siguiente forma: en la Sección 2 se presentan algunas de las características más relevantes de los procesos de negocio en el dominio de la salud; en la Sección 3 se describe un escenario de las tareas principales, uso y explotación de la MP como servicio; en la Sección 4 se presenta una descripción de la estrategia de confidencialidad de la bitácora de eventos propuesta; en la Sección 5 se muestran los detalles de la experimentación y resultados obtenidos; finalmente, en la Sección 6 se presentan las conclusiones y trabajo futuro.

2. Procesos de negocio en salud

Uno de los intereses latentes de las organizaciones y, particularmente, en el sector salud consiste en aplicar MP para conocer la trayectoria de diferentes pacientes desde el momento de su ingreso hasta su alta hospitalaria. Cada visita de un paciente a un hospital constituye una instancia de proceso y los eventos individuales de cada caso se pueden obtener a partir del Sistema de Información. Este último normalmente registra información sobre las actividades logísticas y de tratamiento realizadas para pacientes específicos y el personal del hospital que los realizó. Además, una parte de un proceso en salud puede ser las interacciones con otras instituciones asistenciales y la solicitud de documentación médica previa. Por lo tanto, algunos de estos datos pueden ser compartidos sobre los límites de la organización o entre otras organizaciones y, así, realizar MP con éstos.

Típicamente, los procesos en salud se caracterizan por presentar altos niveles de variación debido a una vasta diversidad de actividades que pueden llevarse a cabo de forma secuencial o paralela por distintos participantes del proceso. Normalmente, los algoritmos de MP en salud pueden ser apoyados por guías y protocolos médicos para dar una referencia del orden de las actividades a seguir dentro del proceso. Sin embargo, en ocasiones los procesos en salud deben

considerar situaciones extraordinarias o de emergencia no previstas, que no necesariamente corresponden con el orden establecido en el proceso, por lo que en la mayoría de los casos el flujo de ejecución de las actividades en el proceso no se cumple adecuada o completamente de acuerdo con el orden previsto.

Los procesos de salud son llevados a cabo por un equipo multidisciplinar (médicos, enfermeras, especialistas, asistentes, etc.), quienes, de manera autónoma e independiente, ejecutan una o varias actividades del proceso y toman decisiones sobre determinadas tareas complejas sin apegarse completamente al proceso clínico establecido y sin limitarse al acceso de la información sensible que pueden manejar. La sensibilidad de los datos es un tema de gran importancia que necesita constantemente tomarse en cuenta, ya que los datos podrían incluir información tal como la condición médica actual del paciente, sus comorbilidades, tratamientos e información personal que no debería ser revelada a ninguna entidad externa o interna que no cuente con un acceso o permiso necesario. En este escenario, las bitácoras de eventos obtenidas a partir de la ejecución de un proceso en salud deberían ser cuidadosamente manejadas debido a la confidencialidad, privacidad, uso y almacenamiento de la información que éstas contienen. Además, el riesgo de privacidad se incrementa con el uso de proveedores de servicios externos, como la nube, para delegarle el costo computacional y de almacenamiento asociado con el uso de algoritmos, como los de MP, y de almacenamiento de las bitácoras de eventos. Sin embargo, las técnicas clásicas de MP no están preparadas para lidiar con problemas relacionados con la confidencialidad de las bitácoras de eventos dentro de un ambiente de trabajo multidisciplinario, así como la externalización de las bitácoras de eventos para tareas de MP en la nube.

3. La Minería de Procesos como servicio

A partir de la MP, muchas organizaciones pueden identificar cuellos de botella, desviaciones, anticipar y diagnosticar problemas de rendimiento y cumplimiento mediante el uso de las bitácoras de eventos y los modelos de procesos. Particularmente, las técnicas MP pueden ser agrupadas en tres tareas principales:

1. El **descubrimiento de modelos de procesos** tiene por objetivo descubrir automáticamente el modelo del proceso asociado a los eventos almacenados en la bitácora de eventos obtenidos a partir de la ejecución de un sistema de información, es decir, construye el modelo de proceso tomando como entrada la bitácora de eventos;
2. La **verificación de conformidad** consiste en reproducir cada una de las instancias o trazas contenidas en la bitácora de eventos en el modelo de proceso para comprobar si lo que se tiene registrado en la bitácora corresponde con la ejecución y orden de las actividades que se muestran en el modelo de proceso y viceversa; a partir de esta tarea es posible identificar desviaciones

que pueden ocurrir en el proceso o en su correspondiente bitácora de eventos;

3. La **mejora del proceso** se enfoca en mejorar el desempeño de los procesos, ya sea cambiando o extendiendo el modelo previo construido.

Con el auge del Big Data, diariamente se generan grandes cantidades de datos de eventos a partir de la ejecución de procesos de distintos dominios de los sistemas de información de las organizaciones, lo que hace necesario contar con técnicas de MP disponibles como un servicio para delegarles tareas de descubrimiento, conformidad y mejora. En este escenario, los datos (bitácoras de eventos) dejan de estar bajo el control del propietario del proceso de negocio y pueden ser accedidos por el proveedor del servicio, lo que puede ocasionar un riesgo de confidencialidad.

La privacidad se centra en el uso y el manejo de los datos personales de los individuos, así como las políticas que garantizan que la información personal de los usuarios se recopile, comparta y utilice de manera correcta [19]. Uno de los enfoques más usados para garantizar la confidencialidad en la bitácora de eventos sin perder utilidad en los datos, es el cifrado, el cual transforma un texto legible en texto ilegible y viceversa, utilizando sus respectivas claves de cifrado y descifrado. A continuación se describe la estrategia de confidencialidad propuesta.

4. Estrategia de Confidencialidad de la Bitácora de Eventos

La Estrategia de Confidencialidad para la Bitácora de Eventos (ECBE) descrita en este trabajo está basada en métodos criptográficos y técnicas de anonimización. Esta estrategia asegura la confidencialidad en los datos de la Bitácora de Eventos, ya que estos se transforman a un formato ilegible que no puede ser interpretado por los usuarios. La confidencialidad de los datos permite que la bitácora sea compartida con un servidor externo, como la nube, para llevar a cabo tareas de MP sin que se pueda revelar información que no esté relacionada con la tarea de descubrimiento del proceso de negocio cifrado. ECBE se asegura de preservar la utilidad de los datos en la Bitácora de eventos mediante técnicas de cifrado determinista, es decir, siempre se genera el mismo texto cifrado para un texto plano y una clave dados, lo cual mantiene la diferenciación en los datos para poder descubrir el modelo del proceso. El escenario de trabajo de la estrategia ECBE se compone de dos entornos, privado y externo (ver Figura 2), donde la bitácora de eventos puede moverse de un entorno privado (seguro) a un entorno externo (inseguro) en su versión cifrada (BE'). Una vez que el algoritmo de descubrimiento de modelos de proceso recibe como entrada una BE', el modelo resultante de esta tarea es un modelo de procesos cifrado (MP') del cual no se puede extraer información del contexto del proceso de negocio en estudio. El objetivo del uso de los entornos en la ECBE es garantizar la protección y determinar los niveles de seguridad que se tendrán en los registros que se tienen

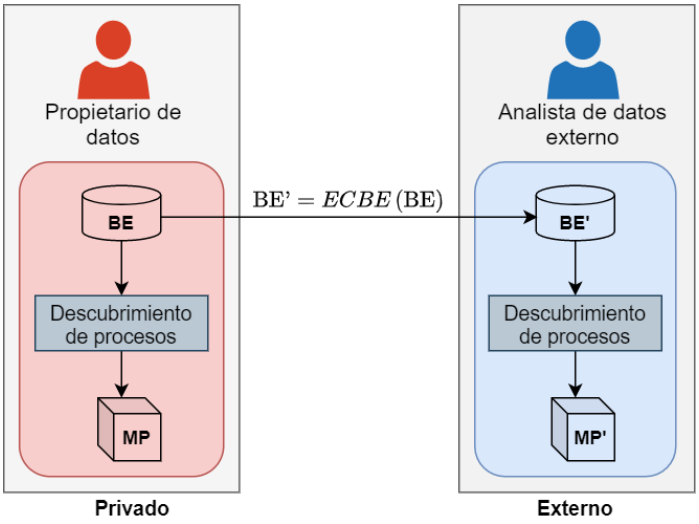


Figura 2: Entornos de seguridad considerados en ECBE.

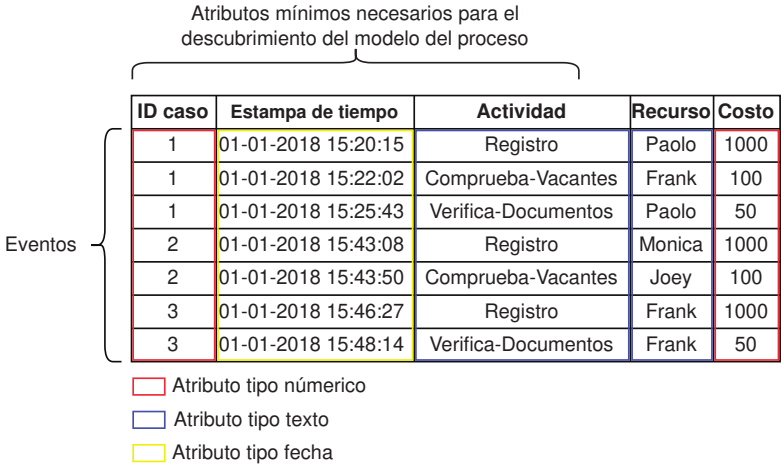
en las bitácoras. En el entorno externo se protege a la bitácora de eventos de analistas externos del proceso y de cualquier otro individuo que requiera hacer uso no permitido de la misma.

La bitácora de eventos cifrada (BE') evita cualquier riesgo de fuga de información que pueda ocurrir, aunque el atacante conozca el contexto en el que ocurre el proceso de negocio. ECBE asegura la bitácora de eventos considerando sus atributos, como puede verse en el ejemplo de la Figura 3, donde se presentan eventos que contienen atributos que almacenan valores del tipo numérico, texto o fecha. Comúnmente, la estrategia ECBE se enfoca en proteger los atributos más relevantes para la tarea de descubrimiento de procesos, como los que se presentan en la Figura 3 (estampa de tiempo, actividad, recurso, costo).

La estrategia de confidencialidad ECBE se compone de cuatro etapas (ver Figura 4), las cuales se enfocan en proteger los datos de los atributos de la bitácora de eventos e información que se puedan derivar de ésta. A continuación se describen cada una de las etapas.

Etapa 1. Filtrado y modificación de la entrada. Esta etapa consiste en preparar los datos relevantes para el análisis deseado. Incluye las siguientes actividades:

1. A partir de la bitácora de eventos L en formato de tabla bidimensional, se retiran los atributos que son irrelevantes al análisis deseado (datos no sensibles). L está conformado por una serie de trazas de la forma $(a_1, a_2, \cdots, a_n)$, donde a_i representa una actividad. Si la traza se repite k veces, se indica como $(a_1, a_2, \cdots, a_n)^k$.



ID caso	Estampa de tiempo	Actividad	Recurso	Costo
1	01-01-2018 15:20:15	Registro	Paolo	1000
1	01-01-2018 15:22:02	Comprueba-Vacantes	Frank	100
1	01-01-2018 15:25:43	Verifica-Documentos	Paolo	50
2	01-01-2018 15:43:08	Registro	Monica	1000
2	01-01-2018 15:43:50	Comprueba-Vacantes	Joey	100
3	01-01-2018 15:46:27	Registro	Frank	1000
3	01-01-2018 15:48:14	Verifica-Documentos	Frank	50

Figura 3: Atributos típicos de una BE.

2. Se realiza un filtrado de trazas removiendo de la bitácora de eventos aquellas trazas con menor frecuencia de acuerdo a un umbral θ. Los casos que formen la misma traza se agrupan, de esta manera se forma un conjunto de trazas únicas. Por ejemplo, sea L una bitácora de eventos con $L = \{(a,b,c,d,e,f)^{10}$,$(a,b,e)^{8}$, $(a,c,b,d,f,e)^{20}$,$(a,c,e,f)^{4}$, $(a,c,f,c,e)^{15}\}$, y $\theta = 10$. Después del filtrado se tendría $L' = \{(a,b,c,d,e,f)^{10}$, $(a,c,b,d,f,e\ )^{20}$, $(a,c,f,c,e)^{15}\}$.

Etapa 2. Cifrado del texto en claro. En esta etapa se proporciona confidencialidad sin pérdida de utilidad a los datos de la bitácora. Para mantener la capacidad de aplicar cálculos matemáticos básicos a los valores numéricos en la bitácora de eventos, se usa el algoritmo de cifrado homomórfico Paillier [15], mientras que para datos de tipo texto, se aplica un algoritmo de cifrado determinista AES [5]. Los atributos *Id Caso* y *Estampa de tiempo* no son cifrados en esta etapa; éstos son protegidos mediante técnicas que les ayudan a conservar su utilidad. La realización de esta etapa se ilustra en la transformación de los atributos "*Actividad*", "*Recurso*" y "*Costo*" de la bitácora de eventos en claro en la Tabla 1 y la bitácora de eventos cifrada (BE') en la Tabla 2.

Etapa 3. Conversión a tiempos relativos. En esta penúltima etapa, el atributo "*Estampa de tiempo*" es modificado a fin de evitar que los periodos de tiempo de la bitácora de eventos sean identificados. Para ello, se selecciona otra fecha (que se mantiene secreta junto a las claves de cifrado) para que todos los eventos sean relativos a ésta. La estampa de tiempo de cada evento en la bitácora de eventos se sustituye por su diferencia con la fecha secreta seleccionada. En las Tablas 1 y 2, en el atributo "*Estampa de tiempo*" se muestra la conversión

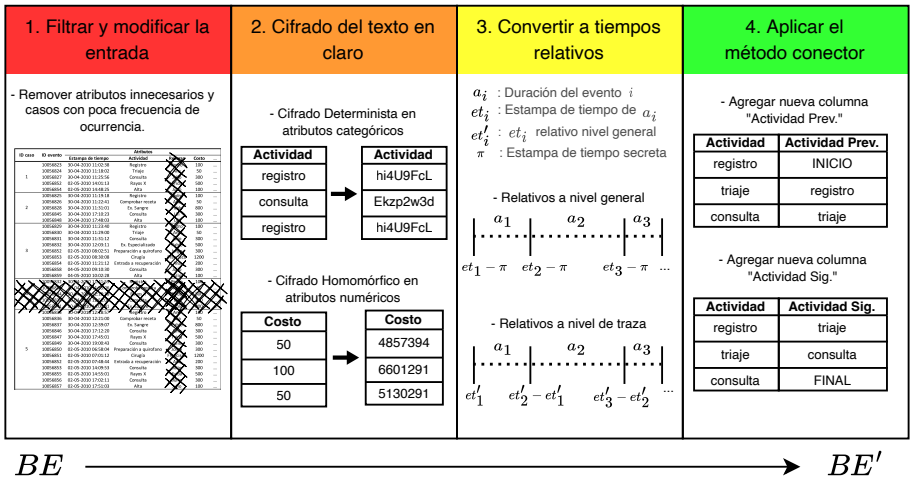


Figura 4: Etapas de la estrategia ECBE.

realizada.

Etapa 4. Aplicación de método conector. La última etapa consiste en aplicar un método que permite de forma segura obtener y extraer la estructura de un grafo dirigido a partir de la bitácora de eventos cifrada (BE'). Esta estructura sirve de entrada de muchos de los algoritmos de MP. En esta misma etapa se realiza la reconstrucción de la bitácora de eventos original una vez que se cuente con las claves de descifrado y los valores de fecha relativos. Las actividades consideradas en el método conector son las siguientes:

1. Se agrega el atributo llamado "*Actividad Prev.*", el cual indica la actividad previa a cada evento en la bitácora de eventos. Esto permite identificar cuáles actividades están directamente conectadas. En el caso del primer evento de cada traza, éste tendrá una actividad previa artificial "Start" cifrada con los mismos parámetros con los que fue cifrado el atributo "*Actividad*".
2. Se asegura la información contenida en el atributo "*Id Caso*" para no permitir el agrupamiento de eventos en trazas, manteniendo la posibilidad de recrear la bitácora de eventos original. Esto es realizado dando a cada evento un ID aleatorio ("*ID*") e indicando el ID previo ("*ID P.*"). Estas nuevas columnas en BE' permiten identificar el evento siguiente a nivel de traza. El "*ID P.*" de la actividad inicial en una traza siempre será 0.
3. Enseguida, los atributos "*ID*" e "*ID P.*" se ocultan, ya que son utilizados solo para reconstruir la bitácora de eventos. Estos se concatenan y se cifran usando el algoritmo de cifrado determinista AES (igual que en la Etapa 2) y se agrega como nuevo atributo llamado "*Conector*" en la bitácora de eventos.

Tabla 2: Bitácora de eventos de la Tabla 1 con ECBE aplicada.

Estampa de tiempo	Actividad	Actividad Prev.	Recurso	Costo	Conector
00-00-0000 00:00:42	y7y4PUi2	NM7Jgoum	XRCDyLgS	59301	5q8aL2at
00-00-0000 00:01:47	UGdnk8fh	y7y4PUi2	hLrq2mYD	46012	KQBindVr
00-00-0000 15:46:27	bvS(28op	UGdnk8fh	4hIDYn0q	98744	CKl07FSq
00-00-0000 00:01:47	y7y4PUi2	NM7Jgoum	tpwUTcAl	58430	N9a1qeto
00-00-0000 00:03:41	jhg!676	y7y4PUi2	XRCDyLgS	81023	XIQ7ZnqA
00-00-0000 15:20:15	y7y4PUi2	NM7Jgoum	XRCDyLgS	59015	M4qAwqqz
00-00-0000 15:43:08	UGdnk8fh	y7y4PUi2	hLrq2mYD	42110	z5Zb56jY

4. Posteriormente, se usa el atributo "*Estampa de tiempo*" para proteger los valores de tiempo en eventos que cuenten con el mismo "*Id Caso*". Éstos se hacen relativos con respecto a la estampa de tiempo precedida, a excepción del primer evento de cada traza, el cual se mantiene con el mismo valor. Esto permite el cálculo de la duración de cada uno de los arcos en un grafo dirigido de seguimiento directo (DFG), pero hace complicado identificar eventos basándose en la parte temporal en la que ocurrieron.
5. Para finalizar, se retira el atributo "*Id Caso*" y se desorganiza el orden de todas las filas y se obtiene como resultado una bitácora de eventos protegida para el entorno externo (BE') (ver Tabla 2), capaz de ser usada para descubrimiento de procesos basándose en las causalidades de los eventos haciendo uso de un grafo DFG (generado de "*Actividad*" y "*Actividad Prev.*").

5. Evaluación de la estrategia ECBE

La estrategia ECBE se implementó en el lenguaje de programación Java, considerado como un lenguaje de alto nivel que permite expresar los algoritmos en un nivel más abstracto del lenguaje máquina. ECBE se evaluó usando datos reales extraídos de sistemas ERP (Enterprise Resource Planning) de tres bitácoras de eventos del campo de salud (ver detalles en la Tabla 3) con una clave de cifrado AES de una longitud de 128 bits. Con estas bitácoras se comprobó que los modelos de proceso DFG resultantes de las bitácoras de eventos cifradas fueran iguales a los resultantes de sus versiones en claro. De esta manera se comprueba que no existe pérdida en la utilidad de los datos al protegerlos con la estrategia propuesta.

En la evaluación de la estrategia propuesta se consideró el correcto descubrimiento de los modelos procesos y se comprobó que la solución propuesta permite obtener modelos con una nula pérdida de precisión, lo que demuestra que la solución propuesta basada en una estrategia criptográfica no produce pérdida de utilidad en los datos de entrada para los algoritmos de descubrimiento de procesos.

Tabla 3: Bitácoras de eventos de datos médicos reales.

Nombre	No. eventos	No. actividades	No. trazas
Sepsis Cases [9]	15,214	16	1,050
BPIC11 [6]	150,291	624	1,143
Hospital Billing [10]	451,359	18	100,000

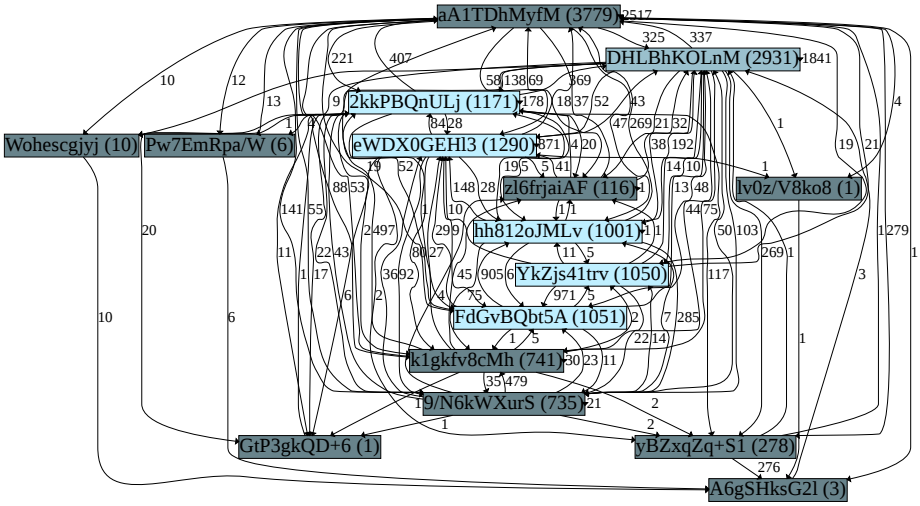


Figura 5: Modelo de proceso DFG de la bitácora de eventos cifrada Sepsis Cases resultante de ECBE.

La Figura 5 muestra el modelo descubierto usando la bitácora de eventos de Sepsis Cases cifrada usando la estrategia ECBE propuesta. Al comparar el modelo de proceso cifrado con el obtenido del modelo de proceso en su versión en claro, ambos modelos presentan la misma cantidad de aristas (137) y el mismo peso en cada una, lo que demuestra que no existe pérdida de utilidad al usar ECBE. Por lo tanto, el modelo obtenido a partir de la bitácora de eventos cifrada (BE') puede convertirse en el modelo que se obtendría usando la bitácora de eventos (en claro) y las llaves de descifrado correspondientes.

6. Conclusiones

Los procesos en salud representan un gran desafío debido a las diversas características que éstos presentan, como altos niveles de variación, diversidad en la ejecución y flujo de las actividades de los procesos, trabajo colaborativo de un equipo multidisciplinar (médicos, enfermeras, especialistas, asistentes, etc), así como el manejo de datos sensibles. Ésto hace necesario contar con algoritmos de MP capaces de trabajar con temas de confidencialidad, privacidad, uso y al-

macenamiento de la información, particularmente de las bitácoras de eventos y modelos de procesos. En este trabajo se describieron algunas de las características más relevantes de los procesos en salud y se presentó una estrategia llamada ECBE para garantizar la confidencialidad de la bitácora de eventos de un proceso de negocio en salud, sin pérdida de su utilidad en tareas de MP. ECBE está basada en la protección de los campos de una bitácora de eventos usando un algoritmo de cifrado determinista y técnicas de anonimización para mantener la diferenciación en los datos a fin de compartirlos con un servidor externo como la nube y, así, poder llevar a cabo tareas de MP sin que se pueda revelar información del proceso de negocio. A partir de la bitácora de eventos cifrada resultante por la estrategia propuesta no es posible relacionar los eventos, permitiendo que no se descubran trazas o secuencias de eventos, manteniendo la capacidad de realizar las tareas de MP. Una de las limitaciones de ECBE es que solo permite la creación de modelos DFG, lo que no permite usar algoritmos de MP basados en consecuencias. Como parte del trabajo futuro se considera el d esarrollo de un mecanismo de control de acceso y compartición segura tanto de la bitácora de eventos como del modelo de proceso para entidades autorizadas.

Referencias

[1] Wil M. P. van der Aalst. *Process Mining: Data Science in Action*. 2.ª ed. Heidelberg: Springer, 2016. ISBN: 978-3-662-49850-7. DOI: 10.1007/978-3-662-49851-4.

[2] Wil M. P. van der Aalst y A. J. M. M. Weijters. "Process mining: a research agenda". En: *Comput. Ind.* 53.3 (2004), págs. 231-244. DOI: 10.1016/j.compind.2003.10.001. URL: https://doi.org/10.1016/j.compind.2003.10.001.

[3] Karim Abouelmehdi, Abderrahim Beni-Hessane y Hayat Khaloufi. "Big healthcare data: preserving security and privacy". En: *Journal of Big Data* 5 (ene. de 2018). DOI: 10.1186/s40537-017-0110-7.

[4] Elisabetta Benevento et al. "Evaluating the Effectiveness of Interactive Process Discovery in Healthcare: A Case Study". En: *Business Process Management Workshops*. Ed. por Chiara Di Francescomarino, Remco Dijkman y Uwe Zdun. Cham: Springer International Publishing, 2019, págs. 508-519.

[5] Joan Daemen y Vincent Rijmen. "The Advanced Encryption Standard Process". En: *The Design of Rijndael: AES — The Advanced Encryption Standard*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, págs. 1-8. ISBN: 978-3-662-04722-4. DOI: 10.1007/978-3-662-04722-4_1.

[6] Boudewijn van Dongen. *Real-life event logs - Hospital log*. 2011. DOI: 10.4121/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54.

[7] EU General Data Protection Regulation. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC*

(General Data Protection Regulation) (Text with EEA relevance). 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[8] Business Process Management Initiative. *BPMN Specification - Business Process Model and Notation*. 2005. URL: https://www.bpmn.org.

[9] Felix Mannhardt. *Sepsis Cases - Event Log*. 2016. DOI: 10.4121/uuid:915d2bfb-7e84-49ad-a286-dc35f063a460.

[10] Felix Mannhardt. *Hospital Billing - Event Log*. 2017. DOI: 10.4121/uuid:76c46b83-c930-4798-a1c9-4be94dfeb741.

[11] Niels Martin et al. "Interactive Data Cleaning for Process Mining: A Case Study of an Outpatient Clinic's Appointment System". En: *Business Process Management Workshops*. Ed. por Chiara Di Francescomarino, Remco Dijkman y Uwe Zdun. Cham: Springer International Publishing, 2019, págs. 532-544. ISBN: 978-3-030-37453-2.

[12] Jorge Munoz-Gama et al. "Process mining for healthcare: Characteristics and challenges". En: *Journal of Biomedical Informatics* 127 (2022), pág. 103994. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2022.103994. URL: https://www.sciencedirect.com/science/article/pii/S1532046422000107.

[13] Blake Murdoch. "Privacy and artificial intelligence: challenges for protecting health information in a new era". En: *BMC Medical Ethics* (ene. de 2021). DOI: https://doi.org/10.1186/s12910-021-00687-3.

[14] Sharyl Nass, Laura Levit y Lawrence Gostin. *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. Feb. de 2009. ISBN: 978-0-309-12499-7. DOI: 10.17226/12458.

[15] Pascal Paillier. "Public-Key Cryptosystems Based on Composite Degree Residuosity Classes". En: *Advances in Cryptology — EUROCRYPT '99*. Ed. por Jacques Stern. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, págs. 223-238. ISBN: 978-3-540-48910-8.

[16] Marco Pegoraro et al. "Analyzing Medical Data with Process Mining: a COVID-19 Case Study". En: *arXiv e-prints*, arXiv:2202.04625 (feb. de 2022), arXiv:2202.04625. arXiv: 2202.04625 [cs.DB].

[17] Eric Rojas et al. "Process mining in healthcare: A literature review". En: *Journal of Biomedical Informatics* 61 (2016), págs. 224-236. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2016.04.007. URL: https://www.sciencedirect.com/science/article/pii/S1532046416300296.

[18] Sudhakar Sengan et al. "Secured and Privacy-Based IDS for Healthcare Systems on E-Medical Data Using Machine Learning Approach". En: *International Journal of Reliable and Quality E-Healthcare* 11 (oct. de 2021), págs. 1-11. DOI: 10.4018/IJRQEH.289175.

[19] H. Smith, Tamara Dinev y Heng Xu. "Information Privacy Research: An Interdisciplinary Review". En: *MIS Quarterly* 35 (dic. de 2011), págs. 989-1015. DOI: 10.2307/41409970.

[20] Zoe Valero-Ramon et al. "Dynamic Models Supporting Personalised Chronic Disease Management through Healthcare Sensors with Interactive Pro-

cess Mining". En: *Sensors* 20.18 (2020). ISSN: 1424-8220. DOI: 10.3390/s20185330. URL: https://www.mdpi.com/1424-8220/20/18/5330.

[21] Alan F. Westin. "Privacy And Freedom". En: *Washington and Lee Law Review* 25 (ene. de 1968), págs. 166-170.

Muyal-Chimalli: Servicio para el acceso seguro y confiable a datos sensibles

Diana E. Carrizales-Espinoza[1[0000-0002-3925-031X]], J.L. González-Compeán[1[0000-0002-2160-4407]], Miguel Morales-Sandoval[1[0000-0003-1702-8467]], y Ricardo Marcelín-Jiménez[2[0000-0002-5355-5830]]

[1] Cinvestav Tamaulipas, Cd. Victoria, México
[2] UAM-Iztapalapa, Cd. México, México
diana.carrizales@cinvestav.mx, joseluis.gonzalez@cinvestav.mx, miguel.morales@cinvestav.mx, rmarcelin@izt.uam.mx

Resumen. La producción de datos ha aumentado de forma exponencial en los últimos años debido, entre otras razones, al incremento de dispositivos IoT (p. ej., electrocardiogramas y sensores) y dispositivos de usuario final (p. ej., celulares y tabletas). De esta forma, los datos generados son almacenados en diferentes ubicaciones geográficas durante todo su ciclo de vida. Lo anterior, provoca una gestión jerárquica que permite producir una respuesta rápida al analizar, preservar o transportar un gran volumen de datos (*big data*). La nube se ha vuelto de vital importancia para gestionar esta gran cantidad de datos; sin embargo, es necesario contar con mecanismos que permitan asegurar, preparar, entregar y recuperar datos sensibles de forma segura, confiable y eficiente en escenarios reales donde estos datos son clave para la toma de decisiones (p. ej., los del ámbito médico, tales como expedientes clínicos y tomografías). En este capítulo se presenta Muyal-Chimalli, una herramienta computacional conformada por un conjunto de servicios que permiten a las instituciones de salud, profesionales de salud, pacientes y/o comunidad científica acceder a los servicios de e-salud y/o sistemas de analítica para asegurar el manejo, preparación y acceso seguro y confiable de datos sensibles. Muyal-Chimalli verifica, automáticamente y de forma transparente, que cada sistema de e-salud observe las normas nacionales e internacionales, garantizando la privacidad, confidencialidad, integridad y disponibilidad de los contenidos, así como estableciendo tolerancia a fallas de servicios/servidores y creando registros inmutables en una red privada (blockchain) para brindar trazabilidad al manejo de los datos.

Palabras clave: Seguridad Informática · Confiabilidad y Eficiencia de Datos · Trazabilidad · Datos Sensibles · Sistemas de e-Salud y Analítica.

1 Introducción

En los últimos años se ha observado un incremento exponencial en la producción de los dispositivos de IoT (p. ej., sensores, electrocardiogramas, tomógrafos, espirómetros) [1]. En consecuencia, el volumen de los datos producidos y gestionados por las organizaciones también ha aumentado [2]. Lo anterior se debe a que los usuarios finales que se encuentran asociados a dichas organizaciones producen, almacenan, intercambian y utilizan los datos constante y continuamente, provocando, así, un efecto de acumulación de datos [3].

En escenarios reales, esta tendencia da como resultado un procesamiento de grandes volúmenes de datos conocido como *big data*. En estos escenarios, grandes repositorios de datos son producidos de forma continua por los dispositivos de IoT (es decir que se genera un gran *volumen* de datos) para obtener información útil que será utilizada como entrada (información *veraz* que tiene *valor* para un grupo de personas y/u organizaciones) en procesos críticos de toma de decisiones (los cuales necesitan realizase con la mayor *velocidad* posible debido a la sensibilidad de dichos procesos) [4], [5].

De esta forma, los datos generados son almacenados en diferentes ubicaciones geográficas durante todo su ciclo de vida. La Figura 1 muestra una representación conceptual del ciclo de vida de los datos, el cual incluye la adquisición, el indexamiento, pre-procesamiento, análisis, uso, compartición y el consumo de los datos.

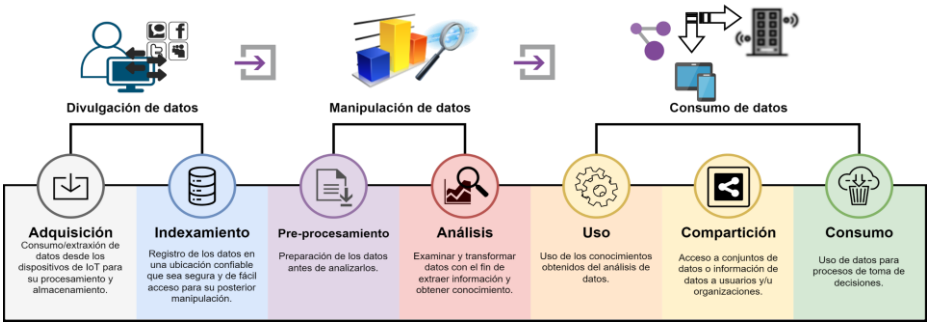


Figura 1. Representación conceptual del ciclo de vida de los datos.

Convencionalmente, el cómputo en la nube [6] ha sido la solución más popular para el procesamiento y almacenamiento datos producidos desde los dispositivos de IoT y de usuario final (p. ej., celulares, tabletas, laptops, estaciones de trabajo, etc.) [7]. De la misma forma, el cómputo en la nube se ha convertido en un soporte para escenarios de big data [8], [9]. No obstante, a medida que los sistemas que almacenan estos datos escalan y la producción de datos aumenta, la recopilación, la gestión y el procesamiento de datos de forma centralizada se vuelve inviable.

Además, es importante considerar que, en los distintos entornos organizacionales, como lo son los hospitales, es necesario procesar, preservar y compartir

datos sensibles con otras organizaciones de forma segura, confiable y rentable. Para dicho fin se han establecido distintos requisitos no funcionales obligatorios, impuestos por las leyes en cada país, a través de normas (p. ej., NOM-024-SSA3-2010 y NIST) y protocolos (p. ej., DICOM/HL7) para el intercambio y preservación de datos sensibles que es necesario que las organizaciones cumplan. Estos requisitos van desde la seguridad de los datos durante su transporte y almacenamiento, hasta el establecimiento de controles de acceso a los datos (es decir, privacidad) y el aseguramiento de la integridad y la confidencialidad de estos, así como su confiabilidad.

Debido a estos factores, ha surgido la necesidad de contar con servicios que permitan mantener, transportar y procesar datos sensibles de forma segura, eficiente y confiable. En este capítulo se presenta Muyal-Chimalli, una herramienta computacional conformada por un conjunto de servicios que permiten el acceso seguro de datos sensibles en servicios de e-salud y servicios de analítica. La Figura 2 muestra una representación conceptual del uso de Muyal-Chimalli en un escenario real de transporte y preservación de datos sensibles.

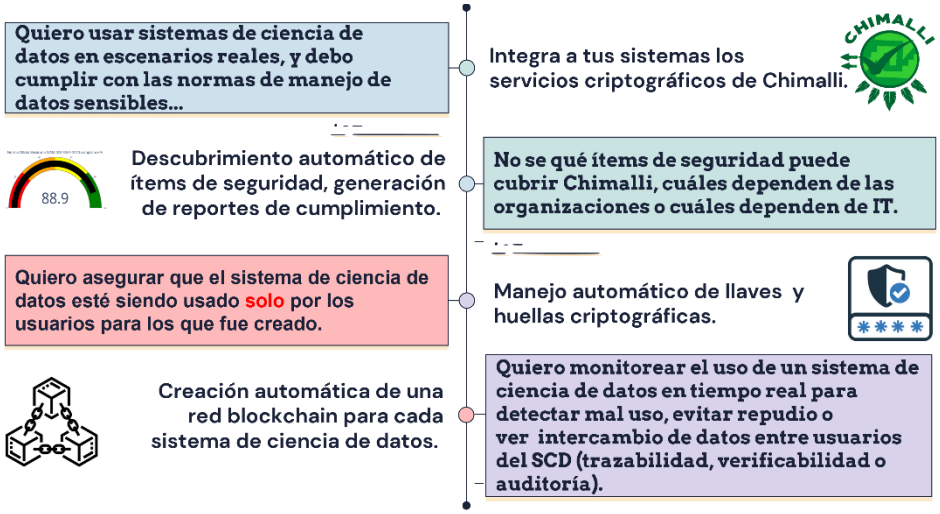


Figura 2. Representación conceptual de la descripción general del uso de Muyal-Chimalli.

Muyal-Chimalli se encuentra compuesto por un conjunto de servicios para acceder, manejar, preparar, transportar y recuperar datos médicos. Estos servicios permiten a las instituciones de salud, profesionales de la salud, pacientes y/o comunidad científica acceder a los datos de forma segura y confiable a través de servicios de e-salud y/o sistemas de analítica para obtener información útil que ayude a mejorar la toma de decisiones en escenarios de salud. Por ejemplo, la Figura 3 muestra una representación conceptual de un escenario donde el médico de un hospital envía una radiografía a un especialista que se encuentra en otro hospital. Dicho proceso debe ser realizado siguiendo los protocolos necesarios para asegurar que los datos sean íntegros, confiables y seguros, además de que sean entregados en el menor tiempo posible.

Figura 3. Representación conceptual de un escenario real de salud crítico para la toma de decisiones.

Las principales características que provee Muyal-Chimalli a los datos son:

- *Confiabilidad y disponibilidad* - garantiza el acceso a los datos, así como a los sistemas de e-salud y analítica en escenarios de fallas de servidores y almacenamiento de datos, así como apagones en los centros de datos; Muyal-Chimalli permite configurar los servicios de confiabilidad y disponibilidad de acuerdo con los recursos disponibles en la organización [10], [11];
- *Eficiencia* - Muyal-Chimalli maneja los datos hasta 10 veces más rápido que opciones disponibles en el mercado; además, permite una reducción del 34% en relación con los costos de utilización de la nube y de un 70% en relación con los costos de almacenamiento; asimismo, Muyal-Chimalli permite compartir sistemas con otras instituciones en minutos y de forma segura (es decir, permite compartir datos de forma intrainstitucional e interinstitucional) [12];
- *Trazabilidad* - crea automáticamente redes de blockchain para auditoría continua durante el intercambio de datos; Muyal-Chimalli elimina los costos derivados de contratar un servicio de blockchain con terceros (p. ej., se estima que se requieren, aproximadamente, $3,400 dólares por 4 meses de uso de la red en la nube) [13];
- *Seguridad* - asegura la *privacidad*, *confidencialidad* e *integridad* de los datos y establece *controles* de *acceso* de forma automática; además, permite la verificación y creación de reportes que muestran el grado de cumplimiento de los protocolos y normas oficiales nacionales e internacionales para el intercambio y almacenamiento de datos sensibles [14], [15].

El objetivo principal de Muyal-Chimalli es proveer, de forma automática y transparente, tolerancia a fallas en TIC (Tecnologías de la Información y la Comunicación), así como privacidad, confidencialidad, integridad, disponibilidad y trazabilidad en datos sensibles para que los sistemas de e-salud y los sistemas de analítica cumplan con las normas nacionales (NOM-024-SSA3-2010, NOM-004-SSA3-2012) e internacionales (NIST, COBIT 5, ISO 27001:2013) para el transporte y preservación de datos.

El resto de este capítulo está organizado de la siguiente forma. La Sección 2 presenta una descripción detallada sobre el servicio de acceso seguro a los servicios de e-salud y/o servicios de analítica; de la misma manera, se describen

los componentes que lo conforman. La Sección 3 describe los principales resultados obtenidos al utilizar Muyal-Chimalli en servicios de e-salud y/o analítica de datos. Finalmente, la Sección 4 concluye este trabajo dando un resumen sobre el servicio presentado en este capítulo.

2 Servicio de acceso seguro a servicios de e-salud y/o sistemas de analítica

La palabra Muyal proviene del glifo maya que significa "nube en el cielo" (ver Figura 4), mientras que la palabra Chimalli proviene del náhuatl y significa *escudo* (ver Figura 4), el cual era un objeto defensivo utilizado por las fuerzas militares prehispánicas mesoamericanas. De esta manera, Muyal-Chimalli hace referencia a un *escudo en la nube*, el cual permite proteger los datos, ya sea para su transporte o preservación.

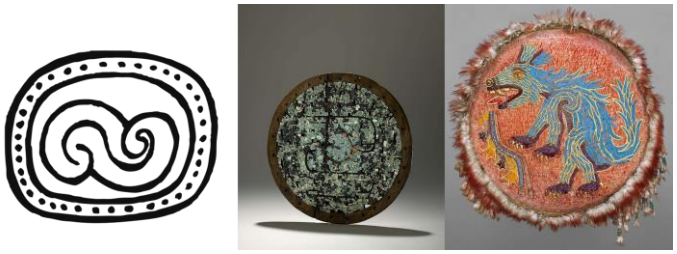


Figura 4. Representación maya de Muyal y ejemplos de dos chimallis utilizados en la antigua Mesoamérica.

Muyal-Chimalli es una herramienta computacional conformada por un conjunto de servicios que permiten a las instituciones de salud, profesionales de la salud, pacientes y/o comunidad científica acceder de forma segura a datos sensibles a través de servicios de e-salud y/o sistemas de analítica. Para este fin, Muyal-Chimalli cuenta con *esquemas* que permiten la *preparación* y *recuperación* de los datos, así como *mecanismos de control de acceso*.

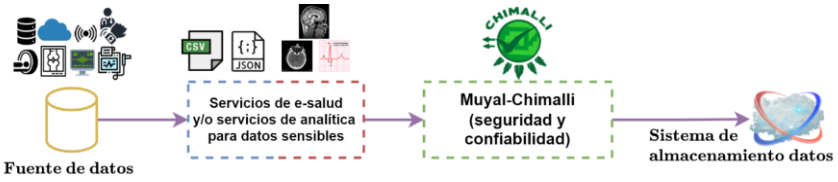


Figura 5. Representación conceptual de un flujo de datos de un sistema de e-salud y/o un sistema de analítica utilizando Muyal-Chimalli para proveer seguridad a los datos.

Muyal-Chimalli garantiza que los datos y los tomadores de decisiones reúnan las condiciones para realizar procesos de análisis. Además, permite la validación y el registro de cualquier operación de compartición de datos que sea realizada dentro de los sistemas de e-salud y/o los sistemas de analítica. La Figura 5

muestra una representación conceptual del flujo de datos de un sistema utilizando los servicios de Muyal-Chimalli para proveer seguridad a los datos.

Muyal-Chimalli permite alcanzar un porcentaje de hasta el 70% en el cumplimiento de las normas y protocolos estandarizados de forma nacional e internacional referentes al manejo, intercambio y transporte seguro y confiable de datos sensibles. Es importante considerar que el porcentaje restante que no es posible cubrir con los servicios de Muyal-Chimalli corresponde a aquellas actividades que necesitan ser realizadas de forma manual por el personal de TI de una organización/institución, o que no pueden ser realizadas por un servicio tecnológico (p. ej., la generación de llaves públicas y privadas, la cual necesita ser realizada por una entidad de confianza). Además, Muyal-Chimalli permite cubrir todas las fases de interconexión establecidas por las normas oficiales (NOM-024-SSA3-2010, NOM-004-SSA3-2012, ISO 27001:2013, COBIT 5 y NIST).

En este sentido, las normas nacionales NOM-004-SSA3-2012 y NOM-024-SSA3-2010 establecen tantos los objetivos funcionales y las funcionalidades que es necesario que los sistemas observen para garantizar la interoperabilidad, procesamiento, interpretación, confidencialidad, seguridad y uso de estándares y catálogos de la información de los registros en salud, así como el establecimiento de los distintos criterios científicos, éticos, tecnológicos y administrativos en la elaboración, integración, uso, manejo, archivo, conservación, propiedad, titularidad y confidencialidad de los expedientes clínicos [16]. Por otra parte, las normas internacionales COBIT 5, NIST e ISO 27001:2013 establecen el marco de trabajo para la gestión de los sistemas de seguridad de la información, así como la gestión de las tecnologías de la información para proporcionar confidencialidad, integridad, seguridad y disponibilidad de forma continua, así como el cumplimiento legal de ésto [17].

Para este fin, Muyal-Chimalli crea un reporte/resumen de seguridad, el cual permite revelar las tareas de ciberseguridad (práctica de proteger los sistemas e información de los ataques digitales) que dependen de actividades realizadas por el personal de la salud para que las instituciones puedan crear un plan para implementarlas. Además, asegura tanto la confidencialidad como el anonimato mediante el uso de técnicas de cifrado, las cuales son aplicadas a los datos entrantes y salientes de los sistemas de e-salud y/o los sistemas de analítica. También, permite detectar si surge alguna alteración en los datos (corrupción de datos), asegurando la integridad de éstos. De la misma forma, permite realizar automáticamente la gestión de contratos inteligentes y de transacciones (los procesos, etapas y usuarios por las cuales pasaron los datos), así como la verificabilidad de estas de forma confidencial.

Los componentes principales de Muyal-Chimalli son los siguientes:

- Servicios para la *preparación* y *recuperación* de datos sensibles - dichos servicios son configurables, e incluyen métodos, algoritmos y técnicas para brindar los requisitos de *seguridad*, *trazabilidad*, *integridad* y *eficiencia*;
- Mecanismos de *trazabilidad y verificabilidad* de datos basados en blockchain - permiten registrar todos los procesos, etapas y usuarios por los que han pasado los datos;

- Mecanismos de *control de acceso* de usuarios - aseguran que solo aquellos usuarios/organizaciones que tengan el debido acceso a los datos puedan acceder a ellos;
- Servicios de *validación* de normas oficiales nacionales e internacionales y protocolos DICOM/HL7 - permiten la generación de reportes de cumplimiento de las normas y protocolos;
- Un servicio que permite la utilización de técnicas de criptografía de siguiente generación [18] para la transformación de datos en objetos seguros - permite mantener la *integridad* y *confidencialidad* de los datos.

2.1 Servicios de preparación y recuperación de datos médicos configurables que proveen seguridad, trazabilidad, integridad y eficiencia a los datos

En escenarios reales de gestión de datos sensibles, es necesario contar con mecanismos que provean requisitos no funcionales (un requisito que especifica los criterios que se pueden utilizar para juzgar el funcionamiento de un sistema, en lugar de comportamientos específicos), tales como seguridad, eficiencia y confiablidad. Estos requisitos son necesarios debido a las normas de gestión de datos sensibles (por ejemplo, las normas oficiales mexicanas NOM-004-SSA3-2012 y NOM-024-SSA3-2010) y a las leyes impuestas por los gobiernos y organizaciones [19], [20].

En esta subsección se presenta una descripción de los esquemas de *preparación* y *recuperación* de datos, los cuales permiten el manejo de los requisitos no funcionales. En este sentido, dicha *preparación* de datos es realizada antes de que los datos sean transportados a través de un flujo de datos (cargados para su preservación o compartidos a través de entornos no controlados, como lo es la nube [21]). Para ello, primero se presenta una descripción de la estructura de procesamiento de tuberías. Dicha estructura permite crear los esquemas de *preparación* y *recuperación* de datos y, posteriormente, se realiza una descripción de los requisitos no funcionales que pueden ser incluidos a las tuberías de procesamiento.

Estructura de los esquemas de *preparación* y *recuperación de datos*. En Muyal-Chimalli, la estructura de los esquemas de *preparación* y *recuperación* de datos está construida como tuberías, las cuales se modelan con base en un grafo acíclico dirigido (*DAG*, por sus siglas en inglés *Directed Acyclic Graph*). Los nodos que componen este *DAG* representan los algoritmos que proveen los requisitos no funcionales, mientras que las aristas representan las entradas requeridas por los algoritmos, así como los resultados producidos por ellos. En este sentido, una tubería puede incluir tantos algoritmos, que permitan proporcionar requisitos no funcionales, como sea necesario. Lo anterior permite dar cumplimiento a las normas, protocolos y leyes nacionales e internacionales para la preservación e intercambio de datos sensibles en una misma organización y entre distintas organizaciones (es decir, de forma intrainstitucional e interinstitucional). De este modo, la ejecución sucesiva de los requisitos no funcionales permite la creación de una tubería de procesamiento de datos. La **Figura 6**

muestra una representación conceptual del intercambio de datos entre dos organizaciones utilizando Muyal-Chimalli. En este ejemplo, los datos son extraídos desde la organización A; posteriormente, son registrados en la base de datos y luego son cifrados. Una vez que los datos han sido cifrados, éstos son enviados al servicio de almacenamiento. Después, la organización B podrá acceder a los datos; para ello, el esquema de recuperación realiza el proceso inverso al esquema de preparación. En este caso, los datos primero son descifrados y, posteriormente, se verifica su integridad. Una vez realizado este proceso, los datos son entregados a un usuario perteneciente a la organización B.

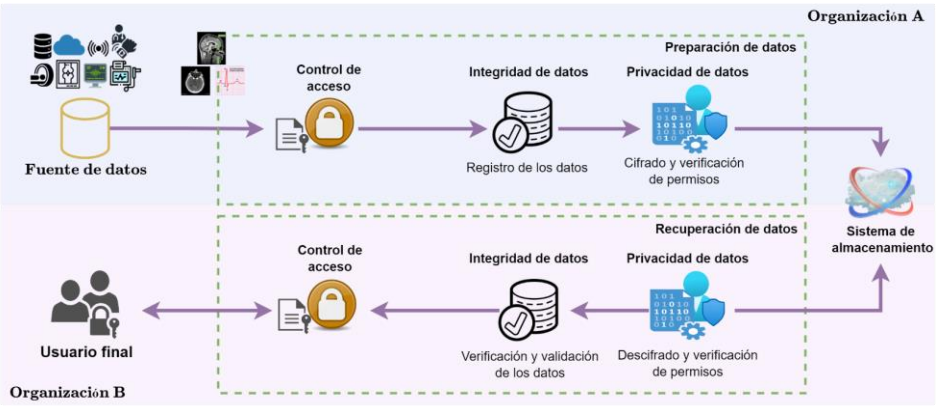


Figura 6. Representación conceptual del intercambio de datos entre dos organizaciones utilizando los esquemas de preparación y recuperación de datos de Muyal-Chimalli.

Los algoritmos que permiten observar los requisitos no funcionales en los flujos de datos son computacionalmente costosos. Lo anterior se debe a que dichos algoritmos añaden retrasos en cada una de las etapas de la tubería que se encuentran definidas en el *DAG*. Para poder mitigar el impacto en el rendimiento que se produce al agregar los requisitos no funcionales para la *preparación* y *recuperación* de datos, los esquemas generados consideran bifurcaciones (es decir, división de tareas en la tubería) para producir paralelismo de datos y procesamiento de tareas concurrentes en cada etapa de la tubería.

En las tuberías, cada etapa invoca a un componente llamado gestor de paralelismo, el cual se encarga de crear clones (trabajadores) de los algoritmos de los requisitos no funcionales a los cuales se les asigna una carga de trabajo. Este gestor también se encarga de desplegar una entidad conocida como balanceador de carga la cual se encarga de mejorar el rendimiento del procesamiento a través de la distribución equitativa (en un sentido estadístico) de los datos/tareas sobre los *trabajadores* habilitados. Este modelo de procesamiento posibilita la producción de un paralelismo implícito que permite la ejecución de los algoritmos de los requisitos no funcionales de forma paralela. Lo anterior, permite reducir el tiempo de preparación y/o recuperación de los datos antes de que sean transportados a través de los flujos de trabajo.

Requisitos no funcionales en los esquemas de *preparación* y *recuperación* de datos. Muyal-Chimalli permite agregar al manejo de datos en los

flujos de procesamiento distintos requisitos no funcionales: i) confiabilidad; ii) seguridad; y iii) costo-eficiencia.

En este sentido, la confiabilidad permite mitigar aquellos problemas causados por deficiencias presentadas en la infraestructura donde los datos se procesan y almacenan [22], [23]. Este requisito se consigue utilizando el conocido algoritmo de dispersión de información (IDA, por sus siglas en inglés *Information Dispersal Algorithm*) [24] . Este algoritmo fue propuesto en 1990 por Michel Rabin y permite dividir/separar/segmentar un archivo en n segmentos (donde n es un número natural), a los cuales se les añade redundancia (es decir, la duplicación o reescritura de información encontrada en el archivo), lo cual permite aumentar la confiabilidad a la hora de recuperar los datos. Dichos segmentos, conocidos como dispersos, son enviados a distintas ubicaciones (de forma distribuida), en donde, para recuperar un archivo, solo se requieren un total de m de ellos (donde m es un número natural y siempre será menor a n) de los n originales. Por ejemplo, con este algoritmo un archivo puede ser transformado en 5 dispersos (n), los cuales serán enviados a 5 ubicaciones distintas. Para recuperar dicho archivo, por ejemplo, solo se requiere acceder a 3 de estas ubicaciones (es decir, permite tolerar 2 fallas) para extraer los dispersos que almacenan (m) y con ellos reconstruir el archivo original. Esto es útil en escenarios distribuidos de datos (como la nube), donde si alguna de las ubicaciones de almacenamiento presenta una falla, el usuario no lo notará, ya que podrá recuperar sus datos accediendo a las demás ubicaciones de almacenamiento donde se colocaron el resto de los segmentos.

Por otro lado, la característica de *seguridad* (en la cual se añade *confidencialidad*, *integridad* y *control de acceso*, CIA, por sus siglas en inglés Confidentiality, Integrity and Access Control), es añadida a los datos para resolver problemas que surjan del transporte y compartición de datos en entornos no controlados y no confiables (por ejemplo, la nube [25], [26]). Esta característica se añade mediante técnicas de criptografía de siguiente generación, como lo es el cifrado basado en atributos.

En estos esquemas, los datos se cifran utilizando el cifrado AES [27] para añadir *confidencialidad*. AES es una función matemática de cifrado conocida como el método de cifrado por bloques (es decir, el cifrado de los datos por medio de algoritmos que permiten convertir un bloque de datos claros en un bloque de datos cifrados a través de una clave) más seguro que existe debido a que, en la práctica, éste no se puede romper, además de que es rápido y eficiente; mientras que, para añadir *controles de acceso*, se utiliza el algoritmo CP-ABE [28], el cual es un cifrado basado en atributos de política de texto cifrado que utiliza árboles de acceso para cifrar datos. En este algoritmo, las claves secretas de los usuarios se generan sobre un conjunto determinado de atributos.

Por otro lado, la *integridad* de los datos se añade generando una huella o firma digital de cada dato. Esta firma/huella es conocida como hash y se centra en lograr dos objetivos: i) la identificación de los contenidos replicados antes de que estos sean enviados al resto de las etapas de *preparación/recuperación* de datos; y ii) la detección de alteraciones en los datos cuando los usuarios descargan/recuperan archivos.

Finalmente, la característica de *costo-eficiencia* se consigue mediante técnicas de compresión y deduplicación de datos. Dichas técnicas permiten reducir el número de contenidos, así como la cantidad de datos a procesar.

Además, estas técnicas permiten reducir el volumen de datos enviados a la nube y, por ende, reducir los costos resultantes de la subcontratación de servicios para realizar las tareas de gestión de datos.

En este sentido, para agregar eficiencia durante el procesamiento de los datos y la ejecución de tareas, se utilizan patrones de paralelismo (es decir, técnicas que permiten realizar tareas o procesar los datos de forma paralela). En Muyal-Chimalli, principalmente se cuenta con dos patrones: i) el patrón manejador/trabajador; y ii) el patrón *divide&vencerás*.

Patrones de paralelismo para la eficiencia en los esquemas para la preparación y recuperación de datos de Muyal-Chimalli. El paralelismo de datos es un paradigma de la programación concurrente. Dicho paradigma consiste en dividir un conjunto de datos y/o tareas de manera que a cada procesador disponible le corresponda un subconjunto de estas tareas/datos.

En este sentido, el patrón *manejador/trabajador* permite procesar los datos en diferentes fases: i) la clonación de tareas; ii) la distribución de tareas y datos; y iii) la supervisión de la ejecución de las tareas.

En la fase de clonación de tareas, el manejador crea instancias de contenedores virtuales (es decir, aplicaciones independientes, empaquetadas con sus dependencias, que pueden ser desplegadas en cualquier entorno sin mayor preparación), los cuales representan clones de una etapa determinada en la tubería del flujo de datos. Los clones creados en esta fase se denominan *trabajadores*.

En la fase de distribución de tareas/datos, el *manejador* lee el contenido almacenado en una fuente de datos y, posteriormente, crea una lista de datos/tareas utilizando un conjunto de rutas. Cada componente de la lista se distribuye de forma balanceada a los *trabajadores* utilizando el algoritmo de balanceo de carga *two choices* [29]. Este algoritmo permite elegir dos trabajadores de forma aleatoria y, posteriormente, escoge el *trabajador* que tenga la menor carga de trabajo [30]. Finalmente, en la fase de supervisión, el *manejador* verifica que los *trabajadores* entreguen los resultados de la tarea realizada a la siguiente etapa en el flujo de los datos.

Por otro lado, el patrón *divide&vencerás* cuenta con tres entidades conocidas como: i) divide; ii) trabajadores; y iii) vences. La entidad *divide* es una instancia de software que se encarga de dividir/segmentar los datos en s segmentos (donde s es un número natural) sin redundancia, los cuales son procesados por los *trabajadores* (similares a los *trabajadores* del patrón *manejador/trabajador*). La entidad *vences* se encarga de consolidar los resultados de cada uno de los *trabajadores* en un único resultado para entregarlo a la siguiente etapa en el flujo de datos.

2.2 Mecanismos de trazabilidad y verificabilidad de datos basados en blockchain de Muyal-Chimalli

La Organización Internacional para la Estandarización (ISO 9001:2008) define la *trazabilidad* como aquella propiedad del resultado del valor de un estándar determinado (donde éste puede estar relacionado con distintas referencias específicas). Usualmente, estos estándares se refieren a normas nacionales o internacionales a través de una cadena continua de comparaciones [31]. Por otro

lado, el comité de seguridad alimentaria de AECOC define la *trazabilidad* como aquel conjunto de procedimientos preestablecidos y autosuficientes que permiten conocer detalles de un producto (como lo es el histórico, la ubicación y la trayectoria) a lo largo de una cadena de suministro en un momento dado, a través de herramientas determinadas [32].

En Muyal-Chimalli, el proceso de *trazabilidad* de datos juega un papel importante dentro de los flujos de trabajo. Lo anterior se debe a que este proceso brinda la posibilidad de identificar el origen de los datos/procesos, así como las distintas etapas por las que han paso los datos a lo largo de todo su ciclo de vida (adquisición, proceso productivo, distribución y logística, hasta llegar al consumidor final) [33], [34].

Este proceso cumple una parte fundamental dentro de Muyal-Chimalli debido a que permite que cualquiera de las entidades involucradas (p. ej., organizaciones, usuarios finales, etc.) puedan acceder a la información de cada una de las etapas por las cuales ha pasado el contenido digital. Lo anterior, permite verificar si se han cumplido con las acciones pactadas, así como conocer si los datos han sido procesados por las entidades correctas. Esto, posibilita aceptar o rechazar los datos con base en la información del flujo de datos (conocida como *traza*), apoyando, de esta forma, la toma de decisiones y mejorando la confianza en el resultado obtenido. En Muyal-Chimalli, este servicio tiene como objetivo asegurar el registro inmutable de cada acción realizada sobre cada activo digital que es procesado en las diferentes cadenas de valor (flujos de datos) generadas a través de los servicios de construcción de sistemas de e-salud y/o sistemas de analítica de datos sensibles.

El mecanismo de *trazabilidad* de Muyal-Chimalli provee las características de *trazabilidad* y *verificabilidad* a cada uno de los productos/datos que se procesan en cualquiera de los servicios de e-salud y/o analítica de datos sensibles. Además, este mecanismo permite realizar *trazabilidad* de forma interna y externa (es decir, dentro de una misma institución, así como colaboraciones entre varias instituciones/organizaciones) de los productos digitales que son procesados y gestionados a través de los sistemas de e-salud y/o analítica.

La tecnología de blockchain de Muyal-Chimalli permite preservar y compartir los datos de forma segura y transparente sin un órgano central de control. Esta tecnología utiliza una base de datos que es segura y que será compartida solo por aquellos usuarios autorizados. De esta forma, es posible comprobar la validez de cada uno de los procesos realizados en el flujo de trabajo en cada una de las etapas.

El servicio de trazabilidad hace uso de las siguientes herramientas para su correcto funcionamiento: i) *Sawtooth* (solución empresarial para construir, implementar y ejecutar redes y aplicaciones de contabilidad distribuida [35]); ii) *Docker* (proyecto de código abierto que automatiza el despliegue de aplicaciones dentro de contenedores de software [36]); y iii) *Docker Compose* (herramienta para definir y ejecutar aplicaciones Docker de varios contenedores, creando una red virtual que permite la comunicación entre ellos de forma eficiente y simple [37]). Además, Javascript es utilizado como lenguaje principal y MySQL como gestor de base de datos.

Lo anterior quiere decir que la blockchain provee diversas ventajas en el sector de cadenas de suministro (conjunto de actividades, instalaciones y medios

de distribución necesarios para llevar a cabo un proceso [38]), el cual se usa como enfoque en Muyal-Chimalli.

2.3 Servicio para la transformación de datos en objetos seguros mediante el uso de técnicas de criptografía de siguiente generación

Debido al crecimiento acelerado de datos que se ha presentado en los últimos años, el cómputo en la nube se ha convertido en el nuevo núcleo de vida de los datos. Incluso, se espera que, en los próximos años, el 49% de los datos se almacenen en la nube pública [39]. Para evitar incidentes o mitigar riesgos que aún ocurren en la nube, como lo son las alteraciones de los datos [40], la pérdida de privacidad, las violaciones de seguridad o los accesos no autorizados a los datos, las organizaciones deben contar con sistemas o servicios que permitan entregar/recuperar los datos de forma segura, confiable y transparente [41].

En este contexto, Muyal-Chimalli cuenta con un servicio para la construcción eficiente de sistemas de seguridad que permiten a las organizaciones compartir, intercambiar y rastrear información en la nube. En este servicio, los sistemas de seguridad y el software de blockchain se convierten en servicios independientes y autónomos. Para mejorar la eficiencia de los servicios de seguridad, así como la experiencia de los usuarios finales, se agregan patrones de paralelismo implícitos junto con técnicas de balanceo de carga a los servicios.

Este servicio permite a las organizaciones respaldar patrones de intercambio de información en línea entre múltiples participantes al acoplar distintos servicios para cumplir con múltiples combinaciones de requisitos de seguridad (p. ej., confidencialidad, integridad, no repudio, autenticación y trazabilidad). Las características principales de este servicio son: i) flexibilidad; y ii) eficiencia.

La *flexibilidad* permite integrar (en un único sistema integral), sobre la marcha y bajo demanda, tantas aplicaciones de seguridad como inquietudes sean expresadas por las organizaciones (etapas en un flujo de datos) y por los participantes de cada flujo de trabajo organizacional. En este método, las aplicaciones de seguridad y el software se convierten en servicios en la nube independientes y autónomos. Este conjunto de servicios se combina para crear sistemas de seguridad en la nube que permiten admitir flujos de trabajo organizacional en línea y que incluyen a varios participantes. En este sentido, un *framework* (estructura que se puede aprovechar para desarrollar un proyecto) basado en este método, permite la creación de sistemas de seguridad en la nube que aseguran el cumplimiento de múltiples requisitos no funcionales de seguridad.

Por otro lado, la *eficiencia* se obtiene utilizando un modelo de programación paralela para la gestión de datos basado en la combinación de patrones de paralelismo, así como de esquemas de balanceo de carga (integrados en las aplicaciones de seguridad). A partir de este modelo se crearon dos esquemas de patrones paralelos llamados: i) *pipeline*; y ii) *overlapped*. El esquema *pipeline* combina dos patrones de paralelismo: *pipe&filter* y manejador/trabajador. En este esquema, el primer patrón es utilizado para organizar las aplicaciones de seguridad en forma de tuberías y el segundo patrón se encarga de desplegar las tuberías creadas en forma de *trabajadores* para ejecutarlas en paralelo. Este esquema se encarga de cifrar/descifrar pequeños conjuntos de datos en paralelo. Por otro lado, el esquema *overlapped* permite el acoplamiento de aplicaciones

de seguridad independientes para que se ejecuten de forma suprapuesta. De la misma forma, permite que aquellas aplicaciones de seguridad que cuentan con algún tipo de dependencia se acoplen en forma de tubería. Este esquema se encarga de cifrar/descifrar grandes conjuntos de datos en paralelo.

Los componentes principales de este servicio son dos: i) un método de seguridad múltiple en la nube para la creación de servicios de gestión de la seguridad de los datos de forma confidencial, flexible, integral y eficiente; y ii) los esquemas de paralelismo (*pipeline y overlapped*) que permiten mejorar el rendimiento de los sistemas de seguridad en la nube, así como el mejoramiento de la experiencia de servicio de los usuarios finales de Muyal-Chimalli.

2.4 Mecanismos de control de acceso de usuarios de Muyal-Chimalli

Para poder cumplir con los requisitos de seguridad en Muyal-Chimalli, se creó un repositorio de aplicaciones de seguridad. Dicho repositorio cuenta con un conjunto de aplicaciones disponibles para que los usuarios finales puedan incluirlos en sus sistemas de seguridad.

Este repositorio cuenta con distintas aplicaciones de seguridad, como lo son los criptosistemas simétricos (estándar de cifrado avanzado, AES [42]), cifrado basado en emparejamiento (firmas cortas) y cifrado basado en atributos (CP-ABE [43]), así como un bloque de trazabilidad que se utiliza para registrar cada transacción de intercambio de información en una cadena de bloques privada [44].

En este contexto, CP-ABE [45] es utilizado para hacer cumplir, criptográficamente, el *control de acceso* de los sistemas de seguridad creados con Muyal-Chimalli. Los criptosistemas basados en emparejamientos, que producen firmas cortas, son utilizados como servicios de autenticación, no repudio e integridad. En Muyal-Chimalli, estos criptosistemas pueden proporcionar cualquier nivel de seguridad equivalente a 128, 192 o 256 bits, los cuales cumplen con la mayoría de los estándares (p. ej., el estándar NIST [46], [47]).

Tanto CP-ABE como las firmas cortas utilizan un emparejamiento bilineal asimétrico que es computacionalmente eficiente. En este sentido, los procedimientos de firma y verificación se implementan utilizando la instancia de la función hash (también conocidas como funciones resumen) SHA3. La combinación de estos algoritmos y funciones da como resultado la adición de distintas propiedades de seguridad a la información.

2.5 Servicios de validación de normas oficiales mexicanas y protocolos DICOM/HL7

En los procesos de preparación/recuperación, transporte y preservación de datos para escenarios reales de gestión de datos sensibles, es necesario considerar distintos requisitos no funcionales (por ejemplo, seguridad, eficiencia y confiabilidad) que ayuden a cumplir con las normas internacionales y nacionales impuestas por los gobiernos y/o las organizaciones que preservan, procesan y comparten estos datos [19], [20].

Para asegurar el cumplimiento de estas normas, regulaciones y leyes, es necesario contar con un servicio que permita la validación de éstos. En este sentido, Muyal-Chimalli cuenta con un servicio que determina el porcentaje de cumplimiento de los flujos de datos creados en los sistemas de e-salud y/o sistemas de analítica con base en las normas internacionales (NIST, ISO 27001:2013 y COBIT 5) y nacionales (NOM-024-SSA3-2010) para la preservación y compartición de datos sensibles. Las principales características de este servicio son: i) preprocesamiento de los datos para leer y manipular los archivos desde el código fuente; ii) identificación de los flujos de trabajo que conforman los servicios de e-salud y/o los servicios de analítica; iii) consulta de las fuentes de información utilizando APIs (interfaz de programación de aplicaciones) para obtener datos y características contextuales de los contenedores, las cuales representan las tareas que ejecuta el contenedor; iv) búsqueda de palabras clave para determinar el cumplimiento de las normas nacionales e internacionales; v) obtención del porcentaje de cumplimiento y generación de un reporte/resumen donde se visualizan los resultados obtenidos durante el análisis del sistema; vi) descubrimiento de los flujos de trabajo asociados con los archivos de configuración del sistema de e-salud y/o sistema de analítica, así como la generación de una representación del DAG obtenido; y vii) verificación de las normas NOM-024-SSA3-2010, NIST, ISO 27001:2013 y COBIT 5 para el intercambio y preservación de datos sensibles.

En este servicio, los archivos de configuración representan el servicio que será desplegado. El programa recibe como entrada los archivos de configuración y determina qué normas nacionales e internacionales cumple dicho servicio. El cumplimiento es mostrado mediante un reporte en donde se especifica el porcentaje de cumplimiento del servicio según las normas correspondientes. Además, el servicio realiza el descubrimiento del flujo de trabajo asociado a los archivos de configuración del servicio de e-Salud y/o servicio de analítica.

Las normas internacionales y nacionales son representadas mediante listas de verificación. Cada lista de verificación contiene los requerimientos de la norma correspondiente. Los requerimientos pueden ser garantizados por el servicio creado (de forma sistemática) o mediante intervención del usuario (es decir, de forma asistida).

El servicio de validación cuenta con cuatro módulos: i) módulo de preprocesamiento para realizar la lectura de los archivos de configuración, los componentes identificados en este módulo son agregados a un diccionario de datos (contiene la lista de claves que cuentan con información sobre los procesos que se realizan en un flujo de datos); ii) módulo de identificación de flujos de trabajo e información de contenedores, el cual permite identificar los flujos, patrones, bloques de construcción, requisitos no funcionales y etapas de procesamientos; iii) módulo de determinación del nivel de cumplimiento de las normas (en este módulo, las normas nacionales e internacionales fueron capturadas de forma manual en una lista de verificación); y iv) módulo de descubrimiento del flujo de trabajo, en el que se genera un DAG, donde las entradas son representadas por los nodos incidentes y las salidas son representadas por los nodos salientes.

3 Principales resultados obtenidos al utilizar Muyal-Chimalli en servicios de e-salud y/o analítica de datos

Muyal-Chimalli provee servicios de seguridad para sistemas de e-salud y/o sistemas de analítica. Estos servicios de seguridad permiten cumplir con los requisitos no funcionales impuestos por las leyes o los gobiernos de cada país, o por las organizaciones donde se realiza este intercambio y preservación de los datos. Se ha comprobado que, mediante el uso de los servicios de seguridad de Muyal-Chimalli, es posible reducir hasta en un 70% los costos del almacenamiento en la nube, así como en el total de datos almacenados al analizar 51 estudios de imágenes médicas. Lo anterior significa una reducción de hasta el 95% del tiempo al transferir los datos a un servicio de almacenamiento y un aumento de 1.52x en la velocidad de transferencia en comparación con una solución tradicional encontrada en el mercado actual (en este caso *Duplicity*, la cual es una solución tradicional para la transferencia de datos que incluye el manejo de duplicados y la compresión de datos) [48], [49].

Los esquemas de costo-eficiencia de Muyal-Chimalli permiten reducir el tiempo de procesamiento de los datos que son enviados a la nube o intercambiados entre distintas organizaciones. Por ejemplo, se comprobó que utilizando los servicios de seguridad de Muyal-Chimalli es posible procesar hasta 100,971 imágenes médicas de formato DICOM en tan solo 42.71 minutos, mientras que procesar esa cantidad de imágenes con una solución tradicional toma alrededor de 432.21 minutos. Lo anterior, implica una reducción del tiempo de procesamiento y un aumento en la velocidad de éste de hasta 10 veces. Además, en Amazon Blockchain (utilizando los servicios de Muyal-Chimalli), por ejemplo, es posible ahorrar hasta un 32% en costos en la trazabilidad de esta cantidad de datos (considerando que, en México, mantener una red de blockchain con un solo empleado y con tan solo cuatro nodos de AWS durante cuatro meses, requiere una inversión aproximada de 8,373.38 dólares) [48].

Los servicios de seguridad informática de Muyal-Chimalli también permiten cumplir hasta en un 70% con las normas NIST, ISO 27001:2013 y COBIT 5, mientras que, sin estos servicios, usualmente los sistemas solo alcanzan hasta un 20% en el cumplimiento de estas normas, ya que solo consideran la tolerancia a fallos. Es importante mencionar que Muyal-Chimalli también permite proporcionar un nivel de seguridad de 256 bits, el cual es el nivel de seguridad más alto recomendado actualmente por el NIST [49].

4 Conclusiones

Los servicios de seguridad informática generados con Muyal-Chimalli permiten a las organizaciones crear servicios de seguridad que se ajusten a sus necesidades y que los ayuden a cumplir con las normas y leyes establecidas por los gobiernos para el intercambio seguro y confiable de datos sensibles. Muyal-Chimalli permite preparar y recuperar los datos, asegurando su confiabilidad, confidencialidad, disponibilidad, anonimato, integridad, trazabilidad y eficiencia. Además, nos permite realizar la gestión automática de contratos inteligentes, así como la verificabilidad de los procesos realizados dentro de un flujo de tra-

bajo de forma confidencial. Debido a lo anterior, Muyal-Chimalli es una herramienta que ayuda a mitigar los riesgos que surgen durante el almacenamiento e intercambio de datos sensibles en procesos críticos para la toma de decisiones.

Agradecimientos

Este trabajo forma parte del proyecto 41756 “Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud” por FORDECYT-PRONACES.

Referencias

[1] M. A. Malik, "Internet of Things (IoT) healthcare market by component (implantable sensor devices, wearable sensor devices, system and software), application (patient monitoring, clinical operation and workflow optimization, clinical imaging, fitness and wellness measu," Global opportunity analysis and industry forecast, 2014–2021, pp. Allied Market Research, 124, 2016.

[2] J. &. R. D. Gantz, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east.," IDC iView: IDC Analyze the future, 2007(2012), pp. 1-16, 2012.

[3] . D. Reinsel, J. Gantz and . J. Rydning, "The digitization of the world from edge to core.," IDC White Paper, 2018.

[4] M. Marjani, . F. Nasaruddin, A. Gani, A. Karim, . I. A. T. Hashem, A. Siddiqa and I. Yaqoob, "Big IoT data analytics: architecture, opportunities, and open research challenges," IEEE Access, pp. 5247--5261, 2017.

[5] M. R. Anawar, S. Wang, M. Azam Zia, A. K. Jadoon, U. Akram and S. Raza, "Fog computing: An overview of big IoT data analytics," Wireless Communications and Mobile Computing, 2018.

[6] B. P. Rimal, E. Choi and I. Lumb, "A taxonomy and survey of cloud computing systems," 2009 Fifth International Joint Conference on INC, IMS and IDC, pp. 44--51, 2009.

[7] A. Botta, W. De Donato, V. Persico and A. Pescapé, "Integration of cloud computing and internet of things: a survey," Future generation computer systems, pp. 684--700, 2016.

[8] V. Mosco, "To the cloud: Big data in a turbulent world," Routledge, 2015.

[9] V. Malik and S. Singh, "Cloud, Big Data \& IoT: Risk Management," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 258--262, 2019.

[10] E. a. A. R. Bauer, "Reliability and availability of cloud computing," 2012.

[11] Y. a. B. C. Izrailevsky, "Cloud reliability," IEEE Cloud Computing, vol. 5, pp. 39--44, 2018.

[12] M. a. A. M. a. F. S. a. N. S. Darbandi, "involving Kalman filter technique for increasing the reliability and efficiency of cloud computing}," in Proceedings of the International Conference on Scientific Computing (CSC), 2012, p. 1.

[13] M. a. M. J. A. a. M.-C. H. M. a. G.-C. J. L. Morales-Sandoval, "Blockchain support for execution, monitoring and discovery of inter-organizational business processes," PeerJ Computer Science, vol. 7, p. e731, 2021.

[14] A. a. A. M. O. a. W. R. a. W. G. Albugmi, "Data security in cloud computing," in 2016 Fifth international conference on future generation communication technologies (FGCT), 2016, pp. 55--59.

[15] G. a. I. M. a. K. F. A. Ramachandra, "A comprehensive survey on security in cloud computing," Procedia Computer Science, vol. 110, pp. 465--472, 2017.

[16] A. a. l. I. y. P. d. D. P. Instituto Nacional de Transparencia, "Recomendaciones para orientar el debido tratamiento de datos personales en los expedientes clínicos de las instituciones de salud pública," 2021.

[17] isotools, "ISO Tools EXCELLENCE," [Online]. Available: https://www.isotools.org/normas/riesgos-y-seguridad/iso-27001/. [Accessed 2022 11 23].

[18] N. A. B. W. J. &. A. R. Gunathilake, "Next generation lightweight cryptography for smart IoT devices:: implementation, challenges and applications.," in In 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), 2019.

[19] H. Mier and T. Delgadillo, "Regulación del acceso al expediente clínico con fines de investigación en México," Revista CONAMED, vol. 22, pp. 27--31, 2018.

[20] Phillips, "International data-sharing norms: from the OECD to the General Data Protection Regulation (GDPR)," Human genetics, vol. 137, pp. 575--582, 2018.

[21] C. B. Tan, M. H. A. Hijazi, Y. Lim and A. Gani, "A survey on proof of retrievability for cloud data integrity and availability: Cloud storage state-of-the-art, issues, solutions and future trends," Journal of Network and Computer Applications, vol. 110, pp. 75--86, 2018.

[22] Gunawi, Hao, S. O, Laksono, Satria, Adityatama and Eliazar, "Why does the cloud stop computing?: Lessons from hundreds of service outages," SoCC, ACM, pp. 1--16, 2016.

[23] Bala and Chana, "Fault tolerance-challenges, techniques and implementation in cloud computing," International Journal of Computer Science Issues (IJCSI), vol. 9, p. 288, 2012.

[24] R. Marcelín-Jiménez, J. L. Ramírez-Ortíz, E. R. De La Colina, M. Pascoe-Chalke and J. L. González-Compeán, "On the Complexity and Performance of the Information Dispersal Algorithm," IEEE Access, pp. 159284--159290, 2020.

[25] Bhushan and Gupta, "Security challenges in cloud computing: state-of-art," International Journal of Big Data Intelligence, vol. 4, pp. 81--107, 2017.

[26] French-Baidoo, Asamoah and Oppong, "Achieving confidentiality in electronic health records using cloud systems," IJCNIS, vol. 10, p. 18, 2018.

[27] Morales, Gonzalez, Diaz and Sosa, "A pairing-based cryptographic approach for data security in the cloud," IJISP, vol. 17, pp. 441--461, 2018.

[28] Odelu, Rao, Kumari, Khan and Choo, "Pairing-based CP-ABE with constant-size ciphertexts and secret keys for cloud environment," Computer Standards & Interfaces, vol. 54, pp. 3--9, 2017.

[29] M. Mitzenmacher, "The power of two choices in randomized load balancing," IEEE Transactions on Parallel and Distributed Systems, vol. 12, pp. 1094--1104, 2001.

[30] P. Morales-Ferreira, M. Santiago-Duran, C. Gaytan-Diaz, J. Gonzalez-Compean, V. J. Sosa-Sosa and I. Lopez-Arevalo, "A data distribution service for cloud and containerized storage based on information dispersal," SOSE, pp. 86--95, 2018.

[31] iso.org, "ISO 22005:2007(es) Trazabilidad en la cadena de alimentos para alimentación humana y animal — Principios generales y requisitos básicos para el diseño e implementación del sistema," [Online]. Available: https://www.iso.org/obp/ui#iso:std:iso:22005:ed-1:v1:es. [Accessed 28 October 2022].

[32] aecoc.es, "AECOC: La Asociación de Fabricantes y Distribuidores," [Online]. Available: https://www.aecoc.es/comite/seguridad-alimentaria/. [Accessed 28 October 2022].

[33] B. a. F. F. a. C. V. a. B. M. a. C. N. a. M. K. Farahani, "Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare," Future Generation Computer Systems, vol. 78, pp. 659--676, 2018.

[34] P. a. A. G. a. G. R. a. C. G. a. F. G. a. L. A. Pace, "An edge-based architecture to support efficient applications for healthcare industry 4.0," IEEE Transactions on Industrial Informatics, vol. 15, pp. 481--489, 2018.

[35] prsarahevans, "Hyperledger Sawtooth: Blockchain para empresas," [Online]. Available: https://prsarahevans.com/cat-guias/hyperledger-sawtooth-blockchain-para-empresas/. [Accessed 2022 11 23].

[36] docker, "Develop faster. Run anywhere.," [Online]. Available: https://www.docker.com/.

[37] D. docs, "docker docs," [Online]. Available: https://docs.docker.com/compose/. [Accessed 2022 11 23].

[38] Banco de México, "Informe Trimestral. Abril-Juno 2021," 2021.

[39] F. Della Rosa, "Worldwide Software as a Service and Cloud Software Forecast, 2020–2024," International Data Corporation (IDC), 2020.

[40] J. a. M. B. a. V. W. Kelly Finnerty and Sarah Fullick and Helen Motha and Navin Shah, Cyber Security Breaches Survey 2019, Department for Digital, Culture, Media and Sport, 2019.

[41] C. H. a. W. C. a. L. Y. a. Y. D. a. S. J. a. Y. T. a. H. C. a. D. Chen, "Toward security as a service: A trusted cloud service architecture with policy customization," Journal of Parallel and Distributed Computing, vol. 149, pp. 76-88, 2021.

[42] J. a. R. V. Daemen, The Design of Rijndael: The Advanced Encryption Standard (AES), Springer Nature, 2020.

[43] M. Morales-Sandoval, J. L. Gonzalez-Compean, A. Diaz-Perez and V. J. Sosa-Sosa, "A Pairing-based Cryptographic Approach for Data Security in the Cloud," Int. J. Inf. Secur., vol. 17, p. 441–461, August 2018.

[44] F. B. a. S. H. a. T. Halevi, "Supporting private data on Hyperledger Fabric with secure multiparty computation," IBM Journal of Research and Development, vol. 63, no. 2/3, pp. 3-8, 2019.

[45] N. a. L. J. a. Z. Y. a. G. Y. Chen, "Efficient CP-ABE Scheme with Shared Decryption in Cloud Storage," IEEE Transactions on Computers, 2020.

[46] D. Giry, "NIST Report on Cryptographic Key Length and Cryptoperiod (2020)," Key length, 2020.

[47] E. a. B. E. a. B. W. a. P. W. a. S. M. a. o. Barker, "Recommendation for Key Management: Part 1-General," National Institute of Standards and Technology, Technology Administration, 2020.

[48] D. S.-G. D. D. G.-C. J. L. &. C. J. Carrizales-Espinoza, "FedFlow: a federated platform to build secure sharing and synchronization services for health dataflows.," Computing, pp. 1-19, 2022.

[49] D. G.-C. J. L. &. M.-S. M. Carrizales-Espinoza, "Zamna: a tool for the secure and reliable storage, sharing, and usage of large data sets in data science applications.," in In 2022 IEEE Mexican International Conference on Computer Science (ENC), 2022, August.

[50] E. a. H. J. J. a. O. A. a. R. D. a. S. G. a. T. H.-Y. Brynjolfsson, "COVID-19 and remote work: An early look at US data," National Bureau of Economic Research, 2020.

Muyal-Nez: Servicios Agnósticos para la Creación de Sistemas de Ciencia de Datos en e-Salud

Dante D. Sánchez-Gallegos[1][0000-0003-0944-9341], J. L. Gonzalez-Compean[1][0000-0002-2160-4407]. Ricardo Marcelín-Jiménez[2][0000-0002-5355-5830] y Ricardo Landa-Becerra[1][0000-0003-2645-0942]

[1] Cinvestav Tamaulipas, Cd. Victoria, México

[2] Universidad Autónoma Metropolitana-Iztapalapa, CDMX, México
{dante.sanchez, joseluis.gonzalez, ricardo.landa}@cinvestav.mx rmarcelin@izt.uam.mx,

Resumen. Los sistemas de ciencia de datos en e-salud se han convertido en una solución popular para que las organizaciones puedan convertir grandes volúmenes de datos en información para agilizar y ayudar en procesos de toma de decisiones críticos tales como diagnósticos, pronósticos e intervenciones de salud pública. Estos sistemas se componen de un conjunto de etapas, en donde se interconectan diferentes aplicaciones de análisis de datos y aprendizaje máquina, con herramientas de big data y almacenamiento/transporte de datos, lo cual permite manejar el ciclo de vida de los datos desde su adquisición (e.g., captura de una tomografía o imagen de rayos X de un paciente) hasta la entrega de los contenidos a los tomadores de decisiones a través de herramientas de visualización y consumo de datos. Sin embargo, construir servicios seguros y eficientes que permitan manejar este ciclo de vida de los datos no es sencillo, dado que múltiples aplicaciones homogéneas deben de ser integradas en un solo sistema que automáticamente realice el transporte y entrega de los datos a los tomadores de decisiones. En este capítulo se presenta Muyal-Nez, un *framework* orientado al diseño para la construcción de sistemas de ciencia de datos para el procesamiento de datos no estructurados. Muyal-Nez se compone de un conjunto de servicios agnósticos de la infraestructura que permiten disminuir la dependencia que las organizaciones pueden generar con proveedores en la nube, haciendo factible la creación de flujos intra e interinstitucionales para la compartición de datos.

Palabras clave: Servicios Agnósticos · Cómputo en la Nube · Sistemas de Ciencia de Datos · Big Data

1 Introducción

Los sistemas de salud inteligentes y conectados [1], [2] ayudan a los tomadores de decisiones (médicos o enfermeras) a agilizar el proceso de recolección de datos de los pacientes para emitir un diagnóstico o pronóstico. En este sentido, este tipo de sistemas se apoyan en tecnologías utilizadas en la ciencia de datos para el manejo y adquisición de datos en crudo, y su posterior transformación en información. Dentro de estas tecnologías se incluyen la inteligencia artificial [3], cómputo en la nube [4], big data [5] y el internet de las cosas

(IoT, por sus siglas en inglés) [6], [7]. La integración de estas tecnologías en un solo sistema de ciencia de datos es clave para implementar herramientas que faciliten tareas tales como el continuo monitoreo y la asistencia de los pacientes [1], [2], [8],[9], así como para soportar procesos de toma de decisiones tales como pronósticos médicos, diagnósticos e intervenciones de salud pública [10], [11], [12], [13].

Un sistema de ciencia de datos convierte, a través de múltiples etapas [5], [14], datos médicos en crudo en información útil y conocimiento accionable. Durante este flujo los datos son extraídos y adquiridos desde una fuente, y se producen grandes volúmenes de contenidos médicos (e.g., tomografías, mamografías, electrocardiogramas y espirometrías). Estos contenidos son procesados por un conjunto de aplicaciones de inteligencia artificial, análisis de datos y aprendizaje máquina [12]. Comúnmente, un sistema de ciencia de datos es materializado en la forma de un flujo de trabajo o estructura de procesamiento, en los cuales las aplicaciones del sistema son organizadas en la forma de un grafo acíclico dirigido (DAG, por sus siglas en inglés) [15]. En un DAG, los nodos representan las aplicaciones utilizadas para el procesamiento de datos en un sistema de ciencia de datos, mientras que las aristas simbolizan las interconexiones que hay entre estos nodos. En un DAG, no hay ciclos y las aristas tienen una dirección dada, de tal forma que no hay un camino que directamente empiece y terminé un mismo nodo.

Los hallazgos e información obtenida de un sistema de ciencia de datos son entregadas a los tomadores de decisiones a través de herramientas de visualización de datos [16], [17]. Comúnmente, los datos que se procesan en un sistema de ciencia de datos pueden ser clasificados en datos estructurados y no estructurados [5]. Los datos estructurados son todos aquellos que se pueden organizar como registros de una tabla e incluyen números, cadenas de texto y fechas. Por otra parte, los datos no estructurados no se pueden organizar en filas y columnas, e incluyen datos tales como imágenes, documentos y audios. Los datos no estructurados suelen ser de mayor tamaño que los estructurados y, en este sentido, se estima que el 80% de los datos almacenados a nivel mundial son no estructurados.

En años recientes, la nube se ha vuelto una solución popular para manejar y desplegar sistemas de ciencia de datos para el manejo y procesamiento de datos no estructurados [18], [19], [20]. Por lo anterior, diferentes herramientas nativas de la nube han sido desarrolladas para que las organizaciones puedan manejar sus datos médicos mediante el acoplamiento de diferentes aplicaciones en la nube [21], [22], [23]. Sin embargo, el uso de la nube, sobre todo de la pública, tiene implicaciones de confidencialidad y seguridad de datos que las organizaciones deben de observar para que sus datos no sean accedidos por terceros. Además, el uso de estas herramientas crea una dependencia entre las organizaciones y los *frameworks* e infraestructura provista por los proveedores en la nube [24] [24]. Sumado a lo anterior, en escenarios reales comúnmente es requerida la participación de múltiples organizaciones médicas para el manejo y procesamiento de datos a través de diferentes infraestructuras (e.g., en un clúster privado, en la nube o en equipos de cómputo en el borde) [25], [26], [27], lo cual no es considerado por estas herramientas. Otro aspecto que debe de ser considerado es el manejo de requerimientos no funcionales. Un requerimiento no funcional describe características de calidad de un servicio o aplica-

ción [28]. Estos requerimientos no son obligatorios de cumplir, dado que son parte de la función principal del sistema. Por ejemplo, estos requerimientos son agregados para hacer frente a cuestiones de eficiencia, seguridad (en términos de confidencialidad, control de acceso e integridad), así como tolerancia a fallos, los cuales son cruciales en el manejo de datos médicos [29], [30], que se consideran datos sensibles.

En el presente capítulo se describe Muyal-Nez[1], un framework orientado al diseño para la construcción de sistemas de ciencia de datos en e-salud. Este framework se compone de diferentes servicios, los cuales son agnósticos de la infraestructura y que permiten a las organizaciones construir sistemas de ciencia de datos mediante el diseño de un grafo acíclico dirigido que incluye el encadenamiento de diferentes aplicaciones de análisis de datos y aprendizaje automático. El diseño de este grafo no requiere niveles avanzados de conocimiento en programación y su despliegue se realiza de forma automática mediante Muyal-Nez, el cual se encargará, en tiempo de despliegue, de realizar el acoplamiento de las aplicaciones y de crear un flujo continuo de datos en tiempo de ejecución.

El resto del capítulo se encuentra organizado de la siguiente manera: en la Sección 2 se presentan diferentes conceptos básicos que serán mencionados a lo largo de este capítulo; en la Sección 3 se describe el diseño de Muyal-Nez, así como sus principales características; finalmente, en la Sección 4 se presentan las principales conclusiones de este trabajo.

2 Conceptos básicos

En la presente sección se describen algunos conceptos claves utilizados a lo largo de este documento.

2.1 ¿Qué es la nube?

La nube es un modelo de cómputo que se ha vuelto muy popular en los últimos años para permitir el despliegue, manejo y almacenamiento de datos a gran escala mediante el uso de recursos virtualizados. En lugar de adquirir servidores y equipos de cómputo físicos, las organizaciones acceden, a través de Internet, a una piscina de recursos de software y hardware virtualizados [31]. En la literatura se pueden distinguir tres principales tipos de nubes: pública, privada e hibrida.

En la Fig. 1 se observa una representación de cómo la nube provee servicios (e.g., transporte de datos, bases de datos, almacenamiento, procesamiento y mensajería) a diferentes usuarios, los cuales acceden bajo demanda a estos servicios, mediante un dispositivo (e.g., teléfonos, laptops, televisiones, tablets o servidores).

[1] El nombre de Nez está inspirado en Nezahualcóyotl, ingeniero de la época precolombina en México que introdujo diferentes técnicas tales como el uso de estructuras de pilares.

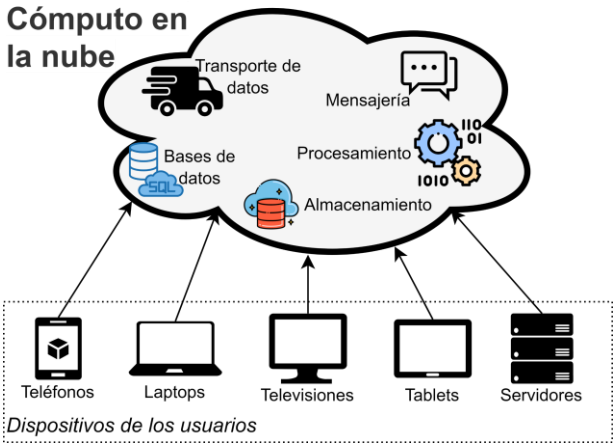


Figura 1. Representación conceptual de cómo la nube provee diferentes servicios de manejo de datos a un conjunto de dispositivos.

Nube pública. La infraestructura en la nube es provista para uso abierto del público en general. Esta infraestructura puede pertenecer y ser manejada y operada por un negocio u organizaciones académicas y gubernamentales, o cualquier combinación de estos. En este sentido, los recursos virtualizados son desplegados en los recursos físicos de un proveedor en la nube (e.g., Amazon EC2, Google Cloud, Microsoft Azure, etc.) [31].

Nube privada. La infraestructura de la nube es provista para uso exclusivo de una sola organización que cuenta con múltiples consumidores (e.g., unidades de negocio). La infraestructura pertenece a la organización que está a cargo de su manejo y operación [31].

Nube hibrida. La infraestructura de la nube es una composición de dos o más infraestructuras en la nube diferentes (privadas, comunitarias, o públicas) que permanecen como entidades únicas, pero que son agrupadas como una sola por una tecnología estandarizada o propietaria, la cual permite la portabilidad de los datos y aplicaciones [31].

2.2 ¿Qué significa ser agnóstico de la infraestructura?

El diccionario de Oxford describe el termino agnóstico como “las creencias de alguien que no cree, o que piensa que es imposible conocer que un dios exista”. En la literatura, y en específico en el ámbito de computación, este término ha sido utilizado para describir una aplicación, sistema o servicio diseñado para ajustar sus parámetros de despliegue (contexto) para que puedan ser desplegados exitosamente en diferentes infraestructuras sin importar sus características.

De forma similar, el agnosticismo de la infraestructura hace referencia a la propiedad de un servicio o aplicación para desplegarse y ejecutarse exitosamente cuando son desconocidos los detalles de la infraestructura donde la aplicación/servicio son desplegadas, así como los detalles de la carga de trabajo entrante [32]. En este sentido, estas aplicaciones/servicios determinan en tiempo de ejecución y de forma automática las limitaciones y capacidades de la infraestructura.

En Muyal-Nez esta propiedad es agregada a los sistemas de ciencia de datos para permitir que estos puedan ser desplegados en múltiples infraestructuras y por múltiples organizaciones, sin que éstas tengan que modificar el código de las aplicaciones a desplegar en el sistema de ciencia de datos.

2.3 ¿Qué es una arquitectura de microservicios?

Una arquitectura de microservicios [33] es un enfoque conceptual en el que las aplicaciones son desarrolladas como un conjunto de servicios pequeños, bajamente acoplados y que pueden conectarse con otros servicios para trabajar juntos. Cada uno de estos servicios son llamados microservicios, los cuales son procesos autónomos que pueden ser programas en cualquier lenguaje de programación y manejados de forma independiente.

En una arquitectura de microservicios sus componentes (microservicios) son distribuidos a través de diferentes infraestructuras e interactúan entre sí a través de mensajes. Cada uno de estos microservicios es independiente y sigue una coreografía en donde se asume que no hay centralización y que utilizan eventos y mecanismos de publicación/suscripción con el fin de establecer una colaboración entre ellos.

2.4 ¿Qué es un contenedor virtual?

Un contenedor virtual es una unidad de software que empaqueta el código y dependencias (e.g., bibliotecas del sistema, paquetes de terceros o archivos de configuración) de una aplicación. Estos contenedores tienen las características de que son ligeros, en términos de espacio de almacenamiento y en comparación con máquinas virtuales tradicionales, así como que se encuentran aislados del resto de aplicaciones en el sistema operativo anfitrión [34].

En la actualidad, la plataforma de contenedores más popular es Docker [35], la cual permite encapsular aplicaciones mediante la escritura de archivos de configuración llamados Dockerfiles, los cuales contienen una serie de comandos para generar una imagen de contenedor. En la Figura 2 se muestra un ejemplo de un Dockerfile para la encapsulación de una aplicación escrita en el lenguaje de programación Python 3.7.6.

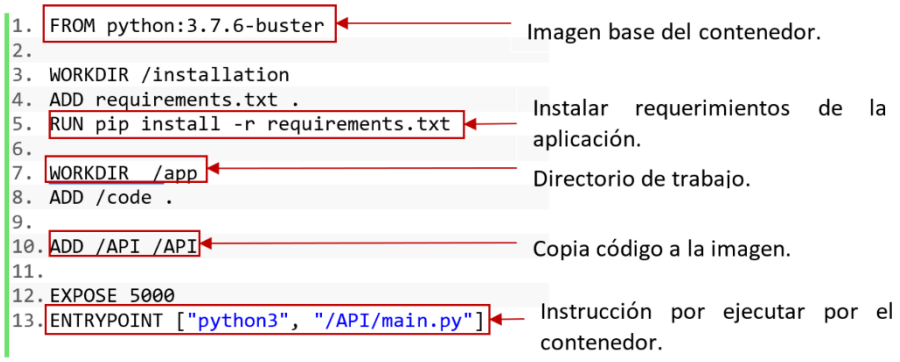


Figura 2. Ejemplo de un archivo Dockerfile.

2.5 ¿Qué es un patrón de paralelismo?

Los patrones de paralelismo permiten que una solución aplicada a un problema en un contexto específico pueda ser aplicada para resolver problemas similares en diferentes contextos. Tradicionalmente, estos patrones se utilizan para procesar grandes volúmenes de forma paralela clonando una aplicación. En estos patrones, el código de las aplicaciones no tiene que ser modificado para paralelizarlas. Ejemplos de patrones paralelismo son el patrón manejador/trabajador, divide&vencerás, y tuberías [25].

Patrón manejador/trabajador. Patrón de paralelismo de tareas el cual es obtenido mediante la creación de n clones de una aplicación los cuales son organizados en la forma de *trabajadores* y que son controlados por una entidad llamada *manejador*. En la Figura 5 se muestra una representación de este patrón, en donde el manejador lee los metadatos de los contenidos en una fuente de datos y distribuye los contenidos entre los n trabajadores utilizando una técnica de balanceo de carga. Los trabajadores procesan los contenidos produciendo nuevas versiones de estos, las cuales son entregados a un repositorio de datos o a otra aplicación que procesará los datos.

Patrón divide&vencerás. Patrón de paralelismo de datos, el cual es obtenido cuando múltiples trabajadores realizan la misma tarea de procesamiento en diferentes segmentos de un contenido. En la Figura 4 se observa la representación conceptual de este patrón, donde cada contenido de entrada es dividido en n segmentos y cada segmento es procesado de forma independiente por un trabajador que puede estar desplegado en un nodo de computación o procesador diferente. En este sentido, en este patrón, un *segmentador* es requerido para generar los segmentos que serán distribuidos a los trabajadores. Además, es requerido un *consolidador* para mezclar las salidas de los trabajadores.

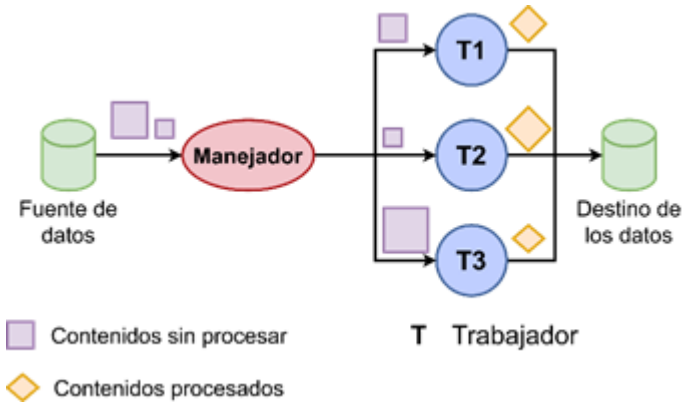


Figura 3. Representación conceptual de un patrón manejador/trabajador.

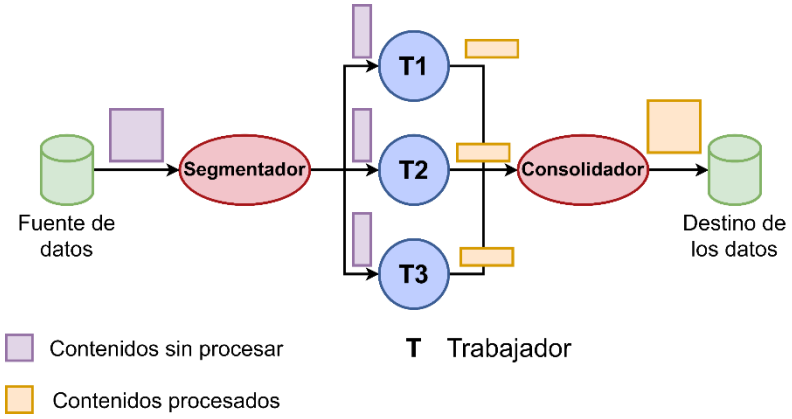


Figura 4. Representación conceptual de un patrón divide&vencerás.

Patrón de tubería. Múltiples aplicaciones independientes son encadenadas y ejecutadas de forma paralela. La salida de una aplicación es la entrada de la siguiente aplicación en la tubería.

3 Construcción de sistemas de ciencia de datos con Muyal-Nez

En la presente sección se describen los principios de diseño de Muyal-Nez, un framework que permite la construcción de sistemas de ciencia de datos en e-salud. Muyal-Nez y las soluciones generadas con éste tienen la característica de que son agnósticos de la infraestructura y, por lo tanto, permiten a las organizaciones interconectar las diferentes aplicaciones consideradas en el manejo del ciclo de vida de sus datos y desplegarlas en diferentes infraestructuras para generar flujos intra e interinstitucionales.

3.1 Bloques de construcción de sistemas de ciencia de datos

En Muyal-Nez, la unidad básica de construcción de un sistema de ciencia de datos de salud es una abstracción llamada bloque de construcción (BC). Un BC es un componente de software genérico y agnóstico de la infraestructura, el cual es construido en la forma de un contenedor virtual que encapsula una aplicación para procesar y manejar datos médicos. En este sentido, un BC incluye el código o binario de la aplicación, sus librerías y bibliotecas, archivos de configuración, así como una versión reducida del sistema operativo que la aplicación requiere para funcionar. En la Fig. 5 se muestra la representación conceptual de un BC utilizado en Muyal-Nez para el manejo de las aplicaciones. Como se puede observar, los BC pueden ser conectados con una fuente y destino de datos utilizando sus interfaces de entrada y salida.

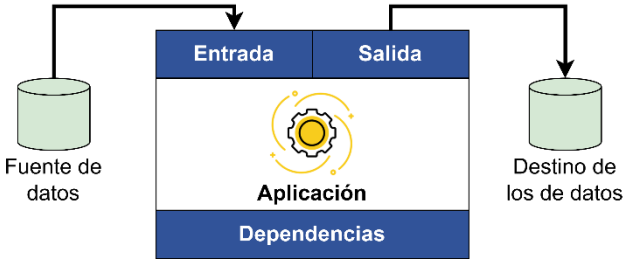


Figura 5. Representación conceptual de un bloque de construcción en Muyal-Nez.

Los BC pueden ser desplegados en diferentes infraestructuras disponibles, por ejemplo, en una computadora personal, un servidor, un clúster de computadoras o en una máquina virtual en la nube.

En tiempo de ejecución, este bloque de construcción es manejado como un microservicio, el cual implementa interfaces de entrada y salida que permiten el intercambio de datos con otros microservicios e incluso con aplicaciones externas o manejadas por terceros.

En Muyal-Nez los desarrolladores pueden crear sus propios bloques de construcción mediante el diseño de archivos Dockerfiles, en los cuales se declara la imagen base de la aplicación (esto puede ser una imagen que ya tenga instalado por defecto alguna versión en específico del lenguaje de programación requerido por la aplicación para funcionar o un sistema operativo base), el código o ejecutable de la aplicación, así como su ubicación dentro del contenedor. Adicionalmente, en el archivo Dockerfile se deben agregar las estructuras de Muyal-Nez, requeridas para que el BC se pueda comunicar con los servicios del framework.

La creación de estos bloques que incluyen todas las dependencias de una aplicación, así como interfaces comunes de entrada y salida, hace factible la creación de sistemas agnósticos de la infraestructura, los cuales pueden ser desplegados en múltiples equipos de cómputo sin modificar el código de las aplicaciones (ver Fig. 6). Lo anterior permite que las organizaciones reutilicen

bloques de construcción previamente creados por otros miembros de su organización o por otras organizaciones.

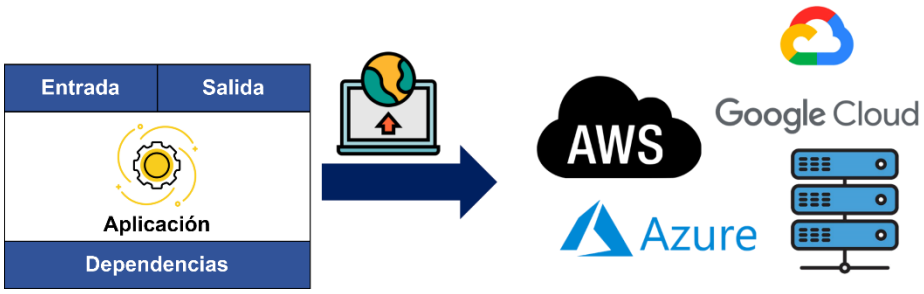


Figura 6. La propiedad de agnosticismo agregada a los bloques de construcción de Muyal-Nez, permite que estos puedan ser desplegados en múltiples infraestructuras.

3.2 Diseño de servicios de ciencia de datos para e-salud

Los bloques de construcción diseñados en Muyal-Nez pueden ser interconectados entre sí para crear estructuras más complejas, tales como servicios de ciencia de datos que permitan el manejo y procesamiento de datos a través de su ciclo de vida, desde su adquisición hasta hacerlos disponibles a tomadores de decisiones (e.g., médicos o enfermeras). En este sentido, la construcción de un servicio de ciencia de datos con Muyal-Nez es realizada mediante la unión lógica de diferentes BC, los cuales pueden estar desplegados en múltiples infraestructuras. Esto es similar a armar un rompecabezas, donde diferentes bloques más pequeños se unen para formar una estructura más compleja.

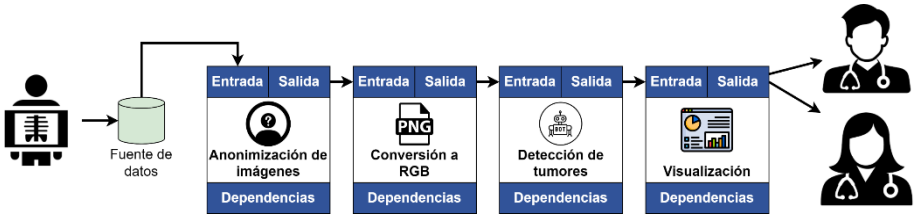


Fig. 7. Ejemplo de un sistema de ciencia de datos construido con Muyal-Nez.

Por ejemplo, en la Fig. 7 se muestra un sistema de ciencia de datos para la detección de tumores en tomografías de pulmón. Este sistema considera cuatro BC para la anonimización de las tomografías, su conversión a una representación RGB en formato PNG, el análisis de los PNG resultantes con una red neuronal para la detección de tumores y, finalmente, los resultados son puestos a disposición en un sistema de visualización, el cual es utilizado por el

tomador de decisiones para observar los hallazgos encontrados por el sistema de ciencia de datos, los cuales le pueden apoyar en la elaboración de un diagnóstico. Observe que, en tiempo de ejecución, el acoplamiento de los BC es realizado de forma automática por Muyal-Nez a partir de un diseño entregado por los diseñadores, el cual es realizado utilizando una interfaz gráfica para elegir e interconectar los BC o mediante el desarrollo de código siguiendo un esquema declarativo y acoplamiento de BC.

En el caso de utilizar la interfaz gráfica de Muyal-Nez, el desarrollador tiene acceso a un catálogo de BC, de los cuales puede elegir aquellos que más le interesan para crear sus sistemas de ciencia de datos. En el último paso de este diseño, el desarrollador deberá ordenar sus BC como un grafo acíclico dirigido, indicando el orden de ejecución de los BC y sus interconexiones.

En tiempo de despliegue, el grafo y los BC elegidos por el desarrollador son entregados a una entidad llamada "lanzador", la cual está a cargo de generar tres archivos de configuración a partir de este grafo. El primero de ellos es un archivo que sigue un lenguaje declarativo y que incluye todos los BC seleccionados por el usuario, así como sus interconexiones. El segundo es un archivo en formato YAML, que incluye la configuración de los contenedores virtuales utilizados para el despliegue de las aplicaciones y que, en tiempo de ejecución, serán manejados como microservicios. Finalmente, en el tercer archivo se incluyen una declaración de las firmas digitales de cada contenedor virtual desplegado, así como la ubicación de sus entradas y salidas.

3.3 Patrones de paralelismo implícitos para el manejo eficiente de datos

Muyal-Nez permite a los diseñadores agregar patrones de paralelismo implícito (e.g., manejador/trabajador o divide&vencerás) a los sistemas de ciencia de datos. El objetivo es permitir el procesamiento eficiente de los datos, replicando aquellos bloques de construcción que puedan generar un cuello de botella en el flujo de datos.

En la Figura 8 se presenta un ejemplo de un sistema de ciencia de datos construido con Muyal-Nez, el cual incluye un patrón manejador/trabajador. En este patrón, el BC para la detección de tumores es replicado tres veces y se agrega una entidad llamada *manejador*, la cual se encarga de distribuir los datos de entrada, producidos por el BC de conversión de imágenes a RGB, entre los trabajadores disponibles.

En tiempo de diseño, los desarrolladores/diseñadores solo tienen que indicar el BC a paralelizar, el tipo de patrón a utilizar (manejador/trabajador o divide&vencerás) y el número de trabajadores en el patrón. En tiempo de despliegue, Muyal-Nez recibirá esta configuración y automáticamente creará los clones de los BC y los organizará como un patrón. Finalmente, en tiempo de ejecución, el manejador automáticamente distribuirá los datos entre los trabajadores. En Muyal-Nez, el manejador incluye un balanceador de carga conocido como *Two Choices* [24], en el cual dos BC del patrón son elegidos aleatoriamente y aquel BC con la menor carga de trabajo será elegido para procesar el contenido de entrada.

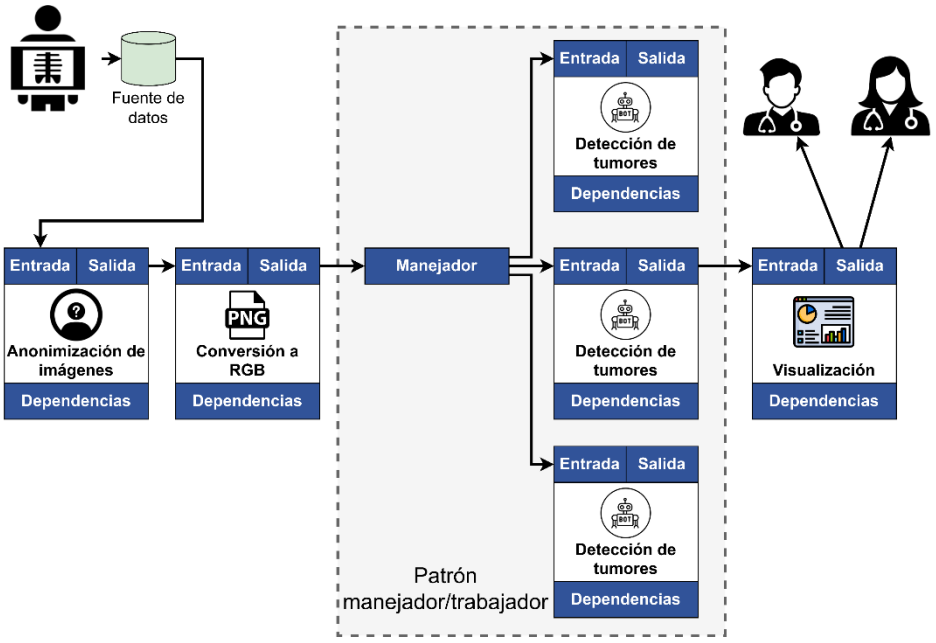


Figura 8. Ejemplo de un patrón manejador/trabajador construido con Muyal-Nez.

3.4 Conexión con servicios de seguridad y transporte de datos

Muyal-Nez es parte de una plataforma llamada Muyal-Ilal[2], la cual provee la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud. En este contexto, Muyal-Chimalli y Muyal-Painal son dos herramientas incluidas en Muyal-Ilal para la gestión del transporte y almacenamiento seguro, eficiente y tolerante a fallos de datos sensibles.

Muyal-Chimalli incluye servicios que permiten agregar confidencialidad, integridad y control de acceso a los datos manejados por las aplicaciones incluidas en los bloques de construcción de Muyal-Nez. Además, Muyal-Chimalli incluye un esquema de trazabilidad que crea de forma automática una red de verificabilidad (blockchain), lo cual permite registrar eficientemente cada operación realizada por Muyal-Nez, permitiendo la identificación de modificaciones no autorizadas sobre los datos. Además, Muyal-Ilal incluye un sistema de verificación del cumplimiento de las normas y protocolos oficiales.

Por su parte, Muyal-Painal permite la construcción de sistemas de almacenamiento eficientes y tolerante a fallos. Estos sistemas permiten a Muyal-Nez la distribución, en forma automática y transparente, de los datos requeridos por los bloques de construcción de un sistema de ciencia de datos. Además, Muyal-Painal incluye un sistema de sincronización automática de datos entre organizaciones y usuarios, lo cual minimiza los costos de envío y almacena-

[2] http://adaptivez.org.mx/e-SaludData/

miento de información, mejorando la experiencia de servicio del usuario final (paciente o profesional de la salud).

3.5 Flujos de datos intra e interinstitucionales para la compartición automática de datos

Muyal-Nez permite la creación de dos tipos de flujo de datos: intra e interinstitucionales. Los flujos de datos intrainstitucionales permiten la interconexión de diferentes participantes dentro de un hospital u organización. Por ejemplo, un departamento de radiología enviando tomografías a un oncólogo. Estos servicios automáticamente realizan el manejo, entrega y procesamiento de los datos al interior de la organización.

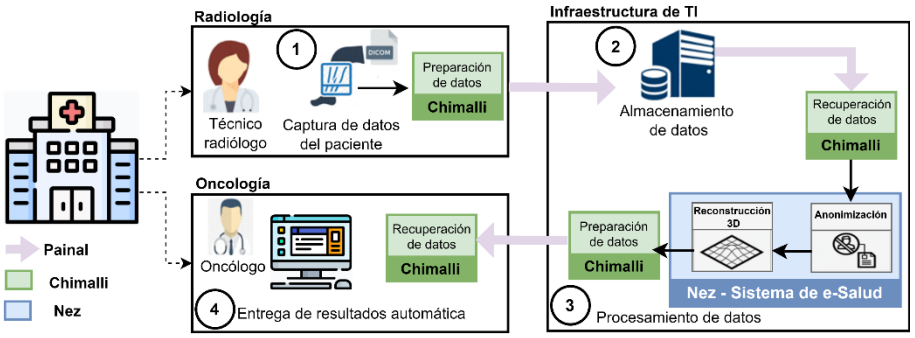


Figura 9. Ejemplo de un flujo de datos intrainstitucional.

En la Figura 9 se muestra un ejemplo de un flujo de datos intrainstitucional, el cual interconecta un departamento de radiología con uno de oncología. Como se puede observar, los datos son asegurados con Muyal-Chimalli y transportados con Muyal-Painal. Además, las imágenes adquiridas son procesadas con BC de Muyal-Nez, los cuales incluyen aplicaciones para la anonimización de las imágenes y la reconstrucción 3D de las mismas.

Por su parte, los flujos inter-institucionales permiten la conexión de múltiples instituciones de salud y hospitales. Las metas principales de estos flujos son:

1. Permitir la compartición de datos entre individuos pertenecientes a diferentes organizaciones;
2. Permitir que las organizaciones puedan compartir recursos para procesar y almacenar datos de forma distribuida.

En la Figura 10 se muestra un ejemplo de un flujo interinstitucional. En este ejemplo, un radiólogo de un hospital captura imágenes de rayos X a los pacientes y las envía a un oncólogo en otro hospital para su análisis y un diagnóstico. Durante el transporte de las imágenes, éstas son procesadas por

un servicio de ciencia de datos para identificar tumores en las imágenes. Además, se incluyen BC que cuentan con mecanismos para asegurar la integridad y confidencialidad de los datos, garantizando que terceros no puedan acceder a las imágenes médicas durante su transporte y almacenamiento.

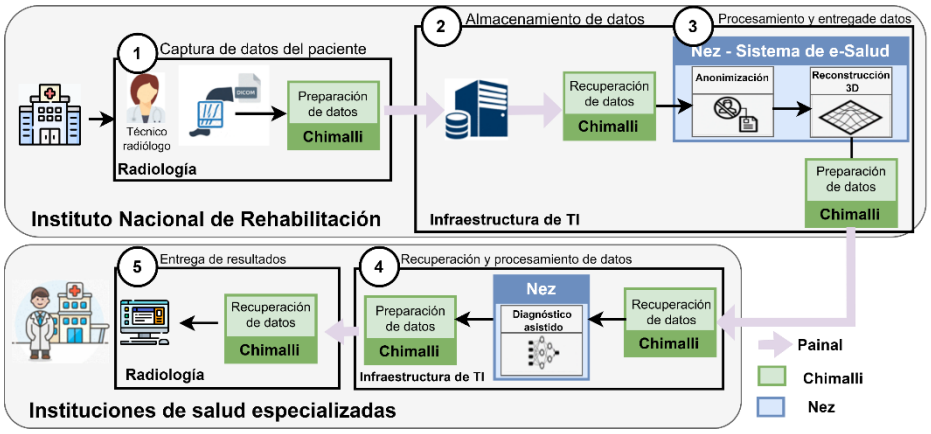


Figura 10. Ejemplo de un flujo de datos interinstitucional.

4 Conclusiones

En el presente capítulo se presentó Muyal-Nez, un framework para la construcción de sistemas de ciencia de datos para el manejo y procesamiento de datos médicos. Este framework se compone de un conjunto de servicios agnósticos de la infraestructura, lo cual permite que las soluciones generadas con éste puedan ser desplegadas en múltiples infraestructuras. Además, Muyal-Nez es un facilitador para que las organizaciones puedan desarrollar sus propios sistemas de ciencia de datos sin utilizar herramientas de terceros, las cuales pueden generan una dependencia con una infraestructura en específico.

Muyal-Nez tiene conexiones con software y herramientas para el transporte seguro y eficiente de los datos. Lo anterior permite crear flujos intra e interinstitucionales que conecten diferentes departamentos al interior de una organización, así como entre distintas organizaciones. Cabe resaltar que Muyal-Nez se puede utilizar de manera independiente al resto de servicios de Muyal-Ilal y que, gracias a sus interfaces de entrada y salida, permite la interconexión de los bloques de Muyal-Nez con otras aplicaciones y servicios externos.

Agradecimientos

Este trabajo forma parte del proyecto 41756 “Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud'” del CONACYT-PRONACES.

Referencias

[1] Leroy, G., Chen, H., & Rindflesch, T. C. (2014). Smart and Connected Health [Guest editors' introduction]. *IEEE Intelligent Systems*, *29*(3), 2-5.

[2] Carroll, N. (2016). Key success factors for smart and connected health software solutions. Computer, 49(11), 22-28.

[3] Prakash, A. V., & Das, S. (2021). Medical practitioner's adoption of intelligent clinical diagnostic decision support systems: A mixed-methods study. Information & Management, 58(7), 103524.

[4] Kuo, M. H. (2011). Opportunities and challenges of cloud computing to improve health care services. Journal of medical Internet research, 13(3), e1867.

[5] Chang, W. L., & Grady, N. (2019). Nist big data interoperability framework: Volume 1, definitions.

[6] Rose, K., Eldridge, S., & Chapin, L. (2015). The internet of things: An overview. The internet society (ISOC), 80, 1-50.

[7] Goad, D., Collins, A. T., & Gal, U. (2021). Privacy and the Internet of Things– An experiment in discrete choice. Information & Management, 58(2), 103292.

[8] Bayoumy, K., Gaber, M., Elshafeey, A., Mhaimeed, O., Dineen, E. H., Marvel, F. A., ... & Elshazly, M. B. (2021). Smart wearable devices in cardiovascular care: where we are and how to move forward. Nature Reviews Cardiology, 18(8), 581-599.

[9] Yi, H., Li, J., Lin, Q., Wang, H., Song, H., Ming, Z., & Nie, Z. (2019). A rainbow-based authentical scheme for securing smart connected health systems. Journal of Medical Systems, 43(8), 1-10.

[10] Bao, Y., Chen, Z., Wei, S., Xu, Y., Tang, Z., & Li, H. (2019). The state of the art of data science and engineering in structural health monitoring. Engineering, 5(2), 234-242.

[11] Cannataro, M., dos Santos, R. W., Sundnes, J., & Veltri, P. (2012). Advanced computing solutions for health care and medicine. Journal of Computational Science, 3(5), 250-253.

[12] Wang, Y., Kung, L., Wang, W. Y. C., & Cegielski, C. G. (2018). An integrated big data analytics-enabled transformation model: Application to health care. Information & Management, 55(1), 64-79.

[13] Purawat, S., Cowart, C., Amaro, R. E., & Altintas, I. (2017). Biomedical Big Data Training Collaborative (BBDTC): An effort to bridge the talent gap in biomedical science and research. Journal of computational science, 20, 205-214.

[14] Sarker, I. H. (2021). Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. SN Computer Science, 2(5), 1-22.

[15] Shi, J., & Lu, J. (2021, April). Performance models of data parallel DAG workflows for large scale data analytics. In 2021 IEEE 37th International Conference on Data Engineering Workshops (ICDEW) (pp. 104-111). IEEE.

[16] Bashir, S., Qamar, U., Khan, F. H., & Naseem, L. (2016). HMV: a medical decision support framework using multi-layer classifiers for disease prediction. Journal of Computational Science, 13, 10-25.

[17] Rasmussen, S. A., & Jamieson, D. J. (2020). Public health decision making during Covid-19—Fulfilling the CDC pledge to the American people. New England Journal of Medicine, 383(10), 901-903.

[18] Sobhy, D., El-Sonbaty, Y., & Abou Elnasr, M. (2012, December). MedCloud: healthcare cloud computing system. In 2012 International Conference for Internet Technology and Secured Transactions (pp. 161-166). IEEE.

[19] Abdelazeem, M., Elamin, A., Afifi, A., & El-Rabbany, A. (2021). Multi-sensor point cloud data fusion for precise 3D mapping. The Egyptian Journal of Remote Sensing and Space Science, 24(3), 835-844.

[20] Griebel, L., Prokosch, H. U., Köpcke, F., Toddenroth, D., Christoph, J., Leb, I., ... & Sedlmayr, M. (2015). A scoping review of cloud computing in healthcare. BMC medical informatics and decision making, 15(1), 1-16.

[21] Zheng, C., & Thain, D. (2015, June). Integrating containers into workflows: A case study using makeflow, work queue, and docker. In Proceedings of the 8th International Workshop on Virtualization Technologies in Distributed Computing (pp. 31-38).

[22] Babuji, Y., Woodard, A., Li, Z., Katz, D. S., Clifford, B., Kumar, R., ... & Chard, K. (2019, June). Parsl: Pervasive parallel programming in python. In Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing (pp. 25-36).

[23] Chard, R., Babuji, Y., Li, Z., Skluzacek, T., Woodard, A., Blaiszik, B., ... & Chard, K. (2020, June). Funcx: A federated function serving fabric for science. In Proceedings of the 29th International symposium on high-performance parallel and distributed computing (pp. 65-76).

[24] Sanchez-Gallegos, D. D., Gonzalez-Compean, J. L., Carretero, J., Marin, H., Tchernykh, A., & Montella, R. (2022). PuzzleMesh: A puzzle model to build mesh of agnostic services for edge-fog-cloud. IEEE Transactions on Services Computing.

[25] Sánchez-Gallegos, D. D., Galaviz-Mosqueda, A., Gonzalez-Compean, J. L., Villarreal-Reyes, S., Perez-Ramos, A. E., Carrizales-Espinoza, D., & Carretero, J. (2020). On the continuous processing of health data in edge-fog-cloud computing by using micro/nanoservice composition. IEEE Access, 8, 120255-120281.

[26] Das, J., Ghosh, S., Mukherjee, A., Ghosh, S. K., & Buyya, R. (2022). RESCUE: Enabling green healthcare services using integrated IoT-edge-fog-cloud computing environments. Software: Practice and Experience.

[27] Mukherjee, A., Ghosh, S., Behere, A., Ghosh, S. K., & Buyya, R. (2021). Internet of Health Things (IoHT) for personalized health care using integrated edge-fog-cloud network. Journal of Ambient Intelligence and Humanized Computing, 12(1), 943-959.

[28] Chung, L., & Prado Leite, J. C. S. D. (2009). On non-functional requirements in software engineering. In Conceptual modeling: Foundations and applications (pp. 363-379). Springer, Berlin, Heidelberg.

[29] Zhang, X., Liu, S., Chen, X., Wang, L., Gao, B., & Zhu, Q. (2018). Health information privacy concerns, antecedents, and information disclosure intention in online health communities. Information & Management, 55(4), 482-493.

[30] Keshta, I., & Odeh, A. (2021). Security and privacy of electronic health records: Concerns and challenges. Egyptian Informatics Journal, 22(2), 177-183.

[31] Mell, P., & Grance, T. (2011). The NIST definition of cloud computing.

[32] Ranabahu, A., Anderson, P., & Sheth, A. (2011). The cloud agnostic e-science analysis platform. IEEE Internet Computing, 15(6), 85-89.

[33] Dragoni, N., Giallorenzo, S., Lafuente, A. L., Mazzara, M., Montesi, F., Mustafin, R., & Safina, L. (2017). Microservices: yesterday, today, and tomorrow. Present and ulterior software engineering, 195-216.

[34] Sharma, P., Chaufournier, L., Shenoy, P., & Tay, Y. C. (2016, November). Containers and virtual machines at scale: A comparative study. In Proceedings of the 17th international middleware conference (pp. 1-13).

[35] Rad, B. B., Bhatti, H. J., & Ahmadi, M. (2017). An introduction to docker and analysis of its performance. International Journal of Computer Science and Network Security (IJCSNS), 17(3), 228.

Xelhua: una plataforma para la creación de sistemas de ciencia de datos bajo demanda

J. Armando Barrón-Lugo[1][0000−0002−9619−8116], José Carlos Morín-García[1][0000−0003−1327−4409], José Lui González-Compeán[1][0000−0002−2160−4407], Ivan Lopez-Arevalo[1][0000−0002−7464−8438]

[1]Centro de Investigación y de Estudios Avanzados (Cinvestav) del IPN Unidad Tamaulipas, Cd. Victoria 87130, Tamaulipas , México
{juan.barron, jose.morin, joseluis.gonzalez, ilopez} @cinvestav.mx

Resumen La ciencia de datos es un conjunto de procesos, técnicas y métodos científicos para la extracción de conocimiento de conjuntos de datos para la toma decisiones. Actualmente, existen múltiples herramientas para extraer información de datos estructurados; no obstante, su uso puede llegar a ser confuso para usuarios menos experimentados, requerir recursos de cómputo dedicados, complicadas instalaciones, o la perdida del control sobre sus datos al usar servicios de proveedores en la nube (*Vendor lock-in*). En este trabajo presentamos *Xelhua*, una plataforma para la creación de sistemas de ciencia de datos bajo demanda y enfocada al diseño que permite a usuarios sin amplios conocimientos en programación realizar análisis de datos estructurados. *Xeluha* permite transformar aplicaciones a servicios y, posteriormente, al despliegue automático de sistemas de ciencia de datos en diferentes infraestructuras de cómputo bajo demanda, mitigando las dependencias con proveedores y evitando escenarios de *vendor lock-in*.

Palabras clave: Manejo de Datos · Ciencia de Datos · Analítica de Datos · Microservicios · Orquestación de Datos.

1. Introducción

Actualmente estamos viviendo en la era de la conexión de información; el uso del teléfono móvil, las redes sociales, nuestras computadoras personales, incluso nuestros coches, todo se encuentra conectado, capturando y almacenando datos para posteriormente analizarlos y producir información y conocimiento útil que ayude en la toma de decisiones [12]. Los resultados de esto se pueden observar en muchos ámbitos de la vida diaria: esa canción recomendada por Spotify, las rutas alternas para llegar a tu trabajo proporcionada por tu teléfono, el pronóstico del clima... todo es resultado de un proceso de análisis de grandes cantidades de datos. Pero, ¿cómo pueden estas aplicaciones procesar tantos datos?

Desde el punto de vista tecnológico, el procesar datos no es una tarea sencilla. Realmente es un conjunto de tareas encadenadas que componen un proceso complejo [6], [3]. Un ejemplo para, inicialmente, comprender la dificultad de analizar grandes cantidades de datos es el intentar abrir un archivo de datos en el formato más comúnmente usado, CSV (*comma separated values*), de 5 GB de tamaño en una computadora personal. En la mayoría de los casos, a menos que se tengan 16 GB o más de memoria RAM, la computadora quedará congelada, incapaz de abrir el archivo, teniendo como única alternativa presionar el botón de apagado para reiniciarla. Esto se debe a que la computadora no tiene la suficiente memoria RAM para cargar el archivo. Para mitigar esta situación se requieren muchos más recursos de cómputo; dependiendo de la complejidad del análisis a realizar se puede requerir más RAM, CPU, almacenamiento o todos juntos [2]. Para el adecuado suministro de recursos de cómputo que permitan el procesamiento de grandes cantidades de datos, se han desarrollado técnicas, estrategias y herramientas. Sin embargo, para una empresa u organización que no tenga un enfoque hacia las Tecnologías de la Información, el uso de estas herramientas y estrategias se puede volver todo un reto, desde la compra y mantenimiento de servidores hasta la instalación y configuración de las aplicaciones a usar. De hecho, ante tal reto, los planes de crecimiento y operación de la organización se ven limitados [1].

Conociendo la necesidad de estas organizaciones, empresas como Amazon o Google ofrecen infraestructura de cómputo ya configurada y han optado por un modelo de negocios conocido como *Software-como-Servicio* (*Software-as-a-Service*) [8], el cual, a grandes rasgos, consiste en el alquiler de software (programas, herramientas, etc.) e infraestructura (servidores, discos de almacenamiento) a las organizaciones, proporcionándoles todo lo necesario para procesar sus datos sin necesidad de instalar software, sin los costos de mantener operativo un servidor y pagando únicamente por lo que usan. En este sentido, los usuarios finales y organizaciones ya no deben preocuparse por detalles técnicos de las herramientas, enfocándose únicamente en el análisis de los datos; sin embargo, el uso de este modelo puede conllevar a generar una dependencia con el proveedor de servicio. Por ejemplo, pueden ocurrir escenarios en los cuales se acumulen grandes cantidades de datos en la nube, lo que conlleva a elevados costos de migración. Otro ejemplo podría ser el acceso, puesto que si no se tiene acceso a internet, luz o el dominio no se encuentra disponible, el acceso a los datos es imposible. A este tipo de problemas por la dependencia a un proveedor de servicio se les conoce como escenarios de *vendor lock-in*.

En este trabajo presentamos a *Xelhua*, una plataforma para la creación de sistemas de ciencia de datos bajo demanda. Esta plataforma está pensada con el objetivo de permitir a usuarios menos experimentados en el área de las tecnologías diseñar y materializar, en cualquier infraestructura de cómputo, sistemas de ciencia de datos para el apoyo a la toma de decisiones, proporcionando un modelo tecnológico para encapsular aplicaciones existentes, desplegarlas y proveerlas para su uso a usuarios finales. En este sentido, *Xelhua* permite a los

usuarios analizar conjuntos de datos estructurados sin necesidad de programar, únicamente seleccionando servicios para análisis que se tienen disponibles.

2. Antecedentes

2.1. ¿Qué son los contenedores virtuales?

El modelo de negocio de empresas como Amazon o Google es posible gracias al concepto de virtualización. La forma más conocida de virtualización son las *máquinas virtuales*, las cuales consisten en la emulación de la infraestructura de una única computadora (RAM, CPU, disco duro y sistema operativo) en una máquina física [7]. En pocas palabras, las máquinas virtuales permiten crear de manera virtual una computadora dentro de otra (anfitrión), la cual tiene su propia configuración y sistema operativo. De esta forma, es posible que en una misma máquina física puedan existir diferentes máquinas virtuales, cada una con recursos (RAM, CPU, disco duro y sistema operativo) distintos entre sí. Asimismo, es posible ofrecer una máquina virtual como un entorno de servidor a múltiples usuarios. Dado que las máquinas virtuales son desplegadas bajo demanda, son portables (ya que al ser virtuales, pueden ser almacenadas y transportadas como un archivo) y están aisladas unas de otras [13]. Si bien las máquinas virtuales han sido una piedra angular en el avance tecnológico de la virtualización y, por ende, de la computación, éstas tienen algunas desventajas, dentro de las que destacan su tamaño y velocidad. Su tamaño, puesto que a cada una se le debe instalar su propio sistema operativo, lo cual hace a su representación física (archivo) muy grande, en el orden de varios GB. Por otro lado, la velocidad de una máquina virtual no es la misma que la máquina física anfitriona, dado que el proceso de virtualización pasa a través de varias capas, lo cual genera latencia.

Una alternativa a las máquinas virtuales son los contenedores virtuales [10]. Éstos siguen un proceso de virtualización similar, pero omitiendo algunas capas, lo que permite realizar tareas similares a las máquinas virtuales, pero siendo más ligeros en tamaño y más rápidos de desplegar. Esta tecnología ha tenido un auge muy importante en los últimos años, puesto que sus propiedades permiten diseñar sistemas muy diversos, adaptables, personalizables y de manera mucho más sencilla [9]. Por ejemplo, es posible instalar un programa o un sistema completo dentro de un contenedor virtual, clonarlo y desplegarlo en una computadora personal, en múltiples servidores o incluso en una infraestructura de nube. Tanto los contenedores virtuales como las máquinas virtuales permiten contener programas y aplicaciones, de tal forma que su instalación y configuración se efectúa una única vez y no cada vez que se despliegue en una computadora o servidor. Algunos ejemplos de servicios desplegados en contenedores que usamos de manera habitual son Gmail, Google Maps, OneDrive, Overleaf, Colaboratory, GitHub, Netflix, Spotify, entre otros. Usualmente, se dice que estas aplicaciones se ejecutan en la nube, algo que se inició con las máquinas virtuales. No obstante, el uso de máquinas virtuales por parte de empresas proveedoras de servicio

ha ido en decremento en los últimos años. Los proveedores han ido migrando poco a poco sus sistemas a contenedores virtuales. Esto último debido a todas las ventajas que proporcionan sobre las máquinas virtuales. Una de las diferencias destacables que es importante mencionar es el consumo energético [4], que últimamente ha atraído mucho la atención por diversos aspectos, el más importante es el impacto que tiene en la reducción de CO2, lo que, a la par, impacta en los costos de mantenimiento de la infraestructura y, en consecuencia en el abaratamiento de los precios hacia el usuario final. Por un lado, los contenedores virtuales tienen una mejor gestión en los recursos de cómputo que se utilizan en una máquina física que su contraparte, las máquinas virtuales, además de que, al contar con menos capas en la virtualización, tienen una menor cantidad de procesos en ejecución.

Los contenedores virtuales tienen mucha aplicabilidad en la vida cotidiana; muchas aplicaciones de software pueden ejecutarse sobre contenedores virtuales sin que el usuario final lo note. De hecho, los contenedores virtuales están pensados para, a bajo nivel, realizar diversas tareas sin que el usuario, en el alto nivel, note cambios en el desempeño de las aplicaciones. Los contenedores virtuales son especialmente útiles en aplicaciones encadenadas (*aplicaciones contenerizadas*) que se ejecutan en la nube. Éstas son aplicaciones *"grandes"*, cuyos componentes ("pequeños") se ejecutan en diferentes contenedores virtuales separados. Estas aplicaciones se apegan a un flujo de trabajo con productos parciales y finales, los cuales se pueden usar en otros procesos como productos parciales o como producto final deseado. Ésto puede verse como un Modelo de procesamiento ETL, Extraer, Transformar y Cargar (ETL $Extract$, $Transform$, $Load$), el cual realiza la adquisición de datos y su posterior transformación ejecutando alguna operación o modificación a los datos y su transferencia a un repositorio destino, el cual puede ser otro contenedor virtual [5]. Por ejemplo, aplicaciones para la limpieza de valores nulos o vacíos en un dataset, detección de outliers, cálculos estadísticos, visualización de datos, etc. Un tipo de estas aplicaciones son aquellas dedicadas al análisis de datos, también denominadas *herramientas de analítica de datos*. Los procesos de análisis de datos en la nube se realizan utilizando tecnologías de virtualización y herramientas de analítica, encapsulando estas últimas en contenedores virtuales y otorgándole acceso a los usuarios finales a éstas mediante una interfaz elegante.

2.2. Plataformas de análisis de datos en la nube

Aun teniendo recursos de cómputo disponibles, el análisis en grandes volúmenes de datos es un proceso complejo. Para un usuario común puede resultar agobiante el tener que aprender a utilizar múltiples herramientas, conceptos y tecnologías de contenedores virtuales, en adición a los conceptos y herramientas de análisis de datos. Las plataformas de análisis de datos son una opción viable para mitigar estos retos. Una de las principales razones para realizar análisis de datos en plataformas en la nube es la facilidad de uso que brindan [14]. Muchas veces las organizaciones no tienen ni el recurso financiero, técnico, ni tiempo

para efectuar todo el trabajo de bajo nivel que conlleva configurar e instalar aplicaciones de análisis de datos. Si bien estas plataformas facilitan y flexibilizan la realización de tareas de análisis de datos, su utilización no está libre de inconvenientes. El empleo continuo de estas plataformas puede generar una dependencia a ellas (*vendor lock-in*), lo que conlleva a problemas a futuro para las organizaciones [11]. Por ejemplo, si una organización tiene sus datos y sistemas en la nube de un proveedor y si este proveedor tiene algún fallo en su dominio (p. ej. no se puede acceder desde su URL), la organización no podrá acceder a sus datos. También, si la cantidad de datos que la organización tiene en la plataforma de análisis es muy grande, los costos de trasladar los datos a la plataforma de otro proveedor o a los servidores de la propia empresa podrían resultar muy altos. Este tipo de escenarios deben ser tomados muy en cuenta por las organizaciones, puesto que, si bien, en un inicio, la facilidad de procesamiento de datos en la nube es muy atractiva, en un futuro puede interrumpir drásticamente la continuidad de la operación de una organización.

3. *Xelhua*, una plataforma de ciencia de datos basada en contenedores virtuales

Considerando las limitantes que se presentan en un escenario de *vendor lock-in*, en Cinvestav Tamaulipas hemos trabajado en el desarrollo de *Xelhua*, una plataforma en la nube para el análisis de datos. *Xelhua* aprovecha la tecnología de contenedores virtuales para el procesamiento y análisis de datos en múltiples recursos de cómputo, todo de manera transparente para el usuario final y sin depender de un proveedor de servicios específico.

3.1. Principios de diseño

Xelhua está basado en un conjunto de modelos matemáticos y esquemas tecnológicos desarrollados por nosotros mismos, que permiten diseñar soluciones de procesamiento de datos sin preocuparse por la infraestructura subyacente donde se ejecute la plataforma, de tal manera que puede ser desplegada en una computadora personal (el ***edge***), en un conjunto de servidores (***fog***), en la nube (***cloud***), o en las tres opciones, a la vez, de manera conjunta. En *Xelhua* pueden existir múltiples aplicaciones y herramientas para procesar datos, cada una encapsulada y aislada de las otras en contenedores virtuales. Mediante la plataforma, es posible crear flujos de procesamiento (*pipelines*) en forma de grafo, uniendo entre sí las aplicaciones requeridas (ver Figura 1).

Xelhua está diseñado con base en el modelo de procesamiento ETL. Este modelo considera que el procesamiento de datos se realiza en 3 etapas: Extracción, Transformación y Carga. Primeramente, los datos se extraen (E) desde una fuente; posteriormente, pasan a una etapa de transformación (T) por medio de algún proceso, aplicación o sistema y, finalmente, se cargan y almacenan (L) los

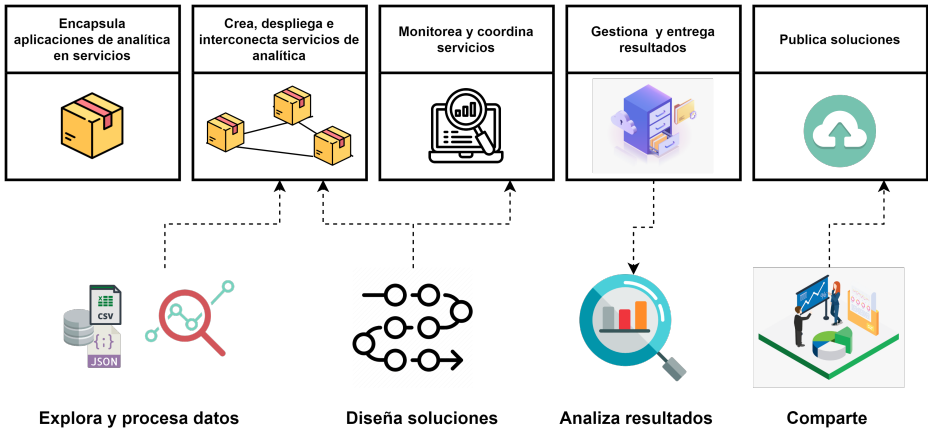


Figura 1: Panorama de actividades posibles de realizar con *Xelhua*.

resultados en un resumidero de datos (ej. una carpeta). En este sentido, podemos ver el proceso de transformación de los datos como si de una caja negra se tratase: un conjunto de datos entra a la caja negra, los transforma y produce resultados. Siguiendo este modelo, diseñamos para *Xelhua* entidades para el procesamiento de datos a las cuales denominamos *ABox*. Los *ABox* tienen como función la abstracción de aplicaciones necesarias para la transformación de datos (por ejemplo, el movimiento de datos, el monitoreo, etc.) en una sola entidad que funcione como una caja negra, para que, posteriormente, pueda ser utilizada como bloque de construcción para la creación de sistemas de ciencia de datos. En otras palabras, los *Abox* permiten encapsular aplicaciones para su consumo como un servicio. Como se puede apreciar en la Figura 2 mediante el uso de contenedores virtuales, los *ABox* encapsulan un conjunto de elementos entre los cuales se encuentran módulos de control de *Xelhua*, archivos de configuración, interfaces de entrada y salida para la comunicación con otros ABox, las aplicaciones a encapsular y las dependencias y librerías necesarias para que la aplicación funcione. Por ejemplo, si la aplicación está desarrollada con el lenguaje de programación Python, las dependencias serían todas aquellas librerías necesarias para ejecutar Python. Una vez que una aplicación se encuentra encapsulada en una *ABox*, se considera un *servicio*, de tal modo que los usuarios finales pueden invocarlo para la construcción de sistemas de ciencia de datos. En este sentido, todas las aplicaciones encapsuladas en *ABox* conforman un repositorio de servicios disponible para usuarios finales. De esta manera, los desarrolladores pueden encapsular aplicaciones para la transformación y análisis de datos y compartirlos en línea con los usuarios.

El encapsulamiento en *ABox* le proporciona las siguientes características a las aplicaciones:

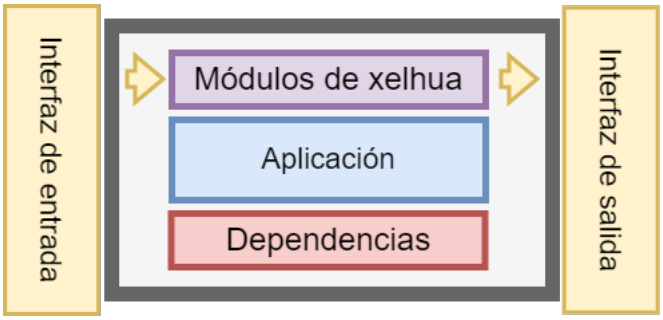


Figura 2: Representación gráfica de los componentes de un *ABox*.

- Agnosticidad a la infraestructura - Al encontrarse encapsuladas con todas sus dependencias, las aplicaciones se consideran auto-contenidas, por lo que pueden funcionar independientemente de la infraestructura en donde se encuentren desplegadas;
- Conectividad - Las interfaces de entrada y salida proporcionan la capacidad de comunicarse con otros *ABox*;
- Manejabilidad - Los módulos de control de *Xelhua* proporcionan un control a la ejecución de las aplicaciones, los datos, resultados y un monitoreo constante del estatus de la aplicación;
- Portabilidad - Al ser auto-contenidos, los *ABox* son paquetes de software portables, capaces de desplegarse en computadoras personales, servidores o en la nube;
- Disponibilidad - Los *ABox* son entidades replicables, por lo que es posible clonarlos y desplegar múltiples copias para asegurar que siempre existan instancias disponibles para los usuarios;
- Confiabilidad - En el caso de fallas en un *ABox* por causas externas (e.g. apagón de un servidor), los servicios tienen componentes de tolerancia a fallos, por lo cual es posible redirigir peticiones de usuarios a una réplica de un mismo *ABox* para no interrumpir los procesos del usuario final.

Por otro lado, *Xelhua* considera un enfoque orientado a microservicios. Los microservicios son un enfoque arquitectónico para el desarrollo de software y consiste en el desarrollo de componentes pequeños denominados servicios. Estos servicios están separados del resto, son independientes entre sí y se comunican unos con otros mediante la red. A diferencia del enfoque monolítico, donde el software es un solo bloque de código, las arquitecturas de microservicio son un conjunto de módulos separados e independientes, lo cual permite que las aplicaciones sean más fáciles de escalar y más rápidas de desarrollar. En *Xelhua* todos los componentes son microservicios, lo cual proporciona diversas ventajas a la plataforma:

- Despliegue distribuido - Al ser componentes independientes, cada uno puede ser ubicado en diferentes contextos, ya sean en servidores públicos (ej. Amazon, Azure, Digital Ocean, etc.), privados, computadores personales, etc.;
- Escalabilidad de servicios de analítica - Los *Abox* que encapsulan las aplicaciones de analítica para el procesamiento de datos son independientes entre sí, por lo que, para desarrollar y añadir más servicios de analítica, no es necesario conocer ni modificar el resto;
- Grado de tolerancia a fallos - Si un servicio de analítica llega a fallar, el resto de servicios no se ven afectados.

Tanto las características de los *ABox*, como de la arquitectura de microservicios, son beneficiosas a la hora de lidiar con los problemas del *vendor lock-in*. Por un lado, la abstracción de los *ABox* permite desplegar los servicios de analítica en cualquier infraestructura de cómputo y conectarlos entre sí para generar flujos de procesamiento sin necesidad de depender de APIs de proveedores de nube, mientras que el desacoplamiento de los microservicios permite la distribución en múltiples servidores y con diferentes características, recuperando, así, el control sobre los datos y aplicaciones sin necesidad de crear una dependencia con proveedores de la nube.

3.2. El ciclo de vida de los datos

Xelhua permite manejar todas las etapas del ciclo de vida de los datos.

- **Adquisición**: Los usuarios finales pueden subir sus fuentes de datos a la plataforma, ya sea directamente desde su computadora (e.g. archivos csv) o mediante el uso de conectores para la adquisición de los datos desde repositorios remotos, bases de datos, etc.
- **Preprocesamiento**: *Xelhua* cuenta con un conjunto de servicios previamente configurados para realizar la preparación de los datos. De manera transparente, los usuarios pueden llevar a cabo la transformación de sus conjuntos de datos, limpiarlos de valores nulos, faltantes u outliers, normalizarlos, estandarizarlos, entre otros procesos de limpieza.
- **Procesamiento y análisis:** La plataforma cuenta con servicios de analítica básica para datos estructurados, así como herramientas de descubrimiento de patrones sobre los datos y Machine Learning. Algunas de estas herramientas son algoritmos de clustering, modelos de regresión, redes neuronales, modelos de clasificación (KNN, Naive Bayes y Máquinas de Vectores de Soporte), cálculos estadísticos, etc.
- **Visualización**: Los resultados producidos por el resto de los servicios pueden ser enviados a herramientas de visualización. Mediante únicamente la selección de parámetros, los usuarios pueden generar diversos tipos de gráficas y mapas que permitan interpretar de manera más cómoda los resultados.

- **Preservación**: Por último, *Xelhua* mantiene un control sobre todas las versiones de los datos y resultados generados en cada una de las etapas, por lo cual todos los resultados son almacenados.

En *Xelhua*, estas etapas son representadas en forma de grafo, donde cada nodo del grafo representa un servicio de análisis del repositorio de servicios de *Xelhua*. Para la creación de este grafo los usuarios utilizan los servicios de este repositorio; posteriormente, el grafo se materializa y se despliega en una infraestructura de cómputo definida. Una vez desplegados los contenedores virtuales (los servicios), comienza la ejecución, partiendo del nodo raíz del grafo y siguiendo la secuencia definida por el usuario. El traslado de los datos producidos por cada etapa es manejado por el modelo de orquestación de *Xelhua* y se enmascaran posibles fallas en los servicios con un modelo descentralizado de réplicas y monitoreo. Por último, los resultados se almacenan y son entregados al usuario final. Este proceso se puede observar en la Figura 3.

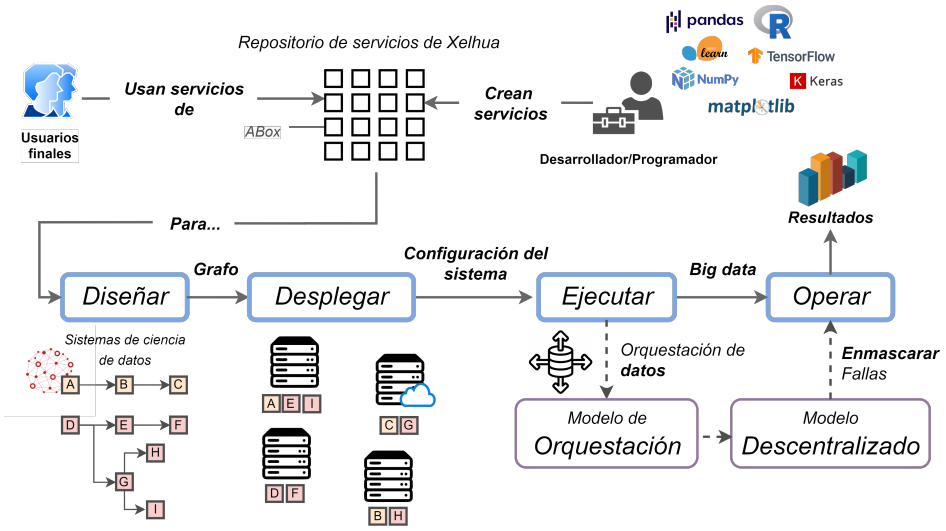


Figura 3: Representación gráfica del proceso de construcción, ejecución y despliegue de un sistema de ciencia de datos en *Xelhua*.

3.3. La interfaz del usuario

Los usuarios acceden a los servicios a través de una interfaz gráfica. La Figura 4 muestra una captura de pantalla de la versión web de *Xelhua*. En ésta se puede observar un área de diseño en la parte derecha y un conjunto de cajas en la parte izquierda. Cada una de estas cajas corresponde a una aplicación de procesamiento o análisis de datos dentro de un contenedor virtual, el cual puede

ser utilizado a necesidad de los usuarios, dentro de un flujo de procesamiento o de manera aislada. Los servicios están divididos en secciones, las cuales están ordenadas de manera descendente siguiendo el ciclo de vida de los datos. Cada caja es configurable, es decir, se adapta a los procesos que el usuario quiera realizar mediante la modificación de parámetros. Por ejemplo, en el servicio con algoritmos de clustering el usuario puede modificar el valor de k (cantidad de grupos a detectar), seleccionar las variables del dataset que se emplearán, el algoritmo de clustering a utilizar, seleccionar si se desea validar el valor de K mediante índices de validación, etc. El usuario final puede realizar el procesamiento de sus datos siguiendo una serie de pasos, como los descritos a continuación, uniendo servicios mediante movimientos *drag-and-drop*:

1. El usuario selecciona la fuente de datos a procesar, puede ser un archivo directo de su computadora o que se encuentre en algún servidor en la nube; hemos experimentado cargando archivos en el orden de los 100 GB de tamaño;
2. Del catálogo de aplicaciones, el usuario debe elegir aquellos que desee usar; por ejemplo, existen servicios para eliminar valores vacíos, normalizar valores, corregir valores, detectar outliers, realizar muestreo, generar estadísticos, crear gráficas, georeferenciar datos en un mapa, etc.; los servicios se unen en forma de grafo en el panel de la derecha, por lo cual se puede elegir el orden en el que estas aplicaciones se van a ejecutar;
3. Para cada tarea, el usuario debe especificar los parámetros necesarios por el algoritmo a emplear; por ejemplo, seleccionar las variables a procesar, los rangos de valores permitidos, la función de similitud a usar, el número de grupos a detectar, etc.;
4. Una vez realizada la configuración preliminar de las aplicaciones, el usuario presiona el botón ejecutar para echar a andar todo el flujo de procesamiento;
5. Finalmente, una vez terminada la ejecución, el usuario puede descargar los resultados, o bien, seguir procesándolos mediante otro flujo de trabajo, o visualizarlos mediante gráficas o mapas dentro de la misma plataforma.

De manera transparente para el usuario, internamente ocurre una serie de procesos que permiten que los resultados pasen de su estado original a ser procesados por las distintas aplicaciones seleccionadas. Una vez que el usuario elige los servicios que desea ejecutar, cada una de estas aplicaciones es desplegada en un contenedor virtual. Se crea un mapa para el traslado de los datos originales a cada una de las aplicaciones seleccionadas por el usuario, en el orden que este mismo las declaró. Los datos son recibidos por las aplicaciones, se procesan, entregan los resultados a la siguiente etapa y, finalmente, este contenedor se desactiva automáticamente. Es decir, los contenedores virtuales se encienden y apagan bajo demanda sin intervención del usuario. Si no existen peticiones por parte del usuario, los contenedores no se activan, reduciendo, así, la cantidad de procesos que se realizan en *Xelhua*.

Una vez que el usuario termina de diseñar su sistema de ciencia de datos, se pueden ejecutar presionando el botón verde. Este botón enviará un conjunto de instrucciones a una entidad coordinadora encargada del despliegue de los contenedores necesarios para el procesamiento. Este proceso es transparente para el usuario, por lo que él solo se entera de la finalización de cada uno de los procesos mediante notificaciones y la entrega de resultados. Al finalizar el procesamiento, el usuario puede elegir finalmente descargar los resultados de cada una de las cajas que seleccionó, visualizar información estadística de los datos procesados, añadir más cajas, eliminar cajas o guardar y compartir el grafo a otro usuario.

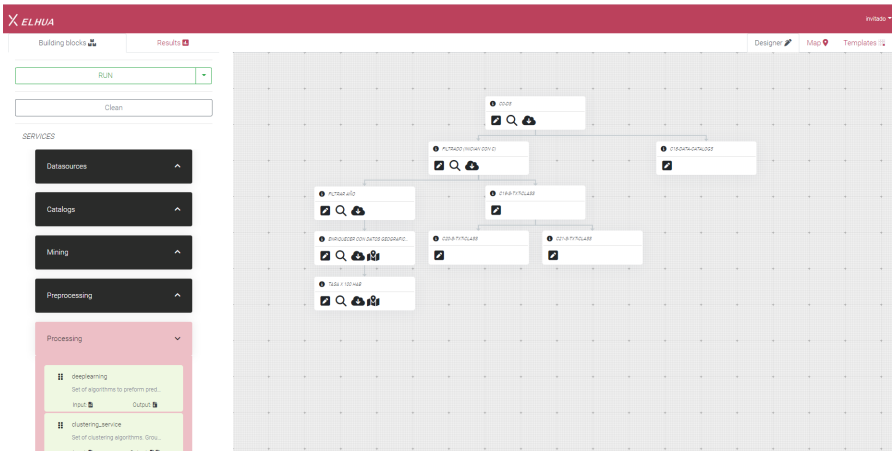


Figura 4: Interfaz gráfica de *Xelhua*. En la parte izquierda se encuentran los servicios arrastrables. En la parte derecha se encuentra el área de diseño de grafos.

4. Caso de uso: Análisis exploratorio de datos de cáncer

Para mostrar la utilidad de la plataforma realizamos un análisis exploratorio de un dataset con conteos y tasas de defunciones ajustadas por rango de edad para distintos tipos de cáncer. El dataset cuenta valores para hombre y mujer de los años 2000 al 2020 para todos los municipios de México. Realizar análisis exploratorios de conjuntos de datos puede llegar a ser un trabajo laborioso, dado que es necesario procesar y analizar múltiples resultados generados con base en combinaciones de parámetros distintos, por ejemplo, los resultados obtenidos de una regresión pueden variar dependiendo de las variables que se utilicen, de la cantidad de datos, o si estos datos pertenecen a un estado en concreto, a una ciudad, o a todo el país. En este contexto, *Xelhua* cuenta con herramientas para el procesamiento de datos en profundidad, de tal modo que los usuarios puedan definir mediante parámetros la forma en que sus datos serán procesados. Para

este caso, definimos una metodología de procesamiento en profundidad para cada tipo de cáncer, estado de la República y año registrado. Como se puede ver en la Figura 5, los usuarios finales diseñan el grafo utilizando los servicios ($ABox$) disponibles. Posteriormente, se definen las variables para definir la metodología de búsqueda en profundidad. En palabras simples, el usuario define variables presentes en el dataset, las cuales servirán para filtrarlo. En nuestro caso se eligieron las columnas que contienen la clave con el tipo de cáncer (Causa_def), la cual fue la causa de defunción, la columna con los nombres de los estados de la República (estado) y la columna con los valores de los años (año). En este sentido, *Xelhua* realizará la división del dataset y generará un subdaset para cada tipo de cáncer (X). Posteriormente, cada uno de estos subdatasets X se dividirán para cada estado (Y) y, finalmente, cada uno de estos subdatasets XY se dividirá para cada año (Z). Posteriormente, a cada subdataset generado se le aplicará el proceso que el usuario haya definido en su grafo. De tal manera, se generarán histogramas, gráficas, modelos de regresión linear y mapas para las diferentes combinaciones, lo cual da un total de 89,033 productos generados (contando los subdatasets). Cada uno de estos productos es consultable, de tal manera que los usuarios finales pueden navegar los resultados e ir centrándose en aquellos que más le interesen. Por ejemplo, si en el mapa de cáncer de mama del todo el país se indica un patrón de tasas elevadas en el estado de Nuevo León, el usuario puede acceder a los productos generados para Nuevo León.

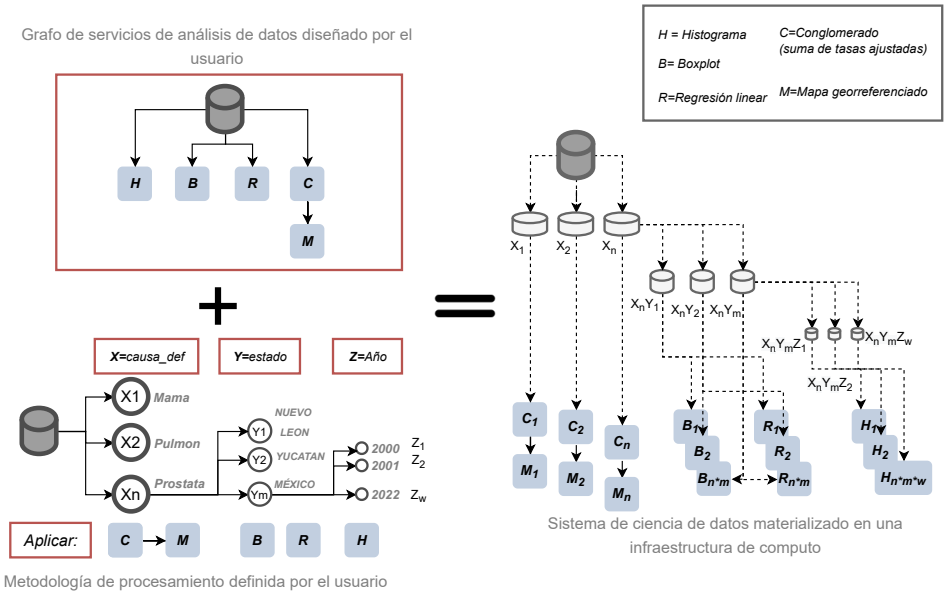


Figura 5: Representación gráfica del un sistema para el análisis exploratorio de datos de cáncer en *Xelhua*.

Para un analista de datos, el generar esta gran cantidad de combinaciones y productos puede conllevar un gran esfuerzo, además de poder perderse fácilmente entre la gran cantidad de resultados generados. *Xelhua* se encarga de facilitar las cosas mediante su despliegue orientado a diseño, proporcionándole al usuario la interfaz gráfica para diseñar un grafo y definir la metodología de procesamiento (recuadros rojos en la Figura 5), mientras que, tras bambalinas, el despliegue de los servicios, la orquestación de los datos y el almacenamiento de resultados es realizado por *Xelhua*. En pocas palabras, *Xelhua* permite materializar el diseño de un usuario en un sistema de procesamiento sobre una infraestructura de cómputo de manera automática, además de permitir mejorar el rendimiento al procesar las combinaciones en procesos en paralelo. Como se puede observar en la Figura 6, al aumentar la cantidad de trabajadores (procesos en paralelo) es posible disminuir el tiempo de procesamiento de los datos, proporcionándole una mejor experiencia al usuario al tener que esperar menos tiempo para obtenerlos todos. No obstante, dado que cada producto generado puede ser accedido por los usuarios, no es necesario esperar a que todos los procesos terminen para poder comenzar a analizar los resultados que se van produciendo.

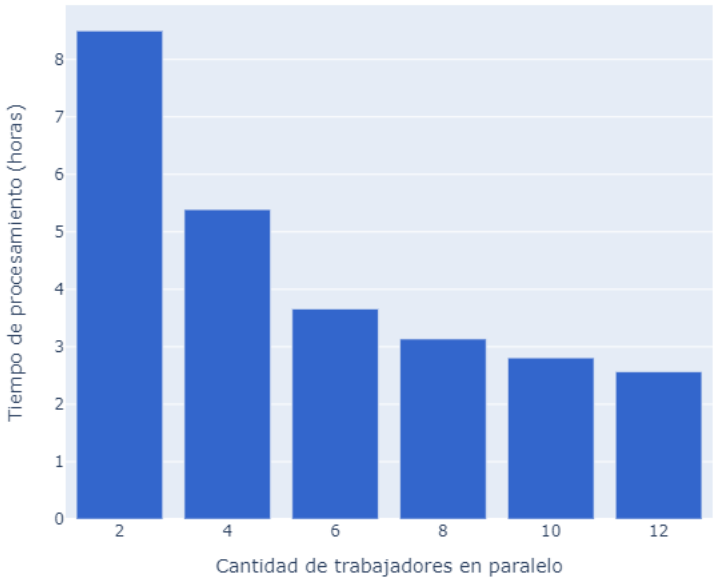


Figura 6: Tiempo de procesamiento al aumentar la cantidad de trabajadores.

5. Conclusiones

Xelhua está pensada como una plataforma para el análisis de datos, pero orientada al diseño de distintas formas de realizar las tareas de análisis, lo cual permite probar distintas versiones de las tareas de análisis que son de interés para el usuario; esta característica es útil para la validación estadística de modelos en el *diseño de experimentos*. *Xelhua* tiene como característica principal el enfoque a la generación de estructuras complejas de procesamiento de datos orientado al diseño, las cuales, de otro modo, serían agobiantes de diseñar y ejecutar para el usuario. Los flujos de procesamiento se forman y ejecutan siguiendo la secuencia de nodos del grafo definido. A diferencia de otras plataformas en la nube, *Xelhua* no depende de algún proveedor de servicios, por lo que es un paquete de software que puede ser instalado y desplegado en diversas infraestructuras sin llegar a depender de la infraestructura de un proveedor específico. Además, ofrece la posibilidad de escalar, tanto en la infraestructura de procesamiento (RAM, CPU, almacenamiento), como en el catálogo de aplicaciones para análisis de datos. Cada aplicación en *Xelhua* es un contenedor virtual con una aplicación y las librerías requeridas dentro. Por tal motivo, es posible añadir más servicios a *Xelhua* simplemente encapsulando las nuevas aplicaciones en contenedores virtuales. *Xelhua* permite la unión, intersección y coordinación de procesos de análisis de datos entre todos los posibles diseños de una tarea de análisis de interés para el usuario, adaptándose a los recursos de cómputo disponibles en la infraestructura sobre la que se esté usando. En resumen, la plataforma permite a los desarrolladores convertir aplicaciones en servicios consumibles bajo demanda, interconectables y altamente disponibles. Estos servicios pueden ser utilizados por usuarios finales (sin grandes conocimientos en programación) para diseñar sistemas de ciencia de datos, los cuales son desplegados bajo demanda de manera automática en una infraestructura definida, manteniendo alta disponibilidad de los servicios para el correcto procesamiento de los datos y con la capacidad de realizar procesamiento en paralelo de manera semi-automatica para mejorar el rendimiento.

Agradecimientos

Este trabajo fue parcialmente apoyado por el proyecto 41756 “Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud” por el FORDECYT-PRONACES de Conacyt (México).

Referencias

[1] Sepideh Bazzaz Abkenar et al. “Big data analytics meets social media: A systematic review of techniques, open issues, and future directions”. En: *Telematics and Informatics* 57 (2021), pág. 101517. ISSN: 0736-5853. DOI:

https://doi.org/10.1016/j.tele.2020.101517. URL: https://www.sciencedirect.com/science/article/pii/S0736585320301763.

[2] Wasim Ahmad Bhat y S.M.K. Quadri. *Big Data promises value: Is hardware technology taken onboard?* 2015. URL: https://www.emerald.com/insight/content/doi/10.1108/IMDS-04-2015-0160/full/html.

[3] Juan José Camargo-Vega, Jonathan Felipe Camargo-Ortega y Luis Joyanes-Aguilar. "Conociendo Big Data". es. En: *Revista Facultad de Ingeniería* 24 (ene. de 2015), págs. 63 -77. ISSN: 0121-1129.

[4] Ismael Cuadrado-Cordero, Anne-Cécile Orgerie y Jean-Marc Menaud. "Comparative experimental analysis of the quality-of-service and energy-efficiency of VMs and containers' consolidation for cloud applications". En: *2017 25th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. 2017, págs. 1-6. DOI: 10.23919/SOFTCOM.2017.8115516.

[5] Papa Senghane Diouf, Aliou Boly y Samba Ndiaye. "Variety of data in the ETL processes in the cloud: State of the art". En: *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*. IEEE. 2018, págs. 1-5.

[6] Reihaneh H Hariri, Erik M Fredericks y Kate M Bowers. "Uncertainty in big data analytics: survey, opportunities, and challenges". En: *Journal of Big Data* 6.1 (2019), págs. 1-16.

[7] Yuzhe Huang et al. "SSUR: An Approach to Optimizing Virtual Machine Allocation Strategy Based on User Requirements for Cloud Data Center". En: *IEEE Transactions on Green Communications and Networking* 5.2 (2021), págs. 670-681. DOI: 10.1109/TGCN.2021.3067374.

[8] Euripidis Loukis, Marijn Janssen y Ianislav Mintchev. "Determinants of software-as-a-service benefits and impact on firm performance". En: *Decision Support Systems* 117 (2019), págs. 38-47.

[9] Markus Murhu. *Containerization and deployment of a virtual learning environment*. 2021. URL: https://aaltodoc.aalto.fi/handle/123456789/110549.

[10] Amit M Potdar et al. "Performance evaluation of docker container and virtual machine". En: *Procedia Computer Science* 171 (2020), págs. 1419-1428.

[11] Seyed Majid Razavian et al. "An analysis of vendor lock-in problem in cloud storage". En: *ICCKE 2013*. 2013, págs. 331-335. DOI: 10.1109/ICCKE.2013.6682808.

[12] Youssra Riahi. "Big Data and Big Data Analytics: Concepts, Types and Technologies". En: *International Journal of Research and Engineering* 5 (nov. de 2018), págs. 524-528. DOI: 10.21276/ijre.2018.5.9.5.

[13] Prateek Sharma et al. "Containers and virtual machines at scale: A comparative study". En: *Proceedings of the 17th international middleware conference*. 2016, págs. 1-13.

[14] Dilpreet Singh y Chandan K Reddy. *A survey on platforms for Big Data Analytics - Journal of Big Data*. 2014. URL: https://link.springer.com/article/10.1186/s40537-014-0008-6.

Un enfoque multidisciplinario hacia la medicina personalizada en México

Gustavo Emilio Mendoza Olguín[1[0000-0001-7164-4987]], María de la Concepción Pérez de Celis Herrero[1[0000-0003-2302-2774]] y María Josefa Somodevilla García[1[0000-0002-1972-2252]]

[1] Benemérita Universidad Autónoma de Puebla, Puebla 08544, México

gustavo.mendozao@alumno.buap.mx, {maria.perezdecelis, maria.somodevilla}@correo.buap.mx.

Resumen. Existe un interés creciente en México por la implementación del Expediente Clínico Electrónico; numerosas propuestas han sido presentadas durante los últimos años por parte de instituciones públicas y privadas. Este expediente, en el caso de poder unificarse para el Sistema Nacional de Salud actual, daría la oportunidad de transitar hacia el paradigma de la medicina personalizada. En este trabajo se hace una evaluación de los retos que las tecnologías de la información podrían enfrentar en nuestro país en el camino hacia el cambio de paradigma de la medicina, tomando en cuenta el papel de los actores involucrados en el proceso. Se puede decir que no es un proceso que dependa solo de las TIC, sino que implica que cada actor sea informado de los beneficios para que evalúen el posible beneficio y, por lo tanto, se fomente su participación.

Palabras clave: Medicina Personalizada · Expediente Clínico Electrónico · Ciencia de Datos.

1 Introducción

El paradigma de la medicina personalizada se fundamenta en el carácter único y variado de las características moleculares, fisiológicas, ambientales, de comportamiento, entre otras, de los individuos. Por esto, las intervenciones de los profesionales de la salud deben ser ajustadas de acuerdo con el contexto individual para proveer el tratamiento indicado para la persona [1]. Es evidente que, para que el sistema de salud nacional se mueva en esa dirección, los profesionales de salud requieren de acceso a toda la información posible referente al paciente y de herramientas computacionales que les permitan convertir toda esa información en indicadores puntuales que agilicen la toma de decisiones y faciliten su trabajo. Es aquí donde las tecnologías de la información adquieren relevancia en la creación de estas herramientas, el diseño de algoritmos más confiables y el desarrollo de metodologías con una perspectiva basada en la generación de conocimiento a partir de los datos.

La Ciencia de Datos es un término acuñado como el enfoque multidisciplinario utilizado para obtener indicadores (conocidos en inglés como *insights*) a partir de conjuntos de datos creados, almacenados y que crecen día con día [2]. Estos indicadores varían dependiendo del nivel de conocimiento que se pretenda obtener de los datos: serán descriptivos si lo que se desea es identificar las características que conforman a los elementos representados en los da-

tos; serán predictivos si lo que se desea es saber qué características podría tener un elemento nuevo que se desee agregar al conjunto de datos, y serán prescriptivos si lo que se desea conocer es cuál es la mejor manera de manejar a un nuevo elemento cuando éste se agregue al conjunto de datos. Si bien las fases de este proceso no están bien definidas por los autores, pueden identificarse cinco etapas en las que la mayoría de los investigadores coinciden, como se muestra en la Fig. 1.

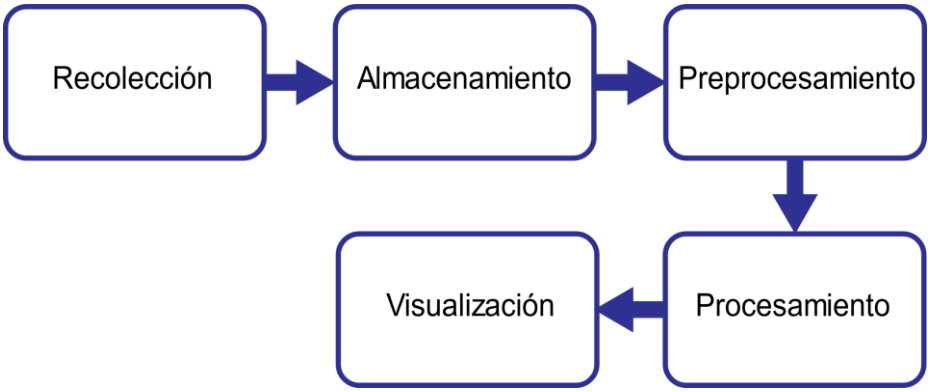


Figura 1. El proceso de Ciencia de Datos. *Fuente:* Elaboración propia.

Si bien el área de la salud se ha visto beneficiada por la inclusión de los avances computacionales, sobre todo en el área de diagnóstico, no deja de ser un reto debido a, entre otras cosas, lo siguiente: las diferentes ramas de especialidad que conlleva; la gran cantidad de información estructurada (por ejemplo, expedientes clínicos, registros personales de salud, estudios de laboratorio, etc.) y no estructurada (por ejemplo, imagenología, radiografías, resonancias, etc.) que debe ser almacenada y procesada: la variedad de predictores significativos para cada una de las áreas; las relaciones desconocidas entre los diferentes sistemas que conforman el cuerpo humano; la unicidad del contexto de cada paciente e incluso la forma en la que los resultados deben presentarse a los profesionales de salud para facilitar la toma de decisiones [3]. Las herramientas computacionales actuales están basadas en un enfoque predictivo, dejando al profesional médico la selección del tratamiento adecuado basándose en sus conocimientos y en su experiencia. El actual paradigma de la medicina basada en evidencias tiene como objetivo la utilización concienzuda, juiciosa y explícita de evidencias científicas para la toma de decisiones en la práctica médica [4], intentando hacer de lado sesgos y dogmas de los profesionales de la salud; sin embargo, tiene el inconveniente de que la toma de decisiones se basa en el criterio de agrupar a los pacientes en clústeres, es decir, si pacientes con determinadas características han respondido positivamente a un tratamiento, este tratamiento, entonces, debe servirle a este nuevo paciente que presenta un cuadro clínico similar. Se pretende en este trabajo presentar cómo las tecnologías de la información pueden ayudar hacia la transición del paradigma actual hacia la medicina personalizada en México, tomando en cuenta la labor de cada uno de los actores involucrados.

2 El paciente empoderado

Ammenwerth define al paciente empoderado como aquel paciente que tiene la habilidad de tomar decisiones autónomas e informadas en una relación de cooperación veraz con su médico acerca de su salud [5], respetando sus decisiones y asumiendo las consecuencias de éstas. Para esto, el autor menciona que deben cumplirse dos condiciones necesarias:

1. Que el profesional de salud establezca un verdadero compañerismo para darle poder al paciente, en vez de considerarlo como la parte "ignorante" de la relación;

2. Que el paciente desee tomar la responsabilidad del manejo de su salud.

La primera condición queda fuera del alcance del área computacional. En la segunda, es en donde las Tecnologías de la información y la Comunicación (TIC) encuentran su campo de acción, ofreciendo a las personas herramientas como Registros Personales de Salud (PHR), cartillas de vacunación electrónicas, dispositivos inteligentes, servicios de internet confiable, entre otros, que ofrezcan a las personas seguridad y confianza en el manejo de su propia información. Adicionalmente, estas herramientas deberán estar diseñadas de manera que, aunque el paciente tenga la libertad de agregar los elementos que considere necesarios, produzcan información relevante para los profesionales de salud en la toma de decisiones.

Este "paciente empoderado" es requerido en la transición debido a que es él quien se encargará de recabar la información necesaria para la generación del contexto personal.

3 El papel de las tecnologías de la información

Las tecnologías de la información tienen un papel relevante en la transición hacia la medicina personalizada. Si bien la percepción general del papel de las TIC en otras áreas de la ciencia es la de desarrollar aplicaciones auxiliares para procesos relevantes propios del área –por ejemplo, realizar simulaciones de procesos complicados, probar teorías, etc.–, esto representa una visión simplista. Se piensa que el Expediente Clínico Electrónico (ECE) es un paso importante hacia la mejora del sistema de salud del país; sin embargo, esto es sólo un paso [6]. En lo referente al manejo de información médica, cada fase del proceso de ciencia de datos tiene actualmente retos que solventar a pesar de los avances que se tienen.

Una parte importante de mencionar antes de adentrarnos en el proceso de ciencia de datos es la existencia de estándares elaborados por grupos de expertos con la finalidad de unificar el significado de un dato dentro de los diferentes contextos en que puede presentarse la información, pues una de las primeras decisiones que deben tomar los desarrolladores es la de seleccionar los estándares que su aplicación tendrá que cumplir para poder facilitar la integración de dispositivos móviles, aplicaciones y un eventual ECE. Por ejemplo, el nombre del paciente puede ser solicitado en una aplicación como "paciente", "nombre del paciente", "*patient's name*", "*user name*", "nombre", etc. Esta variación también puede presentarse no solo en la estructura de la base de

datos, sino también pueden variar el tipo de datos, el rango de valores aceptados e inclusive las reglas de validación, lo que complicaría procesos como el intercambio de información y el almacenamiento. Los estándares facilitan el intercambio de información, al establecer normativas para el manejo, almacenamiento e intercambio de cada dato y se clasifican de acuerdo con el contenido que manejan.

Dentro de estos estándares que contienen terminología y vocabulario médico, se encuentran el CIE10, que contiene el catálogo de enfermedades y desórdenes de salud y que es mantenido por la Organización Mundial de la Salud (OMS); SNOMED-CT, que es una propuesta para estandarizar la terminología médica mantenida por el Colegio Americano de Patólogos; LOINC, que contiene terminología médica para el intercambio de información entre laboratorios; y NANDA, que contiene terminología utilizada para procesos de enfermería.

Por su parte, entre los estándares que contienen terminología referente al intercambio, seguridad y confiabilidad de la información, se encuentran IHE, que es un estándar desarrollado para mejorar la forma en que los sistemas intercambian información; CEN, que es desarrollado por la Comunidad Europea para establecer los requerimientos de seguridad, calidad y desempeño para aplicaciones y aplicaciones que manejen información de salud dentro de la Unión Europea. ISO se encarga de estandarizar los niveles que deben cumplir los servicios médicos, cuidados de calidad, equipos y prácticas médicas.

Una última clasificación de estándares es aquellos que regulan el contenido, vocablos y estructura para otorgar sintaxis. Dentro de estos se encuentra DICOM, que se encarga de estandarizar la forma de transmisión, almacenamiento, recuperación, impresión, procesamiento y visualización de imagenología médica. Finalmente, uno de los más utilizados es HL7, el cual está desarrollado por la *Health Level 7* y se encarga de estandarizar la integración, intercambio y recuperación de información electrónica de salud. Este estándar es el recomendado por la legislación nacional actual como el mínimo que debe cumplir cualquier aplicación utilizada por alguna institución de salud para el intercambio de información [7].

Desde la perspectiva de la recolección de datos, se encuentra la validez y veracidad de la información recolectada. Si bien se menciona en la sección anterior que el paciente debe ser parte activa en esta recolección, los sistemas y dispositivos de recolección de información de salud deben estar construidos basados en las necesidades de los profesionales de salud y ofrecer un conjunto de reglas de validación previamente establecidas por consejos médicos para que la información recabada sea de relevancia para agregarse a los ECE, de manera que tanto el paciente como los profesionales de salud valoren los beneficios que conlleva su utilización. Adicionalmente, es importante recordar la importancia de que toda aplicación que recopile información personal o médica debe ofrecer a sus usuarios un formulario de consentimiento informado que deberá incluir de forma explícita los aspectos técnicos y fundamentación por los que su información será recabada, almacenada y procesada, incluyendo las métricas que se pretenden evaluar y la temporalidad de cada recaudo, estableciendo de forma clara quiénes serán los responsables del almacenamiento, acceso y manejo de la información, así como los procedimientos que el usuario debe seguir para la destrucción de la misma.

Por su parte, el almacenamiento de la información tiene dentro de sus retos el almacenamiento de las grandes cantidades de información requeridas. Preguntas como ¿Cuánta información es necesaria? ¿Cuál es la información relevante para el profesional médico? ¿Qué temporalidad debe almacenarse?, entre otras, deben ser respondidas con fundamento en los diferentes tipos de profesionales de salud que pudieran ser beneficiados. Si bien el almacenamiento en la nube ofrece resolver este problema, surgen otras problemáticas como la seguridad, la disponibilidad, la propiedad y las reglas de acceso de la información.

El preprocesamiento de la información incluye todos los pasos previos que requieren los datos para poder obtener conocimiento significativo a partir de ellos. Esta fase depende de la presentación de la información; los datos estructurados (por ejemplo, expedientes clínicos y resultados de laboratorio y gabinete) requieren de técnicas y pasos diferentes que los datos no estructurados (por ejemplo, la imagenología clínica). A fin de facilitar este proceso, las herramientas diseñadas deben estar orientadas hacia la estandarización de, tanto el vocabulario médico utilizado en el llenado de los expedientes clínicos, como las limitaciones técnicas que las imágenes pudieran tener.

La fase del procesamiento de la información se refiere a la obtención de indicadores a partir de los datos previamente preprocesados en busca de patrones relevantes, novedosos y significativos. Es en esta fase en donde, a través de análisis retrospectivos de la información disponible de los tratamientos para algún padecimiento en específico, aunado a la información adicional que se tenga disponible para cada registro, se categorizan los tratamientos para obtener identificadores descriptivos que permitan conocer si los resultados obtenidos por éstos han sido congruentes con lo esperado teóricamente.

Posteriormente, estos indicadores descriptivos, a su vez, se evalúan mediante un proceso llamado detección de predictores, en el cual se obtiene un modelo matemático que permite representar los resultados del tratamiento X sobre un conjunto de la población con las mismas características. Finalmente, en un nivel más alto de las analíticas, se evalúan los predictores con el fin de determinar cuáles características son susceptibles de ser optimizadas para, utilizando técnicas como simulación, aprendizaje por reforzamiento y árboles de decisión, entre otras, poder sugerir el tratamiento que dará mejores resultados para un paciente con características específicas.

Los retos de esta fase son evidentes; el sesgo de la muestra seleccionada para los estudios puede influir de manera significativa en la determinación del tratamiento óptimo, pues, como se puede observar, las analíticas prescriptivas deben construirse a partir de las predictivas y éstas, a su vez, de las descriptivas. Los profesionales de salud sugieren que los tratamientos sugeridos mediante estos procesos tengan dos características importantes [8]:

- La completa transparencia de cómo fue obtenida la sugerencia;
- Que las sugerencias vayan acompañadas de una métrica de certeza.

Las métricas comunes de evaluación de resultados son necesarias, mas no suficientes, para determinar si la sugerencia que se está ofreciendo es óptima y en qué grado lo será. La determinación y justificación de nuevas métricas

prescriptivas está siendo abordada ya por algunos investigadores, como se establece en [9]. Este proceso se representa en la Fig. 2.

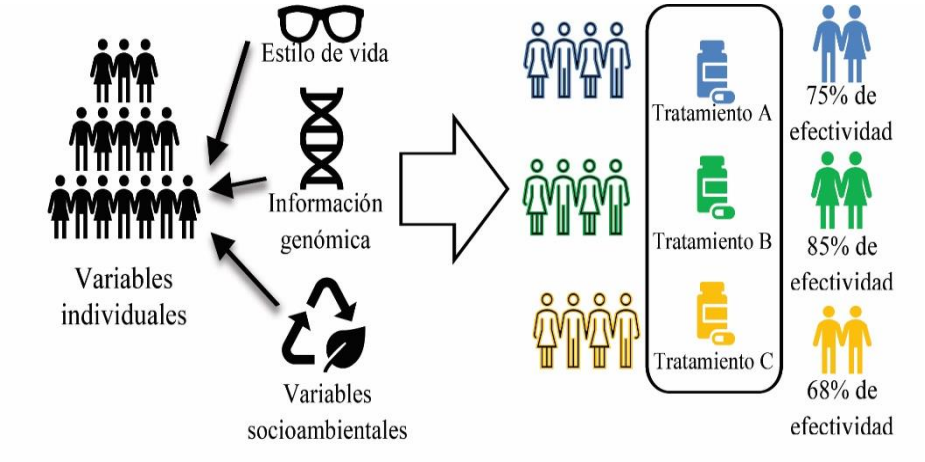


Figura 2. El proceso de medicina personalizada. Fuente: Adaptado y traducido de [10].

Otros retos que deben afrontarse dependen específicamente del campo de la medicina que se pretenda estudiar. La visión por computadora, que es otra de las líneas de investigación de la Inteligencia Artificial, tendría un gran impacto en la obtención de analíticas predictivas o descriptivas a partir de información no estructurada; sin embargo, la mayor parte de las investigaciones realizadas en este campo están siendo realizadas mediante la implementación de soluciones basadas en Redes Neuronales y sus diferentes variables, lo que implica desconfianza por parte de profesionales médicos al no poder explicar lo que pasa dentro de sus capas internas [11]. Otras líneas de investigación, como el Procesamiento de Lenguaje Natural y el Reconocimiento de Entidades Nombradas, también presentan sus propios retos cuando se trata de vocabulario médico en idioma español, al utilizarse para obtener información de expedientes y notas clínicas.

Finalmente, desde la perspectiva de la visualización de los datos, las aplicaciones deberán abordar el reto de mostrar la información relevante para el profesional médico y, si éste lo requiere, cómo es que se está obteniendo un resultado mostrado. Los diseñadores de interfaces de usuario deben lidiar con el hecho de que la mayor parte de las veces estarán trabajando con cantidades muy grandes de información para mostrarla en la pantalla, por lo que técnicas de representación y consolidación de información deben ser investigadas, sobre todo en el caso de las simulaciones en escalas de tiempo que requieren las analíticas prescriptivas para buscar su valor óptimo.

4 El papel de los profesionales de salud

Por su parte, el papel de los profesionales de salud representa la base del conocimiento que se utilizará en cada una de las fases del proceso de ciencia de datos. Son ellos quienes, mediante consensos médicos nacionales, determi-

nen qué información deberá contener el expediente clínico electrónico, qué información debe ser obtenida de los dispositivos personales de los pacientes, la temporalidad del almacenamiento, qué métricas serán relevantes en el contexto nacional, etc. Es importante que los profesionales médicos reconozcan las ventajas que la aplicación de las TIC significan dentro del campo del cuidado de la salud, tomando en consideración los beneficios en tiempo y costos que pueden representar [12]. Desarrollar habilidades para seleccionar, de entre todas las herramientas ofrecidas, aquellas que sean más valiosas para su trabajo, ayudará a los profesionales médicos a enfocarse más en el proceso de toma de decisiones y a reducir tiempo en tareas como la recolección de información, aumentando, con esto, su efectividad y eficiencia.

Uno de los retos más importantes que los médicos deben tener presente es el de garantizar la validez y veracidad de la información, pues ésta servirá para el entrenamiento de las herramientas que se vayan a desarrollar. En Estados Unidos, un consejo de expertos cirujanos está proponiendo la conformación de un enfoque de "*red team*" que evalúe, con una visión adversaria, planes, políticas, sistemas y generalizaciones, con el objeto de eliminar el sesgo existente en los conjuntos de datos [11]. Otros desarrollos para la centralización de la información médica ya incluyen dentro de sus planes la generación de conjuntos de datos anotados por conjuntos de profesionales médicos diseñados específicamente para la investigación [13].

5 Retos externos

No puede dejar de mencionarse una situación que representa un reto en la mayoría de los países: la naturaleza legal de la información médica. Es conocido el caso del manejo de la pandemia de COVID-19 en China, el cual es explicado por Wu y su equipo en [14]. En este trabajo se ejemplifica lo que puede lograrse cuando se tiene acceso a la información personal, médica, de compras, de redes sociales, de viajes, de cámaras en ciudades inteligentes, etc., para poder rastrear a los posibles infectados durante la primera etapa de la pandemia y, entonces, tener conocimiento de qué personas habían estado en una situación de riesgo. Sin embargo, en cualquier otro país del mundo esto hubiera sido alarmante por la calificación de sensible que tiene la información clínica [15].

La información clínica tiene una doble clasificación de privacidad: es información personal y es información sensible. La posesión de información personal por particulares está regulada por la Ley General de Protección de Datos en Posesión de Particulares, la cual establece las obligaciones que adquiere un particular al comenzar a almacenar información personal. Sin embargo, es muy clara en su artículo 7, acerca de que en ningún punto podrá ser información sensible a menos que se obtenga el consentimiento expreso y por escrito de los involucrados.

El tema de la regulación no es trivial; numerosos estudios han sido realizados para evaluar los sistemas regulatorios de diferentes países y los puntos que aún faltan por cubrirse, por ejemplo, Bangladesh [13], la Unión Europea [16], Estados Unidos y Australia [17], entre otros. En México hace falta una legislación clara al respecto de los aspectos del almacenamiento, recuperación y uso de esta información en los procesos de investigación y desarrollo de herramien-

tas auxiliares. Es importante que nuevas leyes y reglamentos sean elaborados basados en una visión que fomente la investigación multicéntrica y multidisciplinaria.

Otro reto externo es el del fomento de la cultura de la salud en el país, el cual debe ser abordado tanto por educadores como por expertos en interacción humano-computadora que permitan a los usuarios discriminar de forma correcta la información que puedan recibir, por ejemplo, de redes sociales u otros medios. Uno de los grandes problemas que se sigue teniendo con las defunciones causadas por COVID-19 un año después de las campañas de vacunación es, precisamente, el surgimiento y fortalecimiento de grupos antivacunas en redes sociales. Se puede prever una situación similar en el momento en que se pretenda obtener la información genómica de las personas.

6 Conclusiones

A simple vista, podría decirse que este trabajo muestra una visión pesimista de que se pueda alcanzar el paradigma de la medicina personalizada en nuestro país. Sin embargo, esto es una idea incorrecta. Lo que se pretende es mostrar al lector el abanico de posibles líneas de investigación que se encuentran vigentes y algunas que en el futuro aparecerán conforme se vaya avanzando hacia un nuevo paradigma en salud. Como es de notarse, no hay una tarea que resuelva una sola ciencia; equipos multidisciplinarios de investigación deben involucrarse para cubrir las necesidades de cada grupo de actores. El cambio de paradigma de salud en el país solo será posible cuando se ofrezca a los actores involucrados las herramientas que requieran para motivar su participación en el proceso y, al mismo tiempo, beneficios a corto, mediano y largo plazo.

Debido a esto, este trabajo se inicia con el enfoque del paciente empoderado. En nuestra opinión, éste es el paso básico para poder llegar a soluciones prácticas: en el sentido de que, cuando el paciente reconozca que su salud es su responsabilidad (y no solo un derecho que debe ser provisto por el estado), entonces entenderá la importancia de otorgar acceso a su información, participar en campañas de salud pública realizadas con fundamento en evidencias y, por tanto, hacer el paradigma de la medicina personalizada cada vez más presente en su vida.

Referencias

[1] Goetz LH, Schork NJ (2018) Personalized medicine: motivation, challenges, and progress. Fertility and Sterility 109:952–963. https://doi.org/10.1016/j.fertnstert.2018.05.006

[2] Poornima S, Pushpalatha M (2020) A survey on various applications of prescriptive analytics. International Journal of Intelligent Networks 1:76–84. https://doi.org/10.1016/j.ijin.2020.07.001

[3] Dalli J, Gomez EA, Jouvene CC (2022) Utility of the Specialized Pro-Resolving Mediators as Diagnostic and Prognostic Biomarkers in Disease. Biomolecules 12:. https://doi.org/10.3390/biom12030353

[4] Junquera LM, Baladrón J, Albertos JM, Olay S (2003) Medicina basada en la evidencia (MBE): Ventajas. Revista Española de Cirugía Oral y Maxilofacial 25:265–272

[5] Ammenwerth E (2018) From eHealth to ePatient: The Role of Patient Portals in Fostering Patient Empowerment. ejbi 14: https://doi.org/10.24105/ejbi.2018.14.2.4

[6] Impulsan creación del Expediente Clínico Electrónico Único. http://comunicacion.senado.gob.mx/index.php/informacion/boletines/43499-impulsan-creacion-del-expediente-clinico-electronico-unico.html. Accessed 18 May 2020

[7] Mendoza GEM, Mendoza IAM, Herrero M de la CP de C, García MJS (2022) Relevancia de los Sistemas Personales de Salud durante la pandemia de COVID-19 en México. Revista de Comunicación y Salud 12:61–81. https://doi.org/10.35669/rcys.2022.12.e287

[8] Pirracchio R, Cohen MJ, Malenica I, et al (2019) Big data and targeted machine learning in action to assist medical decision in the ICU. Anaesthesia Critical Care & Pain Medicine 38:377–384. https://doi.org/10.1016/j.accpm.2018.09.008

[9] Bertsimas D, Orfanoudaki A, Weiner RB (2020) Personalized treatment for coronary artery disease patients: a machine learning approach. Health Care Manag Sci 23:482–506. https://doi.org/10.1007/s10729-020-09522-4

[10] Mosavi N, Santos M (2020) How Prescriptive Analytics Influences Decision Making in Precision Medicine. Procedia Computer Science 177:528–533. https://doi.org/10.1016/j.procs.2020.10.073

[11] Hechi ME, Ward TM, An GC, et al (2021) Artificial Intelligence, Machine Learning, and Surgical Science: Reality Versus Hype. Journal of Surgical Research 264:A1–A9. https://doi.org/10.1016/j.jss.2021.01.046

[12] Musacchio N, Giancaterini A, Guaita G, et al (2020) Artificial Intelligence and Big Data in Diabetes Care: A Position Statement of the Italian Association of Medical Diabetologists. J Med Internet Res 22:e16922. https://doi.org/10.2196/16922

[13] Hassan S, Dhali M, Zaman F, Tanveer M (2021) Big data and predictive analytics in healthcare in Bangladesh: regulatory challenges. Heliyon 7:e07179. https://doi.org/10.1016/j.heliyon.2021.e07179

[14] Wu J, Wang J, Nicholas S, et al (2020) Application of Big Data Technology for COVID-19 Prevention and Control in China: Lessons and Recommendations. J Med Internet Res 22:. https://doi.org/10.2196/21980

[15] Ienca M, Vayena E (2020) On the responsible use of digital data to tackle the COVID-19 pandemic. Nature Medicine 26:463–464. https://doi.org/10.1038/s41591-020-0832-5

[16] Broek T van den, Veenstra AF van (2018) Governance of big data collaborations: How to balance regulatory compliance and disruptive innovation. Technological Forecasting and Social Change 129:330–338. https://doi.org/10.1016/j.techfore.2017.09.040

[17] Thapa C, Camtepe S (2021) Precision health data: Requirements, challenges and existing techniques for data security and privacy. Computers in Biology and Medicine 129:104130. https://doi.org/10.1016/j.compbiomed.2020.104130

Plataforma Tecnológica para la Gestión, Aseguramiento, Intercambio y Preservación de Grandes Volúmenes de Datos en Salud: Muyal-Ilal

J.L. González-Compeán[1[0000-0002-2160-4407]], Diana E. Carrizales-Espinoza[1[0000-0002-3925-031X]], Dante D. Sánchez-Gallegos[1[0000-0003-0944-9341]] y J.A. Barrón-Lugo[1[0000-0002-9619-8116]]

[1] Cinvestav Tamaulipas, Cd. Victoria, México
{joseluis.gonzalez, diana.carrizales, dante.sanchez, juan.barron}@cinvestav.mx

Resumen. Los sistemas de expediente clínico electrónico (SECE) han sido herramientas clave para mejorar los procesos de atención de pacientes. Sin embargo, existen aún áreas de mejora en dos vertientes: i) la primera tiene que ver con el intercambio de contenidos en información médica entre múltiples roles de profesionales de la salud, múltiples niveles de atención e incluso múltiples instituciones de salud o gubernamentales; y ii) la segunda tiene que ver con la posibilidad de que sistemas computarizados de analítica de datos puedan convertir tanto las fuentes de datos de los SECEs (datos históricos), como la información producida por la práctica médica, sensores y dispositivos médicos, en información útil que soporten procesos de toma de decisiones. Para abordar estas dos vertientes es necesario contar con sistemas de ciencia de datos que permitan el manejo, transporte y preservación de datos sensibles de forma segura y confiable. En este capítulo se describe el desarrollo de Muyal-Ilal, una plataforma de servicios eficientes para la gestión, aseguramiento, intercambio, trazabilidad y almacenamiento de grandes volúmenes de datos médicos. En esta plataforma, las instituciones de salud pueden crear servicios seguros de inteligencia artificial que analicen datos médicos (notas médicas, bases de datos, etc.) para asistir procesos de toma decisiones (diagnósticos asistidos y predicciones de riesgo), servicios para visualizar información mediante graficación y mapas espaciotemporales, así como servicios para verificar/asegurar que los SECEs cumplan normas nacionales/internacionales de manejo de datos/contenidos personales.

Palabras clave: Ciencia de Datos · Big Data · Sistemas de E-salud · Almacenamiento de Datos

1 Introducción

En México, la atención médica es crucial para mejorar el bienestar y la calidad de vida de los ciudadanos [1]. Esta práctica profesional produce escenarios de procesamiento, intercambio, y preservación de grandes volúmenes de

datos que son producidos por diversas fuentes heterogéneas (p. ej., sensores, dispositivos médicos, tomografías, etc.) [2]. Estos datos, al ser sensibles (es decir, datos que contienen información personal que puede revelar aspectos como el origen racial o étnico, estado de salud, información genética, etc.), deben ser procesados rápidamente (*velocidad*) por un conjunto heterogéneo de sistemas de expedientes clínicos electrónicos o SECEs (*variedad*), los cuales entregan información útil a diferentes repositorios de datos (*veracidad y valor*) [3].

En este sentido, el Plan Nacional de desarrollo 2019-2024 permite vislumbrar las dimensiones de este escenario, donde, a finales de 2018, el IMSS contaba con 68.5 millones de derechohabientes, tanto el ISSSTE como el IMSS-Secretaría de Bienestar con más de 13 millones, y Sedena, Semar y Pemex con 2 millones. El IMSS, por ejemplo, desplegó su SECE en el 99% de unidades de atención primaria con adopción programada para hospitales de segundo y tercer nivel (en el *segundo nivel* se encuentran los hospitales de referencia, es decir, hospitales que, por su tamaño y calidad asistencial, se especializan y asumen pacientes complejos de otros hospitales del mismo servicio de salud, y el *tercer nivel* es el formado por hospitales de alta tecnología e institutos especializados). Sin embargo, a pesar de los esfuerzos que realizan las instituciones de salud pública, actualmente existe un déficit entre el número de médicos disponibles para atender a los pacientes y la población que requiere atención médica de cualquier tipo. Por ejemplo, hasta 2014 solo existían 220 médicos por cada 100,000 habitantes en México, lo cual representa un porcentaje aproximado del 31% inferior al número de médicos recomendados por esa cantidad de habitantes por la Organización para la Cooperación y el Desarrollo Económicos (OCDE) (*se recomiendan al menos 320 médicos*) [4].

En este contexto, se puede observar que existe una alta dispersión y separación geográfica tanto de las fuentes generadoras de los datos, como de los SECEs, los pacientes y de los profesionales de la salud que intervienen en el proceso de atención médica. Para mitigar los problemas que pueden generarse debido a esta alta dispersión, en los últimos años, dentro del ámbito médico, se ha hecho uso de tecnologías para el desarrollo de la ciencia de datos (el *NIST* define la ciencia de datos como la síntesis empírica de conocimiento procesable a partir de datos sin procesar durante todo el ciclo de vida de los datos) que hacen factible el diseño de sistemas de salud inteligentes e interconectados [5]. Para construir un sistema de este tipo, es necesario que las organizaciones hagan uso de diversas tecnologías y paradigmas, tales como el cómputo y almacenamiento en la nube, el *big data* (grandes volúmenes de datos) [6], [5] y el internet de las cosas (*IoT*, por sus siglas en inglés, *Internet of Things*) [7].

Un sistema de ciencia de datos debe considerar diferentes etapas para permitir la transformación de datos en conocimiento accionable e información útil a través de todo su ciclo de vida (desde su adquisición, hasta su consumo) [6]. De esta forma, un sistema de ciencia de datos incluye desde aspectos del dominio a estudiar (en este caso la *salud*) y de ingeniería, hasta de estadística y aprendizaje máquina [5].

No obstante, la construcción de este tipo de sistemas para el manejo de datos médicos no es una tarea sencilla. Lo anterior se debe a que es necesario adecuar estos sistemas para que cumplimenten con las distintas normas nacio-

nales e internacionales existentes relacionadas con el intercambio y preservación de datos sensibles. Por ejemplo, las normas oficiales NOM-024-SSA3-2010 y NOM-004-SSA3-2012 [8], así como la Ley Federal de Protección de Datos Personales, establecen los requisitos funcionales (manejo de datos y metadatos médicos) que deben ser cubiertos mediante el cumplimiento de distintos estándares internacionales (p. ej., NIST y COBIT 5). Además, estas normas establecen un listado de requerimientos no-funcionales (p. ej., seguridad, confiabilidad, eficiencia, integridad, trazabilidad, etc.), los cuales deben ser garantizados por las instituciones que instalan y operan un sistema para el manejo de datos médicos (p. ej., un SECE), lo cual lo convierte en un gran desafío para las instituciones de salud mexicanas [9].

En consecuencia, tomando en cuenta el gran volumen de pacientes con respecto al número de profesionales de la salud especializados, no solo se requiere que cualquier direccionamiento de pacientes a los profesionales de la salud sea realizado con la mayor seguridad, eficiencia, confiabilidad y prontitud posible, sino que también se les provea a los médicos y especialistas de las herramientas necesarias para mejorar y/o eficientizar los flujos de trabajo y el intercambio de los datos asociados a la prognosis, diagnosis y tratamiento. De la misma forma, es necesario proveer de sistemas de ciencia de datos que agilicen el proceso de toma de decisiones, tales como los sistemas de diagnóstico asistido por inteligencia artificial y sistemas de analítica, para convertir cúmulos de datos en información útil que ayude en los procesos críticos de toma de decisiones (p. ej., el transporte de una radiografía de un técnico radiólogo a un especialista para emitir un diagnóstico al paciente) [10], [11].

La complejidad de este tipo de escenarios aumenta cuando se deben preservar los derechos de los pacientes y/o profesionales de la salud a la privacidad, disponibilidad, integridad, confidencialidad y auditoria de sus datos [12]. La heterogeneidad, tanto de las aplicaciones como de las infraestructuras, y la dispersión geográfica de los participantes, así como el volumen de los datos producidos de forma constante, aumentan considerablemente la complejidad del problema que resulta proveer, tanto a pacientes como a profesionales de la salud, un ambiente seguro, confiable, controlado y eficiente para mejorar la calidad de la atención que se brinda a los pacientes.

Otro aspecto para tomar en cuenta, que impacta a las instituciones prestadoras de servicios de salud, son las dependencias que se presentan cuando se intentan crear estos sistemas de ciencia de datos, así como sistemas para el intercambio de datos [13]. Dichas dependencias se presentan con los desarrolladores de los servicios, quienes tienen tiempos elevados de construcción de los sistemas e imponen el pago constante de licencias de uso, lo cual crea una relación funcional estrecha entre el proveedor y la institución contratante. Esta dependencia también podría presentarse con proveedores de servicio externos (*servicios en la nube*), a quienes las instituciones podrían delegar procesos de almacenamiento, distribución de datos, procesamiento y análisis (p. ej., servicios en línea de inteligencia artificial). Lo anterior, implica que las instituciones se encuentran delegando a terceros el control sobre los contenidos y las aplicaciones desplegadas en las infraestructuras del proveedor. La pérdida de control sobre datos sensibles y/o sistemas críticos para la toma de decisiones puede llegar a derivar en accesos no controlados, violaciones de integridad, pérdida de confidencialidad, privacidad o extravío temporal o permanente de

los datos [14]. Por ende, eliminar las dependencias existentes y mantener el control sobre los datos de carácter sensible resulta crítico para las instituciones de salud.

En este capítulo se presenta el diseño y desarrollo de *Muyal-Ilal*, una plataforma de diseño, gestión y construcción de sistemas de ciencia de datos para el manejo, procesamiento y análisis de grandes volúmenes de datos. Muyal-Ilal permite a las organizaciones construir sistemas de ciencia de datos seguros y propios, en el cual ellos pueden tener el control de dónde y cómo se procesan y almacenan sus datos. Para ello, Muyal-Ilal incluye cinco conjuntos de servicios diferentes: i) Muyal-Nez; ii) Muyal-Chimalli; iii) Muyal-Xelhua; iv) Muyal-Painal; y v) Muyal-Alwa. La Figura 1 muestra una representación conceptual del manejo del ciclo de vida de los datos en un sistema de ciencia de datos utilizando la plataforma Muyal-Ilal.

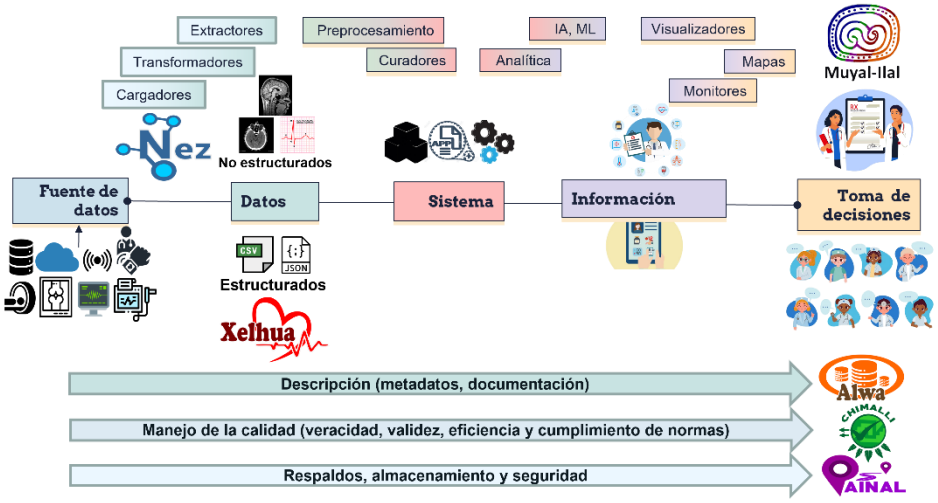


Figura 1. Manejo del ciclo de vida de los datos en un sistema de ciencia de datos utilizando la plataforma *Muyal-Ilal*.

Muyal-Nez permite la creación de sistemas de ciencia de datos para el procesamiento de datos no estructurados. De forma similar, Muyal-Xelhua permite crear sistemas de ciencia de datos para el análisis de datos estructurados. Por otro lado, Muyal-Chimalli, de forma automática y transparente, provee tolerancia a fallas en TICs, así como privacidad, confidencialidad, integridad, disponibilidad y trazabilidad en datos médicos para que los sistemas creados en Muyal cumplan las normas nacionales (NOM-024-SSA3-2010 y NOM-004-SSA3-2012) e internacionales (ISO 27001:2013, COBIT 5, NIST) para el manejo, intercambio y preservación de datos sensibles. Además, Muyal-Painal permite la construcción de sistemas eficientes de almacenamiento y distribución de datos/contenidos médicos para que las instituciones soporten escenarios de intercambio ininterrumpido de catálogos de bases de datos, resultados/información y/o sistemas e-Salud a través de intra/internet. Finalmente, Muyal-Alwa permite la construcción de servicios de repositorios (estandariza-

dos y FAIR) para facilitar el acceso a catálogos publicados por las instituciones de salud.

En la Figura 2 se muestran los principales componentes que provee cada servicio de la plataforma Muyal-Ilal. De forma estratégica, los componentes de la plataforma se han creado para que sean utilizados de forma independiente; lo anterior incluye el software que forma parte de cada servicio.

Figura 2. Características principales de la plataforma Muyal-Ilal.

Muyal-Ilal ha sido utilizado para construir sistemas de ciencia de datos para diagnóstico de cáncer de hueso largo y pulmón asistido por inteligencia artificial, estudios espaciotemporales de enfermedades de alta prevalencia con georreferenciación y calculadoras de medición de riesgo de enfermedades cardiovasculares que producirán bases de datos y repositorios para la comunidad científica e instituciones de salud.

El resto del capítulo se encuentra estructurado de la siguiente manera: la Sección 2 presenta los antecedentes de los trabajos realizados previamente en el grupo de investigación relacionados con el manejo, procesamiento y transporte de datos sensibles: la Sección 3 describe detalladamente el desarrollo y diseño de la plataforma Muyal-Ilal; la Sección 4 presenta los principales resultados obtenidos con la plataforma; finalmente, la Sección 5 concluye este trabajo dando un resumen sobre la plataforma presentada en este capítulo.

2 Antecedentes

Los sistemas de expediente clínico electrónico (SECEs), dispositivos médicos y plataformas para el manejo de datos de salud permiten mejorar los tiempos de respuesta de la atención a los pacientes del sector salud de México. Además, estos componentes posibilitan el fortalecimiento de mecanismos de control de los sistemas de salud y permiten a las instituciones mantenerse en línea con las diversas políticas, normas (por ejemplo, NOM-024-SSA3-2010 y NOM-004-SSA3-2012 así como ISO-270001-13, COBIT5, NIST) y estándares (por ejemplo, DICOM [15] y HL7 [16], [17], [18]) para el manejo, transporte y preservación de datos sensibles. En este contexto, el término "*sistema de e-Salud*" se refiere al uso de las diversas tecnologías de la información, telecomunicaciones y manejo de operaciones para integrar los componentes antes mencionados en un solo sistema coherente y de fácil manejo, cuya operación se base en las normas y protocolos oficiales.

Previamente, en nuestro grupo de trabajo, se han realizado desarrollos tecnológicos con el fin de cumplir con la norma oficial mexicana en su rubro de conservación y almacenamiento de datos clínicos. Específicamente, el Instituto Nacional de Rehabilitación (INR), que forma parte del grupo de trabajo que ha desarrollado la plataforma *Muyal-Ilal*, implantó un sistema llamado PACS-INR, el cual permite manejar, distribuir, almacenar y recuperar imágenes médicas con calidad de diagnóstico en formato DICOM.

En 2015, el INR, en colaboración con la empresa INFOTEC, desarrollaron una integración de PACS-INR con BABEL, un sistema distribuido de almacenamiento tolerante a fallos, el cual almacena a la fecha 82 millones de imágenes de diferentes modalidades (RM, TC, US, MN y RX) distribuidas en 73 TB de almacenamiento físico.

Bajo este contexto, el INR desarrolló funcionalidades de almacenamiento confiable, basado en tecnología nacional, lo cual permitió eliminar costos de licencias y cambiar hardware de gama alta por equipo de cómputo de gama media. Este sistema, sin embargo, actualmente no considera mecanismos de intercambio seguro de datos, la construcción de flujos de trabajo inter/intrainstitucionales, o la incorporación de herramientas de mejora funcional de preprocesamiento o sistemas de diagnóstico asistido por inteligencia artificial. En este sentido, el manejo de datos sensibles (como lo son las imágenes médicas) implica la utilización de estructuras de datos que permitan acceder tanto a los datos como a los metadatos, así como sistemas de colocación y localización que permitan compartir, cargar, descargar, visualizar y/o eliminar datos/metadatos.

El transporte de los datos adquiridos es uno de los grandes problemas que se presentan cuando se manejan datos sensibles. Lo anterior se debe a que, al ser datos sensibles, es necesario garantizar servicios de seguridad tales como integridad, privacidad, confidencialidad, trazabilidad y estrictos controles de acceso. De igual manera, al transportar grandes volúmenes de datos de un sitio a otro, es necesario contar con distintas técnicas que permitan eficientizar dicho proceso. Para hacer frente a este tipo de problema, se deben utilizar herramientas que permitan realizar operaciones de cifrado/descifrado para proveer seguridad a los datos, así como operaciones concurrentes/paralelas para mejorar la utilización de los recursos computacionales.

Esquemas criptográficos experimentales para compartir, de forma segura, contenidos en ambientes organizacionales a través de la nube, que hasta la fecha solo se habían probado conceptualmente y mediante modelos matemáticos, mostraron la factibilidad de proteger datos sensibles en escenarios reales en términos de integridad (mediante firmas digitales), confidencialidad y control de acceso (mediante el cifrado basado en atributos), así como privacidad (mediante el uso de sobres digitales). En nuestro grupo de trabajo también se desarrollaron esquemas para el transporte confiable, seguro, flexible y eficiente de contenidos, los cuales fueron probados en escenarios reales para que agencias compartan datos espaciales e imágenes satelitales y, posteriormente, fueron probados con éxito en el transporte de imágenes de tomografía.

Estas soluciones incluyen mecanismos para reducir el consumo de almacenamiento producido por procesos de tolerancia a fallos, mecanismos de paralelismo basados en tuberías de procesamiento que mejoran la eficiencia en el procesamiento de datos e imágenes, herramientas de transporte de grandes volúmenes de información a través de flujos de trabajo creados por diferentes usuarios y agencias, y mecanismos que garantizan la integridad de los datos transportados de extremo a extremo. En este contexto, las soluciones extremo a extremo permiten a los usuarios proteger sus datos antes de que estos sean enviados a la nube, así como agregar confiabilidad a los mismos (mediante el cifrado basado en atributos) y soportar fallas (mediante técnicas de dispersión de información) que pueden ser provocadas por diversos problemas, tales como interrupciones de los servicios e incidentes de bloqueo por parte de los proveedores.

Además, se han desarrollado esquemas basados en emparejamientos criptográficos, los cuales conforman un servicio de almacenamiento de extremo a extremo para modelos de nube híbrida. Estos servicios permiten realizar el intercambio de archivos en escenarios donde los datos son enviados a la nube y grupos selectos y autorizados de usuarios pueden acceder a los mismos. Los escenarios reales donde nuestras propuestas fueron probadas en el pasado, guardan similitud con los escenarios de movilidad y flujos de trabajo donde se comparten datos clínicos realizados por profesionales de la salud al interior/exterior de instituciones y exhiben marcadas similitudes con los requerimientos de manejo, almacenamiento y transporte de imágenes médicas (seguridad, confiabilidad, eficiencia y confidencialidad).

3 Muyal-Ilal: sistemas de ciencia de datos para el manejo de grandes volúmenes de datos

En esta sección se describe Muyal-Ilal, una plataforma de software de ciencia de datos que habilita a las organizaciones médicas [1], profesionales de la salud y usuarios del sistema de salud público/privado la creación de sistemas portables y eficientes para escenarios intrainstitucionales e interinstitucionales (es decir, de forma interna y externa).

3.1 Construcción de sistemas de e-salud basada en conceptos de software autosimilar y autocontenido

Las arquitecturas tradicionales de cómputo permiten la creación de plataformas de e-salud que, al utilizar herramientas y software de terceros, generan una dependencia con la infraestructura de cómputo (p. ej., Amazon EC2 o Microsoft Azure).

En cambio, la arquitectura de Muyal-Ilal está basada en conceptos de software autosimilar y autocontenido, lo cual permite construir diferentes plataformas agnósticas e independientes de la infraestructura. El concepto de autocontenido hace referencia a la creación de paquetes de software que contienen todos los elementos necesarios para que estos puedan ser desplegados en diferentes infraestructuras con el mínimo esfuerzo. En este sentido, los componentes de Muyal-Ilal se encuentran diseñados para funcionar como objetos autocontenidos, los cuales contienen los modelos descriptivos, de procesamiento, de comunicación y de programación requeridos para que el objeto funcione; eso es, sin importar en que infraestructura se despliegue.

Por otro lado, la característica de autosimilitud hace referencia a que la estructura de estos objectos es idéntica entre sí. En este sentido, para agregar esta característica a un objeto autocontenido, es necesario agregar interfaces de entrada y salida, las cuales enmascaran el funcionamiento de este objeto, haciéndolo parecer una caja negra (es decir, donde el usuario no puede ver el funcionamiento interno del objeto) de cara al usuario y permitiendo que estos objetos puedan ser conectados con otros objetos y componentes de la plataforma Muyal-Ilal.

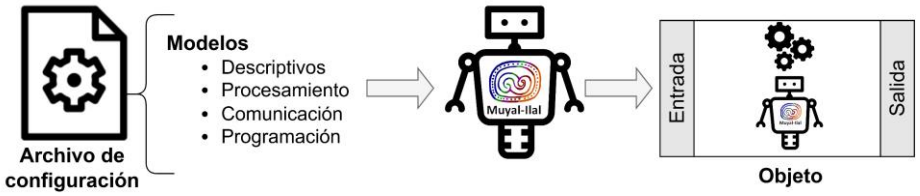


Figura 3. Representación conceptual de la creación de objetos autosimilares y autocontenidos.

La Figura 3 muestra la representación gráfica de la construcción de un objeto autosimilar y autocontenido utilizando la arquitectura de la plataforma Muyal-Ilal. Como se puede observar en la Figura 3, cada objeto funciona como un robot que permite la creación de los diferentes elementos de un sistema de ciencia de datos para e-salud (*procesamiento de datos estructurados y no estructurados, descripción, manejo de calidad y almacenamiento*).

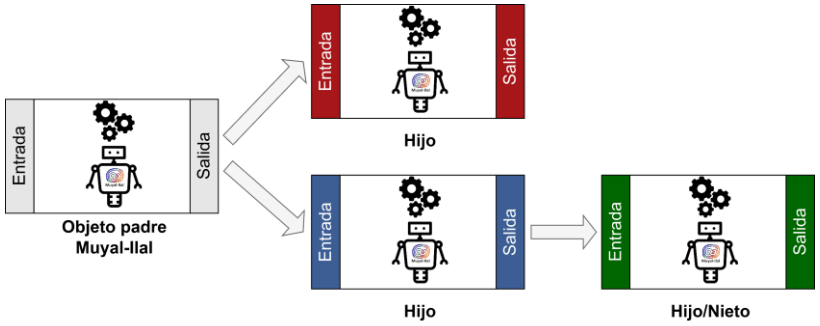


Figura 4. Los objetos construidos con Muyal-Ilal tienen las características de herencia y polimorfismo, lo cual permite crear nuevas aplicaciones heredando el comportamiento de una aplicación base.

Similar a la programación orientada a objetos, los objetos en Muyal-Ilal incluyen distintas características tales como la herencia y polimorfismo. La Figura 4 presenta un ejemplo donde un objeto con una identidad, un comportamiento y un estado es clonado para generar *"hijos"*, los cuales heredan sus atributos del *"padre"*. De la misma forma, los *"hijos"* pueden realizar modificaciones a dichos atributos para que el nuevo objeto implemente un comportamiento diferente al de su *"padre"*. Esto permite crear nuevos objetos a partir de un *molde padre*, el cual que puede ser especializado para realizar alguna tarea dentro de un sistema construido con la plataforma Muyal-Ilal.

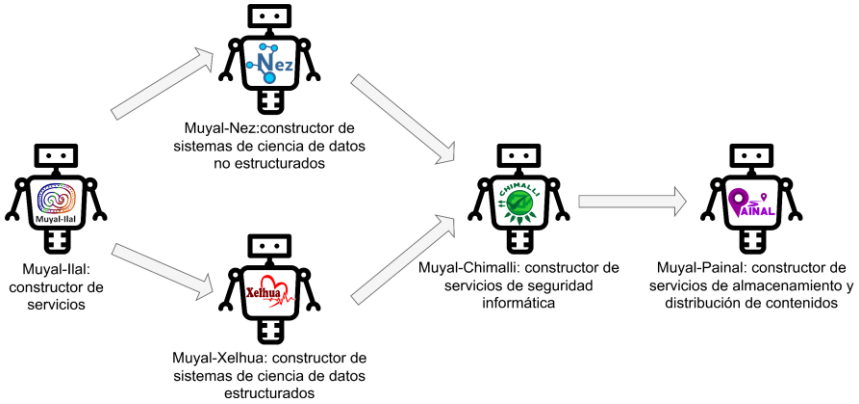


Figura 5. Principales productos construidos con Muyal-Ilal.

Para hacer frente a los retos que surgen al construir un sistema de ciencia de datos (procesamiento de datos estructurados y no estructurados, almacenamiento de grandes volúmenes de datos, preservación de la seguridad de los datos y manejo de metadatos), a partir de un objeto génesis llamado *Muyal-Ilal*, se construyeron cinco plataformas (ver Figura 5): i) Muyal-Nez, para la construcción de sistemas de ciencia de datos no estructurados; ii) Muyal-Xelhua, para la construcción de sistemas de ciencia de datos estructurados; iii) Muyal-Chimall, para la construcción de servicios de seguridad informática; iv) Muyal-Painal, para la construcción de servicios de almacenamiento y distribución de contenidos; y v) Muyal-Alwa, para la creación de servicios de manejo de repositorios de datos.

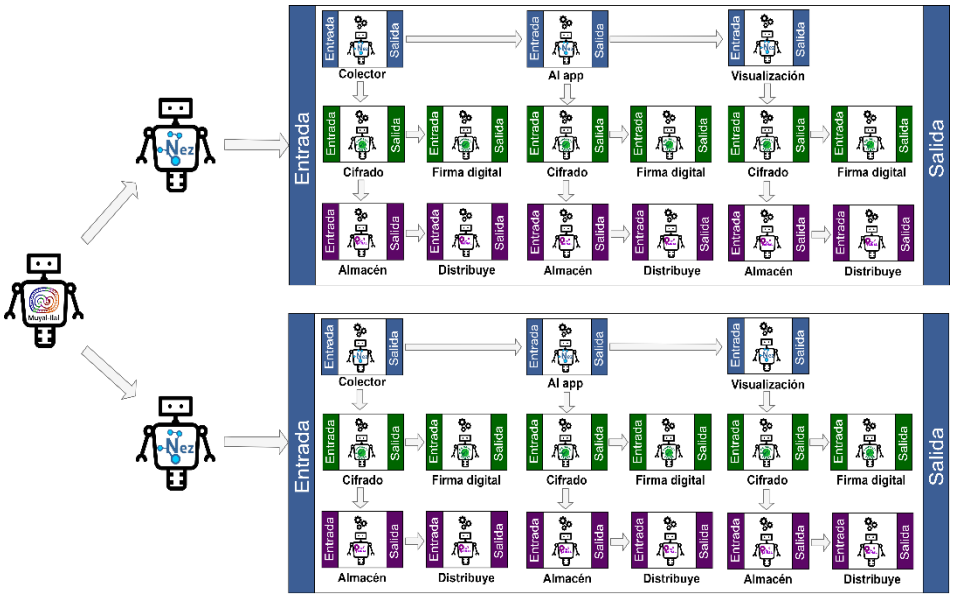


Figura 6. Creación de sistemas complejos de ciencia de datos con Muyal-Ilal.

Las propiedades de autosimilitud y autocontenido de Muyal-Ilal permiten la creación de sistemas de ciencia de datos complejos para el procesamiento y manejo de grandes volúmenes de datos. Por ejemplo, en la Figura 6 se muestra un sistema creado a partir de la creación de clones de sistemas previamente creados con Muyal-Ilal. En este sentido, la creación de clones permite que los datos de entrada sean distribuidos entre los diferentes clones, lo cual reduce la carga de trabajo de los sistemas.

Esta capacidad de Muyal-Ilal para crear clones de sistemas ya existentes permite que una solución pueda ser reutilizada por otra organización para el procesamiento de sus propios datos en la infraestructura de su organización.

3.2 Muyal-Nez: Servicio de construcción de sistemas e-salud para el procesamiento de datos no-estructurados

Muyal-Nez fue desarrollado mediante un esquema declarativo para que el usuario (diseñador o científico de datos) solo se preocupe por dos cosas: i) por elegir, desde el repositorio de servicios, las aplicaciones (existentes en las instituciones de salud o en la comunidad científica) que contendrán sus sistemas de e-salud y ii) por declarar los lugares donde dichas aplicaciones se ejecutarán (p. ej., la nube, servidores del hospital o computadoras de los profesionales de la salud). Muyal-Nez incluye una novedosa arquitectura de microservicios y nanoservicios que encapsula las aplicaciones de los sistemas (p. ej., procesamiento de imágenes) en cripto-contenedores, los cuales son desplegados, manejados y controlados de forma automática. Los principales resultados y beneficios de Nez son: i) los usuarios pueden crear sistemas de e-salud en minutos, reusando aplicaciones existentes, sin programar o realizar configuraciones complejas, lo cual elimina la dependencia tecnológica de las instituciones de salud con los proveedores de software; ii) los sistemas de e-salud pueden ser desplegados en diferentes infraestructuras sin modificar el código de las aplicaciones; iii) Muyal-Nez permite la reusabilidad de las aplicaciones al permitir que diferentes organizaciones (hospitales) y usuarios las utilicen en su infraestructura sin tener que volver a programarlas; y iv) un nuevo esquema de paralelismo y cómputo de alto rendimiento que permite a los sistemas de e-salud maximizar el uso de los recursos de cómputo.

3.3 Muyal-Xelhua: Servicios para análisis de datos estructurados

Muyal-Xelhua es una plataforma para la construcción de sistemas de ciencia de datos estructurados bajo demanda, sin amplios conocimientos en programación y orientada al diseño. Permite utilizar aplicativos para la preparación de datos, análisis y visualización de datos, así como acoplarlos unos con otros y crear flujos de trabajo personalizados. En Muyal-Xelhua, los datos son procesados por cada una de las aplicaciones (encapsuladas en microservicios) que el usuario selecciona. Lo anterior, permite generar nuevos productos derivados (p. ej., datos transformados, gráficas, mapas, estadísticos, o cualquier producto accionable que permita realizar observaciones sobre los datos, o bien, permita tomar una decisión).

3.4 Muyal-Chimalli: Servicios de seguridad y tolerancia a fallos para sistemas de e-salud

Muyal-Chimalli fue diseñado para que los sistemas de e-salud cumplan, automáticamente, con las normas nacionales e internacionales referentes al manejo, preservación e intercambio seguro y confiable de datos sensibles. Muyal-Chimalli provee los siguientes resultados y beneficios: i) seguridad en términos de confidencialidad, integridad, trazabilidad, no repudio y controles de acceso de los datos manejados por las aplicaciones de los sistemas de e-salud. Esto se realiza mediante una suite de algoritmos de criptosistemas de siguiente generación basados en matemática de curva elíptica. Los resultados de estos criptosistemas muestran que los datos asegurados por Muyal-Chimalli serán in-

descifrables por terceros en un período de hasta 30 años en el peor de los casos; ii) un esquema de trazabilidad de bloques criptográficos que permite la creación, de forma automática, de una red de verificabilidad (blockchain) y registra eficientemente cada operación realizada con los datos médicos, permitiendo a los profesionales de la salud, entidades de gobierno y pacientes obtener un reporte de las acciones realizadas sobre sus datos, lo cual reduce costos monetarios (p. ej., el costo de una red blockchain durante 4 meses en la nube es de aproximadamente $3,400 dólares); iii) un esquema de paralelismo basado en patrones recursivos, que hace factible asegurar los datos médicos –debido al esquema de paralelismo, Muyal-Chimalli mejora los tiempos de respuesta hasta 10.22 veces en comparación con los sistemas criptográficos disponibles actualmente–; iv) un servicio de compartición segura de datos sensible que permite a profesionales de la salud crear grupos de intercambio de datos seguros en escenarios intrainstitucionales (compartir datos con personal de la misma organización) e interinstitucionales (con usuarios de instituciones externas); v) un sistema de verificación del cumplimiento de las normas y protocolos oficiales para el manejo, intercambio y preservación de datos sensibles –en Muyal-Chimalli, el cumplimiento de estas normas y protocolos es mostrado mediante un reporte donde se especifica el porcentaje de cumplimiento del servicio según las normas o protocolos correspondientes; además, el servicio realiza el descubrimiento del flujo de trabajo asociado a los archivos de configuración del servicio de e-Salud–; y *vi)* búsquedas cifradas que evitan que el personal de informática, que se supone es honesto, pero podría ser curioso, tenga acceso a los contenidos almacenados en los servidores de las instituciones de salud, lo cual es clave cuando se contratan servicios de cómputo y almacenamiento en la nube.

3.5 Muyal-Painal: Servicio para el intercambio seguro y confiable de datos médicos

Muyal-Painal fue agregado a la plataforma Muyal-Ilal para permitir el intercambio de datos médicos entre organizaciones de forma segura y eficiente. Los principales resultados y beneficios de esta plataforma son: i) la construcción, de forma automática y sin intervención del personal de la salud, de sistemas de almacenamiento eficientes y tolerantes a fallas, los cuales reducen los costos de almacenamiento en la nube hasta en un 70%, mejoran 4 veces el rendimiento de sistemas similares y soportan múltiples fallas mediante un esquema de tolerancia a fallas configurable; ii) creación de sistemas de logística y distribución de datos para los sistemas de e-salud, los cuales reducen la cantidad de datos transmitidos por la red (intranet e intranet) y, además, permiten crear áreas de intercambio federadas para escenarios intra e interinstitucionales; iii) un modelo de publicación/suscripción que permite al personal de las instituciones de salud publicar catálogos de datos médicos que pueden ser suscritos (descargados y accedidos) de forma inter e intrainstitucional; iv) Muyal-Painal incluye un sistema de sincronización automática de datos entre organizaciones y usuarios, el cual reduce hasta en un 48% la transferencia de datos al identificar datos replicados, lo que minimiza los costos de envió y almacenamiento de información, mejorando la experiencia de servicio del usuario final (paciente o profesional de la salud).

Utilizando Muyal-Nez, con el apoyo de Muyal-Chimalli para el manejo de datos, se creó un sistema de e-salud para el diagnóstico asistido de cáncer de hueso largo y pulmón mediante IA (inteligencia artificial). El sistema extrae automáticamente tomografías, identifica lesiones malignas en hueso (95%) y pulmones (100%) y entrega el diagnóstico al profesional de la salud.

3.6 Muyal-Alwa: Servicio de repositorios FAIR

Muyal-Alwa es un sistema de repositorios (FAIR) basado en sistemas estandarizados de exposición de catálogos, el cual permite automatizar la publicación y consumo para uso privado y/o público de datos, componentes o sistemas completos de e-Salud, así como información producida por los usuarios en la plataforma de Muyal-Ilal. Cada contenido agregado/descargado en Muyal-Alwa posee un *"pasaporte"* que es revisado por Muyal-Painal y validado por Muyal-Chimalli previo a su incorporación a los repositorios de Muyal-Alwa o Al consumo por parte de los usuarios a través del internet.

3.7 Creación de múltiples flujos de datos con Muyal-Ilal

Por ejemplo, en la Figura 7 se muestran diferentes flujos construidos con estas plataformas. En estos, Muyal-Nez guía al personal de informática y profesionales de la salud para crear, en minutos y sin programar ni realizar instalaciones, sistemas de e-salud (p. ej., algoritmos de IA o sistemas de analítica o visualización de datos). Muyal-Xelhua permite la creación de sistemas de analítica para convertir grandes volúmenes de datos (históricos o estadísticos) en información útil para la toma decisiones. Muyal-Chimalli garantiza y verifica que los sistemas de e-salud, construidos con Muyal-Nez y Muyal-Xelhua, cumplan con los requerimientos de las normas nacionales (NOM-024-SSA3-2010 y NOM-004-SSA3-2012) e internacionales (NIST, ISO27001:2013 y COBIT5) sobre el manejo, transporte y preservación de datos sensibles.

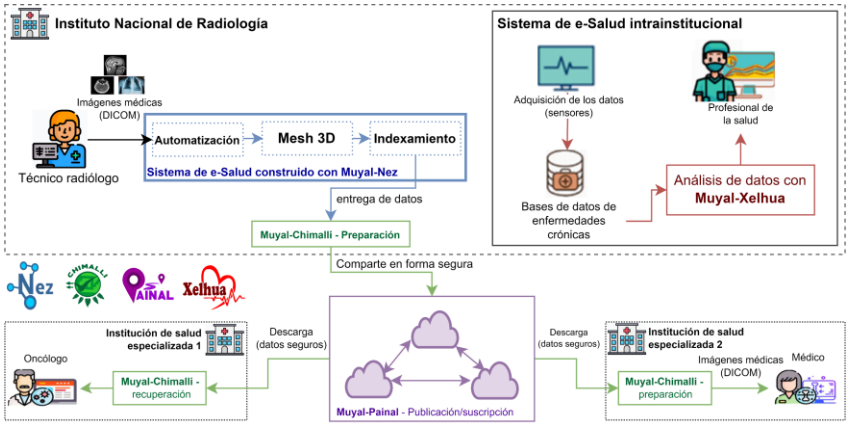


Figura 7. Representación conceptual de múltiples flujos de datos creados con Muyal-Ilal.

Muyal-Painal crea sistemas de logística en la nube para compartir y transportar los datos, así como reutilizar sistemas disponibles entre múltiples instituciones, evitando la dependencia tecnológica con proveedores de servicios de software e infraestructura, permitiendo entregar el control a los dueños de los datos. Finalmente, Muyal-Alwa es un servicio que crea repositorios en línea en el cual convergen los datos, bases de datos, contenidos médicos de entrenamiento y validación, así como reportes y resultados producidos por los sistemas de e-salud.

4 Principales resultados obtenidos

Actualmente, Muyal-Nez permite al INR conectar su PACS-INR con las diferentes estaciones de los profesionales de la salud (p. ej., radiología, consulta y oncología). Se pretende extender a epidemiología y demás departamentos del Instituto, así como con otros institutos u hospitales referentes que envían pacientes al INR. Muyal-Chimalli ha permitido alcanzar un 70% de las regulaciones estandarizadas en forma internacional para el manejo seguro de datos sensibles y cubre todas las fases de interconexión establecidas por las normas oficiales. El 30% de requerimientos no alcanzados se debe a que la mayoría de estos se deben de cumplimentar de forma asistida por los administradores de la infraestructura en donde se desplegarán los sistemas e, incluso, por parte de los usuarios, como en el caso del manejo de sus llaves de acceso al sistema. El resumen de seguridad creado por Chimalli también ha permitido revelar las tareas de ciberseguridad que no dependen de la plataforma Muyal, sino de actividades realizadas por personal de salud, para que las instituciones creen un plan para implementarlas. Muyal-Xelhua ha permitido extender el análisis de contenidos (imágenes del PACS) a datos (históricos y bases de datos del sistema de expediente), y extender el servicio no solo a traumatología y oncología, sino también a epidemiología y departamentos asociados a la toma de decisiones y/o intervención de salud pública. Muyal-Painal les permite compartir, mediante transferencia tecnológica, sus sistemas con otros hospitales, los cuales podrán ser consumidos en la nube o descargados en su infraestructura. Esto reduce el tiempo de obtención de sistemas de e-Salud a minutos u horas, dependiendo de la modalidad de transferencia elegida, lo cual reduce la dependencia con los proveedores de software e infraestructura, así como el pago constante de licencias.

Actualmente, los sistemas que conforma Muyal-Ilal superan a los sistemas actuales experimentales. Muyal-Nez mejora la eficiencia de los procesos de manejo de imagenología en 4.9 veces, en comparación con un sistema tradicional, y reduce el consumo de almacenamiento hasta en un 66% gracias a los algoritmos de compresión de datos y detección de archivos duplicados implementados. Muyal-Chimalli reduce los tiempos de aseguramiento de datos hasta en un 80%, con respecto a sistemas criptográficos similares, y provee cualidades de integridad, trazabilidad, control de acceso, privacidad, confidencialidad y tolerancia a fallos en un solo sistema sin que los usuarios configuren, programen o instalen nada (esto produce un cumplimiento inmediato de hasta el 70% de los requerimientos de los estándares internacionales).

4.1 Flujo para detección asistida para cáncer de huesos largos y detección de nódulos en pulmón

El Sistema de e-salud para el diagnóstico asistido de cáncer de hueso largo y pulmones mediante inteligencia artificial, proporciona un apoyo al especialista, indicándole las imágenes en las que podría existir algún tumor y la región en la que se ubica.

Dado que los tumores presentan diferentes características dependiendo de la parte del cuerpo en la que se desarrollen, es necesario generar un modelo diferente para cada uno. Anteriormente, se utilizaron técnicas en las que se buscaban intensidades de las tomografías que pudiesen indicar, por ejemplo, algún tipo de clasificación en algunos órganos, pero, en el caso particular de cáncer de hueso, estas técnicas no siempre resultan eficientes, ya que el tumor puede presentar intensidades similares al de un hueso sano. Es por ello que, para este trabajo, se optó por utilizar nuevas técnicas de visión artificial que permiten distinguir diferentes patrones más allá de diferentes intensidades.

Se diseñó un flujo para la detección asistida de cáncer de huesos largos y nódulos en pulmón (ver Figura 8). El flujo permite la obtención de las imágenes DICOM, en donde el proceso de extracción de éstas contiene un formato adecuado. Posteriormente, se realiza una partición de éstas de manera aleatoria para formar un conjunto de entrenamiento, pruebas y validación, en los que regularmente se toman el 70%, 20% y 10% de la cantidad total de las imágenes; mismas que generan un TF Records a partir de imágenes y XML´s. Continuando con el flujo, se realiza una tubería de entrenamiento extrayendo las imágenes y entrenando el modelo, seguido de la exportación del modelo y prueba de éste para obtener el conjunto de pruebas para la detección asistida para cáncer de huesos largos con base en marcos de referencia de las normas internacionales (NIST, ISO 27001:2013 y COBIT 5) y nacionales (NOM-024-SSA3-2010). El programa, además, descubre el flujo de trabajo asociado a los archivos de configuración.

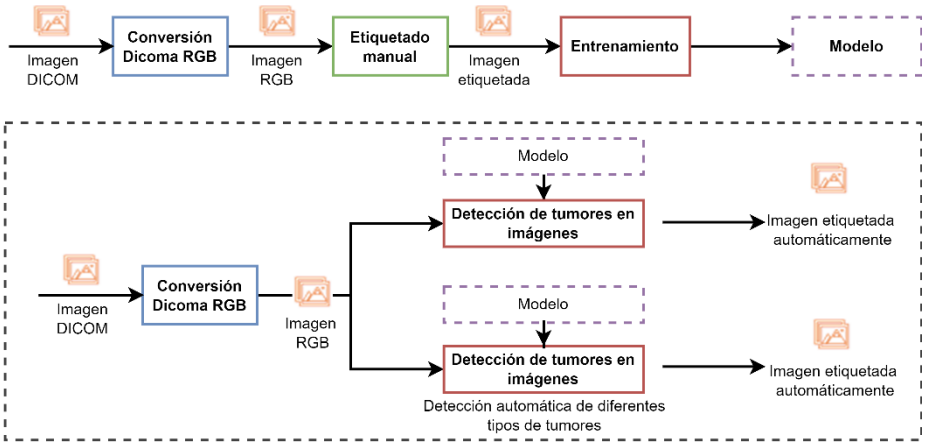


Figura 8. Representación general del flujo para la detección de cáncer de huesos largos y nódulos en pulmón.

5 Conclusiones

En este capítulo se presentó el desarrollo de una plataforma para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud (*big data*) y construcción de un repositorio nacional de servicios de análisis de datos de salud llamada Muyal-Ilal. Esta plataforma se encuentra compuesta de cinco servicios principales: i) Muyal-Nez; ii) Muyal-Xelhua; iii) Muyal-Chimalli; iv) Muyal-Painal; y v) Muyal-Alwa. Muyal-Nez permite a instituciones de salud crear, en minutos y sin programación ni instalaciones complejas, sistemas (e-Salud) intrainstitucionales (para diferentes profesionales de la salud) e interinstitucionales (para múltiples instituciones de salud y/o gubernamentales) para el intercambio ininterrumpido de datos/contenidos médicos, así como asistir diagnósticos con inteligencia artificial. Muyal-Nez automáticamente construye sistemas de almacenamiento y distribución de datos/contenidos a los destinatarios, resolviendo la dependencia tecnológica entre instituciones y proveedores de software y servicios. Muyal-Xelhua, similar a Muyal-Nez, permite a las instituciones crear, en minutos, sistemas de analítica (*big data*) para convertir datos (históricos estadísticos, textos de diagnósticos, etc.) en información útil para procesos de toma de decisiones, así como realizar estudios espaciotemporales y mapas de riesgo. Tanto Muyal-Nez como Muyal.Xelhua permiten crear conexiones con sensores de dispositivos médicos para acceder a esos datos en tiempo real. Por otro lado, Muyal-Chimalli crea automáticamente redes de cripto-contendores y blockchain para verificar y garantizar que los sistemas de e-Salud, creados con Muyal-Nez y Muyal-Xelhua, observen, de forma automática y transparente, las normas nacionales (NOM-024-SSA3-2010 y NOM-004-SSA3-2012) e internacionales (ISO-270001-13, COBIT5, NIST), garantizando privacidad, confidencialidad, integridad, disponibilidad de los contenidos, tolerancia a fallas de servicios/servidores y trazabilidad. Además, Muyal-Painal permite a las instituciones compartir, intrainstitucional e interinstitucionalmente, bases de datos, resultados/información y sistemas de e-Salud mediante un modelo de publicación/suscripción, en privado y/o público, de catálogos a través de intra/internet. Finalmente, Muyal-Alwa, automáticamente, crea servicios de repositorios (estandarizados y FAIR) supervisados por Muyal-Painal y validados por Muyal-Chimalli. Los componentes de Muyal-Ilal fueron evaluados para demostrar su eficiencia en el manejo de datos, así como el cumplimiento de diferentes normas nacionales e internacionales.

Agradecimientos

Este trabajo forma parte del proyecto 41756 “Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud” por FORDECYT-PRONACES.

Referencias

[1] Dantés, O. G., Sesma, S., Becerril, V. M., Knaul, F. M., Arreola, H., & Frenk, J. (2011). Sistema de salud de México. Salud pública de México, 53(suppl 2), s220-s232.

[2] Kruse, C. S., & Beane, A. (2018). Health information technology continues to show positive effect on medical outcomes: systematic review. Journal of medical Internet research, 20(2), e8793.

[3] Leal, H. V., Campos, R. M., Domínguez, C. B., & Sheissa, R. C. (2011). Un expediente clínico electrónico universal para México: características, retos y beneficios. Universidad Veracruzana, 28(06).

[4] Heinze, G., Canchola, V. H. O., Miranda, G. B., Fuentes, N. A. B., & Sánchez, D. P. G. (2018). Los médicos especialistas en México. Gaceta médica de México, 154(3), 342-351.

[5] Chang, W. L., & Grady, N. (2019). Nist big data interoperability framework: Volume 1, definitions.

[6] Lee, R. (Ed.). (2019). Big Data, Cloud Computing, and Data Science Engineering (Vol. 844). Berlin/Heidelberg, Germany: Springer.

[7] Rose, K., Eldridge, S., & Chapin, L. (2015). The internet of things: An overview. The internet society (ISOC), 80, 1-50.

[8] Espinosa López, L. E. (2020). Evaluación del cumplimiento de la NOM 004-SSA3-2012 del expediente clínico, en el servicio de pediatría. Comparación de dos instrumentos de medición (Bachelor's thesis, Benemérita Universidad Autónoma de Puebla).

[9] O'Connor, J., & Matthews, G. (2011). Informational privacy, public health, and state laws. American journal of public health, 101(10), 1845-1850.

[10] Yan, L. (2018). DICOM standard and Its Application in PACS system. Medical Imaging Process & Technology, 1(1).

[11] Sanchez-Pinto, L. N., Luo, Y., & Churpek, M. M. (2018). Big data and data science in critical care. Chest, 154(5), 1239-1248.

[12] Van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., ... & Burke, D. S. (2014). A systematic review of barriers to data sharing in public health. BMC public health, 14(1), 1-9.

[13] Ranabahu, A., Anderson, P., & Sheth, A. (2011). The cloud agnostic e-science analysis platform. IEEE Internet Computing, 15(6), 85-89.

[14] Opara-Martins, J., Sahandi, R., & Tian, F. (2014, November). Critical review of vendor lock-in and its impact on adoption of cloud computing. In International Conference on Information Society (i-Society 2014) (pp. 92-97). IEEE.

[15] Pianykh, O. S. (2012). What is dicom?. In Digital Imaging and Communications in Medicine (DICOM) (pp. 3-5). Springer, Berlin, Heidelberg.

[16] Dolin, R. H., Alschuler, L., Beebe, C., Biron, P. V., Boyer, S. L., Essin, D., ... & Mattison, J. E. (2001). The HL7 clinical document architecture. Journal of the American Medical Informatics Association, 8(6), 552-569.

[17] Duda, S. N., Kennedy, N., Conway, D., Cheng, A. C., Nguyen, V., Zayas-Cabán, T., & Harris, P. A. (2022). HL7 FHIR-based tools and initiatives to support clinical research: a scoping review. Journal of the American Medical Informatics Association, 29(9), 1642-1653.

[18] Setyawan, R., Hidayanto, A. N., Sensuse, D. I., Suryono, R. R., & Abilowo, K. (2021, November). Data Integration and Interoperability Problems of HL7 FHIR Implementation and Potential Solutions: A Systematic Literature Review. In 2021 5th International Conference on Informatics and Computational Sciences (ICICoS) (pp. 293-298). IEEE.

Herramienta de apoyo para el diagnóstico de cáncer de hueso largo (Muyal-Ilal - Casos de estudio)

Miguel Contreras-Murillo[1], José Luis González-Compeán[2]

[1] miguelc297@gmail.com

[2] joseluis.gonzales@cinvestav.mx

Cinvestav Tamaulipas, Cd. Victoria, México

Resumen. Debido a la alta demanda en especialistas de la salud abocados al diagnóstico de tumores y a que por cada paciente se analizan hasta cientos de imágenes por estudio tomográfico, se tiene la necesidad de un sistema de detección de tumores mediante algoritmos de inteligencia artificial. El presente proyecto no tiene la intención de sustituir al personal capacitado, sino proporcionarle una ayuda indicando las imágenes en las que podría existir algún tumor primario en hueso largo o secundario en pulmones en caso de presentar metástasis, y la región en la que se ubica. Dado que los tumores presentan diferentes características dependiendo de la parte del cuerpo en la que se desarrollen, es necesario generar un modelo diferente para cada uno. Anteriormente, se utilizaron técnicas en las que se buscaban intensidades de las tomografías que pudiesen indicar, por ejemplo, algún tipo de clasificación en algunos órganos, pero en el caso particular de cáncer de hueso, estas técnicas no siempre resultan eficientes, ya que el tumor puede presentar intensidades similares al de un hueso sano. Es por ello por lo que, para este proyecto, se decidió utilizar nuevas técnicas de visión artificial que permiten distinguir diferentes patrones más allá de diferentes intensidades. Se entrenaron dos modelos de tipo *Faster R-CNN*, uno para la identificación de lesiones en hueso y otro específico para pulmones. Este tipo de modelos tienen la ventaja de permitir la búsqueda de objetos de cualquier tamaño en cualquier zona de la imagen, de una forma rápida y eficiente.

Palabras clave: Inteligencia artificial · Red Neuronal Convolucional · Cáncer.

1 Introducción

Según datos del INEGI, 90,603 personas fallecieron en 2020 a causa de algún tipo de cáncer, convirtiéndose en la cuarta causa de muerte en México [1]. En México se cuenta con 270,600 médicos; esto equivale a 2.4 médicos por cada 100,000 habitantes, mientras el promedio en países de la OCDE es de 3.5 [2].

De acuerdo con su origen, un tumor puede clasificarse como primario o secundario; los primarios ocurren en las primeras fases de la enfermedad, mientras que los secundarios indican el esparcimiento de la enfermedad en otras regiones y órganos del cuerpo, también conocida como metástasis.

Actualmente, el uso de tomografías permite a los especialistas la detección de tumores y la definición de la etapa en la que se encuentra la enfermedad. Esto, gracias a que las tomografías permiten la visualización de las estructuras internas del cuerpo en diferentes cortes. Cada estudio puede contener hasta cientos de imágenes detalladas del mismo paciente, en las cuales el especialista debe realizar una búsqueda exhaustiva, identificando cualquier estructura con características similares a las de un tumor.

Desafortunadamente, las características de cada tumor dependen del tipo, órgano, y región en la que se encuentre; por lo que las investigaciones de algunos tipos de cánceres, así como las técnicas utilizadas en su detección, podrían no aplicar a otros. Aunado a esto, el cáncer primario de hueso largo es menos frecuente que otros y requiere de especialistas capacitados para diagnosticarlo correctamente. Actualmente, la mayoría de los pacientes con sospechas de este padecimiento son derivados a hospitales especializados en diagnosticarlo, provocando una carga de trabajo excesiva para el personal.

El presente escrito muestra el funcionamiento de un sistema que ayuda a los especialistas a analizar estudios de pacientes con sospechas de la enfermedad, indicándoles regiones de interés con características similares a las de un tumor primario en hueso largo y pulmón, ya que los pulmones son, generalmente, de los primeros órganos del cuerpo en presentar tumores secundarios. De esta forma, aunque este sistema no sustituye al personal calificado para diagnosticar la enfermedad, busca disminuir el tiempo necesario para analizar los cientos de imágenes de cada paciente.

1.1 Inteligencia artificial

El aprendizaje automático o aprendizaje máquina tiene como fin que las computadoras aprendan de forma autónoma a imitar la inteligencia humana por medio de algoritmos matemáticos y estadísticos. La computadora aprende algunos patrones con base en su experiencia o datos históricos. Para lograrlo, se entrena un modelo capaz de generalizar el conocimiento de estos datos sin necesidad de darle explícitamente las reglas para solucionar cada problema [3].

El modelo contiene parámetros e hiperparámetros con los que la computadora toma decisiones autónomamente. Los hiperparámetros son diferentes para cada algoritmo y problema a solucionar, son definidos antes del entrenamiento y contienen algunas características generales del modelo, como tamaño del modelo, velocidad de aprendizaje, métricas, funciones, etc. Por otro lado, los parámetros son los valores internos de cada modelo y se ajustan durante el entrenamiento de acuerdo con los datos históricos disponibles, siendo, así, en ellos donde se almacena el conocimiento adquirido.

Dependiendo de los datos disponibles y el tipo de problema que se tiene, se emplean diferentes técnicas de aprendizaje automático: supervisado, no supervisado y por refuerzo.

Supervisado. Se utiliza principalmente en regresión y clasificación. Este tipo de aprendizaje requiere datos históricos ya etiquetados, esto es, para cada muestra se necesita conocer a qué clase pertenece o el valor esperado por medio de

etiquetas que lo indiquen. El modelo resultante debe ser capaz de generalizar el conocimiento al aprender a identificar las características inherentes de cada clase o valor, pero sin memorizar los datos de entrada para poder utilizarse con nuevas muestras [4], [5].

Se divide principalmente en dos etapas: entrenamiento, y prueba. En la primera etapa el algoritmo genera un modelo con los datos históricos; en caso de ser un algoritmo iterativo, se verifica que los resultados sean similares a los esperados y se calcula el error obtenido para corregir los parámetros del modelo, repitiendo estos pasos hasta mejorar los resultados. Posteriormente, en la etapa de pruebas, se utiliza el modelo generado para evaluar los resultados obtenidos con datos históricos diferentes a los usados durante el entrenamiento. De esta forma se puede tener una idea del comportamiento que se tendrá con datos nuevos y verificar que el modelo no haya memorizado los datos de entrenamiento. La implementación del algoritmo es similar a las pruebas, se utiliza el modelo para obtener resultados de datos nuevos no etiquetados.

No supervisado. Se llama aprendizaje no supervisado, ya que el algoritmo aprende por sí mismo y no requiere de datos etiquetados. Utiliza datos históricos y la computadora se encarga de inferir la relación entre ellos. Es útil para agrupar información catalogando o etiquetando los datos de muestra. El modelado de estos datos puede ser predictivo o únicamente descriptivo. Esto significa que la información resultante puede utilizarse para clasificaciones futuras o ser relevante por sí sola, al mostrar la distribución de los datos con patrones que no pueden ser observados a simple vista [4], [5].

Existen dos categorías de este tipo de algoritmos: agrupamiento y asociación. En el primer caso, se agrupan los datos que comparten ciertas características, generalmente por medio de métricas de similitud o distancia. En el segundo caso, se busca la relación que existe entre las variables de un conjunto de datos.

Por refuerzo. Es el menos estudiado debido a que suele ser usado para aprender la ejecución de diferentes acciones (robótica, videojuegos, etc.). No requiere de conocimiento previo para aprender, ni de un humano para indicarle las acciones que el modelo debe seguir. Aprende de las consecuencias de sus actos y está basado en objetivos, donde la experiencia se obtiene al interactuar con el ambiente. El algoritmo realiza acciones al azar por medio de un simulador y recibe incentivos o castigos por cada acción. Después de varias iteraciones, el algoritmo aprende reglas para realizar una tarea [4]–[7].

Visión artificial. Algunas técnicas de inteligencia artificial que suelen aplicarse a datos en general también pueden aplicarse al análisis de imágenes, incluyéndose las del ámbito médico. Las aplicaciones que se le dé dependen del enfoque que se tenga, dentro de las que destacan la segmentación, detección y clasificación de imágenes. En la segmentación de imágenes se analiza la información disponible para dividir la imagen en los elementos que la componen sin llegar a definirlos, dejando la interpretación al observador o a alguna etapa posterior; en imagenología médica se suelen utilizar estas técnicas para resaltar algunas secciones en la imagen y realizar mediciones o análisis manuales, mientras que la clasificación de imágenes identifica el objeto presente en la imagen de entrada.

La detección permite la localización de algunos elementos buscados en la imagen, definiendo las coordenadas en el objeto. Es posible mezclar algunas de estas técnicas en diferentes etapas de un único algoritmo.

Para la segmentación de imágenes se suelen utilizar técnicas basadas en aprendizaje no supervisado y se suele buscar la relación que existe entre las intensidades (valores de brillo en imágenes en escala de grises) o color de cada píxel, u otras características en la imagen, como texturas. Algunos algoritmos relevantes son: *k-means* [8] el método Otsu y el crecimiento de región por semilla. En el algoritmo *k-means*, las intensidades de cada pixel se agrupan de acuerdo con su similitud con el valor promedio de grupos de intensidades. Los grupos se actualizan iterativamente, sin considerar la ubicación espacial de cada pixel. Aplicado a tomografías, permite al observador notar las diferentes intensidades presentes en la imagen, ya que suelen coincidir con las diferentes estructuras del cuerpo. En el método de Otsu, en un primer paso se calcula el histograma de las intensidades de los pixeles en la imagen, se definen diferentes umbrales calculando la varianza entre grupos y se selecciona el umbral que mejor divide los dos grupos. Aunque genera imágenes en blanco y negro, existen variantes del algoritmo que proporcionan resultados similares a *k-means*. Otro método que permite un control manual en la segmentación de imágenes es el crecimiento de región por semilla. En este método se seleccionan uno o varios pixeles y se busca si los pixeles vecinos comparten características similares. La búsqueda local se repite recursivamente hasta que los bordes de la región no cuentan con vecinos similares y se forma una figura cerrada. Es útil para etiquetar regiones manualmente.

Para la clasificación de imágenes se suelen utilizar técnicas de aprendizaje supervisado, por lo que se requieren datos históricos etiquetados manualmente para la etapa de entrenamiento. Para detectar algunas estructuras del cuerpo humano en tomografías, los humanos suelen buscar algunas características previamente definidas, como el tamaño, textura, bordes, etc. Debido a la complejidad que esto requiere para poder ser definido manualmente de forma matemática, esto no es factiblemente de manera programable. En años anteriores se implementaban técnicas como el uso de plantillas, esto es, se comparaba la imagen de entrada, usualmente binarizada con pixeles en 1 y 0, con imágenes de referencia del mismo tamaño. Entre mayor el número de pixeles iguales, mayor similitud; desafortunadamente, esta técnica es únicamente útil identificando objetos con poca variación en figura, tamaño y orientación, como letras en textos impresos. Debido a esto, en lugar de intentar identificar la imagen completa, se busca identificar algunas formas que conforman cada objeto por medio de filtros, donde cada filtro es una matriz con figuras binarias preestablecidas, se multiplican por la imagen de entrada pixel por pixel y se suman los resultados. Estos resultados forman un vector de características con el que el objeto puede ser identificado por medio de algún algoritmo de clasificación, como máquinas de vectores de soporte, redes neuronales, árboles de decisión, etc. Un inconveniente que puede tener esta técnica es que los filtros, al ser predefinidos, suelen estar limitados a pocas formas y no siempre son los óptimos para identificar todos los objetos; las redes neuronales convolucionales (CNN) aprovechan la propagación hacia atrás del entrenamiento de la red neuronal que clasifica el objeto para generar automáticamente un gran número de filtros específicos para cada aplicación [3], [4], [9], [10].

Una CNN aprende a identificar objetos en imágenes, siendo capaz de identificar patrones similares a los usados durante el entrenamiento sin requerir que sean exactamente iguales en tamaño o rotación. Esto lo logra aprendiendo a identificar las partes que lo componen, en lugar del patrón completo.

Las capas que la componen son las siguientes: capa de entrada, capa convolucional, normalización por lotes, capa de agrupamiento (*pooling*), capa de rectificación (ReLU), capa totalmente conectada, y salida. Es posible tener múltiples capas intercaladas del tipo convolucional, normalización por lotes, de agrupamiento y rectificación, pero siempre en el mismo orden

La capa de entrada tiene un tamaño fijo, por lo que analiza la imagen a través de ventanas deslizantes. Cada una de estas ventanas pueden ser redimensionada al tamaño de esta capa durante el recorrido. La capa convolucional analiza si existen partes de la capa anterior que contienen características similares a los filtros, donde los puntos donde el filtro encuentra similitudes suelen tener valores altos. La Normalización por lotes se requiere, ya que, aunque las entradas de una red se suelen normalizar para que cada característica tenga el mismo peso en la misma escala, los valores crecen muy rápidamente después de la convolución y es necesario normalizar nuevamente los valores entre capas. En la capa de agrupamiento (*pooling*) se disminuye la resolución sin perder información importante, ya que características que sí se encontraron en la convolución tienen valores altos; los valores bajos no son relevantes. Se suele usar *Max Pooling* (mantiene los valores mayores) o *Average Pooling* (calcula los valores promedio) para disminuir la resolución de la imagen, manteniendo únicamente la información relevante. La capa de rectificación (ReLU) es opcional y permite un entrenamiento más rápido, ya que, en caso de que la imagen de entrada esté segmentada, se podría utilizar -1 para denotar los pixeles que pertenecen al fondo de la imagen. Para evitarlo, cualquier valor negativo debe ser convertido en 0. Después de pasar por las capas anteriores, la imagen de entrada reduce sus dimensiones en altura y anchura, pero se incrementan en profundidad y resulta en un vector que puede utilizarse como entrada de una RNA en la capa totalmente conectada. Esta capa tiene el mismo comportamiento que una RNA de tipo *Back Propagation* y realiza la clasificación final de la imagen de entrada. Finalmente, después de realizarse la clasificación se pueden tener una o más salidas con la información del objeto encontrado en la imagen (en caso de que el algoritmo lo haya identificado).

Debido a que, generalmente, el objeto buscado se encuentra en alguna parte de la imagen de entrada, identificarlo requiere de buscarlo y detectarlo localizando las coordenadas en las que se encuentra, lo cual es posible por medio de ventanas deslizantes de tamaño predefinido que recorren la imagen de izquierda a derecha y de arriba abajo. Al ser una búsqueda exhaustiva, no suele ser la forma más eficiente de detección, además de que los objetos están limitados a tamaños específicos. Para contrarrestar estas limitantes, se puede realizar una segmentación previa de la imagen y cada segmento es redimensionado para coincidir con la entrada del clasificador. Se conoce R-CNN cuando se utiliza una CNN para analizar cada una de estas regiones. Para agilizar incluso más este proceso, es posible filtrar cada región y analizar con la CNN únicamente los segmentos con mayor probabilidad de contener el objeto buscado, si la selección

de estas regiones propuestas se realiza por medio de técnicas clásicas de segmentación, se conoce como *fast R-CNN*, y *faster R-CNN* si esta segmentación forma parte de la red completa.

Otra forma de realizar esta detección es por medio de una red neuronal artificial de tipo *hourglass*, llamada así por su forma de reloj de arena. Para su funcionamiento, se disminuye la resolución de la imagen de entrada por medio de filtrados. Posteriormente, se incrementa la resolución, con lo que se sabe si existe algún objeto en la imagen, y se combina con las salidas de las capas análogas anteriores para mantener coherencia con la imagen de entrada. De esta forma se pueden realizar detecciones de objetos y obtener mapas de calor, regiones de interés, segmentos, etc.

2 Materiales y métodos

Para este sistema de ayuda en el diagnóstico de cáncer de hueso largo se requiere entrenar dos modelos, uno para detectar masas en hueso largo y otro para detectar masas en pulmones. Ambos modelos cuentan con la misma estructura y funcionamiento, por lo que únicamente es necesario entrenarlos con los conjuntos de imágenes correspondientes para caso; de hecho, este sistema de detección de tumores en tomografías se basa en sistemas generales de detección de objetos, por lo que con pocas adecuaciones se podría reentrenar y utilizar en otros tipos de imágenes.

Para el entrenamiento del primer modelo, el hospital del caso de uso cuenta con aproximadamente 80 millones de imágenes, aunque no todas corresponden a pacientes con tumores en hueso largo. Adicionalmente, la disponibilidad está en función de lo aprobado por el comité de ética, por lo que solo es posible utilizar estudios con consentimiento del paciente y el comité, y que los pacientes tengan estructuras óseas maduras, por lo que se tiene considerado el uso de 196 estudios de pacientes positivos a tumoraciones en huesos largos, además de un número similar estudios de pacientes negativos a cáncer de hueso como control. De los pacientes con diagnósticos positivos, se estima que entre un 50% y 70%, presenten más de un objeto con intensidades u otras características similares a las de un tumor. Debido a que los modelos de visión artificial usados no dependen únicamente de la intensidad de cada objeto en las imágenes de entrada, se espera que no exista interferencia alguna cuando diferentes objetos con intensidades atípicas similares en rango a las de un tumor, pero bien identificadas (como prótesis), aparezcan en las imágenes; incluso, ayudan al algoritmo a aprender que no corresponden a la clase buscada. Para este estudio, solo las masas en huesos que correspondan a un tumor y estén presentes en cada estudio deben ser etiquetadas en todos los cortes, incluso si el paciente presente tumoraciones múltiples.

Se tiene considerado que imágenes con fracturas de hueso cuentan con características similares a tumores, por lo que podrían resultar en falsos positivos. No se cuenta con información *a priori* sobre si existen estudios positivos que además cuenten con fracturas visibles en el mismo corte, pero, en caso de contar con imágenes de pacientes con fracturas sin tumores, se recomienda usarlas para la evaluación general del algoritmo.

Para el entrenamiento y evaluación del modelo de detección en pulmones se usarán imágenes de bases de datos públicas de 350 pacientes [11], [12].

A diferencia de otros sistemas de detección de objetos, donde cada imagen pertenece a una escena diferente, para este sistema se requiere que las imágenes se mantengan separadas por expedientes y por pacientes, ya que los cortes en las imágenes de tomografías contienen características muy similares entre sí y, de mezclarse indiscriminadamente durante el proceso de entrenamiento-validación-prueba, los resultados podrían ser erróneamente mejores a la realidad.

2.1 Preparación de datos

Conversión de imágenes de DICOM a PNG. Los archivos de almacenamiento de imágenes médicas suelen estar en formato DICOM. Estos archivos contienen, además, otro tipo de información del paciente y de la propia imagen, por lo que es necesario convertirlas a formato PNG, ya que éste es legible por el sistema y contiene únicamente la imagen.

La intensidad de la imagen en formato DICOM utiliza unidades Hounsfield en un único canal, que puede ir de valores negativos correspondientes a densidades muy bajas, a positivos, correspondientes a cuerpos con densidades más altas, como se observa en la Tabla 1. La escala Hounsfield es una escala cuantitativa utilizada en los estudios de tomografía axial computarizada para describir los diferentes niveles de radio-densidad de los tejidos humanos (no aplica en radiografías)[13], [14].

Tabla 1. Valores en escala de Hounsfield de diferentes tejidos y materiales.

		Valor
Aire		-1000
Grasa		-120 a -90
Tejido suave con contraste		100 a 300
Hueso	Esponjoso (interior)	300 a 400
	Cortical o compacto (exterior)	500 a 1,900
Cuerpos extraños	Vidrio	500
	Aluminio	2,100 a 2,300
	Piedra caliza	2,800
	Cobre	14,000
	Plata	17,000
	Acero	20,000
	Oro	30,000

Las imágenes en RGB tienen valores del 0 al 255 (24 bits), mientras que el formato DICOM admite 4,096 valores negativos y positivos, van de -1,024 HU hasta 3,071 HU, con 12 bits. A pesar de que, al usarse escala de grises en RGB, la información pasa de 12 a 8 bits, los valores negativos no aportan tanta información relevante para la detección de tumores en huesos debido a que las tomografías contienen valores muy negativos solo para marcar espacios vacíos, siendo solo útiles los valores negativos cercanos a 0 para detección de lesiones en pulmones.

Debido a que esta escala es incompatible con el formato PNG, que usa valores entre 0 y 255 en tres canales RGB para representar imágenes en color, y que, al igual que otros algoritmos de inteligencia artificial, todas las entradas de datos necesitan estar dentro un rango preestablecido, para evitar que cualquier cambio en la intensidad global de las imágenes pueda interferir en los resultados del sistema y estandarizar las entradas, las imágenes DICOM son truncadas y escaladas de un rango de -100 a 3,200 unidades Hounsfield de 0 a 255 intensidades, y se repiten los datos en los tres canales RGB.

Etiquetado manual. Es necesario etiquetar manualmente cada región de interés en cada muestra para entrenar los modelos. Para esta etapa se cuenta con el apoyo de personal de la salud debidamente capacitado y se utilizan herramientas gratuitas, como LabelImg [15] y CVAT [16].

Cada objeto de interés en la imagen debe ser señalado por medio de un rectángulo o cuadrado de tamaño suficiente para albergar en su totalidad al tumor. A diferencia de algoritmos anteriores, donde se necesitaban etiquetas de un tamaño y figura específica, los algoritmos usados en este sistema pueden recibir rectángulos de diferentes tamaños como entrada, ya que internamente realizan un escalado automático de ser necesario.

Se almacena el resultado en formato PascalVoc, ya que este etiquetado manual será almacenado en un archivo XML, y es el mismo formato con el que se guardan los resultados de la ejecución del sistema ya entrenado; ambos archivos XML se requieren para la evaluación del modelo. Es necesario mantener las imágenes separadas por paciente.

En pruebas preliminares se observó que el modelo puede fallar cuando se usan imágenes de prueba de regiones diferentes a las usadas durante el entrenamiento, detectando estructuras normales del cuerpo como tumores. Este problema se espera solucionar con el uso de más muestras durante el entrenamiento y utilizando muestras negativas durante el mismo. Debido a que las imágenes y etiquetas se compilan en un único archivo para el entrenamiento, se requiere generar un XML sin el campo de área de interés para estas etiquetas negativas. Esta actividad puede realizarse manualmente con el botón *Verify Image* dentro LabelImg, aunque se podría proporcionar un script que genere estos archivos automáticamente antes del entrenamiento para evitar mezclar las etiquetas positivas con las de muestras negativas.

División de datos por paciente en entrenamiento, validación y pruebas. Las redes neuronales utilizadas en ésta y otras investigaciones se basan en aprendizaje supervisado, esto es, requieren de las muestras de entrada etiquetadas manualmente por un especialista para poder aprender. Como entrada se

tienen las imágenes de entrada en formato PNG y las etiquetas XML; en un paso intermedio en el script de entrenamiento proporcionado, estos archivos son compilados.

Se requieren tres conjuntos de entrenamiento con imágenes PNG y etiquetas XML: i) de entrenamiento, requiere estas imágenes etiquetadas para ajustar los parámetros; ii) de prueba, se usan muestras diferentes para la evaluación final; iii) de validación, es útil para validar el modelo durante el entrenamiento cada determinado número de ciclos. El algoritmo debe detenerse forzadamente cuando los resultados de la validación tiendan a empeorar significativamente. Es recomendable que el conjunto de prueba y de validación también sean diferentes. Es necesario dividir y evaluar los datos por paciente, seleccionando algunos pacientes para entrenamiento, otros para validación, y otros más para pruebas.

Durante el entrenamiento, el algoritmo ajusta sus parámetros iterativamente hasta que el algoritmo es capaz de identificar los objetos como se espera. Idealmente, el algoritmo logra generalizar el conocimiento y los resultados de la evaluación son buenos. Desafortunadamente, no siempre es posible visualizarlo únicamente con los datos usados durante el entrenamiento, además de que pueden presentarse otros dos casos adversos, el bajo-ajuste y el sobreajuste, como se observa en la Figura 1. En el primero, es notorio que los resultados del entrenamiento son malos durante el entrenamiento; esto puede deberse a muchos factores, como que el sistema llegó al límite de las iteraciones predefinidas, las muestras son insuficientes, incorrectamente etiquetadas, o algunos hiperparámentros de la red no son adecuados para el problema presentado.

Por otro lado, un sobreajuste se presenta cuando el modelo se entrena más de lo necesario y pasa de generalizar conocimiento a memorizar las muestras de entrada. De utilizarse únicamente un conjunto de datos, no sería posible distinguir entre un entrenamiento óptimo y un sobreajuste, de aquí que se requiera apartar algunos datos antes de iniciar el entrenamiento para poder probar el desempeño real del algoritmo después del entrenamiento con datos nuevos para el modelo. Se observa un sobreajuste cuando los resultados durante el entrenamiento son muy superiores a los resultados de las pruebas. Para evitar sobrepasar el punto óptimo de entrenamiento, se recomienda el uso de un tercer conjunto de datos para validar el modelo durante el entrenamiento cada determinado número de ciclos. Al comparar los resultados de esta validación contra los del entrenamiento se espera que los resultados de ambos sean inicialmente malos para, posteriormente, mejorar en similar tendencia. El algoritmo debe detenerse forzadamente cuando los resultados de la validación tienden a empeorar significativamente a pesar de que los de entrenamiento continúen en mejoría. Posteriormente se puede evaluar el desempeño final con el conjunto de pruebas. Si bien los datos de validación no influyen directamente en el entrenamiento del modelo, es altamente recomendable que el conjunto de prueba y de validación también sean diferentes.

Bajo otras circunstancias se recomendaría mezclar todas las imágenes disponibles para posteriormente dividirlas en estos tres conjuntos de datos, pero, dada la naturaleza de las tomografías, donde diferentes cortes del mismo expediente contienen imágenes con características prácticamente iguales, es necesario dividir y evaluar los datos por paciente; esto es, seleccionar algunos pacientes

para entrenamiento, otros para validación y otros más para pruebas. Posteriormente, todas las imágenes (archivos PNG) y etiquetas (archivos XML) de entrenamiento deberán copiarse en una única carpeta para poder compilarse e iniciar el entrenamiento; y con el mismo proceso deberán separarse las imágenes de validación en su específica ruta. Los datos de prueba etiquetados no requieren este paso. En caso de ser necesario, se podrá proporcionar un script para este paso.

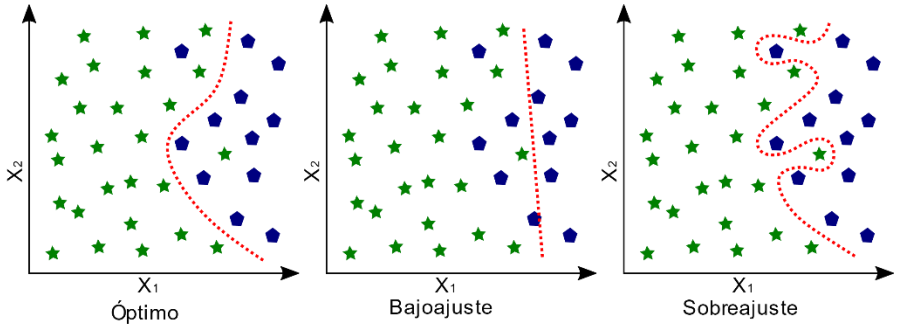


Figura 1. El ajuste óptimo acepta la existencia de muestras incorrectamente clasificadas, ya que pueden ser valores atípicos. Un bajo-ajuste tiene resultados deficientes. El sobreajuste memoriza las muestras.

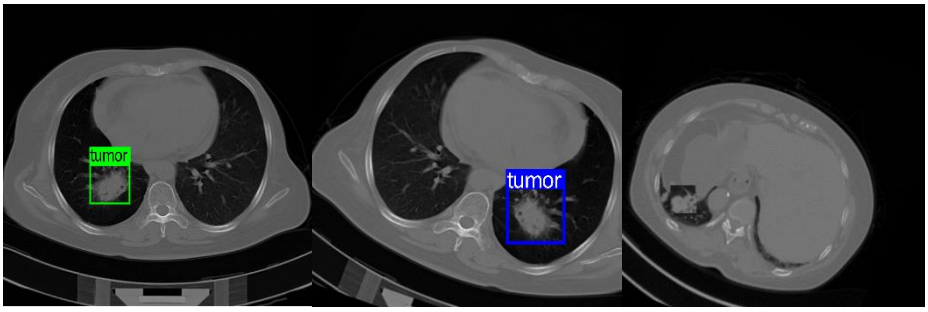
Aumentación de datos. Para evitar un bajo-ajuste, se recomienda utilizar la máxima cantidad de datos posibles para el entrenamiento. Desafortunadamente, no siempre es posible contar con las suficientes muestras. Como alternativa, pueden ser utilizadas algunas técnicas de aumentación de datos para mejorar esta disponibilidad, como se observa en la Figura 2.

Estas técnicas consisten en modificar las imágenes disponibles con transformaciones geométricas como giros, rotaciones, o deformaciones y, aunque no es muy recomendable para este sistema, ya que podrían agregarse características muy diferentes a las de imágenes reales, en las imágenes también pueden modificarse las intensidades, aplicarse filtros o agregar obstrucciones o recortes en la imagen. Internamente, los algoritmos proporcionados utilizan algunas de estas técnicas durante el entrenamiento de forma automática, pero, en caso de que se requiera un mejor control de estas características, es posible generar estas variaciones en las imágenes de entrenamiento.

Otra técnica explorada en esta investigación consiste en copiar la región etiquetada y agregarla en otra imagen del mismo paciente sin etiquetar. En ambas imágenes se deben aplicar las exactamente las mismas transformaciones geométricas. De esta forma se logra que el algoritmo aprenda con imágenes aumentadas de regiones donde generalmente no aparecen tumores del tipo buscado, pero que generan falsos positivos.

Después de generadas las nuevas imágenes aumentadas, es necesario etiquetarlas de nuevo. Aunque un etiquetado manual permite mantener un mejor control de las coordenadas donde se encuentra el objeto en la nueva imagen, debido a la cantidad de nuevas imágenes, es preferible realizar un etiquetado

automático utilizando como base cada etiqueta original y aplicarle exactamente las mismas transformaciones geométricas que se aplicaron en la imagen de entrada para calcular las nuevas coordenadas. Posteriormente, estas imágenes aumentadas se agregan al conjunto de entrenamiento.



a b c

Figura 2. Se aplican diferentes transformaciones a la imagen original (a) para obtener una imagen similar, pero con apariencia diferente (b). Además, es posible recortar la región tumoral y agregar en una imagen diferente del mismo paciente a la que se le aplican las mismas transformaciones aleatorias para obtener una nueva muestra completamente diferente (c) [11], [13].

2.2 Entrenamiento

De forma similar a la que se enseña a los humanos, un especialista le enseña a la computadora con ejemplos. Al ser aprendizaje supervisado, se comparan los resultados obtenidos contra los del especialista y se aprende de los aciertos y errores, ajustando múltiples parámetros. La enseñanza finaliza cuando la computadora obtiene una buena evaluación.

El sistema presentado ha sido entrenado para identificar masas en huesos y pulmones, pero nuevos modelos pueden ser entrenados para identificar otros tipos de tumores. Por defecto, se proporcionan los modelos con los hiperparámetros que mejores resultados han proporcionado.

Con los avances que se han tenido en visión artificial, se han diseñado diferentes variantes de redes neuronales para la detección de objetos. Para evitar rediseñar y reprogramar algún modelo propio, se ha optado por utilizar algunos de los modelos más utilizados para fines generales de detección de objetos que requieren ser probados para saber cuál se ajusta de mejor forma a los fines de esta investigación. Para esta transferencia de aprendizaje se utilizan modelos preentrenados disponibles en el Model Zoo de TensorFlow, los cuales, posteriormente, son reconfigurados y reentrenados para este sistema. El entrenamiento y la validación pueden ser observados con TensorBoard.

Se recomienda el uso de una tarjeta de gráficos para acelerar el proceso, ya que, en comparación, un procesador puede tardar más de 10 veces en realizar el entrenamiento (puede tomar varias horas o días). La ejecución del algoritmo

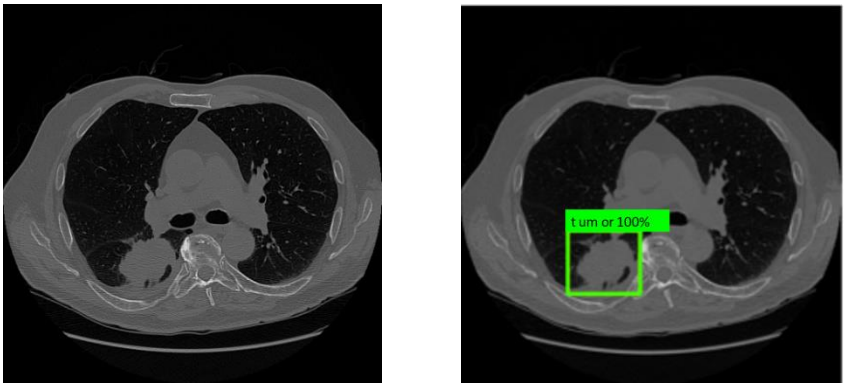
entrenado también puede mejorarse, aunque solo toma algunos segundos o minutos por estudio.

Al usarse varias tarjetas de gráficos también es posible mejorar los resultados, ya que, al contar con más VRAM disponible, es posible aumentar el tamaño del lote durante la normalización por lotes (parte del entrenamiento). Es posible que en el proceso de entrenamiento con GPU la bloquee para su uso exclusivo y ocurran errores al intentar el proceso simultáneo de validación; por tanto, se provee un script que fuerza la ejecución de la validación en CPU.

2.3 Implementación

Después de entrenado el modelo, con la implementación de este sistema, el algoritmo de visión artificial le indica al especialista algunas regiones de interés automáticamente etiquetadas en una imagen en formato PNG (figura 3), y en formato XML si se desea superponer sobre la imagen original con algún visualizador compatible con PascalVoc.

El tiempo de ejecución del análisis puede tomar algunos segundos o minutos por estudio, dependiendo de la cantidad de imágenes por estudio y del hardware que se tenga. Debido a que esta investigación es parte del proyecto Muyal-Ilal, cada modelo es empaquetado en un contenedor y el análisis de varios estudios puede realizarse de forma paralela, además de ser compatible con diferentes plataformas, aunque se recomienda ampliamente el uso de procesadores compatibles con instrucciones AVX y tarjetas de gráficos compatibles. En pruebas preliminares, se han analizado ~1,000 imágenes del estudio de un paciente en ~30 minutos con un procesador no compatible con instrucciones AVX, sin GPU y usando únicamente un núcleo, y en ~6 minutos usando 24 núcleos.



a b

Figura 3. La imagen de entrada (a) es analizada automáticamente por el sistema, y se tiene como resultado la misma imagen etiquetada (b) [11], [13].

3 Experimentos y resultados

El sistema de detección de tumores se basa en diferentes algoritmos de visión artificial, y su desempeño puede ser evaluado de forma similar a cualquier otro sistema de inteligencia artificial. A continuación, se presenta una serie de pasos y consideraciones relevantes consideradas para el entrenamiento y evaluación.

La evaluación puede ser vista en tres secciones. La primera, consiste en preparar los datos que se usarán; aquí es necesario convertir las imágenes, etiquetarlas manualmente y dividir los datos en conjuntos de entrenamiento, evaluación y prueba. En la segunda, los datos de entrenamiento y evaluación sirven para entrenar los diferentes modelos que se pueden usar en producción, para posteriormente probarlos. Finalmente, los resultados de cada experimento son analizados.

Después de entrenado el modelo, las imágenes de prueba ingresan al clasificador. El resultado son nuevas imágenes marcadas y un archivo XML. Estos archivos de salida se distribuyen de la misma forma que los archivos de entrada. Debe existir una carpeta para cada paciente y, dentro de ésta, una para cada estudio. Los conjuntos de datos de prueba, entrenamiento y validación deben elegirse manteniendo todos los expedientes de pacientes elegidos al azar. Los resultados pueden evaluarse por paciente y no únicamente en el total de las imágenes.

Las métricas más importantes para este estudio son las siguientes: IoU, precisión, sensibilidad y Valor F1.

IoU (*Intersection over Union*): Se refiere a cuánto una etiqueta generada automática se sobrepone sobre la etiqueta de referencia generada manualmente. Generalmente, cuando estas dos etiquetas comparten al menos un 50% del área generada por ambas, se puede calificar como un acierto. Es necesario calcularla para definir si el modelo acertó o no y generar una matriz de confusión.

Precisión: Se refiere a la cantidad de muestras etiquetadas automáticamente como positivas que efectivamente eran muestras positivas. Esto es, qué tan correctamente el modelo logra detectar tumores entre las muestras totales etiquetadas de esa forma.

Sensibilidad o exhaustividad: Busca la cantidad de muestras negativas etiquetadas como negativas. Esto es, verificar si el algoritmo no detecta otros objetos como tumores.

F1 Score o Valor F1: Es una métrica que toma en consideración ambas métricas (precisión y sensibilidad), por lo que cualquier error de clasificación puede ser visto con solo una calificación.

3.1 Resultados preliminares

Dado que no se cuenta con una base de imágenes públicas con las características adecuadas para el entrenamiento de este sistema, únicamente se han realizado pruebas preliminares para identificar cuáles son los modelos que mejor desempeño pueden dar y probar el correcto funcionamiento de las diferentes etapas del sistema.

Actualmente, y habiendo obtenido los permisos necesarios para trabajar con imágenes reales del hospital, se ha formado un grupo de expertos de la salud que se encuentra seleccionando y etiquetando manualmente las imágenes disponibles de pacientes diagnosticados con cáncer de hueso. Debido a esto, se han realizado diferentes experimentos preliminares con imágenes públicas de pulmones [11], [13].

A continuación se describen algunos tipos de redes neuronales disponibles para entrenamiento en el sistema, y se muestran los resultados obtenidos de algunas pruebas preliminares únicamente como referencia previa a la evaluación final.

- *EfficientDet*: A pesar de que suele dar los mejores resultados en la literatura, para esta aplicación los resultados preliminares son malos, además de que su alto consumo de recursos requiere de hardware muy potente.
- *SSD*: Su principal característica es su bajo uso de recursos y rápida ejecución, pero los resultados han sido malos.
- *Faster R-CNN:* Este modelo funciona correctamente para este fin; internamente se realiza una segmentación de la imagen de entrada y los segmentos con más probabilidad de contener el objeto buscado entran en la etapa de clasificación.

En 48 de 50 pacientes se encontró el tumor en al menos una imagen. Se identificaron falsos positivos en algunas estructuras normales del cuerpo. Se toma como positivo cualquier etiquetado automático con un *score* resultante superior a .18 (valor por defecto en un rango de 0 a 1), y como positivo verdadero si, además, comparte más de un 50% de IoU.

Aunque no en todas las imágenes fue posible identificar el tumor, como se observa en la Tabla 2, cabe recordar que las tomografías se componen de una serie de imágenes del mismo paciente en un área determinada, por lo que es probable que sea visible en el corte posterior.

Tabla 2. Resultados de prueba de una *Faster R-CNN* para detección de tumores en pulmones.

Métrica	Valor
Precisión	0.83
Sensibilidad	0.73
Especificidad	0.73
F1-score	0.76

- *HourGlass*: en experimentos preliminares ha tenido un buen desempeño, como se observa en las Figuras 4 y 5, donde, a diferencia de la evaluación de la red anterior, para este experimento se definió un umbral inicial de .02 hasta 1, permitiendo un mayor número de falsos positivos; esto, con el fin de, posteriormente, encontrar el mejor umbral a partir del cual se dan los mejores resultados para poder calcular las métricas esperadas como en la Tabla 2. Mientras que en la mayoría de los clasificadores este umbral se

calcula maximizando *F1-score*, dada la naturaleza del problema se prefiere una mayor precisión encontrando la mayoría de los tumores, a pesar de que esto signifique un mayor número de falsos positivos.

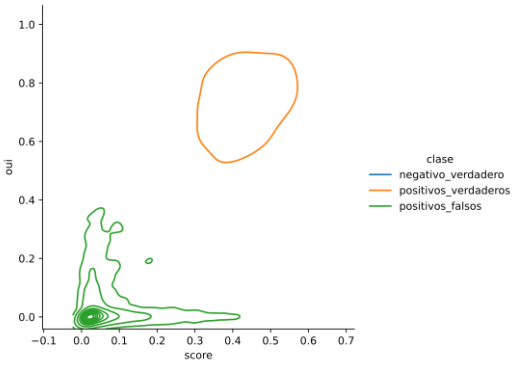


Figura 4. Los valores positivos verdaderos comparten al menos un 50% del área con el etiquetado manual y tienen un score superior que los falsos positivos que tienden a 0.

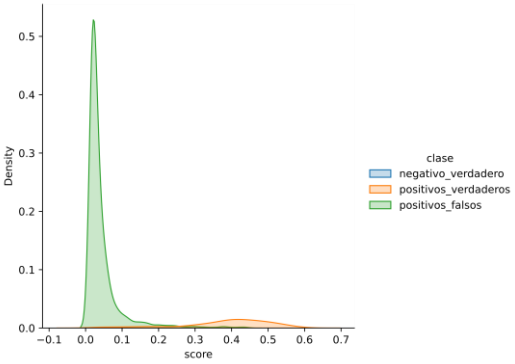


Figura 5. Es necesario definir el umbral de aceptación a partir del cual se reportarán los resultados como posiblemente positivos verdaderos; los valores con un *score* más bajo tienden a ser falsos positivos.

4 Conclusiones

El módulo presentado es parte del proyecto Muyal-Ilal y permite el flujo de imágenes desde un origen hasta el usuario final. Esta herramienta permite el entrenamiento e implementación de modelos compatibles.

Se han realizado algunas pruebas preliminares con las imágenes disponibles a la fecha para determinar cuál será el comportamiento del sistema. Actualmente, se realiza el etiquetado manual por especialistas, por lo que se realizará un nuevo entrenamiento con más imágenes en cuánto estén disponibles para mejorar los resultados.

Los resultados preliminares muestran que la red de tipo *Hourglass* es las más adecuada para esta aplicación, ya que los falsos positivos tienen una calificación

baja en comparación con los verdaderos positivos, aunque no se descarta el uso del tipo *Faster R-CNN*, ya que es más rápida y los falsos positivos ocurren en estructuras del cuerpo fácilmente reconocibles como normales.

5 Agradecimientos

Este trabajo forma parte del proyecto 41756 “Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud” por FORDECYT-PRONACES.

Referencias

[1] “Características De Las Defunciones Registradas En México Durante 2020,” p. 92, 2021.

[2] OECD and The World Bank, *Panorama de la Salud: Latinoamérica y el Caribe 2020*. OECD, 2020. doi: 10.1787/740f9640-es.

[3] I. Vasilev, *Python deep learning: exploring deep learning techniques and neural network architectures with PyTorch, Keras, and TensorFlow*. 2019. Accessed: Feb. 13, 2021. [Online]. Available: http://proquest.safaribooksonline.com/?fpi=9781789348460

[4] G. Zaccone, R. Karim, and an O. M. C. Safari, *Deep Learning with TensorFlow - Second Edition*. 2018. Accessed: Feb. 13, 2021. [Online]. Available: https://www.safaribooksonline.com/complete/auth0oauth2/&state=/library/view//9781788831109/?ar

[5] R. Gopalakrishnan and A. Venkateswarlu, *Machine Learning for Mobile*. Place of publication not identified: Packt Publishing, 2018. Accessed: Feb. 13, 2021. [Online]. Available: https://www.safaribooksonline.com/library/view/title/9781788629355/?ar?orpq&email=^u

[6] S. Ravichandiran, *Hands-On Reinforcement Learning with Python: Master Reinforcement and Deep Reinforcement Learning Using OpenAI Gym and TensorFlow.* Birmingham: Packt Publishing Ltd, 2018. Accessed: Feb. 13, 2021. [Online]. Available: https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5439844

[7] M. Lapan, *Deep reinforcement learning hands-on: apply modern RL methods, with deep Q-networks, value iteration, policy gradients, TRPO,AlphaGo Zero and more*. Birmingham Mumbai: Packt Publishing, 2018.

[8] D. Rothman and Packt Publishing, *Artificial intelligence by example: develop machine intelligence from scratch using real artificial intelligence use cases*. Birmingham: Packt Publishing Ltd., 2018.

[9] C. C. Aggarwal, *Neural networks and deep learning: a textbook*. Cham: Springer, 2018. doi: 10.1007/978-3-319-94463-0.

[10] R. Shanmugamani and S. Moore, *Deep learning for computer vision: expert techniques to train advanced neural networks using TensorFlow and Keras*. Birmingham Mumbai: Packt, 2018.

[11] P. Li, S. Wang, T. Li, J. Lu, Y. HuangFu, and D. Wang, "A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis." The Cancer Imaging Archive, 2020. doi: 10.7937/TCIA.2020.NNC2-0461.

[12] K. Clark *et al.*, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," *J. Digit. Imaging*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013, doi: 10.1007/s10278-013-9622-7.

[13] M. H. Lev and R. G. Gonzalez, "17 - CT Angiography and CT Perfusion Imaging," in *Brain Mapping: The Methods (Second Edition)*, A. W. Toga and J. C. Mazziotta, Eds. San Diego: Academic Press, 2002, pp. 427–484. doi: 10.1016/B978-012693019-1/50019-8.

[14] J. Broder and R. Preston, "Chapter 1 - Imaging the Head and Brain," in *Diagnostic Imaging for the Emergency Physician*, J. Broder, Ed. Saint Louis: W.B. Saunders, 2011, pp. 1–45. doi: 10.1016/B978-1-4160-6113-7.10001-8.

[15] "heartexlabs/labelImg." Heartex, Oct. 31, 2022. Accessed: Oct. 31, 2022. [Online]. Available: https://github.com/heartexlabs/labelImg

[16] CVAT.ai Corporation, "Computer Vision Annotation Tool (CVAT)." Sep. 2022. Accessed: Oct. 31, 2022. [Online]. Available: https://github.com/opencv/cvat

Reflexiones sobre el almacenamiento digital de las organizaciones

Ricardo Marcelín-Jiménez[1][0000−0002−5355−5830], José Luis González-Compeán[2][0000−0002−2160−4407], Hugo G. Reyes-Anastacio[2][0000−0002−9003−6765], and Dante D. Sánchez-Gallegos[2][0000−0003−0944−9341]

[1] Universidad Autónoma Metropolitana (UAM), Iztapalapa, Departamento de Ingeniería Eléctrica, Ciudad de México 09340, México.
rmarcelin@izt.uam.mx

[2] Centro de Investigación y Estudios Avanzados (Cinvestav) del IPN Unidad Tamaulipas, Cd. Victoria 87130, Tamaulipas, México.
{joseluis.gonzalez, hugo.reyes, dante.sanchez}@cinvestav.mx

Resumen A lo largo de este capítulo se presenta un análisis de las diferentes tecnologías que pueden soportar las necesidades de almacenamiento modernas, considerando que no existe una solución que pueda atender todos los requerimientos asociados con el tratamiento de la información, sino que debemos pensar en la construcción de ambientes en los que se interconectan componentes de almacenamiento que atiendan necesidades complementarias, así como los sistemas para recuperar la información con agilidad a partir de sus metadatos o descriptores.

Palabras Clave: Big Data · Ambientes de almacenamiento · Catálogos digitales

1. Introducción

Dejar un registro de nuestras actividades es parte de nuestra naturaleza como seres humanos. Es establecer un diálogo con una versión futura de nosotros mismos o con alguien que recibirá nuestro mensaje, tal vez mucho tiempo después de que nos hayamos ido. Es guardar una nota al interior de una botella para luego arrojarla en el vasto océano del tiempo.

La *Historia*, con mayúscula, comenzó cuando los registros se basaron en la escritura. Entonces, surgió el problema de la gestión de colecciones documentales y la humanidad construyó sus primeras bibliotecas. Se sabe también que en todo tiempo el registro de la información ha estado asociado con la tecnología disponible en la época. En el siglo XXI, por ejemplo, este registro es fundamentalmente de naturaleza digital, esto es, electrónica de signos binarios. En la Fig. 1 se muestran diferentes ejemplos de mecanismos de almacenamiento actuales.

Por su parte, la gestión de las colecciones implica la catalogación y preservación de los documentos. Estas dos funciones van de la mano pero, en nuestro

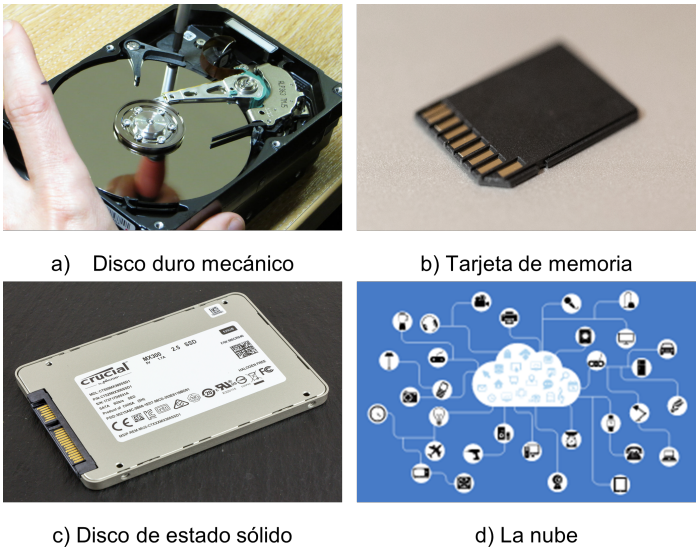


Figura 1: Ejemplos de mecanismos de almacenamiento de datos actuales.

tiempo, el volumen implica nuevos retos. En el 2020, por ejemplo, se generaban, cada minuto del año, 500 horas de video en YouTube, 347,222 publicaciones en Instagram, 28 canciones en Spotify [3]. Se sabe también que el 75 % de la información es producida por individuos, pero el 80 % de la misma está bajo la responsabilidad de organizaciones. Se prevé un crecimiento exponencial en el tamaño del "universo digital", pasando de 33 Zettabytes (ZB) en 2018 a 175 ZB en 2025[3] [8].

Este capítulo es un trabajo basado en la experiencia de los autores. No es un reporte de investigación, sino un artículo de opinión que invita a la reflexión sobre las tendencias en las tecnologías de almacenamiento. Dado que el mayor porcentaje de las operaciones relacionadas con la preservación de contenidos digitales está asociada con las organizaciones, en la sección 2 se revisan los procesos de las mismas que dan lugar a los flujos de información. A partir de ello, se describen ejemplos en los que se utilizan diferentes tipos de almacenamiento a lo largo de la vida útil de los documentos. Lo anterior da pie para que, en la sección 3, se presenten los llamados paradigmas de almacenamiento y se describa cómo obedecen a diferentes necesidades. A continuación, en la sección 4, se presenta un caso de estudio que recupera las ideas expuestas previamente. Por último, en la sección 5 se realiza una serie de consideraciones acerca de las decisiones que deben tomar las organizaciones para abordar el problema del almacenamiento de largo plazo y cerramos reflexionando sobre las limitaciones de las soluciones consideradas, en término de su disponibilidad en el muy largo plazo.

[3] Para mayor referencia, un ZB son mil millones de Terabytes (TB).

2. Los procesos de las organizaciones

En vista de que son las organizaciones las que custodian los mayores volúmenes de documentos digitales, deberíamos entender los ciclos de vida que experimenta la información que tienen bajo su resguardo, si queremos abordar los retos de la gestión de volúmenes masivos de información. A lo largo de su ciclo de vida, el cual está determinado por los objetivos de la organización y sus compromisos, la información puede moverse entre distintas colecciones que pueden alojarse entre distintos repositorios. Se sabe, como en el caso de la imagenología médica, que los documentos pasan por diferentes épocas o momentos, que reflejan un estatus [6], [7], por ejemplo, para reducir el tamaño requerido por su almacenamiento y para brindar características de seguridad como confidencialidad y control de acceso (ver ejemplo en la Fig. 2). Al inicio de su vida se les consulta con mayor frecuencia y en algún momento pueden desecharse o preservarse, si se prevé alguna consulta de carácter histórico.

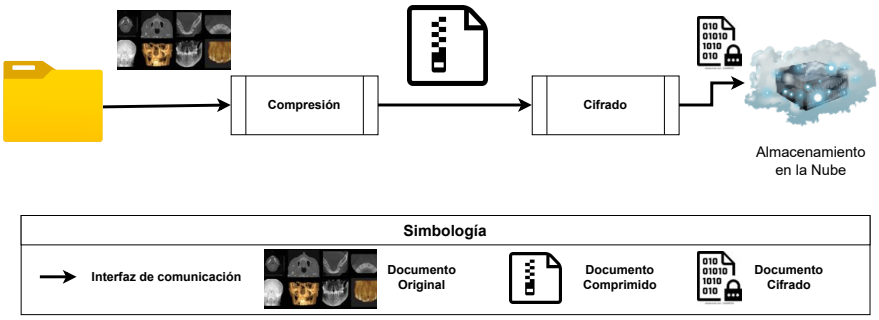


Figura 2: Ejemplo del ciclo de vida de documentos médicos con etapas de compresión y cifrado.

Bajo esta perspectiva, algunos sistemas ya consideran la existencia de diferentes tipos de almacenamiento por los que pueden pasar los documentos digitales de la organización. Al realizar un diseño moderno se debe contemplar un almacenamiento primario, en el que se guardan los documentos que son consultados en el día a día. Por tanto, este almacenamiento debe ser de baja latencia, esto es con bajos tiempos de respuesta, aunque no se espera que sirva para alojar un volumen masivo. Además, debe de existir un almacenamiento secundario en el que se aloja la colección histórica y que, de ser necesario, pueda soportar el movimiento de documentos desde o hacia el almacenamiento primario. En la Fig. 3 se muestra un ejemplo de los tipos de almacenamiento que pueden estar presentes en una organización. Se espera, en cambio, que el almacenamiento secundario pueda acomodar una cantidad masiva de documentos por periodos de tiempo más largos, a costa de una mayor latencia.

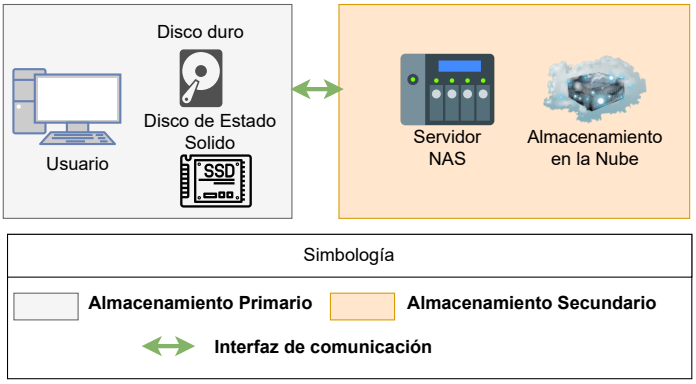


Figura 3: Ejemplo de los tipos de almacenamiento en una organización.

3. Los paradigmas de almacenamiento

Las necesidades para las que se han pensado los diferentes tipos de almacenamiento digital nos obligan a revisar los llamados paradigmas tecnológicos. Se dice que los dispositivos de almacenamiento obedecen a uno de los siguientes paradigmas: por archivos, bloques u objetos. Para entender las diferencias entre estos podemos recurrir a la relación que guardan entre sí los documentos almacenados y los descriptores de dicha colección, a los que podemos denominar sus metadatos.

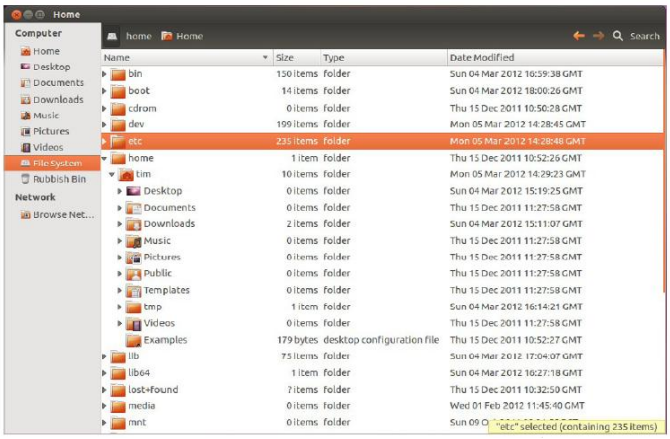


Figura 4: Explorador de archivos en Ubuntu.

En el almacenamiento por archivos tenemos una colección de documentos relativamente pequeña (hasta algunos TB) que *vive* en el mismo dispositivo que

sus metadatos. Por su parte, estos contemplan tanto los detalles de su uso, como los detalles de su alojamiento. Con ello, queremos decir que los metadatos se organizan en un modelo lógico denominado sistema de archivos con una estructura jerárquica o arborescente, en la que se definen carpetas y archivos y se describen los derechos de acceso de aquellos usuarios que pueden trabajar con ellos. Por ejemplo, en la Fig. 4 se muestra el explorador de archivos de Ubuntu, el cual está basado en un sistema de almacenamiento por archivos. Además, también se incluyen los detalles físicos como (en el caso de los discos mecánicos) los cilindros, sectores y pistas que contienen a los documentos digitales de la colección.

Por su parte, en el almacenamiento por bloques, los documentos y sus metadatos se alojan en dispositivos distintos, con la diferencia de que los documentos se fragmentan y, posiblemente, se reparten entre varios dispositivos para aumentar la posibilidad de acceder a ellos de manera concurrente y, con ello, ganar en velocidad de escritura/lectura. Tanto los dispositivos de almacenamiento por archivos como los de bloques se conocen como dispositivos transaccionales. Ello quiere decir que los documentos que se almacenan pueden recuperarse con facilidad para modificarse a lo largo del tiempo.

Finalmente, el almacenamiento por objetos está pensado para acomodar volúmenes de información en la escala de los Petabytes (PB) o más, incluso. La colección de documentos y sus metadatos se almacenan en dispositivos diferentes. Cada documento de la colección pasa por algunas etapas de procesamiento antes de ser almacenado. Entre otras cosas, se le puede fragmentar y luego se genera algún tipo de codificación de redundancia que da lugar a una o varias secuencias de dígitos binarios llamadas objetos. Cada objeto se emplaza dentro de un espacio lógico de almacenamiento que, a diferencia de los modelos jerárquicos, puede entenderse como un espacio plano. Sobre este espacio lógico se mapean las capacidades de los dispositivos físicos que constituyen al sistema. Hay que considerar que, en vista de su escala, la capacidad de almacenamiento se consigue con la participación de un número de dispositivos físicos que crece en la medida en que aumenta la capacidad del sistema en su conjunto.

Un sistema de almacenamiento por objetos debe construirse como un sistema distribuido, definido por software, que soporte una interfaz estándar y optimice el manejo de la redundancia. Las organizaciones deben diseñar sus soluciones de almacenamiento pensando en la disponibilidad en el corto, mediano y largo plazo, lo cual implica una cuidadosa combinación de paradigmas. Esta idea de incorporar diferentes tipos de almacenamiento es a lo que en este trabajo llamamos “un ambiente de almacenamiento”.

Por lo que toca a los sistemas de almacenamiento definidos por software, se trata de sistemas que utilizan mecanismos de software para proporcionar dispositivos virtuales, sobre los que pueden efectuarse las operaciones de almacenamiento y recuperación de información, con independencia de la tecnología que los sustenta. La principal ventaja de este tipo de soluciones es que otorga a los administradores la libertad para elegir o cambiar proveedores, sin perder continuidad en los servicios que se ofrecen o caer en dependencias tecnológicas.

4. El caso del fondo documental de la UAM

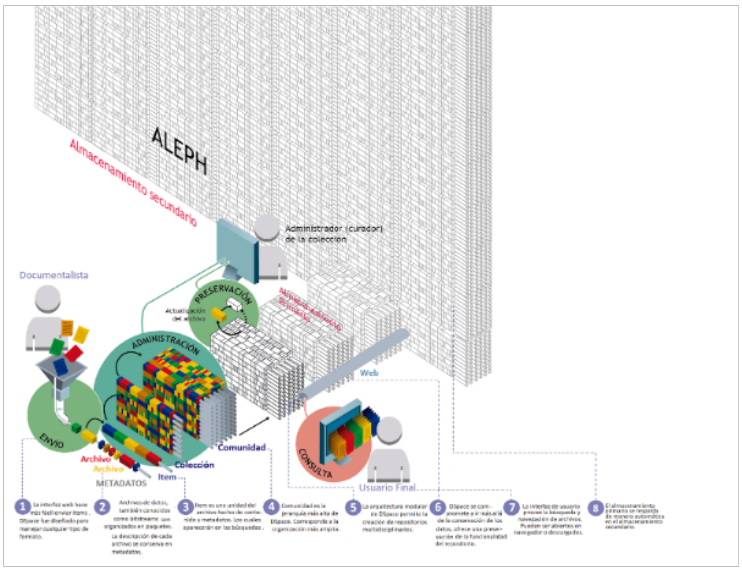


Figura 5: Diagrama del Sistema de Gestión Documental "Fondo UAM".

Presentamos a continuación una descripción del Sistema de Gestión Documental "Fondo UAM". Se trata de un sistema construido para la Rectoría General de la Universidad Autónoma Metropolitana, que contempla la gestión documental de colecciones catalogadas, así como su preservación de largo plazo. La solución integra una plataforma de software denominada *DSpace*, que ha sido conectada con un sistema de almacenamiento por objetos denominado *Aleph*, que es un desarrollo a la medida, el cual ofrece un espacio escalable y de alta disponibilidad (véase la Fig. 5). Para la creación de un repositorio documental se requiere de una plataforma de software capaz de ofrecer una serie de atributos para el manejo de una colección digital catalogada, entre los que podemos mencionar: soporte de normas internacionales para la descripción de los recursos (Dublin Core, o DICOM, por ejemplo), capacidad para interoperar con otros repositorios, definición de roles para la gestión de la colección, definición de atributos de acceso para cada uno de los recursos catalogados (públicos, privados, de grupo), capacidad para crecer el volumen de la colección (escalabilidad), garantías de disponibilidad e integridad de la información. Existe un interés mundial por el uso de este tipo de plataformas, lo que ha dado origen a un conjunto de soluciones, entre las que podemos citar casos tales como: DSpace, OpenRepository, Archimed, Fedora, Eprints, entre otros [9], [2], [1], [5], [4].

De todos los repositorios institucionales en funciones, a nivel mundial, el 48 % de ellos utiliza DSpace como su plataforma de base. Se trata de un software de código abierto que provee herramientas para la administración de colecciones digitales. Soporta una gran variedad de documentos, incluyendo libros, tesis, revistas, fotografías, películas, datos de investigación y otras formas de contenido. Un repositorio institucional soportado por *DSpace* se estructura por comunidades y colecciones. Cada comunidad contiene subcomunidades y/o colecciones y, finalmente, las colecciones contienen elementos (ítems). Por su parte, un elemento puede contener uno o varios archivos digitales. Los usuarios se organizan en cuentas personales y de grupo. Cada usuario tiene definida una serie de permisos y autorizaciones que abarcan desde el acceso restringido en lectura, hasta el acceso sin restricciones para escritura, lectura, modificación y eliminación de registros. Entre algunas de las ventajas de *DSpace* se encuentra el hecho de que es totalmente configurable y personalizable, lo que permite adaptarlo a las necesidades de las instituciones que lo adoptan. El conjunto de organizaciones que ya han adoptado *DSpace* incluye desde pequeños equipos o empresas (públicas o privadas) hasta los gobiernos federales de varios países.

Como hemos argumentado a lo largo de las secciones previas, un sistema de gestión documental requiere de dos tipos de almacenamiento: el primario o de corto plazo, que tiene una capacidad limitada e igualmente una baja latencia de acceso y, por otro lado, un almacenamiento secundario o de largo plazo, que tiene una capacidad que puede crecer, incluso, a escalas de petabytes (PB). Por su diseño, *DSpace* fue pensado solamente para contar con un almacenamiento primario que reside en la computadora donde se instala el servicio de catalogación. En este sentido, nuestra oferta de valor consiste en el desarrollo de una interfaz que conecta a *DSpace* con un almacenamiento secundario que se describe a continuación.

El sistema de almacenamiento *Aleph* es un sistema definido por software que ofrece garantías de escalabilidad y alta disponibilidad, así como un manejo eficiente de la información redundante. Entre sus características más relevantes podemos mencionar un cuidadoso desacoplamiento entre su colección documental y sus metadatos, un middleware que garantiza la consistencia de los metadatos que se manejan como una base de datos replicada, así como sus propios procedimientos para balanceo de carga y almacenamiento de objetos, los cuales se adaptan al número y capacidades de los dispositivos de almacenamiento que componen el sistema.

Para el diseño del Fondo UAM propusimos que cada documento que se recibe en el almacenamiento primario del servidor de *DSpace* se respalde en automático en *Aleph*. De igual manera, todos los metadatos de las colecciones se respaldan regularmente en el propio *Aleph* para habilitar procedimientos de recuperación en caso de desastres. Para optimizar el uso del almacenamiento primario, propusimos un mecanismo de tiempo que elimina todos los documentos que no se han consultado al cabo de un plazo que puede programarse. Si luego de una consulta se requiere un documento que ya no se encuentra en el primario, entonces se recupera del *Aleph* y se vuelve a emplazar en el primario, de manera transpa-

rente para el usuario final. Llamamos a esta nueva organización "el modelo de biblioteca cerrada", porque, como ocurre en algunas bibliotecas, los usuarios no están autorizados para interactuar con todas las colecciones. En su lugar, existe un bibliotecario que almacena y recupera cualquier documento que se almacena en la gran estantería. Hemos construido un bibliotecario automático que es el único autorizado para interactuar con los documentos guardados en el *Aleph*.

Consideramos que este modelo propuesto ofrece algunas ventajas para el manejo de las colecciones: 1) soporta un servicio ágil para un número importante de usuarios concurrentes con diferentes necesidades; 2) es posible manejar diferentes tipos de colecciones y catálogos usando el mismo servidor; 3) la comunicación entre el almacenamiento primario y secundario permite que el primero no se sature sin que se limite el tamaño de las colecciones que son ingresadas al sistema; y 4) al mismo tiempo, el almacenamiento secundario ofrece un almacenamiento de largo plazo.

5. Reflexiones finales

Estamos atestiguando la evolución de las tecnologías de almacenamiento. Entre los implicados inmediatos en esta transformación podemos mencionar al gobierno, los servicios financieros, el sector salud y todas las organizaciones cuyos procesos de negocio están fuertemente vinculados con las tecnologías de la información y comunicaciones (TIC). En los próximos años, los administradores de las TIC (CTO: Chief Technology Officers) deberán tomar importantes decisiones concernientes a las capacidades de almacenamiento de sus organizaciones: ya sea que estas capacidades se soporten basándose en recursos propios, como un servicio provisto por un tercero, o una combinación de diferentes soluciones dentro y fuera de casa, es decir, su infraestructura propia y de terceros. Sin embargo, los cambios impulsados por esta transformación llegarán hasta las pequeñas organizaciones y, aun, a los usuarios particulares.

Por otra parte, estamos observando la construcción de grandes sistemas para el almacenamiento masivo de datos. Sin embargo, puestos en perspectiva, tendríamos que preguntarnos si la información que en ellos hemos depositado podría sobrevivir la prueba del tiempo. La mayor dificultad que podemos reconocer es que se trata de sistemas electrónicos que requieren de energía para mantenerse en operación. Si queremos garantizar un registro de largo plazo tendríamos que pensar en soluciones que no requieran de energía, o bien, que sean capaces de proveerse a sí mismos de esta. Las mejores soluciones son aquellas que se volvieron invisibles y "siempre" han estado ahí. Dicho de otro modo, ¿de qué manera se guarda información en la naturaleza? Al inicio de esta reflexión mencionamos que no existe una solución de almacenamiento que acomode todas las necesidades asociadas al ciclo de vida de la información. Tal vez, en algún momento, los ambientes de almacenamiento deberán incorporar componentes biológicos para la preservación de muy largo plazo, si es que queremos apostar a que el mensaje en la botella nos sobreviva. Reconocemos, sin embargo, que esta última reflexión entra en el terreno de lo especulativo.

6. Agradecimientos

Este trabajo ha sido parcialmente apoyado por el proyecto No. 41756 titulado "Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud" del fondo PRONACES-CONACYT.

Bibliografía

[1] Archimed. *Archimed - Gestion de la connaissance*. Available at: `https://www.archimed.fr/`, Last accessed: 2022-10-27. Oct. de 2022.

[2] Inc Atmire. *Open Repository - premium DSpace hosting*. Available at: `https://www.openrepository.com/`, Last accessed: 2022-10-27. Oct. de 2022.

[3] DOMO. *Data never sleeps*. `https://www.domo.com/learn/infographic/data-never-sleeps-8`. 2019.

[4] University of Southampton Electronics & Computer Science. *EPrints Services*. Available at: `https://www.eprints.org/uk/`, Last accessed: 2022-10-27. Oct. de 2022.

[5] Lyrasis Fedora. *Fedora is the flexible, modular, open source repository platform with native linked data support*. Available at: `https://duraspace.org/fedora/`, Last accessed: 2022-10-27. Oct. de 2022.

[6] Josefina Gutiérrez-Martínez et al. "A software and hardware architecture for a high-availability PACS". En: *Journal of digital imaging* 25.4 (2012), págs. 471-479.

[7] O.S Pianykh. *Digital imaging and communications in medicine (DICOM) Cap 11. DICOM Media and Security*. Springer 2nd Edition, 2012.

[8] David Rydning et al. "The digitization of the world from edge to core". En: *Framingham: International Data Corporation* 16 (2018).

[9] Chris Wilper. *DSpace 4. x Documentation*. `https://wiki.lyrasis.org/display/DSDOC4x`. 2016.

Interoperabilidad de Sistemas Expediente Clínico Electrónico: Modelo para Generación de Repositorio de Datos de Salud

Victor Morales-Rocha[1], Alan Ponce Rodríguez[1], Macario Ruiz Grijalva[2]

[1]Universidad Autónoma de Ciudad Juárez
Av. Del Charro 450 Nte, Cd. Juárez, Chihuahua, México

[2] Tecnológico Nacional de México. Instituto Tecnológico de Ciudad Juárez
Av. Tecnológico 1340, Cd. Juárez, Chihuahua, México.
{victor.morales, alan.ponce}@uacj.mx, mmruiz@itcj.edu.mx

Resumen. En México se realizan más de un millón de consultas médicas diariamente, lo que representa una cantidad enorme de registros médicos. Las instituciones de salud públicas y privadas que hacen uso de algún sistema de expediente clínico electrónico han podido constatar los beneficios de utilizar este tipo de tecnologías; sin embargo, los registros de expediente clínico electrónico que se realizan con dichos sistemas solamente son accesibles dentro de la misma institución, por lo que se desaprovecha, en gran medida, las ventajas de tener la información clínica en formato digital. Al no contar con mecanismos que permitan el acceso generalizado a la información de pacientes, se limita la oportunidad de proveer mejores servicios de salud a nivel nacional y de crear modelos para el análisis de datos a gran escala, necesidad que ha sido evidente a partir de la reciente pandemia. El presente trabajo describe un modelo de interoperabilidad de sistemas de expediente clínico electrónico, así como la plataforma que implementa dicho modelo, a la cual se puede integrar prácticamente cualquier sistema de expediente clínico electrónico con el fin de consultar la información de pacientes que se encuentra distribuida en los diversos sistemas. Además de su evidente utilidad para mejorar la atención a la salud, el modelo de interoperabilidad también permite la recolección de información relevante para conformar un repositorio de datos en salud, con el propósito de ponerlo a disposición públicamente para fines de ciencia de datos en salud.

Palabras clave: Expediente Clínico Electrónico · Repositorio de Datos de Salud · Ciencia de datos · Interoperabilidad.

1 Introducción

De acuerdo con un reporte de la Subsecretaría de Integración y Desarrollo del Sector Salud de la Secretaría de Salud [1], en México se otorgan diariamente 1.2 millones de consultas médicas externas en el sector público. Por su parte, del Informe sobre la Salud de los Mexicanos 2016 [2] se infiere que, del total de consultas externas, el 80% son consultas con médico general y el 20% con especialista. Lo anterior, sin contar las consultas realizadas en entidades privadas. Además, en datos recientes [3] se muestra que las unidades médicas de salud pública a nivel federal que dan atención de primer nivel (consulta externa general) son un total de 15,174, de las cuales 4,576 hacen uso de un sistema de expediente clínico electrónico. Entre las instituciones públicas de salud se encuentran el IMSS, ISSSTE, PEMEX, SEDENA, DIF y SEMAR, así como las unidades médicas a cargo de las secretarías de salud de las 32 entidades federativas.

Por otra parte, en el ámbito privado se utilizan cerca de 100 sistemas de expediente clínico electrónico, y solo una pequeña parte han sido certificados en la Norma Oficial Mexicana NOM-024-SSA3-2012 o se encuentran en proceso de certificación [4].

Estos datos dan una idea aproximada de la cantidad de información de salud que se genera diariamente en formato digital, la cual podría ser aprovechada para la generación

de repositorios para la explotación de la información con fines de investigación clínica y epidemiológica. A medida que se incremente el número de instituciones de salud que hacen uso de un expediente clínico electrónico, se exponenciará la cantidad de información relevante para todo tipo de investigación en el área de salud.

El modelo actual de recolección de información de salud a nivel federal, a través de la plataforma SINBA [5], establece formatos estandarizados para el registro de la información, la cual es capturada en las jurisdiciones sanitarias. En este proceso se pueden presentar una serie de imprecisiones, ya que se requiere la intervención de capturistas que interpretan la información contenida originalmente en formatos en papel o bien en sistemas de información no acreditados por la Norma Oficial Mexicana. Por lo tanto, la transcripción, interpretación y registro de la información por terceros provoca que exista una probabilidad alta de errores. Esto puede ocasionar, entre otras cosas, información tergiversada o duplicada. A consecuencia de esto, se puede concluir que no existe información de salud a nivel nacional precisa y la toma de decisiones entre las autoridades de salud, al basarse principalmente en dicha información, tiene un margen de error amplio.

En las siguientes secciones se describe un modelo de interoperabilidad para sistemas de expediente clínico electrónico, el cual, además de facilitar el acceso a la información clínica de un paciente, sin importar en qué sistema se haya generado, permite la recolección de información relevante para el diseño de modelos de ciencia de datos en salud. Esta información puede incluir, por ejemplo, síntomas, diagnósticos, edad, código postal del paciente, etcétera. Al diseñar un mecanismo para extraer de manera automática la información nacional de salud, desde los sistemas de expediente clínico electrónico, se ofrecerá información más fiable, impactando directamente en las estrategias de salud del país.

2 Interoperabilidad en sistemas de información en salud

Como resultado del desarrollo acelerado de la era digital, se ha detectado la necesidad de que los sistemas compartan la información que generan a fin de optimizar procesos y/o contar con información de manera oportuna. La interoperabilidad puede ser necesaria dentro de una organización, donde diferentes sistemas y dispositivos de información deben compartir datos entre sí. Asimismo, las empresas y todo tipo de organizaciones podrían compartir información con otras para llevar a cabo sus operaciones regulares. Ejemplo de esto en un sistema de banca en línea, que requiere interactuar con otros bancos, con diferentes entidades financieras y con otro tipo de organizaciones para ofrecer cada vez más servicios a sus clientes. Si no existieran protocolos, formatos y estándares de interoperabilidad, tal interacción no sería posible hoy en día. Lo mismo sucede en otro tipo de aplicaciones, tal como el área de la salud, en donde la posibilidad de compartir información permite una mejor atención a los pacientes.

La interoperabilidad puede definirse como "La capacidad de dos o más sistemas o componentes para intercambiar información y utilizar la información que ha sido intercambiada" [6], o como "La capacidad, promovida pero no garantizada por la conformidad conjunta con un conjunto dado de estándares, que permite que equipos heterogéneos, generalmente construidos por varios proveedores, trabajen juntos en un entorno de red" [7]. Debido a que la interoperabilidad implica el uso de la información que se intercambia, y no simplemente la capacidad de transferirla de un sistema a otro, se puede distinguir entre dos tipos de interoperabilidad de sistemas, a saber, la sintáctica y la semántica, como se puede ver en [7] y [8].

En las últimas décadas, las instituciones de salud han optado por incorporar sistemas de información que gestionan la información clínica de los pacientes, lo que ha dado como resultado el Expediente Clínico Electrónico (ECE). En México, una gran cantidad de instituciones de salud hacen uso de algún sistema de gestión de expediente clínico electrónico. Sin embargo, el contar con un expediente clínico electrónico de manera local, o que incluye solamente la información generada en una institución de salud, es solo un primer paso para disponer de manera oportuna de la información clínica.

La diversidad de sistemas de expediente clínico electrónico ha obstaculizado una interacción entre diferentes instituciones de salud, ya que, a pesar de que algunos de los sistemas cumplen con la norma oficial mexicana NOM-024-SSA3-2012, existen dificultades técnicas que inhiben la posibilidad de compartir información del expediente clínico electrónico entre entidades de salud, principalmente debido al formato y a la estructura de la información almacenada en las bases de datos de dichos sistemas, así como a la variedad de tecnologías utilizadas.

Esta falta de interoperabilidad limita en gran medida el acceso universal al expediente clínico electrónico en México, lo cual implica una atención deficiente a los pacientes que acuden a distintas unidades de salud, así como la posible multiplicidad de la información de un solo paciente, reportada a las instancias de salud federales que realizan análisis estadísticos e investigación clínica y epidemiológica.

Actualmente, los esfuerzos internacionales en relación con interoperabilidad de sistemas de información en salud se centran en estándares de registro, almacenamiento, intercambio, extracción y uso adecuado de la información. La tendencia mundial es hacia la integración de la información a través de su registro adecuado en el punto de atención, a través de la conceptualización y jerarquización de la terminología clínica y la estandarización del intercambio de información en el sector salud a través de estándares, para que trabajen de manera integrada sobre interfaces de usuario que faciliten al trabajador de la salud el registro de la información y que disminuyan el esfuerzo y tiempo invertido en el trabajo administrativo de llenado de formatos en el expediente clínico. Asimismo, el punto anterior hará más eficiente el intercambio de información con las plataformas nacionales, disminuyendo la intervención humana, los errores de transcripción, la interpretación de la información y el uso inadecuado de la misma.

En algunos países ya se ha trabajado ampliamente con el desarrollo de esquemas de interoperabilidad del ECE. Tal es el caso del proyecto MedCom [9], en Dinamarca, a través del cual se ha logrado estandarizar e interconectar la información clínica de todos los ciudadanos, a fin de hacer más eficientes las labores de las instituciones de salud. El 100% de los médicos de atención de primer nivel en el país tienen acceso al expediente clínico completo de sus pacientes a través de diversos sistemas de información. Dichos sistemas de información han sido desarrollados por al menos 50 proveedores, lo que da un ejemplo claro de interoperabilidad entre diferentes sistemas de expediente clínico electrónico. El proyecto MedCom también permite a los pacientes el acceso a su propio expediente y les alerta cada vez que un profesional de la salud accede a su información médica.

La Comisión Europea ha anunciado que tiene como objetivo permitir a todos los ciudadanos de cualquier parte de Europa el acceso a su propio expediente clínico. Para lograr esto, se debe contar con los mecanismos de interoperabilidad. En un informe emitido por dicha Comisión [10], se emiten algunas recomendaciones para lograr la interoperabilidad buscada. Entre ellas, se destaca que se tienen que lograr tres aspectos primordiales: interoperabilidad técnica, interoperabilidad semántica y protección de datos personales.

En México, por su parte, se han realizado esfuerzos importantes hacia el desarrollo de un expediente clínico electrónico único, principalmente en instituciones de salud federales o estatales. Tal es el caso de instancias federales como son el IMSS, ISSSTE, PEMEX, SEDENA, DIF Y SEMAR, así como en instancias estatales, como es el caso de las Secretarías de Salud de los estados de Colima, Chihuahua, Zacatecas y Guanajuato, entre otros. Entre los casos mencionados se observa que las instituciones públicas federales han logrado una cobertura importante del uso del expediente clínico electrónico. Esa cobertura implica que en sus unidades de salud se está utilizando el sistema de ECE propio de la institución; sin embargo, en pocos casos se cuenta con una interconexión entre todas las unidades de salud de la propia institución. En el caso de las instancias estatales, Colima es el estado que cuenta con una mayor cobertura de uso. Desde el año 2011 cuenta con el 100% de las unidades de atención primaria, usando el sistema SAECCOL y, actualmente, 400,000 pacientes del estado cuentan con expediente clínico electrónico. El mismo sistema es utilizado en los estados de Tlaxcala y Guanajuato. El éxito del alcance de dicho sistema,

especialmente en Colima, se debe principalmente a la estrategia de implementación gradual, así como del apoyo constante de las entidades médicas y administrativas de la secretaría de salud del estado.

De acuerdo con las experiencias locales e internacionales, no se considera conveniente contar con un sistema único de expediente clínico electrónico en México debido a: i) las dificultades técnicas para procesar y almacenar los expedientes clínicos de una población de alrededor de 126 millones; ii) la centralización de la información; iii) la monopolización de la gestión del expediente clínico electrónico; y iv) el desperdicio en las inversiones ya realizadas tanto en el sector público como en el privado.

Sin embargo, es posible abordar la problemática planteando la interoperabilidad de los diversos sistemas a través de los mecanismos adecuados para la transferencia de datos. Estas experiencias y recomendaciones, además de la normatividad actual mexicana, sirven de guía al trabajar en un modelo de interoperabilidad de sistemas de ECE en México.

3 Modelo de interoperabilidad de sistemas de expediente clínico electrónico

En esta sección se describe un modelo tecnológico que facilita la interoperabilidad de los diferentes sistemas de expediente clínico electrónico en México. La primera tarea para crear el modelo propuesto fue definir una serie de submodelos a manera de capas de información e interacción de elementos. El modelo está representado por 4 capas, como se describe a continuación (ver Figura 1). Aunque cada capa ofrece un nivel específico de interoperabilidad, la finalidad es integrar cada una de ellas en un modelo completo.

Capa léxica. Esta primera capa nos lleva a un "Modelo Lógico de Referencia" basado en el análisis conjunto y cruce de información de tres elementos principales: las Guías de Intercambio de Información en Salud (GIIS); la Norma Oficial Mexicana NOM-004-SSA3-2012; y la Norma Oficial Mexicana NOM-024-SSA3-2012. El Modelo Lógico de Referencia se centra en la integración de los diferentes datos que contiene un expediente clínico electrónico y que pueden ser tratados a nivel de información.

Capa sintáctica. Ésta especifica la forma en que se deben definir y estructurar los diferentes elementos del expediente clínico electrónico. El resultado es un conjunto de plantillas o estructuras en formato JSON que fueron diseñadas para llegar a un "Modelo Técnico" que permita el intercambio de información basado en el estándar HL7-FHIR [11]. La capa sintáctica se enfoca en formar y enviar los elementos del ECE seleccionados para ser compartidos. Dichos elementos se integran en un archivo JSON con el formato establecido.

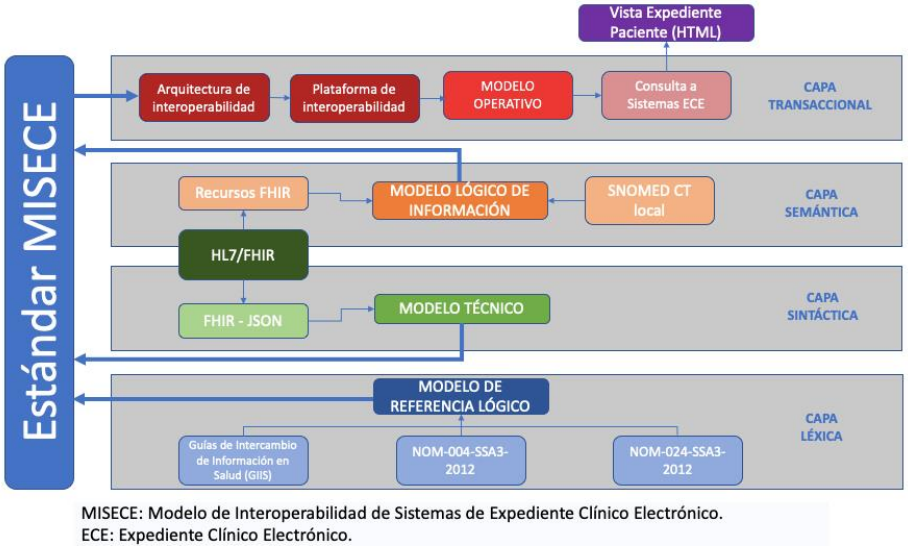


Figura 1. Capas del modelo de interoperabilidad.

Capa semántica. En esta capa se define un "Modelo Lógico de Información". Éste es responsable de interpretar y utilizar los datos que se comparten entre dos o más componentes evitando la ambigüedad. Para este nivel, se ha tomado como base la terminología clínica SNOMED CT [12], específicamente para los diagnósticos y tratamientos. Esta terminología se ha adaptado a una versión para el idioma español de México y brinda una base ontológica que provee la semántica del modelo. El uso de SNOMED CT, junto con los recursos FHIR, que también proveen semántica, hasta cierto punto, debido a su estructura estandarizada, forman el Modelo Lógico de Información.

Capa transaccional. Es la capa de interacción humana y permite presentar la información en un formato uniformizado. Se utiliza un archivo en formato HTML para permitir al médico o paciente visualizar la información contenida en diversos sistemas de ECE. La capa transaccional define un "Modelo Operativo" derivado de una "Plataforma de Interoperabilidad" y ésta, a su vez, derivada de la "Arquitectura de Interoperabilidad".

3.1 Plataforma de interoperabilidad

La plataforma de interoperabilidad que utiliza el modelo descrito forma parte de la capa transaccional. La plataforma funciona como un *middleware* para recibir y procesar solicitudes de consulta de información clínica de un paciente

y para presentar al solicitante los resultados de la búsqueda en un formato estandarizado.

La Figura 2 muestra la arquitectura de la plataforma y la interacción con componentes externos, principalmente sistemas de expediente clínico electrónico, así como el repositorio de datos de salud.

A continuación se describen brevemente los elementos principales de la arquitectura.

- Paciente: Persona titular del ECE que permite que su información clínica sea consultada por un profesional de la salud, a través de la plataforma de interoperabilidad. También tiene la posibilidad de consultar su información clínica almacenada en los diferentes sistemas ECE.
- Solicitante: Personal médico con posibilidad de consultar el ECE de los pacientes a través de la interfaz del sistema ECE de su institución o a través de la interfaz web de la plataforma. Las consultas del ECE requieren la autorización del paciente por medio de un código numérico que se le envía a éste en tiempo real a través de un SMS. El médico también puede realizar una consulta de un paciente sin la autorización de éste. En este caso, solo se puede consultar la información básica que sirva para la atención de emergencia del paciente. El solicitante también puede ser un paramédico que requiera consultar la información básica del paciente. Ejemplo de información básica del paciente puede incluir alergias, tipo de sangre, tratamiento actual o reciente, entre otros.
- Sistema ECE: Sistema de expediente clínico electrónico de una institución médica privada o pública, utilizado para la administración del expediente del paciente. Permite consultar los expedientes que tienen almacenados otros sistemas de ECE a través de la plataforma de interoperabilidad. De igual manera, conceden a la plataforma el acceso a los registros clínicos almacenados en su sistema a fin de compartir dicha información con solicitantes usuarios de otros sistemas.
- Administración de usuarios y consulta de ECE: Interfaz web que forma parte de la plataforma, utilizada para la administración de instituciones y usuarios de dichas instituciones. En esta interfaz es posible dar de alta una nueva institución y agregar usuarios (médicos, pacientes, paramédicos) pertenecientes a la misma para otorgarles acceso a la consulta del ECE. Esta interfaz también permite la consulta de ECE por parte de los diferentes usuarios pertenecientes a una institución que fueron previamente registrados.
- Repositorio de datos clínicos: Repositorio de datos recopilados por la plataforma de interoperabilidad, a través de los puntos de acceso desarrollados por los diferentes sistemas de ECE que forman parte del módulo. Los datos son recolectados, preprocesados y almacenados en una base de datos para su posterior consulta y/o descarga por parte de los investigadores.
- API Gateway: Es una interfaz que ayuda a manejar las solicitudes y respuestas de los sistemas integrados a la plataforma. También es responsable de autenticar todos los sistemas y usuarios que intentan interactuar con el módulo. Además, se encarga de gestionar la conexión entre los diferentes módulos que componen el modelo de interoperabiidad.

- Módulo de intercambio de información: Módulo encargado de recibir las peticiones de consulta para un paciente específico. Además, este módulo se encarga de devolver la respuesta al usuario solicitante de información.
- Módulo de indexación de información y validación: Módulo encargado de mantener la tabla de índice de pacientes a través de mecanismos de actualización para mantener la información consistente. Este módulo también se encarga de validar la autorización del paciente al intentar consultar su expediente.
- Módulo de Procesamiento de la Información: Módulo encargado de procesar los registros que contengan terminología médica del expediente o expedientes recibidos. El procesamiento se lleva a cabo para sugerir los términos preferibles basados en la terminología clínica SNOMED-CT.[1]
- Módulo de registro de eventos: Módulo encargado de registrar las consultas realizadas a través de la plataforma. Estos registros se almacenan en una Blockchain privada. Cada evento es la consulta realizada por un usuario válido. Los datos de la consulta que se guardan son: fecha/hora de consulta, paciente consultado, sistema o interfaz desde donde se realiza la consulta, usuario que realiza la consulta y sistema de ECE que aportó resultados.
- Servicio de autenticación: Servicio de autenticación externo utilizado para la administración y autenticación de los usuarios de la plataforma.
- Red privada de blockchain: Red privada de blockchain compuesta por 6 nodos basada en una implementación de ethereum llamada Hyperledger Besu[2] y utilizada con algoritmo de prueba de autoridad como mecanismo de consenso.
- Ontología local SNOMED-CT: Modelo de referencia lógico conceptual para definir la terminología clínica preferible a utilizar en el módulo de procesamiento para estandarizar los conceptos que se reciben de los diferentes sistemas ECE. Fue desarrollado con base en SNOMED CT y una traducción local creada para el idioma español de México. La traducción se obtiene a partir de un proceso de revisión de conceptos, conceptos preferentes, sinónimos y regionalismos para el español de México.

[1] La terminología SNOMED CT se puede consultar en https://browser.ihtsdotools.org/?

[2] Para más información de Hyperledger Besu se puede consultar https://www.hyperledger.org/use/besu

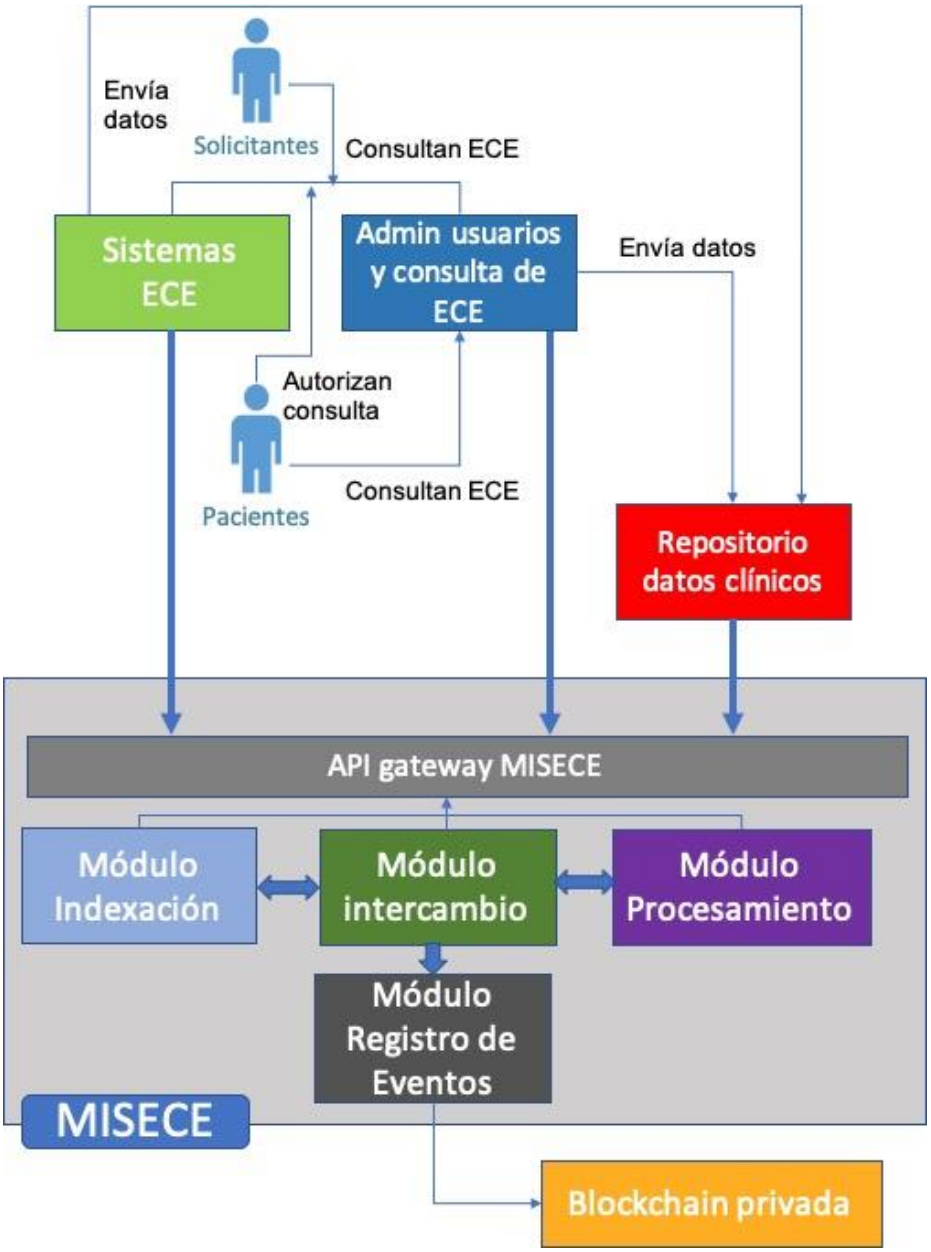


Figura 2. Representación de la arquitectura de la plataforma

Tal como se ha explicado en la sección anterior, el modelo de interoperabilidad hace uso de HL7-FHIR, por lo que se definieron los recursos FHIR, es decir, las estructuras de datos que se utilizan en el MISECE para el transporte de los

datos del expediente clínico electrónico. En este sentido, cada sistema de ECE que se integra al modelo de interoperabilidad previamente debe realizar el mapeo de la estructura de datos definida con la estructura de los campos que utiliza el sistema de ECE. Una vez realizado dicho mapeo, se construye una API que permitirá el accesso a su base de datos, de manera específica a la información contenida en los campos identificados. El transporte de los datos entre los sistemas de ECE y la plataforma se realiza en un archivo JSON que cumple la estructura de HL/FHIR definida.

La integración de los sistemas de un expediente clínico electrónico con la plataforma de interoperabilidad se realizará en base a un proceso de validación y certificación que permitirá determinar el nivel de madurez de dicho sistema para interoperar. Los aspectos a validar serán, entre otros, el cumplimiento mínimo de la normatividad mexicana de expedientes clínicos, el tratamiento seguro de los datos y la capacidad técnica de desarrollar las APIs para el intercambio de datos.

3.2 Repositorio de datos de salud

El repositorio de datos de salud está conceptualizado, dentro de la arquitectura de la plataforma de interoperabilidad descrita previamente, como una aplicación cliente que realiza consultas de registros almacenados en los sistemas de expediente clínico electrónico. La diferencia principal con un cliente común, como lo sería un usuario médico, por ejemplo, es que las solicitudes de información que realiza el repositorio no están relacionadas con un paciente específico, sino con un conjunto de pacientes que hayan tenido alguna actualización en su expediente clínico electrónico en un periodo determinado. A fin de diseñar y crear el repositorio, se realizaron las siguientes actividades: selección de datos para el repositorio; método de extracción de datos; anonimización de registros; pre-procesamiento de datos; almacenamiento de datos; transferencia de datos a repositorio público.

Para la selección de datos para el repositorio, se consultó con expertos en investigación clínica y epidemiológica, ya que era importante determinar cuáles datos, del total de los datos que componen un expediente clínico electrónico, serían relevantes para realizar ciencia de datos.

El método de extracción de datos define el proceso en el que la aplicación del repositorio solicita y obtiene los registros de pacientes que se encuentran almacenados en diferentes sistemas de ECE. Se consideran dos tipos de consultas que se pueden realizar desde el repositorio, siendo la primera una petición bajo demanda. Este tipo de consulta es la utilizada para una carga inicial del repositorio una vez que diversos sistemas de expediente clínico electrónico han sido integrados al modelo. El segundo tipo de consulta es una petición automatizada, la cual se lanza desde el repositorio hacia la plataforma de interoperabilidad en fechas/períodos previamente configurados. Por ejemplo, se puede configurar para que, el último día de cada mes, la plataforma realice la búsqueda de todos los registros que han sido actualizados desde la última consulta automática. Tal como en las consultas de pacientes individuales que puede realizar un usuario, a fin de que la aplicación del repositorio pueda hacer las peticiones de registros globales de pacientes, los sistemas de ECE deben construir una API

que permita el acceso y extracción de los datos que se determinaron relevantes para el repositorio.

Siendo la privacidad de los pacientes una parte importante en el desarrollo del modelo, los registros almacenados en el repositorio no deben revelar la identidad de los pacientes. Para lograr esto, se diseñó un proceso de anomimización de los registros que consiste en aplicar una combinación de funciones criptográficas al identificador del paciente, incluyendo una función hash, la cual es una función irreversible. De esta manera, aunque se pueden identificar los registros individuales de pacientes con el fin de dar seguimiento a través del tiempo a casos de interés, la identidad del paciente se mantiene confidencial.

Una vez que se reciben los datos provenientes de los diferentes sistemas de ECE y que han sido anonimizados, se realiza un pre-procesamiento de los mismos, a fin de almacenar solamente registros "limpios". Debido a la variedad de sistemas de ECE que aportarán información para el repositorio, se estima probable que haya registros nulos, incompletos o bien con información que no corresponde a los campos. Por estos motivos, se contempla una fase de pre-procesamiento que permita almacenar solamente información relevante para ser utilizada en un proceso de investigación.

Una vez que los datos están listos para su explotación, se almacenan de manera temporal en una base de datos intermedia. Posteriormente, serán enviados de manera periódica a un repositorio público de datos de salud administrado por el Consejo Nacional de Ciencia y Tecnología (CONACYT), institución mexicana que promueve la investigación científica.

4 Conclusiones y trabajo futuro

En México, actualmente existen algunos repositorios públicos de datos de salud, los cuales pueden ser aprovechados para realizar investigación; sin embargo, estos han sido diseñados para cumplir un objetivo muy específico. En otros casos, los datos contenidos en repositorios públicos provienen de la recopilación a nivel nacional de registros de salud a través de procesos poco rigurosos, principalmente debido a las diversas fuentes y formatos de los registros existentes.

Los registros digitales que realiza el personal médico a través de sistemas de expediente clínico electrónico son mucho más confiables que los que pueden realizarse por otros medios, por ejemplo, mediante la transcripción o captura por parte de personal de apoyo o administrativo. En este sentido, el contar con un modelo y con las herramientas que permitan la recopilación de los registros de salud, almacenados principalmente en las bases de datos de sistemas de ECE, ofrece beneficios importantes para la ciencia de datos. Al contar con datos más fiables, la analítica de los mismos ofrecerá información más precisa y, por lo tanto, más valiosa para la toma de decisiones. Por otro lado, el modelo propuesto y las herramientas de software desarolladas permiten que la recopilación de datos pueda realizarse en períodos tan cortos como sea necesario, lo cual es de suma relevancia, especialmente para situaciones de epidemias o pandemias.

La información clínica recabada en el modelo de repositorio descrito consiste en un subconjunto muy amplio del expediente clínico electrónico, al considerar

que la mayor parte de la información que se registra en los sistemas de ECE es valiosa para ser analizada; sin embargo, puede resultar de gran utilidad definir subconjuntos de datos más pequeños que sirvan para la creación de modelos de *machine learning* específicos. En este sentido, como trabajo futuro, se pretende definir subconjuntos de datos del expediente clínico electrónico, junto con algunas instituciones públicas mexicanas que realizan investigación en salud, como es el caso del Instituto Nacional de Geriatría y del Instituto Nacional de Salud Pública.

Por otro lado, aunque el modelo presentado ejemplifica el intercambio de datos específicamente entre sistemas de expediente clínico electrónico y el acceso a los mismos para la generación de un repositorio de datos, cualquier otro tipo de sistema de información de atención a la salud o de registro de información clínica, incluyendo dispositivos o sensores, pueden también ser integrados al modelo de interoperabilidad para que la información que almacenan sea compartida para fines de consulta médica o bien para ser utilizada en modelos de ciencia de datos.

Agradecimientos

Los autores dan un agradecimiento especial al Consejo Nacional de Ciencia y Tecnología (CONACYT), institución que ha financiado este trabajo. Se agradece igualmente al Laboratorio Nacional de Tecnologías de Información, sede UACJ, por la infraestructura y soporte técnico proveídos para este proyecto.

Referencias

[1] Atención Primaria de Salud Integral e Integrada, «Subsecretaría de Integración y Desarrollo del Sector Salud,» 2019. [En línea]. Disponible: http://www.sidss.salud.gob.mx/site2/docs/Distritos_de_Salud_VF.pdf. [Último acceso: 30 Septiembre 2022].

[2] Informe sobre la salud de los mexicanos, «Gobierno de México,» 2016. [En línea]. Disponible: https://www.gob.mx/cms/uploads/attachment/file/239410/ISSM_2016.pdf. [Último acceso: 14 Septiembre 2022].

[3] Gomez, Juan Carlos, *Diagnóstico e implementación de ECE,* Ciudad de México, 2020.

[4] Dirección General de Información en Salud, «Sires Certificados en la NOM-024-SSA3-2012,» Secretaría de Salud, 2021. [En línea]. Disponible: http://www.dgis.salud.gob.mx/contenidos/intercambio/sires_certificacion_gobmx.html. [Último acceso: 19 Julio 2022].

[5] Secretaría de Salud, «Sistema Nacional de Información Básica en Salud,» 2022. [En línea]. Disponible: https://sinba.salud.gob.mx/. [Último acceso: 18 Septiembre 2022].

[6] J. Radatz, A. Geraci y F. Katki, «IEEE Standard Glossary of Software Engineering Terminology,» *Standards Coordinating Committee of the Computer Society of the IEEE,* pp. 1-84, 1990.

[7] M. Braustein, «Healthcare in the Age of Interoperability: The Promise of Fast Healthcare Interoperability Resources,» *IEEE Pulse,* vol. 9, nº 6, pp. 24-27, 2018.

[8] N. Ide y J. Pustejovsky, «What Does Interoperability Mean , Anyway? Toward an Operational Definition of Interoperability for Language Technology,» de *2nd International Conference on Global Interoperability for Language Resources*, Hong Kong, 2010.

[9] A. Kushniruk, E. Borycki y M.-H. Kuo, «Advances in Electronic Health Records in Denmark: From National Strategy to Effective Healthcare System Implementation,» de *EFMI Special Topic Conference*, Reykjavik, 2010.

[10] Commission Recommendation on Cross-border Interoperability of electronic health record systems, «European Comission,» 2019. [En línea]. Disponible: https://digital-strategy.ec.europa.eu/en/library/recommendation-european-electronic-health-record-exchange-format?pk_source=ec_newsroom&pk_medium=email&pk_campaign=dae%20Newsroom. [Último acceso: 23 Agosto 2022].

[11] E. Madrigal y L. Phi Le, «Digital media archive for gross pathology images based on open-source tools and Fast Healthcare Interoperability Resources (FHIR),» *Modern Pathology,* nº 34, pp. 1686-1695, 2021.

[12] R. Kate, «Automatic full conversion of clinical terms into SNOMED CT concepts,» *Journal of Biomedical Informatics,* vol. 111, 2020.

Muyal-Painal: Servicio para el transporte y almacenamiento de datos médicos

José Luis González-Compeán[1 [0000−0002−2160−4407]], Victor J. Sosa-Sosa[1 [0000−0001−5465−0410]], and Hugo G. Reyes-Anastacio[1 [0000−0002−9003−6765]]

Centro de Investigación y Estudios Avanzados (Cinvestav) del IPN Unidad Tamaulipas, Cd. Victoria 87130, Tamaulipas, México.
{joseluis.gonzalez, vjsosa, hugo.reyes}@cinvestav.mx

Resumen *Muyal-Painal* es un conjunto de servicios y sistemas desarrollados para que las organizaciones de salud y la comunidad científica puedan: i) almacenar, distribuir y localizar sistemas o servicios de procesamiento a través de catálogos de servicios; ii) generar soluciones que permitan brindar rentabilidad costo-beneficio del almacenamiento y transporte de datos; iii) almacenar, publicar y transmitir repositorios de datos de manera local (intra-institucional) y federada (inter-institucional) utilizando un modelo de publicación/suscripción interconectando la infraestructura de TI privada con servicios de nube (pública o híbrida). *Painal* permite crear catálogos de servicios para que las organizaciones coloquen sus sistemas, servicios o aplicaciones para que otras instituciones de la federación puedan descargarlos y utilizarlos. También, es posible crear catálogos de datos para el almacenamiento persistente y compartición segura de los mismos entre múltiples organizaciones mediante técnicas de compresión, seguridad y paralelismo que permiten reducir el espacio de almacenamiento requerido, brindar control de acceso y confidencialidad a los datos, así como reducir el tiempo de respuesta de las soluciones.

Palabras Clave: Big Data · Catálogos · Almacenamiento Federado · Almacenamiento Seguro

1. Introducción

El volumen de datos producidos y gestionados por las organizaciones ha ido creciendo en los últimos años; esto, debido a que los usuarios, de manera individual o asociados a una organización, producen, almacenan y utilizan datos de manera constante y continua, por ejemplo crean o recolectan nuevos datos (fotos, documentos, audio, vídeo, etc.), ocasionando un efecto de acumulación de datos [19]. Los usuarios y las aplicaciones contratan los servicios de almacenamiento en la nube para solucionar su problema de almacenamiento a través de un modelo de negocio denominado *pago por uso* (*pay-as-you-go*) [8]. A pesar de que estos servicios se construyen utilizando sistemas distribuidos, una acumulación constante de datos crea, de manera gradual, una colección centralizada de

datos en dichos servicios de almacenamiento. Esto no solo da lugar a un único punto de fallo en los escenarios de interrupción [13], es decir, si el proveedor deja de responder las peticiones de los usuarios, éstos no tendrán acceso a los datos. Adicionalmente, se produce una dependencia con el proveedor de estos servicios [23], haciendo que el usuario dependa de sus herramientas y dificultando la migración de sus datos a otro servicio. A esta dependencia se le denomina *vendor lock-in* [24]. La gestión del almacenamiento de datos resulta un proceso clave para que los usuarios y organizaciones reduzcan la saturación y centralización de sus datos en un solo proveedor de almacenamiento en la nube, así como para reducir el tiempo de respuesta de los intercambios de datos en línea con sus socios (inter-institucionales) y otros usuarios (intra-institucionales). En un escenario de intercambio de datos, los aspectos de rendimiento y gestión del almacenamiento resultan críticos, sobre todo en enfoques jerárquicos con distintos niveles de acceso utilizando infraestructuras heterogéneas, es decir, de entornos donde los recursos tienen características diferentes [1], [26].

En este documento se presenta el servicio *Muyal-Painal* o *Painal*, el cual se compone por un conjunto de servicios que permiten realizar el transporte, almacenamiento y gestión de datos médicos a través de catálogos digitales que son accedidos mediante un modelo de publicación/suscripción (Pub/Sub). *Painal* permite realizar la compartición segura de servicios y datos entre diferentes instituciones con el objetivo de mejorar los procesos de intercambio de datos sensibles a través de un entorno regulado (federado).

Painal puede ser utilizado por personal de la salud para compartir datos de un paciente a través de las diferentes etapas del ciclo de vida de los mismos. Por ejemplo, si a un paciente le realizan una tomografía, las imágenes y textos asociados con la misma pueden ser colocados en un repositorio digital, al que llamamos catálogo, el cual es una carpeta que posiblemente se ubica en un servidor del hospital donde se guardan la información relacionada con el paciente en cuestión. Este catálogo podrá ser compartido, de manera transparente, a través del servicio que proporciona *Painal*. Una vez compartido el catálogo, un médico con acceso al mismo podrá visualizar la tomografía, de manera casi inmediata, desde su computadora, gracias a que cuenta con *Painal*. En caso de ser necesario, el catálogo también podrá ser compartido con un especialista, posiblemente de otro hospital, que cuente también con el servicio de *Painal*. En este escenario, solo el paciente y los médicos que dan seguimiento a su caso tienen acceso a los datos, aún cuando existan otros médicos, técnicos y especialistas que pudieran estar accediendo a *Painal*. Lo anterior, gracias a que *Painal* cuenta con un servicio de control de acceso y asignación de permisos que se hace responsable de verificar que los usuarios solo accedan a los datos que les corresponden, proporcionando, así, confidencialidad a los mismos.

1.1. Definiciones

En esta sección se proporcionan algunas definiciones que permitirán al lector comprender los principales componentes y servicios que proporciona *Painal*.

- Cómputo en la nube (*Cloud computing*): El Instituto Nacional de Estándares y Tecnología (NIST, por sus siglas en inglés) define el cómputo en la nube como "un modelo que permite el acceso ubicuo, conveniente y bajo demanda a un conjunto de recursos computacionales configurables que pueden ser rápidamente provistos y lanzados con el mínimo esfuerzo de manejo o interacción con el proveedor de servicio" [21]. Dicho modelo cuenta con cinco características esenciales: autoservicio bajo demanda, amplio acceso a la red, conjunto de recursos heterogéneos, rápida elasticidad y un servicio medido. Los modelos de servicio a los que se tiene acceso a través del cómputo en la nube son Software como Servicio (SaaS, del inglés Software as a Service) e Infraestructura como Servicio (IaaS, del inglés Infraestructure as a Service), los cuales se despliegan en cuatro modelos de nube diferentes: pública, comunitaria, privada e híbrida [14]. La nube pública es la que se encuentra fuera de las instalaciones de la organización y es completamente manejada por el proveedor de servicios, en donde el usuario final solo accede a los recursos a través de internet. La nube comunitaria es aquella donde un grupo de organizaciones se integran para compartir los recursos, los cuales pueden gestionar en un ambiente tipo federado. En el caso de la nube privada, la organización cuenta con las instalaciones y equipo necesario para desplegar un entorno de nube en donde tienen completo control sobre los recursos y de cómo se manejan los procesos. Por último, la nube híbrida es una combinación entre la privada y la pública, en donde se establecen y gestionan qué operaciones se hacen en la organización y cuáles son delegadas a un proveedor.
- Arquitectura de malla: Es la manera en la que se le conoce a un conjunto de nodos de almacenamiento que trabajarán en conjunto para guardar información de manera segura y confiable. Los nodos de almacenamiento se interconectan creando una red (a la que se le conoce como red P2P sin servidor). La idea básica es ofrecer a las organizaciones la posibilidad de construir soluciones de almacenamiento basadas en sus recursos disponibles, locales o remotos; estos últimos pudieran, por ejemplo, estar alojados en la nube. Además, las organizaciones pueden elegir los parámetros a alto nivel (es decir, el número de nodos de almacenamiento necesarios para una solución determinada) y, de esta manera, *Painal* generará un servicio de almacenamiento confiable sin servidor, conocido también como SeRSS, por sus siglas en inglés.
- Infraestructura heterogénea: Se refiere a un conjunto de recursos de cómputo utilizados para desplegar los clientes de los servicios, que tienen características de hardware (capacidad de almacenamiento, memoria o procesador), software (sistema operativo) o infraestructura (nube pública, privada o híbrida) diferentes.
- Mecanismo de publicación/suscripción: En *Painal*, representa un mecanismo utilizado para la publicación/compartición de datos a través de catálogos de documentos o servicios, el cual es responsable de preparar los datos previos a ser mandados a través de la red, y que serán consumidos por otro servicio de suscripción/adquisición de datos, el cual reconstruye los datos para que sean

visibles para el usuario final. Es complementado con técnicas de seguridad, integridad y control de acceso para que los datos sean veraces y que solo los usuarios con los permisos correspondientes puedan acceder.

- Metadato: En *Painal* se le denomina metadato a las características que describen a los datos, archivos o contenido digital que es compartido a través de él, por ejemplo, el nombre del archivo, el tamaño original, la ruta del servidor donde se encuentra almacenado, así como de los archivos que contienen las credenciales de control de acceso requeridas para extraer el contenido de los archivos.
- Multi-hilo: En *Painal*, las aplicaciones ejecutadas en multi-hilo son generadas mediante paralelismo basado en tareas, el cual clona una aplicación y los clones son ejecutados en *cores* o *hilos* de procesamiento independientes administrados por *Painal*. Este proceso es transparente para el usuario final, ya que el proceso cliente se encarga de analizar el equipo en el que es desplegado y utiliza los recursos disponibles con el objetivo de reducir el tiempo de respuesta de la aplicación, es decir, proporcionar al usuario un mejor desempeño al momento de intercambiar datos entre los equipos que comparten catálogos mediante *Painal*.
- Requerimiento no funcional: Se le conoce como requerimiento no funcional a la parte del sistema que describe el *cómo* un software debe de realizar una tarea, por ejemplo, los requisitos de rendimiento del software, los requisitos de la interfaz externa del software, las restricciones de diseño del software y los atributos de calidad del software. Los requisitos no funcionales son difíciles de probar, por lo tanto, generalmente se evalúan de forma subjetiva [6].
- Patrón *peer-to-peer* (P2P): En *Painal*, cuando un nodo puede representar las funciones tanto de Cliente como de Servidor, se le conoce como *peer*. En el patrón P2P, los nodos pueden tanto guardar información que otros nodos le envían, como solicitar información de otros nodos. De esta manera, la información no se encuentra concentrada en un solo punto. Debido a esto, si falla un *peer*, es posible seguir utilizando el patrón de forma normal. Además, la información no se solicita directamente a un mismo *peer*, sino que se encuentra distribuida, por lo que los *peers* no se saturarán de peticiones.
- Tablas hash distribuidas (DHT, por sus siglas en inglés): Son estructuras que permiten el acceso a datos en un ambiente distribuido de manera eficiente. Representan el índice de un sistema de almacenamiento de datos confiable, escalable y de área amplia, que permite a los programadores reducir las complicaciones de construir un sistema distribuido. Las DHT almacenan bloques de datos en cientos o miles de computadoras conectadas a internet, replican dichos datos para mayor confiabilidad y permiten ubicarlos rápidamente a pesar de ejecutar enlaces de área amplia y alta latencia. Las DHT abordan los problemas de localización de datos y fiabilidad, que son comunes en muchos sistemas distribuidos, sin trabajo adicional por parte de la aplicación.

Las DHT proporcionan una interfaz genérica, lo que facilita que una amplia variedad de aplicaciones adopten las DHT para almacenamiento [7].

- Modelo RESTFul: Refiere a un modelo de interacción entre sistemas de manera abierta y escalable, utilizando Interfaces de Programación de Aplicaciones (API, por sus siglas en inglés) que se ajusta a los límites de una arquitectura de ambientes distribuidos conocida como REST [10].

2. Servicio de publicación/suscripción (pub/sub) para el manejo de catálogos, fuentes y repositorios

El servicio de *Painal* permite crear catálogos de datos para el almacenamiento y compartición de *datos* (intra e inter-institucional) de manera segura y transparente para el usuario.[1] Dentro de los catálogos de datos se pueden almacenar datos en crudo (sin procesar, por ejemplo, una tomografía que no está convertida a imagen) o los resultados obtenidos por algún tipo procesamiento (por ejemplo, imágenes procesadas por un algoritmo de inteligencia artificial que permite etiquetar posibles casos de cáncer para asistir al especialista). Cuenta con un mecanismo de publicación/suscripción que permite generar catálogos de datos, agregar nuevos datos a dichos catálogos, transmitirlos a través de la red y descargarlos en otro equipo de manera segura. Todo el proceso de publicación y suscripción se basa en la utilización de tokens de acceso que permiten verificar la identidad de las organizaciones de la federación y los permisos que tienen para acceder a los catálogos.

Painal cuenta con un sistema de distribución de contenidos federado (FedCDS, por sus siglas en inglés) empleado para crear servicios de sincronización de datos médicos que permite a las organizaciones desplegar sus servicios en una nube privada, pública o híbrida que se comunica con el *gestor de Painal* para el manejo de metadata y acceso a los servicios de publicación y suscripción. La Fig. 1 muestra una representación conceptual del FedCDS que permite compartir datos (p. ej., tomografías, mamografías y resonancias magnéticas) de forma sincrónica entre tres hospitales (puede ser cualquier tipo de organización) a través de una red de entrega de contenidos (CDN, por sus siglas en inglés) [15], [11]. La motivación para el uso de una red de servicios federada es que las organizaciones puedan tener un gobierno sobre sus servicios, datos, infraestructura y los requisitos no funcionales (NFR) mediante la inclusión de métodos para cumplirlos. Además, los miembros de la federación pueden generar un conjunto de servicios que permita a los participantes compartir recursos y datos con otras instituciones. Los servicios incluyen la preparación y recuperación de datos [12]. Estos servicios son descritos en *Muyal-Chimalli*, el cual es también conocido como *Zamna* [2]. Su tarea es preprocesar los datos antes de distribuirlos a otros participantes de la

[1] Con transparente nos referimos a que el usuario no ve los procesos que son utilizados para agregar estas características ni interactúa de forma directa con ellos, pero estos siempre se ejecutan

federación o a la nube, aplicando diferentes filtros que cumplan con los requisitos no funcionales requeridos por una organización. Al combinar los catálogos de servicios con la compartición de recursos y datos se obtiene una alta fiabilidad, distribución de la carga, integridad de los datos, confidencialidad de los datos y la independencia del proveedor de servicios de nube.

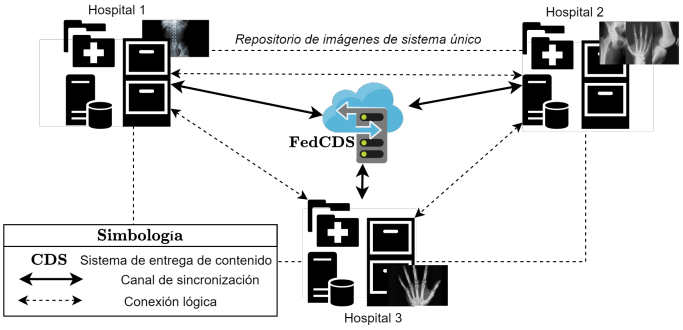


Figura 1: Ejemplo de un sistema de distribución de contenidos federado para tres organizaciones [3].

2.1. Sistema para la distribución de contenidos y administración de requerimientos no funcionales para una organización

El servicio de pub/sub para una organización de *Painal* divide la entrega de contenidos en dos capas: Patrón de Pub/Sub y dispersión de información. La primera capa permite devolver el control de los metadatos al propietario del contenido. Esta capa establece los siguientes roles:

- *Editores* - Son los usuarios responsables de producir nuevos datos o contenido digital;
- *Usuarios finales* - Clientes externos o consumidores que se suscriben a los contenidos con el objetivo de descargarlos en su equipo;
- *Editores/administradores* - Son los usuarios de la organización responsables de aceptar/rechazar tanto las publicaciones como las suscripciones.

La segunda capa se basa en el esquema de dispersión de información descrito en [20], el cual actúa sobre una plataforma de almacenamiento multi-nube con la que el servicio de pub/sub consigue un uso eficiente del espacio de almacenamiento (se reduce la cantidad de datos almacenados) y una alta fiabilidad (los datos pueden ser descargados aun cuando uno o varios nodos de almacenamiento se encuentren fuera de línea). La Fig. 2 (izquierda) presenta el flujo de trabajo de publicación: (1) se lee el contenido a transmitir; (2) la aplicación del *Editor* (o productor de contenido) procesa el contenido utilizando un algoritmo de

dispersión; y (3) el *Editor* transmite un conjunto de archivos llamados dispersos (que son redundantes y anónimos, utilizando procesos descritos en Zamná [2]) a diferentes proveedores de nube (p. ej. Google Drive, One Drive, Amazon S3, etc.). De esta manera, se garantiza que un determinado proveedor no recibe suficientes dispersos para reconstruir el contenido original, garantizando la confidencialidad.

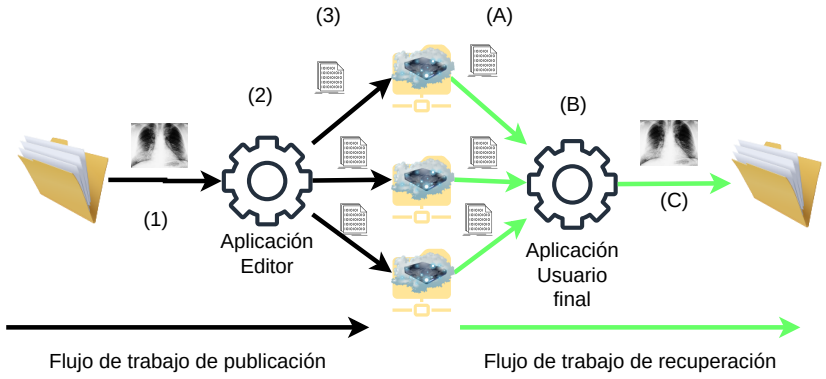


Figura 2: Flujos de trabajo requeridos por los procesos de publicación y recuperación de datos mediante Muyal-Painal.

Por otro lado, la Fig. 2 (derecha) presenta el flujo de trabajo de recuperación: (A) la aplicación *Usuario final* (o consumidor de contenido) recupera un subconjunto de bloques del contenido que se desea recuperar; (B) ejecuta la aplicación que reconstruyen el contenido; y (C) almacena el contenido en la carpeta del usuario final. Debido a la técnica de redundancia aplicada, la recuperación puede obtener contenidos incluso cuando algunas ubicaciones de almacenamiento en la nube no estén disponibles. Como resultado, se reducen los riesgos de la dependencia del proveedor y permite a la organización externalizar el almacenamiento de contenidos de forma controlada a proveedores de servicios de nube públicos.

La Fig. 3 muestra las capas de manejo de metadatos (a través del gestor de *Painal*[2]) y la capa de pub/sub, ejemplificando la entrega de contenidos mediante un esquema de colaboración. Para este ejemplo, un *editor* envía una publicación de contenido ($|C|$) a la capa de flujo de metadatos (Pub). Cuando un *editor/administrador* de la organización autoriza la publicación de este contenido, el *editor* se encarga de dispersarlo a múltiples ubicaciones de almacenamiento en la nube, mediante un flujo de trabajo de entrega (etiqueta Dy) que se ejecuta

[2] El gestor de *Painal* es el componente encargado de controlar los flujos pub/sub en la capa de metadatos y de coordinar el almacenamiento de contenidos en la capa de flujo de contenidos.

en la computadora personal del editor (la aplicación no envía todo el contenido $|C|$ a una única ubicación de almacenamiento en la nube, sino que envía bloques codificados con nombres anónimos a diferentes nubes). El contenido se añade al catálogo y los usuarios finales autorizados ya pueden suscribirse al contenido publicado y dispersado (Sub). Los *usuarios finales* con suscripciones autorizadas se encargan de recuperar los contenidos mediante flujos de trabajo de recuperación ($Retr$). Esta superposición permite al servicio de pub/sub minimizar los riesgos de las situaciones de bloqueo del proveedor y los escenarios en los que hay una falta de control de los procedimientos críticos.

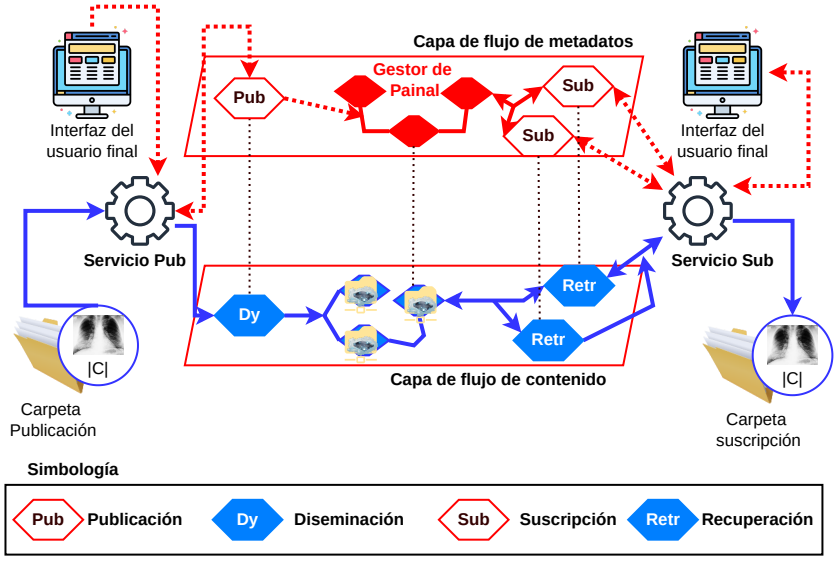


Figura 3: Capa de metadatos y Pub/Sub sobrepuestas.

Las aportaciones del servicio de pub/sub se presentan a continuación.

1. *Entrega colaborativa de contenidos basada en una superposición de pub/sub*: Se desarrolló un modelo $RESTful$ de pub/sub que permite a los usuarios finales y a los editores colaborar en el proceso de entrega de contenidos. Simplifica la gestión de los recursos y el control de las etapas de la entrega de contenidos. Los flujos de trabajo generados se basan en el procesamiento multinúcleo (las aplicaciones se clonan para procesar más de un contenido a la vez) y el transporte de flujo continuo para mitigar la sobrecarga producida en el lado del *usuario final* y del *editor*.
2. *Una estrategia de diversificación basada en la gestión de riesgos para el almacenamiento multi-nube*: Esta estrategia virtualiza las múltiples ubicaciones de almacenamiento disponibles en una organización como una plataforma de almacenamiento multi-nube unificada, que es más sencilla de gestionar

que las ubicaciones separadas. Esta diversificación se basa en un método de colocación de contenidos que permite al servicio pub/sub realizar la asignación y localización de contenidos en esta plataforma para evitar poner todos los huevos/contenidos en la misma cesta/ubicación de almacenamiento en la nube. Este método distribuye los contenidos utilizando una política de evaluación de riesgos, que define el nivel de riesgo para cada cesta/ubicación de la plataforma, así como un conjunto de acciones de respuesta para mitigar un conjunto reducido de riesgos expresados por los editores y las organizaciones sobre la gestión de contenidos en el sistema de entrega de contenido (CDS, por sus siglas en inglés).

3. *Niveles de enmascaramiento de fallos*: Se incluyen tres niveles de tolerancia a fallos del sistema. El primer nivel tolera las fallas del servicio de los usuarios finales y los editores. Este nivel se consigue gracias a las técnicas de dispersión de información aplicadas a los flujos de trabajo de entrega y recuperación (descritos en Zamná [2]). El segundo nivel tolera el desastre en el sitio de la organización desde los usuarios finales. Este nivel se basa en un esquema de federación en el que un conjunto de socios absorbe la carga de la organización durante su interrupción. El último nivel tolera los efectos secundarios de los retrasos de la geodiversidad de todos los usuarios del servicio pub/sub mediante el almacenamiento de los contenidos en caché y su traslado a una ubicación cerca de los usuarios finales.

En *Painal*, los contenidos se asignan y localizan utilizando catálogos; como resultado, solo los contenidos añadidos a un catálogo pueden ser publicados o recuperados. Cuando una organización crea un catálogo, también define los atributos de un conjunto de *editores* que pueden añadir contenidos al catálogo, así como los grupos de *usuarios finales* que pueden suscribirse a los contenidos. La interfaz de usuario final es un componente que permite a los *editores* añadir contenidos fabricados a los catálogos y dispersarlos a la plataforma de almacenamiento multi-nube. También permite a los *usuarios finales* suscribirse y recuperar los contenidos publicados. Los agentes Pub/Sub y el gestor de flujo de trabajo son los componentes que se encargan de recibir y servir las peticiones de las aplicaciones cliente.

Componentes del servicio de pub/sub: Agentes y clientes. El gestor de *Painal* para el servicio de pub/sub se compone de un gestor de metadatos y otro de contenidos. El gestor de metadatos se encarga de administrar los flujos entre los principales componentes del servicio de pub/sub de *Painal* y los clientes/agentes, mientras que el gestor de recursos que se encarga de la asignación/ubicación de contenidos y la gestión de recursos en las infraestructuras de la federación.

A continuación se listan los componentes desarrollados para *Painal*.

1. *Aplicación cliente* - Los servicios de *Painal* pueden ser invocados a través de una aplicación que contiene diferentes módulos, en donde cada módulo corresponde a un rol (administrador, editor o usuario final). Como un usuario

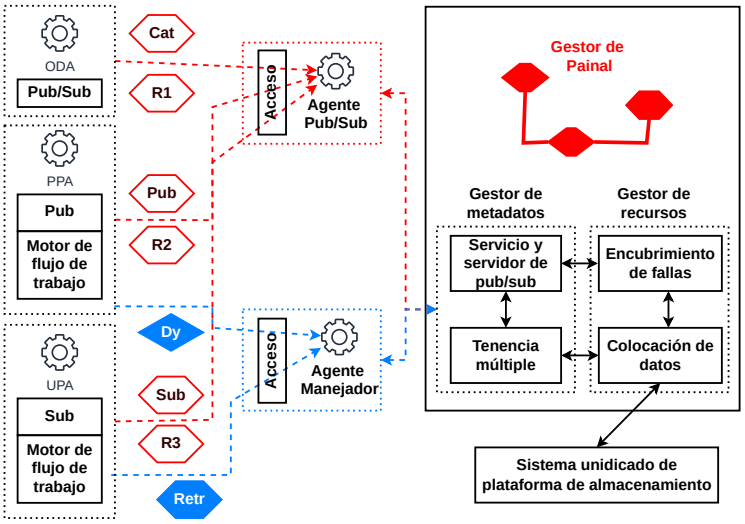


Figura 4: Los componentes y subcomponentes del servicio de Pub/Sub de *Painal*.

puede tener más de un rol, la aplicación de *Painal* habilita los módulos de acuerdo al usuario. A continuación se describen de manera breve los módulos desarrollados:

- *ODA* - Es una aplicación para la configuración de los catálogos de una organización que permite a los usuarios *administradores* crear catálogos, asignarlos a editores y definir las políticas de suscripción que deben cumplir los usuarios finales para acceder a los datos;
- *PPA* - Aplicación que permite a los *editores* obtener el listado de catálogos a los que tiene acceso y permisos de publicación; por cada catálogo encontrado se crea una carpeta y en estas carpetas el editor puede colocar los datos que desea transferir;
- *UPA* - Es la aplicación que usan los *usuarios finales* para suscribirse a los catálogos públicos o acceder a los que el *administrador* de la organización les dio acceso; esta aplicación descarga el contenido de los catálogos (asigna una carpeta por catálogo) y los sincroniza de acuerdo con el contenido que los *editores* coloquen en sus catálogos.

2. *Gestor de metadatos* - Es el responsable de la anonimización de los datos, así como del manejo de la metadata de todos los archivos, procesos, dispositivos de almacenamiento, catálogos y usuarios de *Painal*. Este gestor se compone de los siguientes módulos:
   - *Subsistema de servicios* - Este componente se encarga de recibir las solicitudes de tokens[3] de autorización, así como de recibir las publicaciones y

[3] Token (informática), también llamado componente léxico, es una cadena de caracteres que tiene un significado coherente en cierto lenguaje de programación. Token de seguridad, utilizado para facilitar el proceso de autenticación de usuarios.

los pedidos de suscripción procedentes de los clientes/agentes del servicio de pub/sub de *Painal*; este subsistema es responsable de registrar las transacciones realizadas por los editores, organizaciones y usuarios finales (ver la Fig. 4);

- *Subsistema de publicación y suscripción* - Este módulo se encarga de ejecutar las órdenes de publicación y suscripción autorizadas por el subsistema de servicios (ver la Fig. 4, en donde se utiliza una política basada en atributos de rol); este módulo permite al servicio de pub/sub de *Painal* aceptar o rechazar ordenes de publicación y suscripción; también, incluye un servicio de alerta que envía notificaciones a los editores de contenidos y a los administradores de las organizaciones sobre las suscripciones de contenidos, el cual también notifica al usuario final sobre nuevas publicaciones;
- *Subsistema de tenencia múltiple (multi-tenant)* - Este módulo gestiona las propiedades de los catálogos y contenidos, así como las cuentas de los usuarios, mediante consultas a una base de datos; asegura que los contenidos de cualquier editor, usuario final o administrador están aislados y permanecen invisibles para otros usuarios; y envía las solicitudes de asignación o localización al gestor de recursos para un determinado contenido asociado a un determinado catálogo.

3. *Gestor de recursos* - Es un componente crítico que se instala en una instancia de la nube situada en una infraestructura de nube privada. La Fig. 4 muestra un ejemplo de los flujos de metadatos entre los agentes, el servicio, los subsistemas Pub/Sub y Multi-tenant. También, muestra cómo el gestor de recursos atiende las solicitudes de asignación/ubicación enviadas por los agentes y autoriza a los gestores de flujo de trabajo a transportar datos a un sistema unificado (ULS). Detallaremos el gestor de recursos una vez que el gestor de flujo de trabajo sea descrito en la siguiente sección.
4. *Gestor de flujos de trabajo* - Es el responsable de administrar los flujos de trabajo de entrega para transportar los contenidos desde los ordenadores de los editores a la plataforma de almacenamiento multi-cloud. Adicionalmente, permite administrar los flujos de trabajo de recuperación para transportar los contenidos desde la plataforma de almacenamiento a los ordenadores de los usuarios finales. Los flujos de trabajo se basan en tuberías de procesamiento que incluyen dos fases básicas. La primera fase es la codificación/decodificación de los contenidos, que se basa en algoritmos de información dispersa (IDA) [20] y la segunda etapa es la distribución de los contenidos codificados/codificados, que se basa en el streaming[4] continuo y paralelo [15]. Este gestor de flujos de trabajo es descrito de manera más profunda en la documentación de *Muyal-Zamna* [2].

[4] El *streaming* es un servicio que permite a los usuarios consumir un contenido en línea sin tener que esperar a que se descargue.

Flujos de trabajo de Painal. Como se ha descrito a lo largo de este documento, cada archivo publicado por un editor es procesado y transmitido a diferentes nodos de almacenamiento (publicación) para después ser descargado y reconstruido por un usuario final (recuperación). A este comportamiento se le conoce como flujo de trabajo y, en el caso de *Painal*, se pueden identificar uno para la entrega y otro para la descarga. A continuación se describen de manera breve los dos flujos de trabajo.

- *Flujo de trabajo de publicación* - Es el encargado de leer los contenidos que un *editor* desea transmitir mediante un catálogo. Es responsable proporcionar los requisitos no funcionales: reducir el contenido a procesar/almacenar (mediante compresión); agregar control de acceso (mediante la validación de atributos y tokens); brindar confidencialidad a los datos (mediante un algoritmo de cifrado); proporcionar tolerancia a fallos (agregando redundancia a los datos); y ejecutar las aplicaciones de manera eficiente (mediante técnicas de paralelismo). Este flujo de trabajo incluye dos flujos de metadatos (publicación y gestión) y un flujo de contenidos (canal de dispersión).

  La Fig. 5 presenta el flujo de trabajo de publicación de contenidos digitales, en donde se observan las etapas y procesos realizados. (1) La aplicación del *editor* detecta un nuevo contenido; (2) se verifican las claves de acceso del *editor* con el gestor de *Painal* concediendo o negando el acceso; (3) se invocan las aplicaciones que se encargarán de cumplir con los requerimiento no funcionales (NFR), para este sistema son los procesos de compresión, cifrado y dispersión; (4) los n archivos dispersos resultantes de la aplicación de dispersión deben de ser transferidos por la aplicación del *editor*, la cual se comunica con el gestor de *Painal* para obtener las n URLs relativas que se mapean a n ubicaciones de almacenamiento en la nube (en dónde los dispersos serán almacenados); y (5) la aplicación del editor inicia la transferencia de los datos utilizando las URLs obtenidas.
- *Flujo de trabajo de recuperación* - Está diseñado para que los usuarios finales recuperen contenidos de las ubicaciones de almacenamiento en la nube (mediante un procedimiento de decodificación multi-hilo que es ejecutado de manera transparente para el usuario). En la Fig. 6 se muestran los pasos realizados por este flujo: (1) la aplicación del *usuario final* se identifica con el gestor de *Painal* y, en caso de detectar nuevos datos (agregados por un *editor*) en los catálogos, invoca al siguiente paso; (2) se obtienen las URLs de los dispersos necesarios para reconstruir el contenido ($|C|$), como se aplicó un algoritmo de dispersión con redundancia solo se requiere una m cantidad de los n dispersos almacenados (esta configuración usa $n = 5$ y $m = 3$); (3) se inicia el proceso de descarga de los m dispersos; (4) se ejecutan las aplicaciones que harán cumplir los requerimientos no funcionales en el orden inverso del utilizado en el proceso de carga (dispersión, cifrado y compresión); y (5) el contenido recuperado es colocado en la carpeta seleccionada por el usuario final.

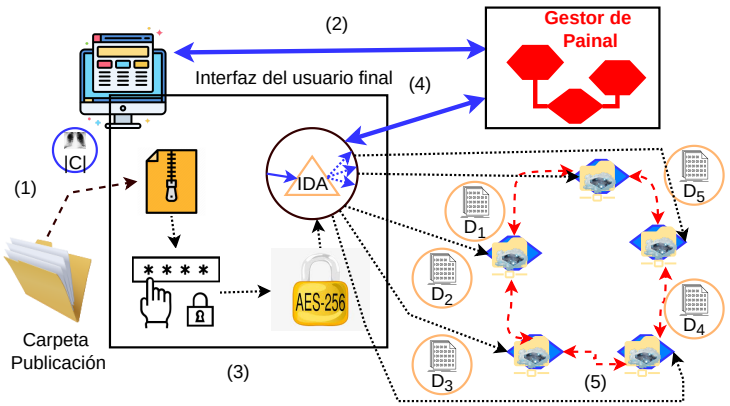


Figura 5: Flujo de trabajo de publicación.

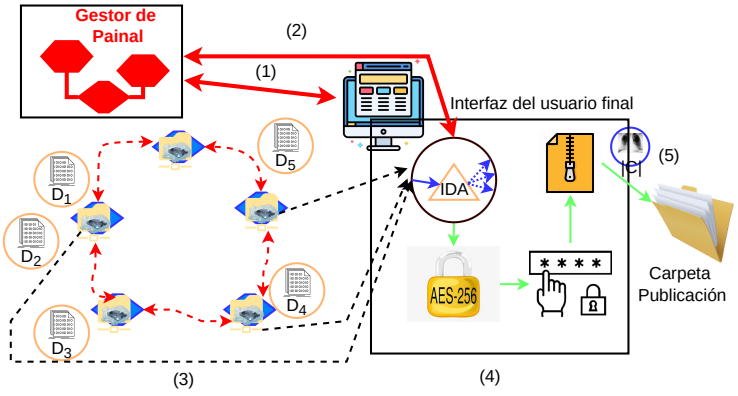


Figura 6: Flujo de trabajo de recuperación.

La gestión de los procesos de los requerimientos no funcionales para los flujos de trabajo se basa en la utilización de patrones de procesamiento, y la etapa de transporte se implementó utilizando las bibliotecas Curl [17], las cuales permiten mover contenidos utilizando el protocolo de la web (HTTP). La implementación multi-hilo de nuestros gestores de dispersión/recuperación mejora el rendimiento de las tareas de codificación y decodificación al aprovechar los múltiples núcleos que se encuentran habitualmente en los dispositivos actuales. Esta técnica permite que el gestor reduzca la sobrecarga de codificación, lo que hace factible la introducción de un esquema tolerante a fallos en el lado del editor de contenidos/usuario final. La implementación del proceso de entrega de contenidos como flujo continuo permite al gestor evitar la escritura de bloques en los discos locales.

2.2. Sistema federado para la distribución de contenidos y administración de requerimientos no funcionales

En esta sección se describen los principios de diseño de FedCDS, la arquitectura y los principales componentes del sistema, así como los esquemas basados en patrones de paralelismo mediante contenedores para preparar/recuperar los datos cuando se cargan/descargan a través del FedCDS y el gestor de *Painal*. Estos esquemas gestionan los requisitos no funcionales seleccionados por las organizaciones para cumplir con las regulaciones impuestas por las organizaciones y los gobiernos.

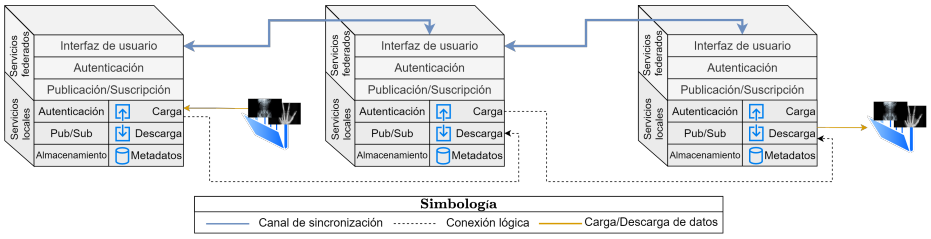


Figura 7: Arquitectura detallada del sistema federado [3] para la distribución de contenidos con Muyal-Painal.

Arquitectura del sistema. La Fig. 7 muestra la arquitectura detallada y los principales componentes del FedCDS de *Painal* para un escenario compuesto por tres hospitales diferentes (organizaciones), lo cual permite observar la comunicación entre los servicios apilados de cada organización. Cada hospital (organización) despliega dos tipos de servicios: servicios federados y servicios locales. El objetivo de distinguir estos dos tipos de servicios es que cada organización tiene el control de sus datos y recursos, pero, al mismo tiempo, algunos servicios permiten a las organizaciones compartir recursos y datos de forma transparente y segura. A través de estos servicios, las organizaciones tienen acceso a los

recursos y datos disponibles en otras organizaciones (imágenes médicas o historias clínicas). Para ello, los servicios federados para cada organización son los siguientes:

- Interfaz de acceso (*Front-End*) - Capa a través de la cual los administradores, editores y usuarios finales de una organización y participantes de la federación pueden acceder a los servicios; dispone de una interfaz de programación de aplicaciones (API) que permite la interacción con otros participantes, así como la validación de tokens y un proxy[5] para redirigir las peticiones a los servicios de los niveles inferiores o disponibles en otras organizaciones de la federación;
- *Autenticación* - Invoca al gestor de *Painal* para validar el acceso de los usuarios finales, que son miembros de la federación, o una organización específica;
- *Pub/Sub* - Servicio encargado de gestionar las solicitudes de publicación y suscripción de datos (imágenes o historias clínicas o historias clínicas), catálogos, fuentes y repositorios de los usuarios pertenecientes a la federación.

Los servicios locales se encargan de gestionar el sistema dentro de cada organización, incluida la gestión de los usuarios internos, así como los datos que se manejan en la organización. La idea de estos servicios es que cada organización pueda gestionar sus catálogos de datos de forma independiente a los publicados en la federación. En el caso de que los usuarios quieran publicar sus datos en la federación, podrán hacerlo con solo cambiar el estado del catálogo de "local/privado" a "federado" o, igualmente, dando acceso solo a un determinado grupo de usuarios de la federación. Dentro de los servicios locales se incluyen tres capas para gestionar la autenticación de usuarios y tokens de la organización, gestionar las publicaciones y suscripciones y gestionar el almacenamiento, así como los metadatos asociados a estos servicios.

El intercambio de mensajes y metadatos entre los miembros de la federación se efectúa a través de API REST, mientras que los datos se transportan utilizando una red de entrega de contenidos del servicio de pub/sub de *Painal*. Los esquemas de preparación de datos [12] se despliegan para añadir los NFR a los datos antes de ser transportados. Estos esquemas procesan los datos para añadir propiedades como la rentabilidad, la fiabilidad y la seguridad para ser compartidos con los miembros de la federación.

3. Mecanismo de usabilidad costo-beneficio para el almacenamiento y transporte de datos

La arquitectura de malla para el almacenamiento de datos se basa en una estructura de almacenamiento que se asocia para gestionar nodos de almacenamiento (*SN*, por sus siglas en inglés). Los *SN* son sistemas de almacenamiento

[5] Es un servidor o programa informático que sirve de intermediario en las peticiones de recursos que realiza el usuario final y el servidor fuente, almacenando una copia caché de los mismos para acelerar su suministro.

tradicionales que incluyen sistemas como la entrega/recuperación de archivos, balanceadores de carga, distribución de datos y los sistemas de colocación, así como los microservicios para atender los requisitos no funcionales (*NFR*). Los componentes de los *SNs* se encapsulan en contenedores virtuales, brindándoles la característica de portabilidad. En los servicios de almacenamiento confiables sin servidor (SeRSS) de *Painal*, estos componentes se gestionan utilizando una estructura de almacenamiento, que convierte las *SNs* en microservicios[6] en forma de redes de almacenamiento descentralizadas P2P. En los microservicios de fiabilidad SeRSS, *Painal* implementa una técnica de codificación/decodificación de datos basada en el algoritmo IDA [22], [20], descrita a mayor profundidad en [2]. Los microservicios de seguridad implementan servicios de criptografía para garantizar cualquier integridad, confidencialidad o autenticación, mientras que los microservicios de patrones paralelos mejoran el rendimiento de los microservicios descritos anteriormente.

La idea básica es que las organizaciones puedan elegir los parámetros de almacenamiento a alto nivel (es decir, el número de nodos de almacenamiento necesarios para una solución determinada) y los parámetros *NFR*. De esta manera, se puede utilizar esta información para construir de manera automática un sistema de almacenamiento fiable en forma de sistema P2P sin servidor que puede ser consumido por los usuarios finales como un servicio que puede ser desplegado en infraestructuras heterogéneas.

3.1. Servicios de almacenamiento confiables sin servidor: Principios de diseño

En esta sección se describen los componentes de una estructura de almacenamiento que se asocian a los *SN*, así como un conjunto de microservicios que proporcionan características no funcionales como la fiabilidad seguridad y eficiencia. Por último, se presentan los componentes para construir sistemas de almacenamiento en forma de un sistema P2P.

Construcción de estructuras de almacenamiento confiables. Las estructuras de almacenamiento gestionadas en los servicios SeRSS de *Painal* permiten a las organizaciones almacenar y gestionar los datos de manera eficiente y confiable. Cada estructura de almacenamiento se despliega como un contenedor virtual. Esto añade la característica de portabilidad a las estructuras construidas, permitiendo a las organizaciones mover la estructura de almacenamiento a través de diferentes infraestructuras. La Fig. 8 muestra la representación en pila de una estructura de almacenamiento. Estas estructuras implementan dos componentes principales para gestionar los datos dentro de una organización: i) un sistema de almacenamiento; y ii) un esquema de entrega.

El sistema de almacenamiento se encarga de gestionar la colocación de los datos a través de los recursos de almacenamiento de una organización. En este

[6] Piezas de software independientes y portátiles que pueden acoplarse a otros microservicios.

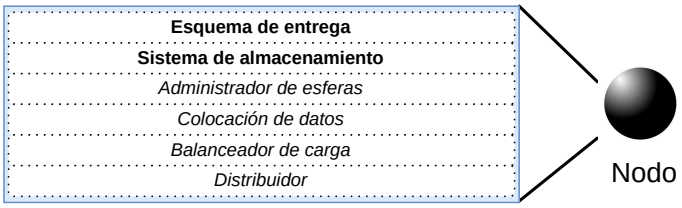


Figura 8: Representación conceptual de una estructura de almacenamiento.

contexto, el método de colocación de datos que se ha implementado se basa en la metáfora de las esferas en los contenedores (*balls-into-bins*) [22], donde los datos se gestionan como esferas que se asignan a un conjunto de contenedores (recursos de almacenamiento). Para lograr este objetivo, el sistema de almacenamiento implementa un gestor de esferas encargado de gestionar los metadatos de los datos que llegan a la estructura de almacenamiento. Los datos se entregan a un sistema de colocación de datos, que llama a un servicio de balanceo de carga basado en el algoritmo de dos opciones (*two choices*) [1], [22] para crear una distribución justa de los segmentos entre las ubicaciones de almacenamiento. Un distribuidor se encarga de asignar los segmentos producidos en las ubicaciones de almacenamiento.

El esquema de distribución permite el intercambio de datos entre diferentes estructuras de almacenamiento desplegadas en diferentes entornos. Estos esquemas imponen el cumplimiento de los NFR (por ejemplo, la disponibilidad y la eficiencia) en la gestión de datos impuestos por los gobiernos y las organizaciones para la gestión de datos sensibles (i.e., NIST [16] , COBIT5 [9], ISO27001-7 [5], y normas mexicanas de gestión de datos [18], [25]). Los esquemas de entrega se manejan como microservicios de NFR basados en el algoritmo de IDA para la *tolerancia a fallos*, esquemas de preparación para resolver problemas de *seguridad* (por ejemplo, la verificación de la integridad usando el algoritmo SHA3) y patrones paralelos para mejorar la *eficiencia* de los componentes de las estructuras de almacenamiento. El microservicio de *confiabilidad* codifica los datos utilizando el conocido algoritmo IDA descrito en [22], [20]. Además, se aplican microservicios de seguridad a los datos para verificar su integridad, mecanismos de control de acceso para permitir que solo los usuarios autorizados accedan a los datos (por ejemplo, utilizando técnicas de sobres digitales) y confidencialidad mediante el cifrado de los datos. Los componentes de la estructura de almacenamiento, incluidos los microservicios de NFR, pueden desplegarse como patrones paralelos para mejorar la *eficiencia* en el procesamiento de los datos. Estos patrones paralelos se implementan como patrones de contenedores virtuales y pueden configurarse en cada esquema de entrega. Estos patrones son descritos a profundidad en el servicio de la plataforma Zamna, descrita en [2].

La creación de sistemas de almacenamiento P2P sin servidor en una malla para el almacenamiento de datos permite a las organizaciones la posibilidad de construir soluciones de almacenamiento basadas en sus recursos disponibles en

cualquiera de los recursos locales, comunitarios o la nube. Además, las organizaciones pueden elegir los parámetros a alto nivel (es decir, el número de nodos de almacenamiento necesarios para una solución determinada) y el servicio de SeRSS de *Painal* construye automáticamente el sistema de almacenamiento P2P sin servidor. En este contexto, la infraestructura de la organización se gestiona como una malla de recursos. La Fig. 9 muestra una representación conceptual de las soluciones de almacenamiento desplegadas con esta arquitectura de malla. Como se puede observar, sobre la malla se despliegan diferentes soluciones de almacenamiento. Las soluciones de almacenamiento se construyeron desplegando un conjunto de estructuras de almacenamiento (SS, por sus siglas en inglés) como P2P. Estas soluciones de almacenamiento utilizan tablas hash distribuidas (DHT). Como se puede observar en la Fig. 9, se pueden desplegar diferentes soluciones sobre la misma malla de recursos. Estas soluciones pueden diferir entre sí en cuanto al número de nodos de la estructura, los requisitos no funcionales gestionados por las estructuras de almacenamiento y los recursos utilizados.

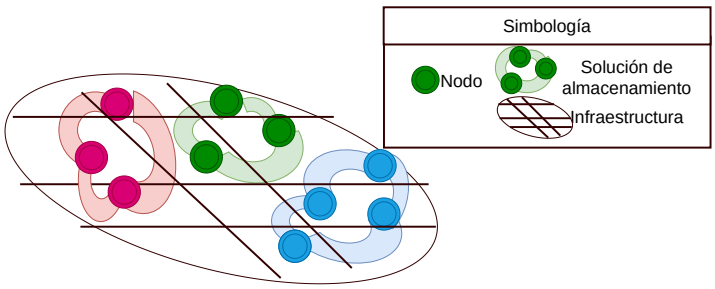


Figura 9: Representación conceptual de diferentes mallas de almacenamiento utilizadas para gestionar datos de manera confiable.

La malla de recursos se construye partiendo los nodos de almacenamiento físico (ps, por sus siglas en inglés) en un conjunto de particiones virtuales (vp, por sus siglas en inglés). Estas particiones virtuales son básicamente divisiones lógicas de los nodos de almacenamiento con una porción limitada de los recursos (memoria, CPU y capacidad de almacenamiento) de la infraestructura física (ps). En esta arquitectura, la solución se despliega en una malla de nodos de almacenamiento (M). Esta M está compuesta por un conjunto de nodos de almacenamiento físico ps (eje horizontal x de la malla), donde cada ps está dividido en y particiones denominadas almacenamiento virtual vp (eje vertical y de la malla) y se define como la malla presentada en la Fig. 10.

En donde q es el número de particiones virtuales en los nodos de almacenamiento y p es el número de nodos de almacenamiento físicos. Así, el número de recursos disponibles en la malla (M) es $p \times q$. Cada celda de la malla es un recurso de almacenamiento disponible, donde un SN virtual puede desplegarse

$$M = \begin{bmatrix} ps_1vp_1 & \dots & ps_pvp_1 \\ \vdots & \ddots & \vdots \\ ps_1vp_q & \dots & ps_pvp_q \end{bmatrix}$$

Figura 10: Ejemplo de malla de almacenamiento.

como un contenedor virtual que contiene una estructura de almacenamiento. Un SN en la malla puede adoptar cualquiera de los siguientes roles:

- *Rol de servicio (ser)* - para atender las solicitudes realizadas por usuarios finales;
- *Rol de gestor (mgr)* - se encarga de ejecutar las tareas de almacenamiento (asignación, localización y equilibrio de carga) y de añadir los requisitos no funcionales a los datos;
- *Rol de nodo (node)* - Se encarga de preservar datos.

Tenga en cuenta que al menos un servicio *ser* y un gestor *mgr* deben estar asignados en la malla. La selección de los nodos de servicio y gestor son realizados durante el diseño y construcción.

Descripción de los componentes utilizados para el desarrollo del prototipo. La arquitectura de malla y los componentes de la estructura de almacenamiento se implementaron como un prototipo desarrollado principalmente en lenguaje de programación C. El componente para la dispersión de datos (IDA) utilizado en la estructura de almacenamiento también se ha implementado utilizando dicho lenguaje. El intercambio de datos entre los nodos de almacenamiento virtual se realiza a través de una red de entrega de contenidos del servicio de publicación/suscripción de *Painal*. En la malla, los nodos de almacenamiento virtual de una solución de almacenamiento se gestionan como una red P2P mediante un algoritmo de tabla hash distribuida (DHT), llamado Chord, implementado en lenguaje Python. Los nodos de almacenamiento virtual se despliegan como contenedores virtuales utilizando la plataforma de contenedores Docker. Estos contenedores virtuales añaden portabilidad a los nodos, ya que, en los componentes del almacenamiento, la estructura se encapsula en el contenedor junto con sus dependencias (es decir, bibliotecas, paquetes, SO, variables de entorno), lo que permite el despliegue de los nodos en diferentes infraestructuras.

4. Conclusiones

El presente capítulo describió la herramienta *"Muyal-Painal: Servicio para el transporte y almacenamiento de datos médicos"*, que forma parte del Proyecto ProNacEs-Pronaii número 41756 titulado "Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos

en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud". *Painal* permite a los profesionales de la salud compartir datos, información y sistemas de e-salud en forma inter-institucional (local) e intra-institucional (un conjunto federado de organizaciones). Permite crear, de forma automática y sin intervención de personal de la salud, sistemas de logística para el almacenamiento y distribución de datos de salud, imagenología y datos de sensores y áreas de intercambio federadas intra/inter-institucionales. Las evaluaciones realizadas y publicadas en [3], [4] muestran que *Painal* permite la sincronización automática de datos, reduciendo los tiempos de respuesta, así como los costos de envió y almacenamiento de información, mejorando la experiencia de servicio para los usuarios finales.

5. Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto No. 41756 titulado "Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud" del fondo PRONACES-CONACYT.

Bibliografía

[1] Roberto Beraldi, Hussein Alnuweiri y Abderrahmen Mtibaa. "A power-of-two choices based algorithm for fog computing". En: *IEEE Transactions on Cloud Computing* 8.3 (2018), págs. 698-709.

[2] Diana Carrizales-Espinoza, JL Gonzalez-Compean y Miguel Morales-Sandoval. "Zamna: a tool for the secure and reliable storage, sharing, and usage of large data sets in data science applications". En: *2022 IEEE Mexican International Conference on Computer Science (ENC)*. IEEE. 2022, págs. 1-8.

[3] Diana Carrizales-Espinoza et al. "A Federated Content Distribution System to Build Health Data Synchronization Services". En: *2021 29th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*. 2021, págs. 1-8. DOI: 10.1109/PDP52278.2021.00011.

[4] Diana Carrizales-Espinoza et al. "SeRSS: a storage mesh architecture to build serverless reliable storage services". En: *2022 30th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*. 2022, págs. 88-91. DOI: 10.1109/PDP55904.2022.00022.

[5] Sanskriti Choubey y Astitwa Bhargava. "Significance of ISO/IEC 27001 in the implementation of governance, risk and compliance". En: *International Journal of Scientific Research in Network Security and Communication* 6.2 (2018), págs. 30-33.

[6] Lawrence Chung et al. *Non-functional requirements in software engineering*. Vol. 5. Springer Science & Business Media, 2012.

[7] Frank Frank Edward Dabek. "A distributed hash table". Tesis doct. Massachusetts Institute of Technology, 2005.

[8] Chun-Ping Deng et al. "Organizational agility through outsourcing: Roles of IT alignment, cloud computing and knowledge transfer". En: *International Journal of Information Management* 60 (2021), pág. 102385.

[9] Yusuf Durachman et al. "IT security governance evaluation with use of COBIT 5 framework: A case study on UIN Syarif Hidayatullah library information system". En: *2017 5th International Conference on Cyber and IT Service Management (CITSM)*. IEEE. 2017, págs. 1-5.

[10] Roy T. Fielding et al. "Reflections on the REST architectural style and "principled design of the modern web architecture"(impact paper award)". En: *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2017, Paderborn, Germany, September 4-8, 2017*. Ed. por Eric Bodden et al. ACM, 2017, págs. 4-14. DOI: 10.1145/3106237.3121282.

[11] José Luis González et al. "SkyCDS: A resilient content delivery service based on diversified cloud storage". En: *Simulation Modelling Practice and Theory* 54 (2015), págs. 64-85.

[12] JL Gonzalez-Compean et al. "Fedids: a federated cloud storage architecture and satellite image delivery service for building dependable geospatial platforms". En: *International journal of digital earth* 11.7 (2018), págs. 730-751.

[13] Haryadi S Gunawi et al. "Why does the cloud stop computing? lessons from hundreds of service outages". En: *Proceedings of the Seventh ACM Symposium on Cloud Computing*. 2016, págs. 1-16.

[14] B Hayes. *Cloud Computing (CC) Communications of the ACM, 51 (7)*. 2008.

[15] Daniel Higuero et al. "HIDDRA: a highly independent data distribution and retrieval architecture for space observation missions". En: *Astrophysics and Space Science* 321.3 (2009), págs. 169-175.

[16] Ahmed Ibrahim et al. "A security review of local government using NIST CSF: a case study". En: *The Journal of Supercomputing* 74.10 (2018), págs. 5171-5186.

[17] The multiprotocol file transfer library. *libcurl - the multiprotocol file transfer library*. Available at: http://curl.haxx.se/libcurl, Last accessed: 2022-10-27. Sep. de 2022.

[18] SUBSECRETARÍA DE INTEGRACIÓN DEL SECTOR SALUD MAKI ESTHER ORTIZ DOMINGUEZ. "NORMA Oficial Mexicana NOM-024-SSA3-2010, Que establece los objetivos funcionales y funcionalidades que deberán observar los productos de S istemas de Expediente C línico Electrónico para garantizar la interoperabilidad, procesamiento, interpretación, confidencialidad, seguridad y uso de estándares y catálogos de la información de los registros electrónicos en salud. Al margen un sello con el Escudo Nacional, que dice: Estados Unidos Mexicanos.-Secretaría de Salud." En: ().

[19] M Malik. "Internet of Things (IoT) Healthcare Market by Component (Implantable Sensor Devices, Wearable Sensor Devices, System and Software), Application (Patient Monitoring, Clinical Operation and Workflow Optimization, Clinical Imaging, Fitness and Wellness Measur". En: *Allied Market Research* (2016).

[20] Ricardo Marcelín-Jiménez et al. "On the complexity and performance of the information dispersal algorithm". En: *IEEE Access* 8 (2020), págs. 159284-159290.

[21] Peter Mell, Tim Grance et al. "The NIST definition of cloud computing". En: (2011).

[22] Pablo Morales-Ferreira et al. "A data distribution service for cloud and containerized storage based on information dispersal". En: *2018 IEEE Symposium on Service-Oriented System Engineering (SOSE)*. IEEE. 2018, págs. 86-95.

[23] Justice Opara-Martins, Reza Sahandi y Feng Tian. "Critical review of vendor lock-in and its impact on adoption of cloud computing". En: *International Conference on Information Society (i-Society 2014)*. IEEE. 2014, págs. 92-97.

[24] David Reinsel-John Gantz-John Rydning, J Reinsel y J Gantz. "The digitization of the world from edge to core". En: *Framingham: International Data Corporation* 16 (2018).

[25] SECTOR SALUD, SUBSECRETARIA DE PREVENCION Y PROMOCION DE y CONSEJO DE SALUBRIDAD GENERAL. "NORMA OFICIAL MEXICANA NOM-024-SSA3-2012, SISTEMAS DE INFORMACION DE REGISTRO ELECTRONICO PARA LA SALUD. INTERCAMBIO DE INFORMACION EN SALUD CONSIDERANDOS". En: ().

[26] Dante D Sánchez-Gallegos et al. "From the edge to the cloud: A continuous delivery and preparation model for processing big IoT data". En: *Simulation Modelling Practice and Theory* 105 (2020), pág. 102136.

Metodología para el Desarrollo de un Chatbot para Detección de Depresión Utilizando Inteligencia Artificial

José Miguel Morales Salazar, María del Carmen Santiago Díaz, Ana Claudia Zenteno Vázquez, Yeiny Romero Hernández, Judith Pérez Marcial, Gustavo Trinidad Rubín Linares

Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, Av. San Claudio y 14 sur, Col. San Manuel, Puebla México

miguel.perezxi@alumno.buap.mx, {marycarmen.santiago, ana.zenteno, judith.perez, yeiny.romero, gustavo.rubin}@correo.buap.mx

Resumen. Debido a la pandemia SARS-COV2, la Organización Mundial de la Salud reportó una paralización en los servicios de salud mental en un 93%. La depresión es uno de los trastornos mentales más peligrosos, ya que puede dejar a las personas discapacitadas o, incluso, llevarlas al suicidio. Es por ello que se propone desarrollar un asistente virtual de primera atención a la salud emocional que orientará una conversación de estrecha interacción, aplicando el inventario de depresión de Beck. Este sistema identifica y califica el nivel de depresión en el usuario, incorporando reconocimiento de voz con el modelo mixto Markov Oculto - Redes Neuronales para tomar los datos de entrada del usuario, un modelo de procesamiento de lenguaje natural para dar contexto a los datos de entrada del sistema y un diagrama de árbol de decisión para orientar la interacción con el usuario, dando como resultado la integración de estos subsistemas en una plataforma con un diseño amigable que favorece la usabilidad y la interacción voz a voz con los usuarios.

Palabras clave: Depresión · Asistente virtual · Reconocimiento de voz · Modelo Oculto de Markov · Redes Neuronales.

1 La Depresión

Debido a la contingencia sanitaria ocasionada por la pandemia de SARS-CoV-2, la depresión se ha vuelto un término muy utilizado. La Organización Mundial de la Salud (OMS) reporta, de un estudio considerando 130 países, que la pandemia ha paralizado los servicios de salud mental en un 93%. Antes de la pandemia, a nivel Nacional la inversión realizada en salud destinaba menos del 2% para la salud emocional [1].

La depresión es un trastorno mental muy frecuente, en la que destaca la falta de energía, pérdida de interés, alteraciones de sueño y de humor, no poder disfrutar u obtener placer, que a la larga puede provocar incapacidad, igual que una enfermedad crónica [3][4]. Tener un software a la mano que diagnostique el nivel de este problema y dé un resultado preliminar será de mucha ayuda, ya que, al identificarlo, se podrá actuar de manera oportuna y, así, esta enfermedad no tenga consecuencias graves.

1.1 Detección del trastorno depresivo

Un análisis de investigación no sistemático sobre los instrumentos de tamizaje menciona que entre los mejores resultados a nivel internacional están la escala de depresión del centro de estudios epidemiológicos (CES-D), con una validez del 0.001 y una confiabilidad del 0.90, el Inventario de depresión de Beck (BDI), con una validez del 0.9 y una confiabilidad del 0.69 y el cuestionario sobre la salud del paciente (PHQ-9), con una validez del 0.75 y una confiabilidad del 0.83 [5].

Entre las distintas pruebas destaca el inventario de depresión de Beck, específicamente en su versión BDI-IA en español, donde se ha concluido que tiene una buena capacidad discriminatoria, sensible y específica, lo cual nos sirve para apoyarnos en la probabilidad de cuando encuentre la presencia del trastorno depresivo o cuando no se encuentre un rastro de ella.

1.2 Cómo se enfrenta la tecnología a este problema

La mayoría de estas tecnologías no se enfocan específicamente en el trastorno depresivo, sino en brindar una ayuda emocional más general. Tal es el caso de What's Up?, el cual utiliza métodos de terapia cognitiva-conductual (CBT) y terapia de aceptación y compromiso (ACT) para ayudar a sobrellevar la depresión, la ansiedad, la ira, el estrés y otras [7]. Otra alternativa es Woebot, que consiste en un entrenador personal para los altibajos de la vida [8] y, finalmente, Replika: My AI Friend, que es un chatbot con inteligencia artificial con la misión de levantar el ánimo. Esta aplicación, de acuerdo con sus fabricantes, puede ser tu mentor, amigo o incluso tu pareja [9]. Es importante mencionar que ELIZA [10] es considerada el primer chatbot de la historia y, aunque no fue programada para brindar ayuda emocional, algunos usuarios revelaban sus problemas psicológicos [11]. Con estos antecedentes, queda clara la necesidad de un sistema tipo chatbot para brindar ayuda emocional, enfocado en la detección del trastorno depresivo.

2 Psicólogo Chatbot

Oracle define al chatbot como un simulador de conversaciones humanas, escritas o habladas que tienen como objetivo ser un puente de comunicación entre el humano y un sistema, de manera que el humano sienta que interactúa con otro ser humano [12].

Para empezar, se necesita tener una estructura muy bien definida para el sistema de chatbot, en la cual tendremos las intenciones (las demandas o

requerimientos que pide el usuario), las entidades (distinciones relativas de una intención) y el diálogo (el conjunto de preguntas o frases para interactuar con el usuario) [11].

2.1 Reconocimiento de voz

Un reconocimiento de voz es un sistema que toma como entrada las señales acústicas del ambiente, procesa la información obtenida y decodifica los datos a un lenguaje entendible para el sistema [14]. A continuación, se presentan las etapas de análisis espectral, conversión (sonido a texto), modelo mixto de Markov-Redes Neuronales Artificiales (HMM-ANN), Procesamiento del Lenguaje Natural (PNL) y árboles de decisión, etapas por las cuales pasa un sistema de reconocimiento de voz de tipo chatbot.

2.2.1 Análisis Espectral

Analizando las características de la voz con la que interactúa el usuario con el chatbot, podemos obtener información del contexto en el que está respondiendo el usuario. Características como el tono, intensidad, duración o la entonación de la voz son de gran ayuda para los expertos en la salud emocional [15].

2.2.2 Conversión (De sonido a texto.)

Existen varias técnicas para descomponer el audio en sonidos individuales y convertir estos sonidos en formatos digitales. Cuando las personas hablan, se generan vibraciones en el aire y, por medio de un convertidor, se transforman en datos binarios que el sistema puede entender. Esta información binaria se compara con la correspondiente a sonidos pregrabados de una base de datos y, así, se obtiene un posible resultado [20].

El lenguaje humano no es tan simple; muchas personas hablan con diferentes acentos que nos ha heredado la zona en la que nos desarrollamos, jergas y malas pronunciaciones. Estas variaciones no se encuentran en el diccionario de la computadora; por esto se utiliza el modelo oculto de Markov (HMM), el cual es un modelo perfecto para seguir un discurso y, junto con las redes neuronales artificiales (ANN), quienes, con sus entrenamientos, hacen una distinción muy eficiente de las diferentes sílabas que lo componen, generando una solución a estos inconvenientes [21].

2.2.3 Red Neuronal Artificial (ANN)

En este sistema, los datos de entrada reciben estímulos externos, con los cuales se realizará un cálculo interno y generará un valor de salida. Internamente, la neurona hará una suma ponderada de estos valores. La ponderación de cada entrada viene dada con el peso que se le asigna a cada uno de los valores de entrada. Los pesos de entrada sirven para definir la intensidad con la que afecta la entrada a la neurona; veámoslo como un regulador de datos.

A todo esto, se le sumará un sesgo o vías para tener mayor control con los datos de salida y, para finalizar, se le agrega una función de activación que evaluará el resultado de los datos procesados. En la Fig. 1 se muestra una ilustración de este sistema [22].

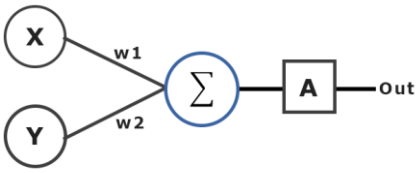


Figura 1. Arquitectura de la neurona artificial [22]

El problema que presentó este sistema fue que las soluciones son muy limitadas, solo dan solución a pocos problemas. Por este motivo, se agregaron más neuronas a un solo sistema para abarcar más problemas a resolver; esto trajo como consecuencia las redes [22].

Al juntar las neuronas de forma vertical, a este conjunto le denominaremos capas; entonces, las neuronas que se encuentran en la misma capa recibirán la información de la capa anterior y los cálculos que realicen los pasarán a la siguiente capa. De esta manera se separa en la primera capa (capa de entrada), la capa de en medio (capa oculta) y la última capa (capa de salida), creando un sistema de redes neuronales artificiales (ANN). En la Fig. 2 se muestra el esquema ilustrativo de este modelo [23].

Figura 2. Arquitectura de una red neuronal artificial (ANN) [24]

2.2.4 Modelo Oculto de Markov (HMM)

Los modelos ocultos de Markov (HMM) son sistemas autómatas abstractos de estados finitos donde se modelan procesos estocásticos. Aquí, los estados se asocian con una distribución de probabilidad y las transiciones de un estado a otro se gobiernan por un conjunto de probabilidades [25]. De esta manera, se fija un patrón de inicio que se tenga en el espectrograma con significado específico para ir agregando probabilidades y, así, poder predecir cuál es la siguiente letra obtenida en los datos de entrada. En la Fig. 3 se muestra un esquema de representación del modelo oculto de Markov para formar palabras [26].

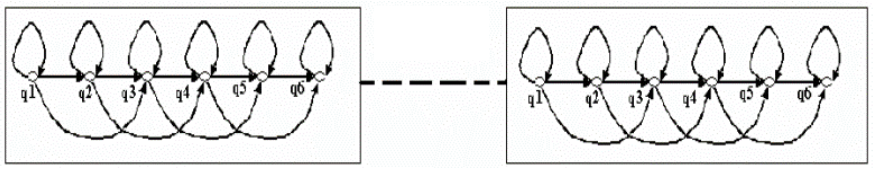


Figura 3. Diagrama del modelo de Markov para la creación de palabras [26]

2.2.5 Modelo mixto HMM-ANN

Para la implementación de éste, se empezará definiendo los estados *N*, en los cuales contendrá nuestras letras, palabras o sílabas, dependiendo del modelo. Se definirá a *M* para los distintos símbolos que pueden observarse. Una matriz de transición *A= {aij}* donde *aij = P(qt+1=j | qt=i),* siendo *qt* el estado actual. Hay que observar que, si uno de los *aij* es definido cero, permanecerá como cero durante todo el proceso de entrenamiento. En este modelo mixto el entrenamiento viene de la forma $y(x) = g(\sum_{i=0}^{n} w_i x_i - \theta_i)$ donde $x = (x_1 \ldots x_n)$ son los input, *y(x)* es el output, $w = (w_1 \ldots w_n)$ son los pesos y θ_i es el umbral.

El problema de HMM es que se encuentran limitados por el número de los diferentes estados que son capaces de distinguir; si sus distribuciones no están bien diferenciadas, la probabilidad acaba solapándose. Este problema se soluciona con las redes neuronales artificiales (ANN), ya que, al potenciarlos con estos modelos, resultan mucho más eficaces. Con una muestra de entrenamiento grande, estos mecanismos son capaces de distinguir una gran cantidad de salidas diferentes. Para lograr esto utilizan las ANN's entrenadas específicamente para cada uno de los estados de HMM. De esta manera, cada estado vendrá determinado por los pesos de las redes neuronales. En la Fig. 4 se muestra la representación del modelo planteado [21].

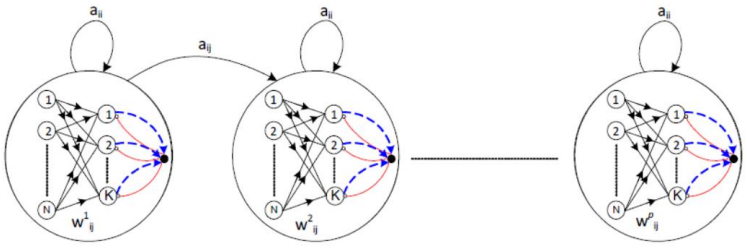


Figura 4. Representación de los estados del modelo HMM-ANN [21]

De esta manera, ya sabemos cómo es que el reconocimiento de voz entiende los sonidos que salen de nuestra boca. Ahora, ya solo falta un paso, darle el contexto.

2.2.6 Procesamiento del lenguaje natural

El Procesamiento del Lenguaje Natural (PLN) es una rama de la Inteligencia Artificial, específicamente de la lingüística computacional. Asimila en un lenguaje de programación definido para poder comunicarse con el ser humano en su lenguaje natural [27].

El PLN tiene varios objetivos; entre ellos, está facilitar la comunicación de la computadora con usuarios que no están especializados con éstas, modelar procesos para la comprensión del lenguaje y, así, diseñar sistemas que realicen tareas lingüísticas como la traducción del lenguaje, resúmenes de textos,

recuperación de información o, en nuestro caso, aplicación de test, entre otros [28].

2.2.7 Árbol de decisión

Anteriormente se vieron algunos tests que pueden detectar el trastorno depresivo. Para la aplicación de estos utilizaremos los árboles de decisión. Estos son modelos bayesianos de inteligencia artificial en donde se representan esquemas gráficos de izquierda a derecha o de arriba a abajo hacia las posibles alternativas que puede tomar el sistema [29]. Entre estos, existen modelos descriptivos, que sirven para distinguir entre objetos de diferentes clases, y los modelos predictivos, que pueden predecir la clase a la que pertenece el objeto conociendo sus características anteriores [30]. Éste es el uso que se le da a la clasificación.

3 Plataforma Web Hálito

Como propuesta, se desarrolló una metodología para el diseño de una plataforma tecnológica que utilice inteligencia artificial con el objetivo de dar una primera atención a la salud emocional (en el enfoque depresivo) de jóvenes y adultos. Como en esta plataforma se iba a contar con un asistente virtual tipo chatbot, se propuso el nombre de Alaia, para que el usuario lo pudiera relacionar de mejor forma. Alaia se apoya en este subcampo importante de la IA, el reconocimiento de voz, a fin de guiar una conversación y, así, aplicar la prueba auto aplicable, el inventario de depresión de Beck en su versión (BDI-II) [5], [6].

Para la interacción con el sistema se tienen diferentes actores que fungirán un papel específico, ya que un usuario promedio no puede acceder a los datos sensibles de todos los que aplicaron con Alaia o a un experto en el área de la salud no le es útil acceder a la codificación del sistema. A continuación se presenta una definición de los actores.

- Usuario: Se trata de un actor que accede a la plataforma en busca de ayuda sobre el trastorno depresivo.
- Expertos: Se trata de los actores capacitados para llevar a cabo el seguimiento del trastorno depresivo (personal de la salud, psicólogos o psiquiatras).
- Programador: Desarrollador del sistema, soporte técnico y mantenimiento.

En la Fig. 5 se presentan los distintos casos de uso que puede tener la plataforma Hálito.

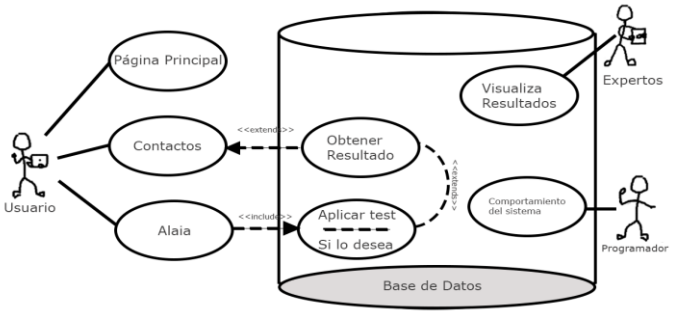


Figura 5. Casos de uso del sistema Hálito

Hálito se compone de 4 distintas interfaces principales, algunas con sus propias ramificaciones. En estas encontramos:

1. Página principal - punto de entrada del sitio web;
2. Contactos - acceso a otras alternativas que puedan abordar su problema;
3. Alaia - asistente virtual con IA que aplica el test de Beck por medio de un chatbot de interacción voz-auditiva;
4. Acceder - acceso a los expertos para valorar la interacción con Alaia y a los programadores para ver el comportamiento del sistema.

3.1 Alaia

¿Cómo funciona Alaia? Anteriormente ya se había mencionado un poco del funcionamiento de. reconocimiento de voz; entonces, Alaia toma este mismo modelo para generar una interacción con el usuario y, así, aplicar el BDI-II. En la Fig. 6 se muestra un diagrama ilustrativo del modelo creado para Alaia.

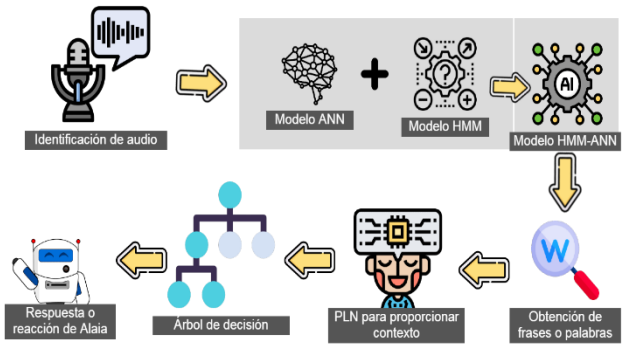


Figura 6. Modelado arquitectónico de Alaia

Al crear el sistema se buscó una compatibilidad con la mayor cantidad de dispositivos posibles. La solución a este objetivo fue una plataforma web que junto, con CSS (hoja de estilos fácil de manipular) y JavaScript (lenguaje que añade características interactivas a los sitios web), resultó en el desarrolló el sistema de Alaia. Ésta, integra una biblioteca de funciones de reconocimiento de voz (Annyang) [31]. En la *identificación de audio* ejecuta una función de *callback* para verificar la entrada de sonidos. Esto se controla por medio de una bandera inicializada en *false*. Una vez activa con los sonidos obtenidos por el micrófono, entra al *modelo de entrenamiento* para que, así, nos pueda proporcionar posibles respuestas. De este resultado salen las frases o palabras. Ahora, con el modelo de *procesamiento del lenguaje natural* se identifica el contexto de la información obtenida. Ëste pasa por el *árbol de decisiones* para que, así, Alaia pueda responder o reaccionar a los datos de entrada.

4 Resultados y Conclusiones

El producto de esta plataforma fue una interfaz con un ambiente cálido para las personas de prueba. En estas pruebas se tuvo al programador como observador, guía para registrar errores y problemas de uso. Aquí participaron 11 personas, 4 varones y 7 mujeres, de edades variadas entre la adolescencia, la juventud y la adultez.

La interacción con Hálito fue positiva en función del diseño y usabilidad. Al interactuar con Alaia, se observó que los primeros minutos de la charla fluía bien, pero, a medida que avanzaba el tiempo, Alaia perdía el hilo de la conversación. Se identificaron dos posibles puntos a estos problemas. El primero fue en la interacción; algunos usuarios responden aun cuando Alaia no terminaba de hablar, lo cual registraba una combinación de palabras entre las preguntas de Alaia y la respuesta del usuario. Una solución a este problema sería desactivar el micrófono cuando el sistema detecte la entonación de Alaia. El segundo problema se debe al entrenamiento, al ser un sistema que se maneja con el lenguaje natural humano, los vocabularios y frases de cada persona son muy distintos. Esta característica posee un lado positivo, al dejar que el usuario se pueda expresar más. La solución a esto es entrenar con más datos y pruebas de charla al sistema para lograr una mayor eficiencia. De esto, podemos concluir que:

- Aún hay camino por recorrer, pero estamos en la dirección correcta; la sustitución del personal médico de enfermedades mentales aún no es posible, pero, con un mayor entrenamiento y añadiendo un receptor visual al sistema para conocer las expresiones faciales que tiene el usuario (factor que da información al aplicar el análisis del trastorno), podemos obtener resultados muy prometedores;
- Existe la posibilidad de que el usuario se exprese más abiertamente al estar interactuando con un robot y no una persona;
- Las pruebas de detección del trastorno depresivo (en este caso aplicado por un chatbot) que niegan el bienestar o confirman el buen estado mental de la persona y pueden ayudar a prevenir futuros incidentes en la población joven y adulta;

- El reconocimiento de voz es una gran alternativa para una primera atención a la salud emocional y, así, dar apoyo a los profesionales de salud emocional.

Referencias

[1] Brunier, A., & Drysdale, C. (2020). Los servicios de salud mental se están viendo perturbados por la covid-19 en la .mayoría de los países, según un estudio de la oms Recuperado el 28 de junio de 2022, de https://www.who.int/es/news/item/05-10-2020-covid-19-disrupting-mental-health-services-in-most-countries-who-survey

[2] Migala. (2020). Podcast migala 22: la depresión. Migala. Recuperado el 28 de junio de 2022, de https://www.youtube.com/watch?v=upd6mixy8us&t=4600s

[3] Alarcón, R., Gea, A., Martínez, J., Pedreño, J., & Pujalte, M. (2003). Guía de práctica clínica de los trastornos depresivos. Recuperado el 28 de junio de 2022, de https://consaludmental.org/publicaciones/gpctrastornosdepresivos.pdf

[4] Navas, W., & Vargas, M. (2012). Abordaje de la depresión: intervención en crisis. Recuperado el 28 de junio de 2022, de https://www.binasss.sa.cr/bibliotecas/bhp/cupula/, 19–35.

[5] Bustos, V., Galvis, E., & Rojas, S. (2015). Instrumentos de tamizaje de depresión para niños, adolescentes y adultos, revisión narrativa de la literatura. pontificia universidad javeriana.

[6] Beltrán, M., Freyre, M., & Hernández, I. (2012). El inventario de depresión de Beck: su validez en población adolescente. https://www.scielo.cl/scielo.php?script=sci_serial&pid=0718-4808&lng=en&nrm=iso, 30(1), 5–13. recuperado el 28 de junio de 2022, de https://www.scielo.cl/scielo.php?script=sci_arttext&pid=s0718-48082012000100001&lng=en&nrm=iso&tlng=en

[7] Jackson Tempra. (2021). what's up? - mental health app. Google play. Recuperado el 28 de junio de 2022, de https://play.google.com/store/apps/details?id=com.jacksontempra.apps.whatsup

[8] Daniels, J. (2017). Woebot. Facebook. Recuperado el 28 de junio de 2022, de https://www.facebook.com/hiwoebot/?ref=page_internal.

[9] Replika: my ai friend. (s/f). Google play. Recuperado el 28 de junio de 2022, de https://play.google.com/store/apps/details?id=ai.replika.app

[10] Fernández, Y. (2017). Así era eliza, el primer bot conversacional de la historia. xataka. Recuperado el 28 de junio de 2022, de https://www.xataka.com/historia-tecnologica/asi-era-eliza-el-primer-bot-conversacional-de-la-historia

[11] Romero, M., Casadevante, C., & Montoro, H. (2020). Cómo construir un psicólogo-chatbot. https://www.papelesdelpsicologo.es/, 41(1), 27–34. recuperado el 28 de junio de 2022, de https://www.papelesdelpsicologo.es/pdf/2920.pdf

[12] ¿Qué es un chatbot? (2022). Oracle. Recuperado el 28 de junio de 2022, de https://www.oracle.com/mx/chatbots/what-is-a-chatbot/

[13] Santander universidades. (2021). Test de Turing: ¿pueden las computadoras sustituir a los humanos? santander|becas. recuperado el 28 de junio de 2022, de https://www.becas-santander.com/es/blog/test-de-turing.html

[14] Laríos, A. (1999). Diccionario español/inglés para el aprendizaje de vocabulario utilizando una interfaz de voz [universidad de las américas]. recuperado el 28 de junio de 2022, de http://catarina.udlap.mx/u_dl_a/tales/documentos/lis/ahuactzin_l_a/capitulo1.pdf

[15] Ruiz, P. (s/f). Clasificación vocal. Recuperado el 28 de junio de 2022, de http://www.voz-profesional.com/wp-content/uploads/2014/02/clasificacion-vocal.pdf

[16] Vozalia. (2020). La voz humana. Decibelios y frecuencia de la voz humana. Vozalia. Recuperado el 28 de junio de 2022, de https://www.vozalia.com/voces/la-voz-humana-decibelios-y-frecuencia-de-la-voz-humana/#:~:text=frecuencia%20de%20la%20voz%20humana,-la%20frecuencia%20de&text=la%20frecuencia%20media%20de%20la,137%20hz%20a%20634%20hz.

[17] Dbelectronics. (2018). Intensidad del sonido en decibelios. Dbelectronics. Recuperado el 28 de junio de 2022, de https://www.dbelectronics.es/intensidad-del-sonido-en-decibelios/

[18] Pinillos, M. (s/f). ¿Hablas rápido? aprende a manejar la velocidad de habla para una comunicación exitosa. Marta pinillos. Recuperado el 28 de junio de 2022, de https://www.martapinillos.com/velocidad-de-habla-en-el-ritmo-del-discurso/

[19] Acyv. (2022). ¿Cuántas palabras por minuto puede hablar de media una persona? el confidencial. Recuperado el 28 de junio de 2022, de https://www.elconfidencial.com/alma-corazon-vida/2022-06-07/palabras-por-minuto-habla-de-media-persona_3435387/

[20] Acadaimy. (2021). How does speech recognition work? learn about speech to text, voice recognition and speech synthesis. youtube. recuperado el 28 de junio de 2022, de https://www.youtube.com/watch?v=6altvgtof9s&list=ll&index=21&t=199s

[21] Tornero, J. (2017). Machine learning: modelos ocultos de markov (hmm) y redes neuronales artificiales (ann) [universidad de barcelona]. Recuperado el 28 de junio de 2022, de http://diposit.ub.edu/dspace/bitstream/2445/122446/2/memoria.pdf

[22] Dot, C. S. V. (2018). ¿Qué es una red neuronal? parte 1 : la neurona | dotcsv. youtube. Recuperado el 28 de junio de 2022, de https://www.youtube.com/watch?v=mriv2iwftpg&t=116s&ab_channel=dotcsv

[23] Dot, C. S. V. (2018b). ¿Qué es una red neuronal? parte 2 : la red | dotcsv. youtube. Recuperado el 28 de junio de 2022, de https://www.youtube.com/watch?v=uwbhopp9xkc&t=27s&ab_channel=dotcsv

[24] Qué son las redes neuronales y sus funciones. (2019). Atria innovation. Recuperado el 28 de junio de 2022, de https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/

[25] Maldonado, L. (2012). Los modelos ocultos de Markov, mom. Recuperado el 28 de junio de 2022, de https://www.redalyc.org/articulo.oa?id=99324907003, 14(3), 433–438.

[26] Oropeza, J. (2006). Algoritmos y métodos para el reconocimiento de voz en español mediante sílabas. Scielo, 9(3), 270–286. Recuperado el 28 de junio de 2022, de http://www.scielo.org.mx/pdf/cys/v9n3/v9n3a7.pdf

[27] García, L. (2017). Asistente virtual tipo chatbot [universidad católica de colombia]. recuperado el 28 de junio de 2022, de https://repository.ucatolica.edu.co/bitstream/10983/17726/1/asistente%20virtual%20tipo%20chatbot_final.pdf

[28] Hernández, M., & Gómez, J. (2013). Aplicaciones de procesamiento de lenguaje natural. revista politécnica, 32(1), 87–96. Recuperado el 28 de junio de 2022, de https://core.ac.uk/download/pdf/18586869.pdf

[29] Hurtado, H. (2019). Contribución del teorema de Bayes a la toma de decisiones gerenciales acertadas en salud [universidad nacional abierta y a distancia]. Recuperado el 28 de junio de 2022, de https://repository.unad.edu.co/bitstream/handle/10596/26083/hhurtadob.pdf?sequence=1&isallowed=y#:~:text=en%20lenguaje%20matem%c3%a1tico%2c%20el%20teorema,uniones%20entre%20los%20puntos%20aristas.

[30] Expósito, C., Expósito, A., López, I., Melían, B., & Moreno, J. (s/f). Árboles de decisión.

[31] Tal ater. (2018). Annyang! npm. Recuperado el 28 de junio de 2022, de https://www.npmjs.com/package/annyang

Procesamiento de datos médicos cualitativos para el análisis y modelado en aprendizaje automático

Edwin Aldana-Bobadilla[1][0000−0001−8315−1813]
Alejandro Molina-Villegas[2][0000−0001−9398−8844]
Hiram Galeana-Zapién[3][0000−0002−8449−9077]
Karina Gazca-Hernández[3]

[1] Conacyt-Centro de Investigación y de Estudios Avanzados del I.P.N. (Cinvestav), Victoria 87130, Mexico
edwyn.aldana@cinvestav.mx
[2] Conacyt-Centro de Investigación en Ciencias de Información Geoespacial (Centrogeo), Mérida 97302, Mexico
amolina@centrogeo.edu.mx
[3] Centro de Investigación y de Estudios Avanzados del I.P.N. (Cinvestav), Victoria 87130, Mexico
{hgalena,karina.gazca}@cinvestav.mx

Resumen La coyuntura tecnológica actual ha hecho posible almacenar, distribuir y procesar grandes y variados tipos de datos procedentes de muchas actividades del acontecer humano en diferentes dominios. El dominio clínico es, sin lugar a dudas, un escenario en el cual este hecho se hace latente y en el que, por su naturaleza, conviven diferentes fuentes de datos, tanto estructurados como no estructurados, alrededor de los pacientes. Muchos de estos datos no son susceptibles de análisis numérico directo –dada su naturaleza cualitativa–, por lo que siempre es requerido realizar tareas de transformación que permitan llevarlos a un espacio en el que dicho análisis sea posible. Este trabajo hace un recorrido por los diferentes tipos de datos no numéricos en el contexto clínico y las alternativas para lograr una transformación que permita extraer su valor informativo en aras de realizar análisis numérico para el apoyo a la toma de decisión. Dada la importancia de dicha transformación en la efectividad de los modelos obtenidos, este trabajo tiene como objetivo servir de guía para ayudar a la selección de las transformaciones más adecuadas en función de los tipos de datos del problema. Se presentan varios casos de estudio en los que se aplican diferentes técnicas de transformación y extracción de características para texto, imágenes y señales fisiológicas que, al ser utilizadas por modelos de predicción, mostraron un buen desempeño.

Palabras Clave: Datos Categóricos · Datos no Estructurados · Extracción de Características.

1. Introducción

Los datos del historial clínico de un paciente son un recurso fundamental en la práctica clínica y en el ámbito de ciencias de la salud e investigación médica. Dichos datos son recolectados por los profesionales de la salud durante los procesos de ingreso hospitalario, diagnóstico y atención del paciente hasta la finalización del tratamiento. Dicho historial clínico está conformado generalmente por diversos tipos de datos, entre los que se encuentran: a) datos recopilados por el médico a través de preguntas al paciente; b) datos recabados por el médico como resultado de una exploración física, medición de signos vitales y resultados de análisis de laboratorio; c) resumen de síntomas y diagnósticos posibles identificados con base en la exploración previa; y d) tratamiento recomendado al paciente. En este contexto, la evolución de las tecnologías y ciencias computacionales ha permitido el desarrollo de sistemas como el expediente clínico electrónico (EMR, por su siglas en inglés) con el fin de organizar y almacenar grandes volúmenes de datos clínicos en un ambiente hospitalario. Como su nombre lo indica, el EMR es un sistema de gestión de datos hospitalarios que se desarrolla bajo criterios y recomendaciones internacionales que buscan garantizar la disponibilidad, confidencialidad e integridad de los datos (e.g. norma oficial mexicana NOM-024). Entre los datos incluidos en un EMR se tienen datos administrativos y demográficos, diagnósticos, tratamientos, prescripciones, resultados de análisis de laboratorio, datos de monitoreo fisiológico, entre muchos otros. Muchos de los análisis en la práctica médica están asociados a variables numéricas continuas en aras de proveer evaluaciones precisas y confiables acerca del estado o condiciones de un paciente. Por ejemplo, en [20] se argumenta el impacto positivo de los probióticos en el control metabólico de los pacientes con diabetes tipo 2 con base en variables numéricas, como el índice de masa corporal, los niveles de colesterol total, colesterol bueno y malo, triglicéridos, glucosa plasmática en ayunas, niveles de insulina en ayunas, presión arterial sistólica y diastólica. En [18] se diseñan modelos de regresión para determinar la relación de prevalencia entre la agresividad tumoral y los cambios en la composición corporal, considerando el porcentaje de masa magra, masa grasa, ángulo de fase, resistencia, reactancia, la leptina plasmática, cambios en la composición corporal entre otros.

Sin embargo, la práctica médica también puede incluir variables que representan: 1) la ocurrencia de un evento (e.g. infección, enfermedad, reingreso hospitalario, deceso) y el grado de afectación de una enfermedad de acuerdo con un conjunto de categorías predefinidas (e.g. alto, medio, bajo); 2) la pertenencia del paciente a uno de dos o más grupos mutuamente excluyentes (e.g. género, etnia, escolaridad, etc.); y 3) la transformación de una variable continua para establecer niveles predefinidos que simplifiquen su análisis (e.g. rango de edad, niveles de insulina, niveles de presión, etc.). Estas variables son ampliamente conocidas como *variables categóricas*, las cuales representan diferentes niveles de medición –nominal y ordinal– con importantes implicaciones en su interpretación y análisis. Estas variables hacen parte comúnmente de datos que corresponden a estructuras predefinidas que representan entidades u objetos de dominio (e.g. pa-

ciente, consultorio, médico), por lo que reciben el nombre de *datos estructurados*. Por otro lado, la tecnología computacional (hardware y software) ha permitido la generación, almacenamiento, transmisión y procesamiento de imágenes, documentos de texto, videos, audios, entre otros, los cuales no están supeditados a una estructura tabular que permita su análisis numérico de manera directa. En virtud de permitir dicho análisis, se ha recurrido a diferentes técnicas de áreas como el procesamiento digital de imágenes [47], el procesamiento de señales [44] y el procesamiento de lenguaje natural [36]. El dominio clínico es, sin duda alguna, una fuente creciente de este tipo de datos denominados *no estructurados*.

El resto del capítulo está organizado de la siguiente manera. En la Sección 2 se presenta una descripción de los datos categóricos y las alternativas para codificarlos en valores susceptibles al análisis numérico. Posteriormente, en la Sección 3 se presentan las principales estrategias existentes para el procesamiento de datos no estructurados. Finalmente, las conclusiones del capítulo se presentan en la Sección 4.

2. Distinguiendo entre datos categóricos y numéricos

Cuando los símbolos (números, letras o cadenas de caracteres) que representan el valor de una variable del paciente se usan exclusivamente para clasificarlo, se dice que dicha variable es una medición en su nivel más débil. Estos números o símbolos constituyen lo que comúnmente se denomina *escala nominal* [34]. Esta escala se presenta en situaciones como aquellas en las que un diagnóstico médico identifica a un paciente como diabético o hipertenso, el paciente es asignado a un grupo étnico particular o cuando dicho paciente es asignado a una nacionalidad o área geográfica en virtud de su origen. Si se usan números $(1, 2, 3, ..., n)$ en una escala nominal, es incoherente sumar o restar sus valores, ya que dicha operación no tiene sentido en el contexto de lo que éstos representan: clases o grupos. Por lo tanto, la única medida estadística descriptiva admisible para este tipo de escala es la moda y la única relación entre dos posibles valores es la igualdad $(=)$. En las secciones subsiguientes se presentan diferentes técnicas para procesar y analizar datos en escala nominal.

Puede suceder que los valores (símbolos) de una variable asociada al paciente no solo representen categorías, sino que, además, exista una *relación de orden* o precedencia entre ellas. Por ejemplo, una variable representando el nivel de escolaridad de un paciente con valores del conjunto {*primaria, secundaria, preparatoria, licenciatura, posgrado*} induce una relación de orden. Este tipo de variable es conocida como *escala ordinal* [26]. La relación de igualdad $(=)$ entre dos posibles valores se mantiene, pero ahora es posible definir una relación denotada por $(>)$ para indicar que un valor 'supera' a otro o $(<)$ para indicar que un valor 'no supera' a otro. Por ejemplo $posgrado > licenciatura > preparatoria$ o $preparatoria < licenciatura < posgrado$. Si cada valor es asignado a un valor numérico o puntuación donde el orden se preserve, el estadístico más apropiado para este tipo de escala será la mediana.

Cuando el valor de la variable induce un orden pero, además, la diferencia entre dos valores tiene sentido, estamos ante una escala más fuerte que la ordinal, usualmente conocida como *escala de intervalo* [34]. Un ejemplo clásico de este tipo de escala es, por ejemplo, la temperatura corporal de un paciente medida en grados Celsius ($^{\circ}C$) o Fahrenheit ($^{\circ}F$). En este tipo de escala, el valor cero es relativo, esto significa que el cero no siempre representa la ausencia de la propiedad que se está midiendo. Por ejemplo, la temperatura de congelación de 0 $^{\circ}C$ no significa ausencia de temperatura; este cero es relativo a la unidad de medida, ya que el mismo valor de temperatura puede ser expresado como 32 $^{\circ}F$. Una propiedad importante de esta escala es que las razones de las diferencias de valores (intervalo) son independientes de la unidad de medida y del punto cero, como por ejemplo $\frac{(30^{\circ}C-10^{\circ}C)}{(100^{\circ}C-30^{\circ}C)} = \frac{(86^{\circ}F-50^{\circ}F)}{(212^{\circ}F-86^{\circ}F)} = 0{,}28$. Comparada con la escala nominal y ordinal, la escala de intervalo es formalmente cuantitativa, por lo que puede ser descrita a través de estadísticos como la media y la desviación estándar. La escala de intervalo es una condición necesaria, pero no suficiente, para realizar pruebas paramétricas.

Para valores cuyas unidades de medida poseen un cero absoluto, la razón entre cualesquiera dos valores es independiente de la unidad de medida. Por ejemplo, si la razón del peso de dos pacientes es $\frac{65kg}{70kg} = 0{,}93$, esta razón se conserva cuando el peso se expresa en libras (lb) $\frac{143{,}3lb}{154{,}3lb} = 0{,}93$. Así mismo, la multiplicación de un valor de la escala por un número c conserva sus proporciones e interpretación independientemente de la unidad de medida: $2(65kg) = 2(143{,}3lb)$. Lo anterior no es posible para valores ordinales y de intervalo. Este tipo de escalas admite todas las operaciones aritméticas y puede ser descrita a través de estadísticos como la media, la mediana y la desviación estándar.

Las anteriores consideraciones son muy importantes a la hora del análisis, especialmente cuando se plantean modelos que incluyen operaciones entre diferentes variables. Como hemos mencionado anteriormente, el dominio clínico incluye frecuentemente datos en escalas nominales y ordinales que, por su naturaleza, no pueden ser usados ni interpretados como números de manera directa. En las siguientes subsecciones se describen algunas técnicas que permiten transformar datos nominales u ordinales con el propósito de inducir algunas propiedades numéricas en ellos.

2.1. Codificación

Una variable de escala nominal u ordinal induce k categorías que pueden ser transformadas en una secuencia de $k-1$ variables. Retomando el ejemplo de la escolaridad del paciente, esta variable es codificada como una secuencia de cuatro variables binarias, como se ilustra en la Tabla 1. Estas variables son denominadas *variables dummy* y pueden ser usadas como variables predictoras en modelos de regresión o clasificación [13]. Los valores de dichas variables (0,1) inducen un espacio ortogonal en el que la suma, la resta y el cálculo de distancia son posibles. Obsérvese que no es necesaria la variable *primaria*, ya que con las cuatro variables restantes codificadas siempre será posible inferirla (cuando las

Tabla 1: Ejemplo de codificación de una variable categórica con $k = 5$

Valor original	**Variables dummy**			
	secundaria	preparatoria	licenciatura	posgrado
primaria	0	0	0	0
secundaria	1	0	0	0
preparatoria	0	1	0	0
licenciatura	0	0	1	0
posgrado	0	0	0	1

cuatro variables son cero). Matemáticamente, es conveniente hacerlo así porque, de lo contrario, se estará introduciendo una variable que queda explicada por las otras, generando un problema de colinealidad que podría inducir matrices singulares que impedirían la solución del problema.

Es de notar que esta codificación pudiera resultar computacionalmente costosa en función del número de categorías o niveles de la variables (e.g. nombres de estados de un país, catálogo de enfermedades, etc.), en cuyo caso es posible recurrir a métodos como aquellos basados en hashing que mapean cada valor de la variable a un valor entero (hash), el cual es, a su vez, transformado en un valor entero que representa un código más compacto de la característica usando aritmética modular [46]. Por ejemplo, podemos aplicar una función hash (e.g. md5) al valor de una categoría así: $md5(diabetes) = c35b49da1b6dd679a48d67f456bc7aaf$. Si nuestro catálogo de categorías es, por ejemplo, del orden de cientos, lo ideal sería mapear ciertas enfermedades a una misma categoría con el fin de obtener un catálogo más compacto. Sea $r = 16$ el número deseado de categorías en nuestro nuevo catálogo; entonces, $c35b49da1b6dd679a48d67f456bc7aaf$ **mod** r será el código nuevo asignado a la categoría de diabetes que, en este caso, corresponde a 3. Aunque este método reduce el espacio de las variables codificadas, éste puede inducir otros problemas como aquellos causados por posibles colisiones en la generación de los valores hash.

Existen otros enfoques de codificación como aquellos basados en los datos de salida o variable a predecir, razón por la cual reciben el nombre de *métodos de codificación supervisados*. Este tipo de codificación es apropiada cuando la variable a ser codificada tiene muchos valores posibles o, inclusive, cuando aparecen nuevos valores después de la fase de obtención del modelo (entrenamiento). Los códigos postales son un buen ejemplo, por ejemplo, en el caso de México hay aproximadamente 33,000 códigos postales. Ante esta cardinalidad, es obvio que una codificación basada en variables dummy resulta computacionalmente ineficiente. Adicionalmente, si el número de valores permitidos es muy grande, es probable que algunos de los valores menos comunes no aparezcan en los datos de entrenamiento en virtud de que estos se obtienen a través de un proceso de muestreo. En [27], [28], [48] se presentan algunos trabajos sobre este enfoque.

Si se desconocen los valores de salida, existen otras formas de abordar el problema de la codificación, por ejemplo en [23] se presenta un algoritmo de agrupamiento aglomerativo basado en similitud para variables mixtas (categóri-

cas y numéricas) propuesto por Goodall en [11] para problemas de taxonomía biológica. La idea central que exponen para definir la medida de similitud entre dos objetos es otorgar mayores pesos a las coincidencias de valores de características poco comunes; asimismo, no asumen las distribuciones subyacentes de los valores de las características. En [1] se presenta un método de agrupamiento basado en el algoritmo *k-means* que es capaz de codificar de manera conjunta las propiedades numéricas y categóricas gracias a una función de costo y medida de distancia basada en la *co-ocurrencia* de valores. En [16] se presenta una nueva función de distancia que toma en cuenta ambos términos, la distancia euclidiana y una medida de similitud ponderada en atributos categóricos. Similarmente, en [25], los autores sugieren un enfoque basado en el concepto de acumulación de evidencia, cuya principal función es combinar en una sola partición los resultados de múltiples agrupaciones tomando las co-ocurrencias de pares de patrones en el mismo grupo. Otras alternativas de codificación, como la planteada en [15], propone un modelo de mapa auto-organizado generalizado basado en el modelo de Kohonen [21], que ofrece un método para expresar la similitud entre valores no numéricos a través de jerarquías de distancia permitiendo el proceso de valores categóricos en el entrenamiento. Con esto se unifica el cálculo de distancia de variables numéricas y no numéricas. En [24] se propone un método que codifica cada valor categórico con base en su valor informativo en términos de la entropía de Shannon, resultando en un enfoque computacionalmente eficiente.

Las diferentes formas de abordar el problema de codificación inducen una transformación de los valores cualitativos a un espacio métrico donde las relaciones de similitud entre estos pueden ser expresadas en términos de operaciones aritméticas y, en consecuencia, es posible describirlos a través de estadísticos de centralidad y dispersión como la media y desviación estándar. Es importante, antes del análisis y diseño de modelos matemáticos, asegurarnos de que los datos cualitativos sean codificados a través de alguno de los enfoques señalados.

3. Datos no estructurados

Hasta aquí hemos señalado las diferencias entre los datos numéricos y categóricos, haciendo énfasis en las tareas de transformación de estos últimos para hacer posible su análisis cuantitativo. Usualmente, tanto los datos numéricos como los categóricos representan atributos de entidades u objetos de algún dominio que pueden ser organizados bajo las pautas de un modelo de datos [7]. Sin embargo, existen otros tipos de datos que requieren de una transformación que habilite su análisis desde el punto de vista cuantitativo. En muchos dominios, entre ellos el dominio médico, se tienen datos en otros formatos, como texto o imagen, que requieren ser procesados con el fin de extraer información en la forma de propiedades cuantitativas o numéricas. Estos datos son típicamente conocidos como *no estructurados* dado que no están enmarcados bajo ninguna estructura o modelo. En la Tabla 2 se presentan algunas diferencias puntuales entre datos estructurados y no estructurados.

Tabla 2: Diferencias entre datos estructurados y no estructurados

	Datos estructurados	**Datos no estructurados**
Características	Modelos predefinidos	Modelos no predefinidos
	Usualmente solo texto	Diversos formatos
	Fácil de buscar	Difícil de buscar
Ubicación	Base de datos relacionales	Aplicaciones
	Data warehouses	Bases de datos NoSQL
		Data warehouses
		Data lakes
Ejemplos	Fechas	Archivos de texto
	Números telefónicos	Reportes
	Números de seguro social	Mesajes de texto
	Números de tarjeta de crédito	Videos
	Direcciones	Imágenes

En esta sección se presentan algunas tareas de procesamiento de datos no estructurados en el contexto clínico que permiten extraer propiedades cuantitativas susceptibles de ser analizadas y modeladas a través de técnicas estadísticas y de aprendizaje automático.

3.1. Texto

La información médica en formato textual tiene dos variantes que deben tratarse de manera distinta para su correcto procesamiento. Por un lado, hay que considerar que uno de los tipos de datos básicos de la mayoría de los lenguajes de programación (y de las bases de datos) es el tipo de dato cadena (string o text en documentación técnica) que, para efectos de procesamiento, debe ser utilizado como datos categóricos. Ejemplos de este tipo de información médica los podemos encontrar en los resultados de un análisis clínicos para medir el ácido úrico. La información de este examen podría contener la leyenda "Valores normales", indicando que se encontró un rango de 3.5-7.2 (en varón). En este caso, la información de la leyenda es ciertamente de tipo textual, pero su procesamiento mediante algoritmos debe hacerse de acuerdo con lo señalado en la Sección 2 Datos Categóricos. Esto es debido a que los datos textuales no corresponden a una narrativa o una descripción, lo cual nos lleva al segundo caso, que se discute a continuación.

Las notas médicas, al ser descripciones de información destinadas a ser leídas, comprendidas e interpretadas por médicos, deben ser tratadas mediante algoritmos de inteligencia artificial (IA) para extraer conocimiento. En particular, toda información textual que represente una narrativa puede ser insumo de algoritmos de procesamiento de lenguaje natural (PLN). La información textual proveniente de notas médicas, descripciones, valoraciones, artículos científicos, reportes técnicos, entre otras, representa una fuente vasta en conocimiento, pero su correcto aprovechamiento representa también grandes e interesantes retos tecnológicos que han motivado el estado del arte en el procesamiento de este

tipo de información. En efecto, la IA y el PLN son áreas de investigación que han mostrado importantes progresos en la generación de modelos y la aplicación de algoritmos para el procesamiento de información médica. Así, se han desarrollado varias líneas de investigación en PLN con enorme potencial en el campo de la medicina, de las cuales hemos seleccionado algunas por ser de interés al contar con algún desarrollo tecnológico en procesamiento de texto en español de México.

Reconocimiento de entidades nombradas La primera línea de investigación que presentaremos se conoce como reconocimiento de entidades nombradas (NER, por sus siglas en inglés). El NER refiere a la detección y clasificación automática de entidades nombradas en documentos de dominio específico. Es decir, un módulo de NER debe procesar bloques de texto para luego producir un bloque de texto anotado con las entidades detectadas. Un ejemplo básico de texto anotado sería:

```
<PERSON>Jim</PERSON>, originario de <LOC>Seúl</LOC>,
compró 300 acciones de <ORG>Acmé Corp.</ORG> en <TIME>2006</TIME>.
```

En el ejemplo anterior han sido detectados y clasificados el nombre de una persona, un nombre de ciudad, un nombre de compañía (dos tokens o unidades léxicas) y una expresión temporal.

Actualmente, los sistemas de reconocimiento de entidades para el inglés tienen un rendimiento cercano al humano, pero cabe mencionar que, para el español, hay cierto rezago debido, en parte, a que hay menos corpus disponibles y los analizadores léxicos suelen ser menos sofisticados. No obstante, esta particular línea de investigación del PLN ha cobrado bastante relevancia gracias al potencial uso, no solo en el ámbito médico, sino prácticamente en cualquier área de conocimiento: Química [39], Biología [35], Historia [42], Geología [43], Geografía [31] y por supuesto medicina [45].

Aunado a lo anterior, las técnicas actuales tienen, hoy por hoy, un alto desempeño en la detección de entidades de dominio específico, y también contribuye el hecho de que existe una amplia gama de herramientas de cómputo disponibles en el mercado. Las más exitosas en implementaciones de software utilizan modelos de máxima entropía (MaxEnt), campos aleatorios condicionales (CRFs, por sus siglas en inglés) y modelos neuronales de aprendizaje profundo (*Deep Learning*).

En el marco MaxEnt, la probabilidad de las etiquetas NER para una secuencia de palabras se modela mediante la máxima entropía. Para este efecto, se define un conjunto de funciones de características arbitrarias que deben ponderarse utilizando un solo parámetro. Las funciones de características podrían considerar aspectos léxicos como recuento de palabras, uso de mayúsculas, prefijos y

sufijos, funciones basadas en diccionario, entre otras características dependientes del idioma.[4]

Un CRF es un modelo estocástico general comúnmente utilizado para etiquetar y segmentar datos secuenciales; proporciona un marco general para construir modelos de datos secuenciales. En el NER, una secuencia observada durante la etapa de entrenamiento es la secuencia de tokens que se ajustan a una oración o a un documento, y la secuencia de estados corresponden a las etiquetas de entidad proporcionadas durante esta etapa. Dado que los idiomas difieren de las convenciones que utilizan para las entidades nombradas, las características deben ser específicas para cada idioma. Por lo tanto, una restricción de este enfoque es que su efectividad es limitada y puede variar de un idioma a otro.[5]

En el enfoque de aprendizaje profundo se utiliza una red neuronal tanto para el aprendizaje de características como para la clasificación de entidades. Las palabras de una oración se tokenizan y luego se dividen en características y se agregan en un vector representativo llamado *Word Embedding*. Luego, este vector se alimenta a una red neuronal convolucional que hace una clasificación basada en el peso asignado a cada característica dentro del texto [41]. La etapa de entrenamiento requiere una gran cantidad de datos etiquetados manualmente para NER.[6]

Otro impacto positivo que han tenido los sistemas NER es que la detección de entidades nombradas se puede conectar con otros procesos, a manera de *pipeline*, con el objetivo de realizar tareas más complejas tales como la extracción de coordenadas a partir de textos o Geoparsing [2], la traducción automática, entre otras. En este sentido, el reconocimiento de entidades nombradas se convierte en la piedra angular en varios tipos de procesamiento de información y de ahí su relevancia.

Aunque los primeros sistemas de reconocimiento de entidades nombradas estaban basados en reglas léxicas [40], los métodos más modernos recaen en los métodos matemáticos antes mencionados. No obstante, cabe mencionar que los métodos basados en léxico se desempeñan muy bien en documentos de áreas de especialidad y, particularmente, en ciencias biológicas. Como muestra de esto, consideraremos la Tabla 3, extraída de [14]. En la Tabla 3 se replican los resultados de reconocimiento de entidades nombradas para un sistema médico usando dos diferentes métodos: el primer método, basado en léxico y expresiones regulares (REGEX) y el segundo método basado en una red neuronal (NN, por sus siglas en inglés) artificial entrenada con literatura médica. La evaluación presentada en el texto original se dividió en tres sub-tareas: reconocimiento de síntomas de una sola palabra (Monomio), reconocimiento de síntomas de dos

[4] La biblioteca Apache OpenNLP proporciona un modelo NER basado en MaxEnt. `https://opennlp.apache.org`

[5] El Stanford CoreNLP ofrece software NER basado en CRF. `https://stanfordnlp.github.io/CoreNLP/`

[6] El paquete Spacy proporciona un módulo NER basado en CNN. `https://spacy.io/usage/training`

palabras (Binomio) y reconocimiento de síntomas de más de dos y hasta siete palabras (n-grama). Las métricas utilizadas fueron *precision* y *recall*. Los resultados mostrados en la Tabla 3 nos revelan aspectos muy interesantes de cara al procesamiento de textos médicos. Se observa que, aunque el método REGEX está limitado a detectar únicamente los términos en un lexicon pre-definido, es muy preciso al ser utilizado en literatura especializada, como es el caso de la medicina. Sin embargo, su nivel de *recall* cae rápidamente al aumentar la complejidad de los términos a detectar. El método NN, por el contrario, tiene una *precision* ligeramente menor, pero su cobertura es mejor detectando entidades de mayor complejidad. Esto se debe a que, a diferencia de REGEX, el método NN no depende de un léxico fijo de términos y esto lo hace robusto en el reconocimiento de patrones que aprende durante la fase de entrenamiento. Incluso, otra propiedad interesante del método NN es que es capaz de detectar entidades con errores ortográficos.

Tabla 3: Resultados de reconocimiento de entidades nombradas para un sistema médico usando dos diferentes métodos.

	Precision Monomio	**Recall** Monomio	**Precision** Binomio	**Recall** Binomio	**Precision** n-gram	**Recall** n-gram
REGEX	1.0000	0.6821	1.0000	0.5748	0.9117	0.4033
NN	0.9120	0.6171	0.9545	0.7608	0.8379	0.5669

Una buena alternativa para extender las bondades del método REGEX en textos médicos se puede encontrar en el proyecto covidminer[7], el cual utiliza ontologías para detectar conceptos relacionados al COVID-19, síntomas, menciones de muestreos, comorbilidades, a partir de información obtenida de expertos y de Wikidata. El software se usó para analizar notas médicas que fueron proporcionadas por autoridades médicas durante el transcurso de la pandemia de COVID-19 y que requerían soluciones precisas y rápidas de implementar. A partir de un primer léxico médico especializado durante el desarrollo de la pandemia en México, se establecieron las entidades de interés a encontrar en las notas médicas. Usando este recurso, se elaboró una ontología *ad hoc*, en la cual los términos a encontrar fueron relacionados con nuevos términos y unificados mediante un identificador único. Por ejemplo, entre el léxico de los síntomas de interés, el término *disnea*, de la Figura 1, puede ocurrir de esta manera, pero también puede ser referido como *dificultad respiratoria*. Así, para poder llevar a cabo una detección y conteos efectivos de las menciones de este síntoma, existe un ID $Q188008$, el cual refiere al síntoma como un concepto, independientemente de la manera en que es mencionado en los textos, tal y como se muestra en el siguiente extracto. Usando esta metodología, se puede contar el número de veces que los pacientes refirieron haber tenido algún síntoma de COVID-19 y contras-

[7] El repositorio covidminer proporciona un módulo completo para NER en Español basado en REGEX y ontologías. `https://github.com/alemol/covidminer`

tar con las menciones en los diagnósticos médicos, pues ambos se relacionan a través del concepto con un ID como se muestra en la Figura 1.

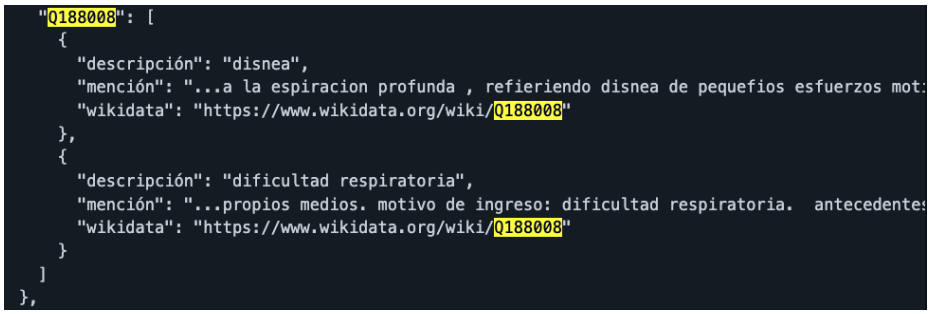


Figura 1: Ejemplo de detección de síntomas de COVID-19 mediante ontologías. Obtenido con el software covidminer: *A Text Mining Emergent Library for Information Extraction from Medical Notes in Spanish during COVID-19 pandemic*, disponible en GitHub.

Para finalizar esta sección, mencionaremos que, hoy en día, los métodos de NER más utilizados están basados en modelos matemáticos. Por su relevancia, en la sección siguiente nos centraremos en modelos neuronales de aprendizaje profundo, a partir de los cuales no solamente es posible caracterizar entidades nombradas en el contexto médico, sino que también son, por sí mismos, modelos del lenguaje que permiten una gran variedad de aplicaciones en el procesamiento de textos.

Word Embeddings. Los Word Embeddings son representaciones vectoriales de palabras que se pueden extender a representaciones de documentos. Estos vectores se generan con el propósito de codificar información textual mediante espacios métricos construidos para modelar la semántica del lenguaje. Desde el punto de vista del aprendizaje automático, los Word Embeddings son útiles porque, a través de vectores densos de números reales, son capaces de representar características de los textos. Esto permite encontrar relaciones mediante operaciones de álgebra lineal, precisamente con la intención de poder modelar la semántica del lenguaje natural a través de una abstracción del significado. Además, desde su origen, los Word Embeddings han sido pensados como un método de codificación que facilita a ciertos algoritmos a reconocer patrones, particularmente los algoritmos de las redes neuronales.

Si bien la generación de Word Embeddings tiene fundamentos similares, independientemente de la implementación, hay diferentes formas de obtenerlos. Así, podemos agrupar las maneras específicas de generar Word Embeddings observando las tareas que resuelven, así como los retos específicos del dominio para el que se construyen.

Existe un primer grupo de Word Embeddings pre-entrenados de propósito general, la gran mayoría de estos basados en modelos neuronales. Por citar algunos mencionamos: *word2vec* [29], *Glove* [37] y *fastText* [6]. Una variante más reciente, también de propósito general, son los llamados Word Embeddings contextuales, basados en modelos de aprendizaje profundo Deep Learning [5], entre los que figuran: *ELMo* [38] o *BERT* [9].

Otro grupo corresponde a la construcción de Word Embeddings específicos para resolver un problema de aprendizaje como la clasificación. Entre estos, es de interés citar trabajos de investigación en los cuales se han creado Word Embeddings específicos para el español de México y que han servido en proyectos como la clasificación automática de solicitudes de atención del servicio de locatel de la Ciudad de México[8] [32] y la clasificación de mensajes con discursos de odio en redes sociales [30], [8], [3]. En este grupo, es de especial interés también el trabajo propuesto en [33], donde se procesaron textos de transcripciones médicas de pacientes con necesidad de realizar una cirugía y pacientes que simplemente iban a una consulta médica regular para luego clasificarlos de manera automática usando una representación espectral para asociar características de imágenes médicas.

Por último, también específicamente en el área medicina, pero exclusivamente para inglés podemos mencionar el proyecto BioBERT [22], el cual es un modelo de Word Embeddings contextuales de textos médicos, el cual fue entrenado con resúmenes provenientes de PubMed y textos completos provenientes de revistas biomédicas. También, el modelo ClinicalBERT [4], el cual fue entrenado con notas médicas extraídas de los datos de MIMIC-III database [17].

3.2. Imágenes

El procesamiento de imágenes con algoritmos de inteligencia artificial tuvo una gran evolución con la incorporación del aprendizaje profundo. La arquitectura de red neuronal que lo cambió todo fue, sin duda, la red neuronal convolucional (CNN, por sus siglas en inglés). Las redes neuronales convolucionales procesan imágenes con el objetivo de extraer sus características. Esta red es ampliamente utilizada en aplicaciones como el reconocimiento de objetos en imágenes [12]. Las CNNs se componen de dos partes: la parte convolucional y la parte densamente conectada. La parte convolucional consiste en realizar operaciones teniendo en cuenta la naturaleza matricial de una imagen.

Una imagen está constituida por píxeles, los cuales representan un determinado color a partir de valor un numérico, generalmente entre el rango de 0 a 255. Cada uno de estos píxeles colorea una región de una imagen y, al tener mayor cantidad de píxeles, es posible representar imágenes más complejas. Como se

[8] El 0311 Locatel es el sistema de reportes de servicios urbanos y atención de no emergencia de la Ciudad de México. Una Inteligencia Artificial desarrollada por un equipo de CentroGeo fue transferida e incorporada al servicio de la CDMX. `https://311locatel.cdmx.gob.mx`

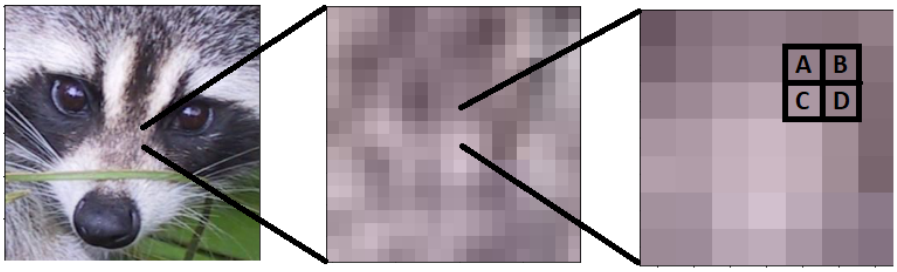


Figura 2: Estructura de una imagen.

puede observar en la Figura 2, los píxeles A, B, C, D son solo una pequeña parte de una imagen mucho más grande.

Operaciones convolucionales. Una convolución hace referencia al proceso de extraer características de una imagen aplicando un filtro. El filtro recorre la imagen realizando operaciones sobre píxeles contiguos, con el objetivo de extraer patrones.

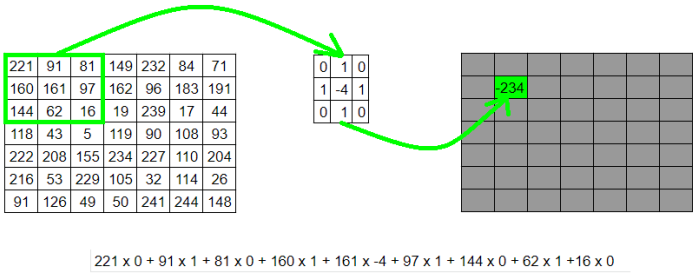


Figura 3: Ejemplo de convolución en una matriz de dimensiones 7×7.

En la Figura 3 se puede observar cómo se realiza una convolución a una matriz 7×7. En la región resaltada se aplica el filtro, es decir, se va a realizar el producto punto entre los valores de la región y los valores del filtro. Eso, como resultado, nos dará un equivalente convolucional de dicho píxel, el cual ha extraído características de los valores adyacentes. Los patrones que se obtienen dependen del filtro que se haya utilizado. Algunos filtros son capaces de encontrar características especiales, como eliminar fondos con el objetivo de solo fijarse en el objeto enfocado, resaltar las líneas verticales de la imagen, resaltar las regiones con mayor contraste, entre otros. Al aplicar un filtro a la imagen, ciertos rasgos se pueden resaltar, por ejemplo, el color negro en la imagen original, dando como resultado una mayor representación del pelaje característico del mapache, como los son sus bigotes y el pelaje que rodea los ojos.

Junto con estas operaciones también se realizan los procesos de *Pooling* y de *Flattening*. El proceso de *Pooling* consiste en reducir la cantidad de características, con el objetivo de acelerar tiempos de procesamiento y reducir el uso de memoria. En el caso de *Flattening*, los algoritmos de aprendizaje automático reciben como entrada vectores unidimensionales, por lo cual se hace necesario convertir las imágenes que son bidimensionales a un vector de una sola dimensión.

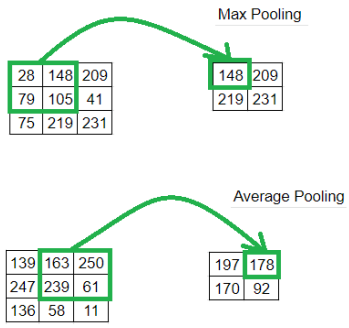


Figura 4: Ejemplo de la operación de *Pooling* en una red de convolución.

En la Figura 4 se puede observar cómo funciona el método de *pooling* en dos instancias diferentes. En el caso *Max Pooling* se selecciona una región de la matriz original y de esa región se obtiene el mayor valor, esta operación se repite hasta recorrer toda la matriz original. Otra forma de obtener estos valores es a través de *Average Pooling*, donde se calcula el promedio de la región seleccionada. De igual forma, se repite este proceso hasta recorrer toda la matriz original. El resultado final es generalmente una imagen de dimensiones mucho más pequeñas que la imagen original, pero que conserva la información necesaria para realizar operaciones de aprendizaje automático, reduciendo costos computacionales.

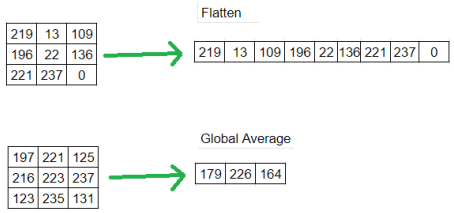


Figura 5: Ejemplo de *Flattening* en una red de convolución.

En la Figura 5 se puede observar cómo funciona el *flattening* para dos instancias diferentes. En el caso de aplanamiento sencillo (*Flatten*) simplemente se concatena cada fila de la matriz, una detrás de la otra, obteniendo la representación de todos los valores de la matriz en un vector unidimensional. Otra forma de convertir la matriz a un vector es obteniendo el valor promedio de cada una de sus columnas (*Global Average*), donde se calcula el promedio de cada columna, con el objetivo de obtener un vector representativo de la matriz de tamaño equivalente a su cantidad de columnas. Este proceso se realiza puesto que los algoritmos de aprendizaje automático utilizan vectores como parámetros de entrada, por lo cual una imagen debe ser transformada a un vector para que el algoritmo de aprendizaje pueda procesarla.

Procesamiento de imágenes médicas para apoyo al diagnóstico. En [33] se detalla un enfoque capaz de extraer las características de varias modalidades (tipos de dato) y obtener una representación unificada que pueda ser usada para obtener un modelo que describa el problema de manera conjunta. Es relevante mencionar que, para la modalidad de imagen, se usaron imágenes médicas como caso de estudio, para lo cual se utilizó un conjunto de datos de 1619 instancias de imágenes colorrectales [19], donde se tienen pruebas visuales de tejido del colon de pacientes, las cuales son examinadas con el objetivo de encontrar signos de cáncer colorrectal.

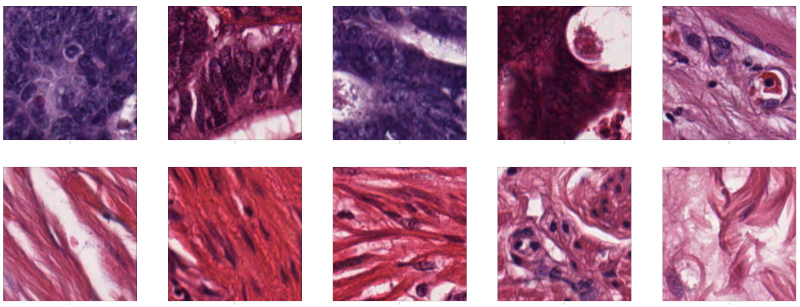


Figura 6: Instancias de imágenes colorrectales.

Como se puede observar en la Figura 6, la primera fila hace referencia a tejido con señales de tumores, mientras que en la segunda fila se muestra tejido sano. El modelo de aprendizaje automático presentado clasifica cada imagen con su correspondiente etiqueta. Para lograrlo, utiliza el modelo *Inception*, una CNN para extraer las características de cada clase y usarlas como un vector representativo. Luego, el vector representativo es utilizado como una de las modalidades que permiten a un clasificador determinar signos de cáncer colorrectal.

3.3. Series de tiempo de signos vitales

Los signos vitales son la cuantifación de acciones fisiológicas, como la frecuencia cardiaca (FC), la frecuencia respiratoria (FR), la temperatura corporal (TC), la presión arterial (PA) y la saturación de oxígeno en sangre (SpO2), que indican la calidad del funcionamiento orgánico de una persona. La PA puede analizarse en su valor sistólico o diastólico, PA_s y PA_d, respectivamente, así como por su valor medio (MAP, por sus siglas en inglés). La monitorización continua de los signos vitales de un paciente es un procedimiento habitual en la práctica clínica, lo cual permite obtener una serie de tiempo por cada signo vital observado. Una serie de tiempo es un conjunto de muestras obtenidas mediante instrumentos de monitorización, las cuales se encuentran ordenadas de forma sucesiva. De manera formal, sea X un conjunto de muestras X_t de un signo vital donde cada una ocurre de forma aleatoria en un tiempo específico t.

Las series de tiempo permiten a los profesionales de la salud comprender la condición que guarda un paciente en términos de su estado fisiológico, permitiendo reconocer *episodios adversos* asociados al deterioro de su salud. En la Tabla 4 se muestran episodios adversos para los signos vitales anteriormente mencionados. En particular, dado que se conocen los valores de referencia considerados como normales para cada signo vital, es posible definir umbrales que pueden ser un indicio de una posible afectación fisiológica. A fin de ilustrar la representación y modelado de series de tiempo, a continuación se considera el caso de episodios agudos hipotensivos (AHE, por sus siglas en inglés), los cuales son episodios graves que ocurren cuando la MAP decae por debajo de un umbral no deseado durante un periodo de tiempo.

Representación de series de tiempo de MAP. Desde el punto de vista médico, la presión sanguínea se define como la fuerza que ejerce la sangre contra las paredes de las arterias y es medida con un esfigmomanómetro. El comportamiento de este signo vital puede verse influenciado por factores como edad, género, hormonas, medicamentos, fiebre y hemorragias, los cuales pueden afectar a dicha variable fisiológica. Tomando en cuenta que la actividad del corazón se compone de distintos momentos, usualmente la medición de la presión sanguínea se realiza considerando sus valores sistólicos y diastólicos. El primer tipo se refiere a la presión dentro de las arterias cuando el corazón se contrae y bombea sangre a través del cuerpo; en adultos sanos el valor de referencia comúnmente considerado es 120 mmHg. La presión diastólica es la presión dentro de las arterias cuando el corazón está en descanso (es decir, es la presión entre los latidos del corazón) y su valor de referencia para adultos sanos es 80 mmHg. Además de lo anterior, en la práctica clínica resulta de mayor valor informativo emplear la MAP como indicador de posibles eventos adversos en la presión sanguínea. En particular, mediante la medición de la presión arterial sistólica y la diastólica es posible calcular un valor medio de la presión arterial como $\text{MAP} = \frac{\text{PA}_\text{s}+2*\text{PA}_\text{d}}{3}$. Los valores de MAP superiores a 60 mmHg son considerados como normales para mantener los órganos de una persona funcionando correctamente. Por tanto, valores de MAP inferiores a dicho valor observados durante un periodo apro-

ximados de 30 minutos puede inducir un AHE u otros episodios adversos que pueden ocasionar daños irreversibles en los órganos del paciente.

Tabla 4: Ejemplos de episodios adversos en series de tiempo de signos vitales. Notación: lpm (latidos por min), rpm (respiraciones por min), mmHg (milímetros de mercurio).

Signo vital	Episodio adverso	Descripción
FC	Taquicardia	FC ¿100 lpm
	Bradicardia	FC ¡60 lpm
TC	Fiebre o hipertermia	TC ¿38°
	Hipotermia	TC ¡35.5°
FR	Bradipnea	FR ¡12 rpm
	Taquipnea	FR ¿20 rpm
	Apnea	Sin respiración al menos 20 min
PA	Hipertensión	PA_s ¿120 mmHg
		PA_d ¿80 mmHg
	Hipotensión	PA_s ¡120 mmHg
		PA_d ¡80 mmHg
SpO2	Desaturación leve	SpO2 entre 93-95 %
	Desaturación moderada	SpO2 entre 88-92 %
	Desaturación grave	SpO2 ¡88 %

El análisis de series de tiempo presenta desafíos debido a que las observaciones realizadas al signo vital guardan un orden temporal y natural. Dicho análisis se puede aplicar al caso de datos continuos de valores reales, datos numéricos discretos o datos categóricos discretos. A continuación, se describen las propiedades generales a observar en las series de tiempo de signos vitales.

- *Tiempo discreto.* Las muestras del signo vital son obtenidas en intervalos de tiempo regulares, expresadas en segundos, minutos, etc.
- *Orden natural.* Representa el orden cronológico de las muestras, el cual debe preservarse durante el análisis y procesamiento debido a que éste establece una relación entre las muestras recabadas durante un intervalo determinado.

A manera de ejemplo, en la Figura 7 se ilustra una serie de tiempo de MAP obtenida de la base de datos MIMIC-II. Esta serie de tiempo representa la evolución de la MAP durante un periodo de 2 horas y 10 minutos, considerando un muestreo a intervalos regulares de un minuto. En el periodo indicado en la gráfica en color amarillo, aproximadamente el 90 % de las muestras obtuvo un valor de MAP$\leq$ 60 mmHg. Por lo tanto, es importante preservar la propiedad del orden de las muestras durante el procesamiento de las series de tiempo.

En [10] se describe un método de representación que permite abstraer ciertas características de interés de las series de tiempo de MAP para el modelado y predicción de AHEs. La representación empleada se ilustra en la Figura 8,

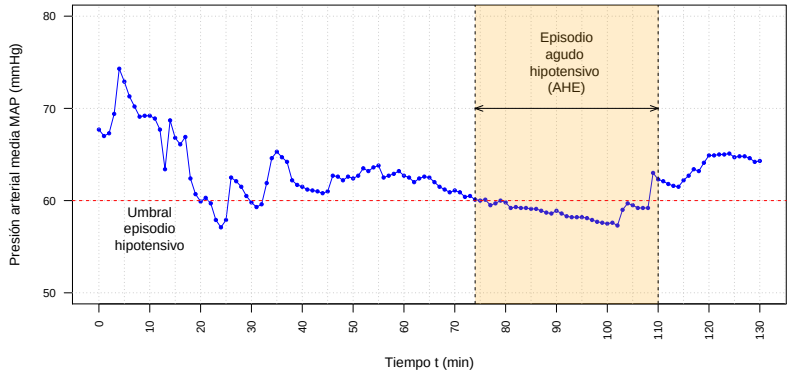


Figura 7: Ejemplo de una serie de tiempo de MAP.

considerando una serie de tiempo de $n = 90$ muestras u observaciones obtenidas cada minuto. Cada muestra $i = 1, \ldots, n$ sufre una transformación binaria B_i de acuerdo con lo siguiente: se asigna un valor de 1 si la muestra de MAP es inferior o igual al umbral del episodio hipotensivo (es decir, $B_i = 1$); o bien se asigna 0 ($B_i = 0$) en caso de que el valor de MAP se encuentre por encima del umbral hipotensivo. La gráfica en la parte superior de la Figura 8 corresponde a la serie de tiempo original, mientras que la gráfica en la parte inferior es el resultado de haber aplicado la representación descrita. De esta forma es posible obtener una señal discreta que contiene valores binarios únicamente. La representación resultante posibilita la obtención de una secuencia de estados asociados a la presencia/ausencia de un AHE.

4. Conclusiones

Los procesos de atención médica son generadores de datos que registran diferentes escenarios y estados de salud de los pacientes, típicamente en la forma de variables fisiológicas y socio-económicas supeditadas a estructuras tabulares y relacionales propias de la institución médica que las genera. A partir de estas variables es posible realizar análisis numéricos y estadísticos en aras de encontrar modelos que apoyen el proceso de toma de decisión. En este escenario, son de especial atención aquellas variables en las que, por su naturaleza, es imposible realizar dicho análisis de forma directa y que requieren, por lo tanto, un proceso de transformación. En las secciones previas se presentaron algunos enfoques ampliamente usados para lograr dicha transformación.

Alrededor del proceso de atención también se generan datos en otros formatos provenientes, por ejemplo, de estudios de laboratorio, imagenología, notas médicas, entre otros. La variedad de formatos y la naturaleza no estructurada de éstos hacen necesaria la ejecución de tareas de procesamiento que permitan

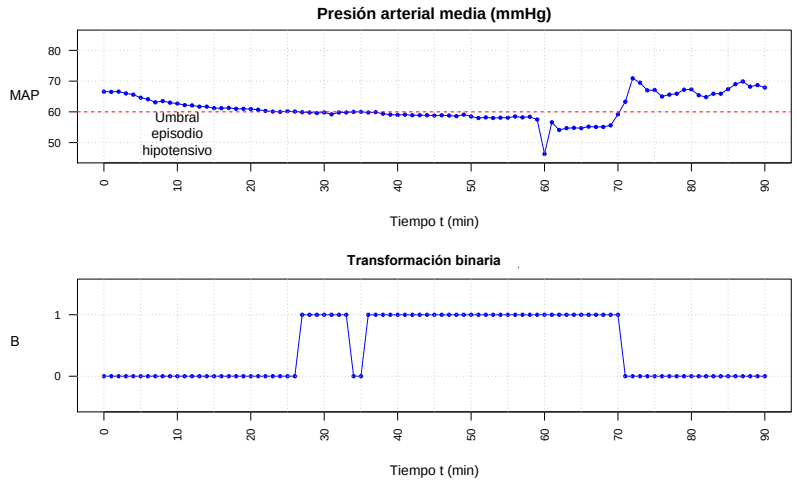


Figura 8: Transformación de serie de tiempo.

extraer propiedades susceptibles de análisis numérico con el fin de diseñar modelos que permitan apoyar el proceso de toma de decisiones en diferentes contextos y problemas de estudio relacionados con la salud de los pacientes. En las secciones previas se presentaron algunos ejemplos de extracción de características en texto, imágenes y series de tiempo que pretenden vislumbrar algunos escenarios posibles en el procesamiento de datos clínicos no estructurados.

Los procesos de transformación presentados están correlacionados con la efectividad de los modelos analíticos. Por ejemplo, una exhaustiva y contundente extracción de características de imágenes de un órgano blanco guiará el proceso de aprendizaje y obtención de modelos de predicción con alta efectividad de predicción.

Referencias

[1] Amir Ahmad y Lipika Dey. "A k-mean clustering algorithm for mixed numeric and categorical data". En: *Data & Knowledge Engineering* 63.2 (2007), págs. 503-527.

[2] Edwin Aldana-Bobadilla et al. "Adaptive Geoparsing Method for Toponym Recognition and Resolution in Unstructured Text". En: *Remote Sensing* 12.18 (2020), pág. 3041.

[3] Edwin Aldana-Bobadilla et al. "A language model for misogyny detection in Latin American Spanish driven by multisource feature extraction and transformers". En: *Applied Sciences* 11.21 (2021), pág. 10467.

[4] Emily Alsentzer et al. "Publicly available clinical BERT embeddings". En: *arXiv preprint arXiv:1904.03323* (2019).

[5] Yoshua Bengio et al. "A Neural Probabilistic Language Model". En: *J. Mach. Learn. Res.* 3 (2003), págs. 1137-1155.
[6] Piotr Bojanowski et al. "Enriching Word Vectors with Subword Information". En: *Transactions of the Association for Computational Linguistics* 5 (2017), págs. 135-146.
[7] Piter Henry Escobar Callejas y Jorge Lu's Bilbao Ramirez. *Gu'a Metodol—gica para la Investigaci—n Cient'fica: para grado y Posgrado*. Lulu. com.
[8] Thomas Cattin et al. "The Geopolitical Repercussions of US Anti-immigrant Rhetoric on Mexican Online Speech About Migration: A Transdisciplinary Approach". En: *International Conference on Geospatial Information Sciences*. Springer. 2022, págs. 41-51.
[9] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". En: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, jun. de 2019, págs. 4171-4186.
[10] Jaime Edwin Arciniegas García. *Método de predicción de episodios hipotensivos basado en una codificación de series de tiempo de presión arterial media y cadenas de Markov*. 2019.
[11] David W Goodall. "A new similarity index based on probability". En: *Biometrics* (1966), págs. 882-907.
[12] Ian Goodfellow, Yoshua Bengio y Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: `http://www.deeplearningbook.org`.
[13] Melissa A Hardy. *Regression with dummy variables*. Vol. 93. Sage, 1993.
[14] Gandhi Hernández-Chan et al. "Medic-Us: Advanced Social Networking for Intelligent Medical Services and Diagnosis". En: *Current Trends in Semantic Web Technologies: Theory and Practice*. Springer, 2019, págs. 213-232.
[15] Chung-Chian Hsu. "Generalizing self-organizing map for categorical data". En: *IEEE transactions on Neural Networks* 17.2 (2006), págs. 294-304.
[16] Zhexue Huang. "Clustering large data sets with mixed numeric and categorical values". En: *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*. Citeseer. 1997, págs. 21-34.
[17] Alistair EW Johnson et al. "MIMIC-III, a freely accessible critical care database". En: *Scientific data* 3.1 (2016), págs. 1-9.
[18] Rute Mattos Dourado Esteves Justa et al. "La agresividad tumoral está asociada a las alteraciones de la integridad celular en las mujeres que sobreviven al cáncer de mama: estudio de seguimiento". En: *Nutrición Hospitalaria* 39.1 (2022), págs. 138-146.
[19] Jakob Nikolas Kather et al. "Multi-class texture analysis in colorectal cancer histology". En: *Scientific reports* 6 (2016), pág. 27988.
[20] Tícia Kocsis et al. "Probiotics have beneficial metabolic effects in patients with type 2 diabetes mellitus: a meta-analysis of randomized clinical trials". En: *Scientific reports* 10.1 (2020), págs. 1-14.

[21] Teuvo Kohonen. "The self-organizing map". En: *Proceedings of the IEEE* 78.9 (1990), págs. 1464-1480.

[22] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". En: *Bioinformatics* 36.4 (2020), págs. 1234-1240.

[23] Cen Li y Gautam Biswas. "Unsupervised learning with mixed numeric and nominal data". En: *IEEE Transactions on knowledge and data engineering* 14.4 (2002), págs. 673-690.

[24] Ivan Lopez-Arevalo et al. "A Memory-Efficient Encoding Method for Processing Mixed-Type Data on Machine Learning". En: *Entropy* 22.12 (2020). ISSN: 1099-4300. DOI: 10.3390/e22121391. URL: https://www.mdpi.com/1099-4300/22/12/1391.

[25] Huilan Luo, Fansheng Kong y Yixiao Li. "Clustering mixed data based on evidence accumulation". En: *International Conference on Advanced Data Mining and Applications*. Springer. 2006, págs. 348-355.

[26] Naresh K Malhotra. *Investigación de mercados: un enfoque aplicado*. Pearson educación, 2004, págs. 65-66.

[27] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman y Hall/CRC, 2020.

[28] Daniele Micci-Barreca. "A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems". En: *ACM SIGKDD Explorations Newsletter* 3.1 (2001), págs. 27-32.

[29] Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". En: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. por Yoshua Bengio y Yann LeCun. 2013. URL: http://arxiv.org/abs/1301.3781.

[30] Alejandro Molina-Villegas. "La incidencia de las voces misóginas sobre el espacio digital en México". En: *Jóvenes, Plataformas Digitales y Lenguajes: Diversidad Lingüística, Discursos e Identidades*. Página Seis, 2022, págs. 39-61.

[31] Alejandro Molina-Villegas et al. "Geographic named entity recognition and disambiguation in Mexican news using word embeddings". En: *Expert Systems with Applications* 176 (2021), pág. 114855.

[32] Alejandro Molina-Villegas et al. "Incorporating Natural Language Processing models in Mexico City's 311 Locatel". En: *LatinX in Natural Language Processing Research Workshop at NAACL 2022*. Seattle: North American Chapter of the Association for Computational Linguistics (NAACL), 2022.

[33] Hernán Guillermo Dulcey Morán. *Modelo de aprendizaje multimodal aplicado al diagnóstico de padecimientos clínicos*. 2021.

[34] Mohammad Naghi Namakforoosh. *Metodología de la investigación*. Editorial Limusa, 2000, pág. 223.

[35] Nhung TH Nguyen, Roselyn S Gabud y Sophia Ananiadou. "COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature". En: *Biodiversity data journal* 7 (2019).

[36] Charlene Jennifer Ong et al. "Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports". En: *PloS one* 15.6 (2020), e0234908.

[37] Jeffrey Pennington, Richard Socher y Christopher D Manning. "Glove: Global vectors for word representation". En: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, págs. 1532-1543.

[38] Matthew E Peters et al. "Deep contextualized word representations". En: *arXiv preprint arXiv:1802.05365* (2018).

[39] Tim Rocktäschel, Michael Weidlich y Ulf Leser. "ChemSpot: a hybrid system for chemical named entity recognition". En: *Bioinformatics* 28.12 (2012), págs. 1633-1640.

[40] Satoshi Sekine y Chikashi Nobata. "Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy." En: *LREC*. Lisbon, Portugal. 2004.

[41] Joan Serrà y Alexandros Karatzoglou. "Getting Deep Recommenders Fit: Bloom Embeddings for Sparse Binary Input/Output Networks". En: *Proceedings of the Eleventh ACM Conference on Recommender Systems*. RecSys '17. Como, Italy: Association for Computing Machinery, 2017, 279–287. ISBN: 9781450346528.

[42] David A Smith y Gregory Crane. "Disambiguating geographic names in a historical digital library". En: *Research and Advanced Technology for Digital Libraries*. Springer, 2001, págs. 127-136.

[43] NV Sobhana, Pabitra Mitra y SK Ghosh. "Conditional random field based named entity recognition in geological text". En: *International Journal of Computer Applications* 1.3 (2010), págs. 143-147.

[44] Ali Hassan Sodhro et al. "An energy-efficient algorithm for wearable electrocardiogram signal processing in ubiquitous healthcare applications". En: *Sensors* 18.3 (2018), pág. 923.

[45] Lorraine Tanabe et al. "GENETAG: a tagged corpus for gene/protein named entity recognition". En: *BMC bioinformatics* 6.1 (2005), pág. 1.

[46] Kilian Weinberger et al. "Feature hashing for large scale multitask learning". En: *Proceedings of the 26th annual international conference on machine learning*. 2009, págs. 1113-1120.

[47] Kyi Pyar Win et al. "Computer-assisted screening for cervical cancer using digital image processing of pap smear images". En: *Applied Sciences* 10.5 (2020), pág. 1800.

[48] Nina Zumel y John Mount. "vtreat: a data. frame Processor for Predictive Modeling". En: *arXiv preprint arXiv:1611.09477* (2016).

Sistemas Inteligentes de e-Salud

Roberto Conte Galván[1], Alejandro Galaviz-Mosqueda*[,2], Salvador Villarreal-Reyes[1], Jose Lozano-Rizk[1] y Raúl Rivera-Rodríguez[1]

*agalaviz@cicese.mx

[1] Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California, Carr Ensenada-Tijuana 3918, Zona Playitas, 22860, Ensenada, B.C., México.
[2] Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California Unidad Monterrey, Apodaca 66629, Nuevo León, México

Resumen. Los sistemas de e-Salud son sistemas que incluyen las Tecnologías de la Información y la Comunicación (TICs) para la provisión de servicios de salud. Los sistemas de e-Salud pueden contribuir a mejorar los servicios de salud y abordar retos como la falta de personal especializado. A través del uso de TICs es factible ampliar la cobertura y el acceso efectivo a servicios de salud, tanto básicos como de especialidad. De manera adicional, en conjunto con algoritmos de inteligencia artificial, los sistemas de e-Salud pueden ofrecer nuevos servicios para detección temprana, seguimiento y diagnóstico asistido de diversas enfermedades. Sin embargo, para poder desplegar los sistemas inteligentes de e-Salud es importante abordar diversos retos científicos y tecnológicos; entre los más importantes se encuentran los relacionados con las redes de datos, seguridad e inteligencia artificial. Se discutirán aspectos relevantes de sistemas y redes de telemedicina para desplegar un sistema de e-Salud ubicuo. En cada sección se proveen referencias que se consideran relevantes y que pueden servir para que se amplíe la exploración de los temas tratados

Palabras clave: Telemediciona · e-Salud · Sistemas Inteligentes.

1 Introducción

Las personas nacen, viven, trabajan, descansan y viajan a todo tipo de lugares y climas alrededor del mundo. Nuestra naturaleza activa e inquisitiva nos lleva a los lugares más extremos del mundo y nuestro comportamiento gregario nos permite convivir tanto en comunidades aisladas como superpobladas. Independientemente de nuestra ubicación geográfica, nuestro cuerpo lleva nuestra salud y nuestras enfermedades a todos esos lugares y, en caso de que surja algún accidente o problema de salud, generalmente esperamos ser atendidos de manera rápida y eficiente. La tecnología de la telemedicina actual brinda acceso a diagnósticos, tratamientos y supervisión remotos de médicos y hospitales calificados en varias regiones del mundo. Los dispositivos,

servicios y sistemas de atención médica digital están actualmente disponibles en muchos entornos urbanos y suburbanos. Sin embargo, para los numerosos grupos de población desfavorecidos de todo el mundo, ya sean indígenas, rurales, pobres, ancianos o gravemente enfermos, a menudo es difícil llegar a instalaciones sanitarias adecuadas, así como acceder a personal sanitario y médico calificado.

Las tecnologías de telesalud y telemedicina han demostrado un gran potencial para mejorar las condiciones de salud en las poblaciones marginadas de todo el mundo. Inherentemente, las soluciones de telesalud y telemedicina se basan en los desarrollos y avances tecnológicos alcanzados por las tecnologías de la información y comunicaciones (TICs), habilitando los sistemas de e-Salud. Específicamente, en este capítulo de libro, la definición de los sistemas de e-Salud se toma de Novillo-Ortiz [39]: uso coste-efectivo y seguro de las TICs en apoyo a la salud y de los ámbitos relacionados con la salud, incluyendo los servicios de atención sanitaria, vigilancia de la salud, literatura y educación, conocimiento e investigación. Prácticamente, todas las soluciones basadas en TICs (incluidas las tecnologías de e-Salud, telesalud y telemedicina) utilizan en algún momento el espectro radioeléctrico y la infraestructura de telecomunicaciones desplegada actualmente. A nivel mundial, la Unión Internacional de Telecomunicaciones (UIT) busca armonizar a las TICs y los servicios que se derivan de su despliegue, con el fin promover su adopción y el acceso equitativo a los beneficios que dichas tecnologías pueden proveer. Cabe mencionar que la UIT es el brazo TIC de la Organización de las Naciones Unidas (ONU) [49]. La Organización Mundial de la Salud (OMS) encargó en 2016 una encuesta global sobre e-Salud con el objetivo de explorar los desarrollos globales en el área. Esta encuesta indicó que la cobertura sanitaria universal no podría ser posible sin el apoyo de e-Salud [59]. Basado en estas iniciativas de la OMS y la UIT, este capítulo enfatiza el papel de los desarrollos de tecnología digital nuevos y emergentes para proponer un marco de atención médica universal viable y completo.

Teniendo en cuenta el desafío global para lograr los 17 objetivos de desarrollo de la ONU, es urgente y fundamental establecer un marco para brindar atención médica a largo plazo eficiente, efectiva y asequible a toda la población mundial. Sin embargo, dado que muchos factores no relacionados con la salud específicos de cada país afectan los resultados de salud, los enfoques actuales tienden a exacerbar el escenario intrincado de brindar servicios de atención médica para toda la humanidad. Estos factores han creado una "brecha de atención médica" humana entre la población mundial que tiene y la que no tiene atención médica. Son muchos los problemas a los que se enfrenta una adecuada atención sanitaria en todo el mundo. Estos problemas son particularmente agudos para sectores de población en estratos socioeconómicos bajos, los cuales, en muchos casos, no tienen acceso efectivo a algún servicio de salud. Adicionalmente, los flujos migratorios actuales presentes, en todos los estratos socioeconómicos, hacen que sea un reto el dar un seguimiento continuo a trastornos de salud crónicos, ya que todos requerimos atención sanitaria a lo largo de nuestra vida dondequiera que estemos, creando un desafío muy complejo para los diversos sistemas de salud en todo el mundo. Por ejemplo, las patologías comúnmente diagnosticadas y tratadas en algunas partes del mundo pueden ser desconocidas para los médicos en otros lugares diferentes, lo que, a menudo, coloca al personal de atención médica competente en

situaciones difíciles debido al turismo extranjero o a los problemas de población migrante. La Organización para la Cooperación y el Desarrollo Económicos (OCDE) ha expresado su preocupación por el importante movimiento de trabajadores de la salud desde el cambio de siglo. La migración y el flujo de médicos, dentistas, enfermeros, farmacéuticos y cuidadores de ancianos dependientes entre los países de la OCDE, así como entre la OCDE y otros países [41], ha creado presiones sobre las tasas de capacitación de la fuerza laboral médica y de la salud en varios países.

El propósito principal de cualquier sistema de salud es brindar un servicio médico y de salud eficiente a un gran número de personas y sus familias inmediatas o parientes, a través de una red de profesionales de la salud (médicos, enfermeras, administradores de salud, trabajadores sociales, etc.). Idealmente, estos profesionales de la salud deberían trabajar en instalaciones altamente eficientes y conectadas, con la ayuda de un número cada vez mayor de sistemas y dispositivos electrónicos, informáticos y de comunicaciones, una gran parte de ellos ya conectados a Internet. Si bien esto se podría cumplir en países desarrollados, en los países en vías de desarrollo dicha infraestructura solo se encuentra disponible en grandes centros urbanos, dejando amplios sectores de la población sin acceso a dicho tipo de servicios. Por lo tanto, se necesita la transformación digital del sistema de atención médica existente para lograr la integración y el funcionamiento continuo de la infraestructura y los sistemas de atención médica para un servicio ubicuo y global.

La OCDE recomienda una densidad de médicos de 2.3 trabajadores de la salud como promedio (2.2 médicos y 2.6 enfermeras) por 1,000 habitantes, mientras que la Organización Mundial de la Salud (OMS) recomienda una densidad de médicos de 3.4 trabajadores de la salud calificados por 1,000 [57], y un mínimo de 2.3 [58]. Si bien dichas cifras son una recomendación que idealmente debería ser cubierta, existen diferentes criterios en cuanto a la educación y formación en salud profesional y vocacional en todo el mundo. También, es necesario considerar las diferencias entre países respecto a las normas, actividades y tareas requeridas para cada trabajador de la salud. En este contexto, es difícil encontrar datos estadísticos de salud precisos respecto al número de profesionales de la salud disponible en cada país, ya que cada uno maneja sus propias definiciones, datos y cifras de personal de salud de manera diferente.

Cualesquiera que sean las definiciones y las cifras de la fuerza laboral médica y de la salud que se utilicen, es común encontrar que, en países en desarrollo, e incluso desarrollados, puede haber un tipo similar de irregularidades geográficas con respecto a la distribución del personal médico. Por ejemplo, puede haber ciertas áreas geográficas urbanas en países desarrollados donde la distribución promedio de médicos locales sea mucho mayor que las recomendaciones de la OCDE o la OMS. En contraste, también pueden existir otras áreas dentro del mismo país donde la distribución de médicos sea mucho menor que las cifras encontradas en países en desarrollo. Las cifras de densidad de médicos publicadas por las autoridades sanitarias de cada país son siempre una media y no siempre están actualizadas. Por lo tanto, las irregularidades solo aparecen cuando se realiza un análisis más detallado. Aunque las cifras publicadas por la OMS, la OCDE y el Banco Mundial pueden diferir, la tendencia al desplazamiento de la fuerza laboral de la salud es real, incluso consi-

derada seriamente como una crisis amenazante [41]. Bajo esta condición, la disponibilidad de TICs (principalmente infraestructura de comunicaciones móviles e inalámbricas) con calidad aceptable entre ambos extremos de un enlace de telemedicina se vuelve significativa. También es crucial el contar con equipos eléctricos, electrónicos y computacionales, así como de personal calificado capaz de adquirir, transmitir, procesar y salvaguardar la información médica digital.

Los sistemas de e-Salud pueden contribuir a mejorar los servicios de salud y abordar retos como la falta de personal especializado. En particular, a través del uso de TICs es factible ampliar la cobertura y el acceso efectivo a servicios de salud tanto básicos como de especialidad. De manera adicional, en conjunto con algoritmos de inteligencia artificial, los sistemas de e-Salud pueden ofrecer nuevos servicios para detección temprana, seguimiento y diagnóstico asistido de diversas enfermedades. Sin embargo, para poder desplegar los sistemas inteligentes de e-Salud es importante abordar diversos retos científicos y tecnológicos; entre los más importantes se encuentran los relacionados con las redes de datos, seguridad e inteligencia artificial.

En este capítulo se discutirán aspectos relevantes de sistemas y redes de telemedicina para poder desplegar un sistema de e-Salud ubicuo. En cada sección se proveen referencias que se consideran relevantes y que pueden servir para que el lector amplíe su exploración de los temas tratados.

El capítulo se estructura de la siguiente forma: en la sección 2 se discute el proceso de transformación digital para la transición de sistemas de salud tradicionales hacia los sistemas de e-Salud; la sección 3 presenta una discusión sobre los sistemas de e-Salud, apoyándose en la definición de telemedicina y telesalud; en la sección 4 se presenta una discusión sobre las redes para el despliegue de sistemas de e-Salud, haciendo énfasis en las redes de telemedicina y telesalud y las tecnologías de comunicación, las alternativas y algunos ejemplos de proyectos alrededor del mundo; la sección 5 presenta una discusión sobre las herramientas de Inteligencia Artificial, Big Data y su aplicación a salud; finalmente, en la sección 6 se presentan las conclusiones finales.

2 Transformación Digital Para el Cuidado de la Salud

La transformación digital es un proceso destinado a reducir las diferencias entre el mundo físico y el mundo digital a través de la creación de nuevas acciones, procedimientos y culturas de la industria (o modificando y adaptando los existentes), utilizando tecnologías digitales, para seguir los cambios cada vez más dinámicos en negocios, mercados y experiencias de clientes, en todo tipo de industrias. A nivel mundial, la industria de servicios de salud se encuentra en diferentes etapas de atención e implementación en cada país. Así, diferentes países tienen diferentes niveles de acceso a los avances médicos-técnicos con diferentes niveles de cobertura poblacional, de acuerdo con sus recursos y políticas particulares. Para reducir esta brecha, la industria del cuidado de la salud (como muchas otras industrias en todo el mundo) ha ido evolucionando cada vez más desde sistemas médicos analógicos, mecánicos y electrónicos antiguos, hacia sistemas de información médica y de cuidado de la salud impulsados digitalmente a un ritmo acelerado. Desafortunadamente, la fuerza laboral médica, de enfermería y de atención médica administrativa no

se está moviendo tan rápido, lo que limita las ventajas de servicio al paciente que pueden brindar estas tecnologías disruptivas. Un artículo de Evans [17] menciona algunos de los problemas típicos encontrados cuando el Expediente Médico Digital (Electronic Health Record - EHR) comenzó a integrarse en los sistemas de salud digitales durante la década de 1990, varios de los cuales todavía se encuentran en los países y sistemas de salud que adoptaron recientemente a esta tecnología. Algunas de las ventajas evidentes de la transformación digital que se han mencionado recientemente [20] se pueden agrupar dentro de 5 áreas de crecimiento inmediato, en particular:

1. Encontrar el médico adecuado, por ejemplo, a través de la búsqueda web, para unir pacientes con especialistas locales/remotos de cualquier otro proveedor de atención médica;
2. Accesibilidad; los pacientes pueden tener acceso a los médicos a través de sus teléfonos inteligentes u otras tecnologías digitales;
3. Responsabilidad, por ejemplo, revisión del paciente en tiempo real y comentarios sobre el estado de calificación del médico o proveedor de atención médica;
4. Transparencia financiera; es factible determinar si existe una necesidad real de los procedimientos recomendados, sus tasas de éxito y tarifas de precios;
5. Compromiso; acceso en tiempo real del médico a la historia médica y de salud histórica del paciente, ayudándole a vivir vidas más saludables y cambiando comportamientos no saludables.

Así, se prevé que la transformación digital de la salud requerirá una constante innovación sanitaria, médica y tecnológica, poniendo al paciente en el centro del servicio sanitario. Sin embargo, hay más en la transformación digital de la atención médica.

El rendimiento de las redes de telemedicina depende de la disponibilidad de Tecnologías de la Información y Telecomunicaciones (TICs) eficientes. El objetivo de dichas redes es la prestación de servicios sanitarios digitales a distancia en forma de datos médicos por parte de los profesionales sanitarios. Para ello, se requiere del uso de un enlace de comunicaciones (el cual es una parte fundamental de cualquier red de telemedicina) entre el profesional sanitario que atiende al paciente remoto y el profesional sanitario distante en el centro médico o asistencial. Según la OMS, en 2018 la 71.ª Asamblea Mundial de la Salud reconoció el potencial de las tecnologías digitales para desempeñar un papel importante en la mejora de la salud pública, instando a sus estados miembros, a través de su Resolución WHA71.7, a priorizar las tecnologías digitales en la salud para promover una cobertura de salud universal [60]. Uno de los ejemplos más exitosos de transformación digital en la atención médica se puede encontrar en el Sistema de Salud de la Universidad Nacional (National University Health System - NUHS) de Singapur, a través de los Servicios de Salud de Jurong (referido como Jurong Health). Este es un grupo reciente de atención médica pública de Singapur, que considera la integración de la capacidad de 700 camas en el Hospital General de Ng Teng Fong y 286 camas del Hospital Comunitario de Jurong [29]. Desde su apertura en 2015, el siste-

ma Jurong Health ha estado centrado en el paciente y fue diseñado para permitir a los cuidadores, pacientes, visitantes y personal usar la tecnología como un habilitador. Este sistema ha "incrustado espacios de enseñanza en clínicas y salas para facilitar la educación continua para los profesionales y estudiantes clínicos, de enfermería y de la salud afines". Además, considera la preparación para una pandemia, lo que le permite evaluar y clasificar a los pacientes durante una pandemia, aislando camas en cada piso y tratarlos con equipos de descontaminación [28], [13]. Dado que la inclusión de TICs fue una parte integral del diseño del edificio del hospital desde la etapa inicial, la automatización derivada del uso de TICs ayuda a Jurong Health a administrar grandes cargas de trabajo, agilizar los procesos de trabajo y mejorar la eficiencia y la productividad de las operaciones, a fin de brindar una mejor atención al paciente [8].

La transformación digital de los servicios de salud es esencial para habilitar los sistemas de e-Salud. De acuerdo con un reporte de la Organización Panamericana de la Salud [42], la e-Salud puede ser entendida como "el uso costeefectivo y seguro de las Tecnologías de la Información y Comunicación (TIC) en apoyo de la salud y de los ámbitos relacionados con la salud, incluyendo los servicios de atención sanitaria, vigilancia de la salud, literatura y educación, conocimiento e investigación". La transformación digital de los sistemas de salud permitirá generar datos que serán la base para las aplicaciones inteligentes de salud. En este sentido, es importante cuidar la privacidad de los pacientes y la seguridad de los datos. Por lo tanto, es importante considerar las normativas enfocadas en proteger la seguridad y privacidad de la información de los pacientes. Aunque el análisis de la información puede ser realizado de forma automatizada, solo los usuarios autorizados pueden tener acceso a los datos recolectados y a los análisis generados de estos datos.

Para preservar la privacidad de la información, las entidades gubernamentales han generado normativas que regulan el uso de información de salud y protegen la privacidad de ésta. Entre las normativas más conocidas a nivel internacional está la ley Health Insurance Portability and Accountability Act (HIPAA) [10] de Estados Unidos, la cual protege la información sensible de salud para que no sea distribuida o divulgada sin el conocimiento o consentimiento del paciente. Para el cumplimento de esta ley en la práctica se debe seguir la Regla de Privacidad HIPPA, para que las entidades autorizadas usen la información dentro de los supuestos permitidos.

En México también existen normas que protegen el uso de la información recabada por los sistemas de e-Salud. En particular, la norma oficial mexicana "NOM-0024-SSA3-2012, Sistemas de información de registro electrónico para la salud. Intercambio de información en salud" [51], de acuerdo con la Secretaría de Salud de México, las principales directrices bajo las cuales se generó la NOM-0024-SSA3-2012 son, textualmente:

- Intercambio de Información entre Sistemas de Información de Registro Electrónico para la Salud (SIRES);
- Creación de un Marco Técnico para el Intercambio de Información en Salud entre SIRES mediante una Arquitectura de Referencia;
- Definición de datos mínimos para la identificación de personas;

- Especificación de documentos técnicos denominados Guías de Intercambio de Información en Salud para escenarios concretos;
- Seguridad de la información y protección de datos referenciada a estándares y disposiciones aplicables en la materia;
- Definición de catálogos y vocabularios mínimos;
- Procedimiento de Evaluación de la Conformidad, que tiene por objeto establecer los requisitos para la Evaluación de la Conformidad para certificar el cumplimiento de los SIRES por parte de los obligados señalados [52].

Observar dichas normas no es trivial desde el punto de vista de la tecnología. Es importante que los sistemas de e-Salud se puedan adaptar a las normativas aplicables a cada escenario. Y, además, deben ser interoperables, de tal manera que puedan compartir información entre los distintos niveles de atención que incluyen el ámbito estatal y el federal.

3 Sistemas de Telesalud y Telemedicina

Los sistemas de e-Salud permiten la oferta de servicios de la salud y médicos, así como la atención a pacientes remotos por parte de profesionales de la salud a distancia, por lo que los conceptos de telesalud y telemedicina son fundamentales. La aplicación de servicios de telesalud y telemedicina se puede explicar siguiendo con atención las etapas que se muestran en la Fig. 1, la cual muestra un sistema básico de telesalud/telemedicina. En esta figura se define un enlace típico de comunicaciones de telemedicina entre el usuario (paciente remoto) y el profesional de la salud (especialista principal), enfatizando las diferentes etapas requeridas para establecer la transferencia de información médica y de salud entre ambos extremos del enlace de comunicaciones.

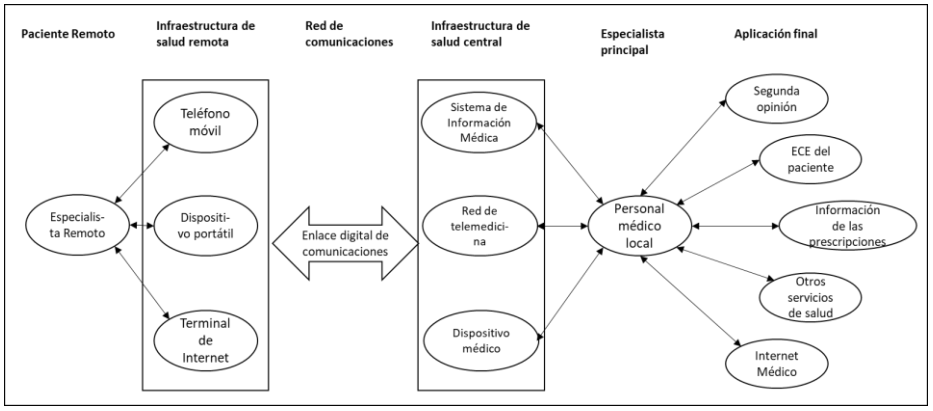


Figura 1. Etapas de la comunicación en un sistema de telemedicina siguiendo un modelo de salud ubicua

Un sistema de e-Salud típico puede integrarse por un conjunto de tecnologías que comprenden componentes comunes, interoperables y compatibles, así como interfaces, protocolos y sistemas previamente desarrollados y desplegados (tanto en presentaciones de hardware como en software). De este modo, un sistema de e-Salud eficiente, de bajo costo y de alta calidad puede desarrollarse a través del uso de redes y sistemas de telemedicina actuales y futuros. Como cada país tiene sus propios sistemas de cuidado de la salud (los cuales consideran aspectos médicos, técnicos, regulatorios y económicos particulares de cada región), los sistemas de telemedicina deben diseñarse y desplegarse a partir de una tecnología global versátil, modular, interoperable y escalable orientada al sector médico. En este sentido, es conveniente seguir un camino similar a lo realizado mediante el modelo de Interconexión de Sistemas Abiertos (Open Systems Interconnection, OSI) para el desarrollo de Internet durante los años 80 y 90; sin embargo, en este caso se debe tener presente el bienestar del paciente en todo momento.

Un informe técnico visionario del 2009 describió este nuevo sistema global de asistencia de la salud basado en el uso de tecnologías de la información, dispositivos clínicos móviles y tecnologías habilitadoras de RFID. Este informe menciona a varios de los primeros impulsores de esta visión, como lo fue un Hospital Infantil en Australia, un proveedor regional de servicios de la salud en España, un proveedor de atención médica en Florida, EUA, y un importante hospital universitario en Noruega [22]. Como se ha mencionado en el presente capítulo, los principales puntos en común por parte de los usuarios finales se abordaron con una "visión estratégica a largo plazo respecto a cómo los diversos componentes de una solución de movilidad podrían unirse para crear un entorno de atención más eficiente y eficaz, y el compromiso organizacional de utilizar la tecnología como un impulsor de ese cambio". éste es otro de los puntos de partida para impulsar un Modelo de Salud Ubicua, como el abordado en esta sección, el cual funcionaría de la siguiente manera (ver Fig. 1):

1. Utilizando uno o más dispositivos con sensores (los cuales pueden ir desde cámaras y micrófonos hasta equipo médico más sofisticado como ultrasonido), se recolectan datos médicos o de salud del paciente remoto;
2. Cuando el trabajador de atención médica remoto recibe los datos médicos o de salud digitalizados en la unidad de atención médica remota, éstos son transmitidos hacia una unidad central de atención médica a través de la red de telecomunicaciones que esté disponible;
3. Los datos transmitidos son procesados en la unidad central de atención médica para realizar diagnósticos, ser almacenados en el expediente clínico electrónico, solicitar segundas opiniones, etcétera.

Un problema en común que se presenta en los dispositivos dedicados y médicos utilizados en los centros de atención médica, tanto locales como remotos, es que dichos dispositivos fueron diseñados con fines médicos específicos para operar en entornos clínicos. Esto los hace costosos y difíciles de usar por parte del personal no médico, como lo menciona [37]. Sin embargo, la tecnología actual que se puede encontrar en las pulseras, teléfonos inteligentes y dispositivos portátiles, ya incluye una serie de sensores no invasivos, así como componentes y aplicaciones que permiten al personal no médico adquirir, manejar

y utilizar datos médicos y de salud básicos. Como referencia, *Telemedicine and Electronic Medicine* [16] es un excelente libro que proporciona bases sólidas sobre tecnologías en electrónica para medicina y de telemedicina.

En un escenario de salud ubicua futuro, un dispositivo de la Internet de las Cosas (Internet of Things - IoT) médicas (m-IoT) con un sensor inteligente (que adquiere datos relacionados a la salud con procedimientos sencillos y claros para cualquier persona que necesite utilizarlos) será capaz de:

- Analizar dichos datos en librerías almacenadas tanto localmente como de manera remota;
- Ubicar al paciente/usuario a través del uso del Sistema de Posicionamiento Global / Sistema Satelital de Navegación Global (GPS/GNSS);
- Usar cualquier red inalámbrica como Bluetooth/WiFi/Celular disponible a través de tecnología cognitiva;
- Encontrar la instalación central de atención médica más cercana;
- Enviar a través de la red o redes de comunicaciones disponibles la identificación del paciente junto con las variables de salud y los signos vitales medidos en tiempo real;
- Podrá solicitar el diagnóstico más cercano a esta instalación central;
- Podrá solicitar la intervención de un sistema inteligente, biblioteca o un especialista en atención médica para un diagnóstico más preciso;
- Podrá solicitar una confirmación del tratamiento, al instante y de forma segura, por sí solo, con pocos o ningún profesional médico/de la salud involucrado.

Este proceso reducirá tiempo, costos e inconvenientes para el paciente. También reducirá la probabilidad de un error médico remoto, al proporcionar al destino final un archivo encriptado vía blockchain que contiene la información personal del paciente, así como sus datos de salud, médicos y de emergencia, la identificación de los doctores que realizaron el diagnóstico en ambos puntos del enlace, y una estampa de tiempo con los signos vitales del paciente, el diagnóstico y su ubicación. Para conformar el EHR del paciente, este proceso también debe incluir el resultado del diagnóstico, los medicamentos recomendados y los procedimientos de tratamiento que, a su vez, sean compatibles con cualquier otro sistema médico o de atención médica válido en el mundo, para que solo los revise un especialista médico o de atención médica certificado.

Un proceso similar sucedería en el caso de equipos médicos altamente especializados, donde un profesional de la salud o un asistente médico colocaría una serie de sensores y dispositivos, así como material quirúrgico si fuera necesario, para obtener cualquier signo vital, variable biométrica, una muestra de tejido o fluido, órgano o implantes, entre otros, que puedan requerirse para evaluar al paciente. Dependiendo de su necesidad médica o de salud, se generaría una representación digital del síntoma, enfermedad o padecimiento medido u observado, o una evaluación de sus variables patológicas o de condición médica obtenidas mediante el uso de diferentes dispositivos, transductores y procesos. Una vez que cualquier enfermedad, condición, síntoma, dolencia,

imagen, sonido, movimiento observado, temperatura o cualquier otra variable biológica o física sea representada como información digital (datos), entonces será posible adquirir, procesar, transmitir, imprimir o visualizar, mediante una gran cantidad de equipos, sistemas o cualquier otra tecnología de salud digital interoperable y compatible al brindar servicios médicos y de salud. Como ejemplo, en el artículo "Development, Implementation, and Multicenter Clinical Validation of the TeleDICOM" [19] se describe la implementación exitosa de teleconsultas ecocardiográficas interactivas para 918 pacientes de cardiología remotos por medio del software TeleDICOM, utilizando máquinas ecocardiográficas de cuatro fabricantes diferentes, y donde también se realizaron varias conferencias médicas en vivo.

En un escenario futuro, una colaboración entre sensores y dispositivos de la salud y biomédicos utilizará lo que ahora se denomina como Tecnologías Emergentes Digitales (Digital Emerging Technologies - DET), que incluiría al IoT, Redes Inalámbricas de Sensores (WSN), Big Data, Inteligencia Artificial (IA), la tecnología de Blockchain, el GPS/GNSS y las Tecnologías Cognitivas, entre otras tecnologías disruptivas. Estas tecnologías deberán interoperar sin problemas con cualquier otra tecnología inalámbrica y cableada, tanto horizontalmente (mismos dispositivos, aplicaciones, redes y sistemas) como verticalmente (diferentes tecnologías interactuando efectivamente entre sí). Este escenario procurará la supervisión médica, técnica y regulatoria más rigurosa y segura desde el inicio, considerando su diseño, desarrollo, implementación, pruebas y uso operativo en general, ya que, finalmente, será una red completa de salud ubicua.

4 Redes de Telesalud y Telemedicina

Los sistemas de e-Salud tienen el potencial de mejorar el acceso efectivo de la población a los servicios de salud. Para esto, es necesario tener sistemas y redes de comunicación que den servicio de manera generalizada. Las redes de sistemas de telesalud y telemedicina pueden realizar esa tarea, dado que obtienen, almacenan, comparten y transmiten información médica entre profesionales de salud, equipos médicos e instituciones, de una forma privada y segura. Las redes de telemedicina representan una herramienta fundamental para habilitar sistemas de e-Salud, dado que permiten que los servicios de salud se desplieguen tanto en áreas urbanas como rurales, así como en sitios que no son accesibles a través de transporte convencional. Sin embargo, habilitar estas redes es un reto tecnológico y científico, aun con las tecnologías de comunicación actuales, tanto inalámbricas (p.ej., la red celular) como cableadas (p. ej., fibra óptica). Si bien estas tecnologías tienen el potencial de cubrir los requerimientos para los sistemas de e-Salud, no necesariamente son interoperables.

Las redes de telemedicina deben ofrecer servicios y aplicaciones confiables, seguras y escalables, con precios accesibles en áreas extendidas. Actualmente es posible establecer servicios multimedia entre lugares distantes utilizando diferentes tecnologías y redes de telecomunicaciones. Asimismo, es posible transferir grandes cantidades de información entre diferentes ubicaciones. Ejemplos de sistemas de comunicaciones que permiten realizar esto son ADSL, enlaces dedicados, redes de fibra óptica, enlaces satelitales, etc., combinados

con dispositivos tecnológicos de uso diario como teléfonos celulares, tabletas, computadoras móviles, etc.

Estos dispositivos y servicios son cada vez más accesibles para diferentes sectores de la población. Y continúan mejorando en términos de cobertura, calidad, precio, número de servicios y usuarios. Es posible tener acceso a diferentes servicios considerando que cada dispositivo es parte de una red que, bajo ciertas condiciones, permitirá establecer un enlace de comunicaciones hacia otros dispositivos e intercambiar información entre ellos. Actualmente es posible interconectar nuevas tecnologías de comunicación cableadas e inalámbricas con diferentes coberturas y capacidades, lo cual aumenta la capacidad de los proveedores de servicio y su área de cobertura, incrementando, a su vez, la capacidad de usuarios a los que se puede brindar servicio.

La Internet actualmente provee conectividad global, por lo que puede ser utilizada como ejemplo del tipo de interconexión requerida para las redes de sistemas de e-Salud. Internet es un sistema global que interconecta redes de computadoras y dispositivos electrónicos, formando la red de comunicación global más utilizada. A través del Internet se transmite información digital de cualquier tipo posible alrededor del mundo [33], por lo que comúnmente es llamada la red de redes. La Internet es una red estandarizada, pública y abierta que permite el acceso externo de diferentes redes para interconectarse con la "World Wide Web" (o simplemente la Web), el cual es uno de los servicios más populares. La mayoría de las aplicaciones y servicios que usamos día a día, como correo electrónico, Voz sobre IP, navegación Web, intercambio de archivos, mensajería instantánea, etc., funcionan sobre la Internet, por lo que comúnmente también es llamada la "carretera de la información". En este sentido, como en el caso de viajar hacia cualquier ubicación a través de caminos y carreteras públicas, no solo se puede conducir a cualquier lado, en cualquier momento, y tampoco se debe subir a cualquier persona que se encuentre en el camino. Dada esta analogía, en la Internet siempre existe la posibilidad de encontrar personas peligrosas (i.e., hackers, phishers), acabar en lugares peligrosos (enlaces no seguros, redes públicas, sitios dudosos), o en situaciones peligrosas (infectar los dispositivos con virus, malware, spyware, etc.).

Además de tomar precauciones básicas usando el sentido común, cuando se usa la Internet se debe tener protección eficiente para cualquier aplicación con acceso, sin importar si maneja información sensible o no. Esto requiere servicios de seguridad (autenticación, encriptación, antivirus) a través de muros de fuego (firewalls) y protección del software a cada nivel, en conjunto con el cumplimento de políticas de ciberseguridad. Esto debe ser especialmente cuidado en el caso de las aplicaciones de salud y debe considerarse en todas las aplicaciones de salud y los servicios relacionados: farmacia en línea, infraestructura médica y seguridad de red o interacción remota con pacientes en casa. Esto, con el objetivo de proteger a los usuarios de riesgos potenciales para ofrecer servicios seguros y eficientes [57].

A diferencia de la Internet, las Intranets son redes de comunicación privadas basadas en la interconexión de computadoras y otros dispositivos compatibles con el protocolo TCP/IP dentro de la misma organización, comúnmente sin acceso (o acceso altamente restringido) hacia sitios externos para mantener la privacidad y el control de la información interna [40], [62]. Actualmente, las Intranets pueden proveer diversas herramientas colaborativas con acceso segu-

ro y específico a recursos internos como foros de discusión, motores de búsqueda, blogs, calendario compartido y aplicaciones móviles.

Ambos tipos de redes, Internet e Intranets, interactúan a través del uso de Firewalls que analizan los datos tanto de salida como de entrada para identificar amenazas y eliminarlas para proteger la información y la infraestructura. En conjunto, la Internet y las Intranets pueden ofrecer el *backbone* requerido para poder desplegar los servicios de e-Salud a través de las redes de telemedicina. Una vez que se encuentren desplegados y funcionales, los servicios de e-Salud podrán beneficiarse tanto de la cobertura global de la Internet como de la información médica, la operación, la seguridad y la eficiencia de las Intranets.

Una de las características importantes de las redes de comunicaciones es el área geográfica que cubren, la cual determina en buena medida su impacto y utilidad para diferentes aplicaciones. Las tecnologías actuales para salud ofrecen una gran mejora en los servicios públicos y privados de salud en el área geográfica que son desplegadas, gracias a la mejora en los indicadores de calidad médica y de salud. Algunos de estos indicadores se describen en reportes como: "2008 outcomes report commissioned by the World Health Organization's Global Observatory for eHealth" [50]; "2013 report on Health IT Quality Measurements commissioned by the U.S. Agency for Healthcare Research and Quality" [44]; y "Policy Brief 25 from the ICARE4EU project issued by the European Observatory on Health Systems and Policies" [6]. Estos reportes describen la introducción de nuevos productos y servicios basados en el desarrollo y uso de TICs, particularmente en el área de redes y sistemas de comunicación y su área de cobertura, orientadas a aplicaciones y servicios de salud.

Los sistemas de telemedicina usualmente operan dentro un área de cobertura limitada por las características de la tecnología de comunicaciones [18], [55]. Cuando se utiliza más de un enlace de comunicaciones para interconectar varios dispositivos para el intercambio de datos médicos, usando un canal de comunicaciones, se crea una red de telesalud o telemedicina. Existe un gran número de aplicaciones de telemedicina, enlaces y redes operando alrededor del mundo desde 1970, con diferentes tecnologías y áreas de cobertura, usando todo tipo de canales que dependen de la tecnología utilizada. Es importante hacer énfasis en que, además del área de cobertura, existen otros aspectos importantes relacionados con la capacidad y la movilidad. Algunas de las principales tecnologías para proveer servicios de salud utilizando sistemas y redes de telecomunicaciones son descritos a detalle en [18], el cual es una es una muy buena introducción de las TICs para la salud y da una descripción general de la tecnología, sus principales usos y ventajas, en aspectos clave que aplican específicamente al sector salud. La tecnología ha avanzado de forma importante, por ejemplo, WiFi (Wireless Local Area Networks, IEEE 802.11), WiMAX (Worldwide Interoperability for Microwave Access, IEEE 802.16) y las redes de telefonía móvil (4G y 5G), entre otras, son tecnologías de comunicación que actualmente se utilizan para proveer servicios de telemedicina. En este sentido, es importante mencionar que cada sistema está limitado por sus principales características, como área de cobertura, capacidad y movilidad.

En las comunicaciones móviles inalámbricas, el usuario se comunica a través de un enlace de radio, desde un teléfono móvil o terminal remota, a la estación base o punto de acceso dentro de cada celda. Desde este punto cen-

tral, la información es enrutada a través de enlaces de cobre o fibra óptica hacia el otro extremo del enlace punto a punto con otro usuario. Estas redes son llamadas redes inalámbricas homogéneas. La mayoría de las redes inalámbricas y celulares son desplegadas utilizando redes homogéneas entre estaciones base y terminales de usuario con tecnología similar. Sin embargo, existe una necesidad creciente de interconectar terminales de usuario y estaciones base con tecnologías heterogéneas, como se describe en [5]. En [31] se presenta también una descripción detallada y concisa de diferentes tecnologías inalámbricas y estándares utilizados en redes de comunicación para salud. La descripción presentada en (Keikhosrokiani 2015) cubre la mayoría de las tecnologías actuales y las aplicaciones médicas y de cuidado de la salud a través de sistemas inalámbricos, móviles y satelitales. Para interconectar estas (y muchas otras) tecnologías de forma transparente para el usuario, es necesario incluir mecanismos de traspaso (referido en inglés como *hand-off* o *handover*), creando, así, redes heterogéneas de comunicación. El traspaso permite que cuando un usuario se mueve de una celda a otra durante una llamada no sufra una interrupción del servicio. Los traspasos entre celdas utilizando la misma tecnología (p.ej., WiFi a WiFi, satélite a satélite) son llamados traspasos horizontales (HHO), mientras que los traspasos entre celdas con diferentes tecnologías (p. ej., WiFi a Bluetooth) son llamados traspasos verticales (VHO).

En el trabajo realizado por Yew, Supriyanto, Satria y Hau [61] se describen las tecnologías utilizadas en redes de telemedicina, incluyendo algunos de los problemas relacionados con la movilidad. Por ejemplo, se establece que, para que el sistema tenga una cobertura amplia y se garantice la calidad del servicio, un sistema móvil de telemedicina debe tener la capacidad de acceder a múltiples redes inalámbricas para que el usuario se conecte a la mejor red basado en los requerimientos del servicio. Además, basándose en esta premisa, también se describe una propuesta de mecanismo de VHO, requerido para preservar la calidad del servicio de telemedicina en ambientes móviles. En [46] se mencionan diferentes tecnologías inalámbricas utilizadas para servicios de salud que consideran tanto HHOs como VHOs. Además, se describen varios casos que ya consideran la provisión de calidad de servicio (QoS) para aplicaciones de telemedicina.

Para un sistema de e-Salud ubicuo, todos los tipos de tecnologías inalámbricas deberían cubrir todo tipo de aplicaciones médicas, por ejemplo: dispositivos vestibles basados en Bluetooth (p. ej., relojes inteligentes, medidores de actividad); aplicaciones de medicina prehospitalaria a través de 4G-LTE; sistemas de monitoreo remoto en casa basado en WiFi; supervisión y cuidado en un hospital utilizando redes inalámbricas de área local; incluso utilizando estaciones terrenas satelitales para comunicar instalaciones de salud remotas para aplicaciones como tomografía computarizada y la transmisión de otro tipo de imágenes médicas, como parte de una red satelital de área amplia (SWAN por sus siglas en inglés). La Fig. 2 muestra las distintas tecnologías inalámbricas, de telemedicina y telesalud que pueden conformar un sistema de salud ubicuo.

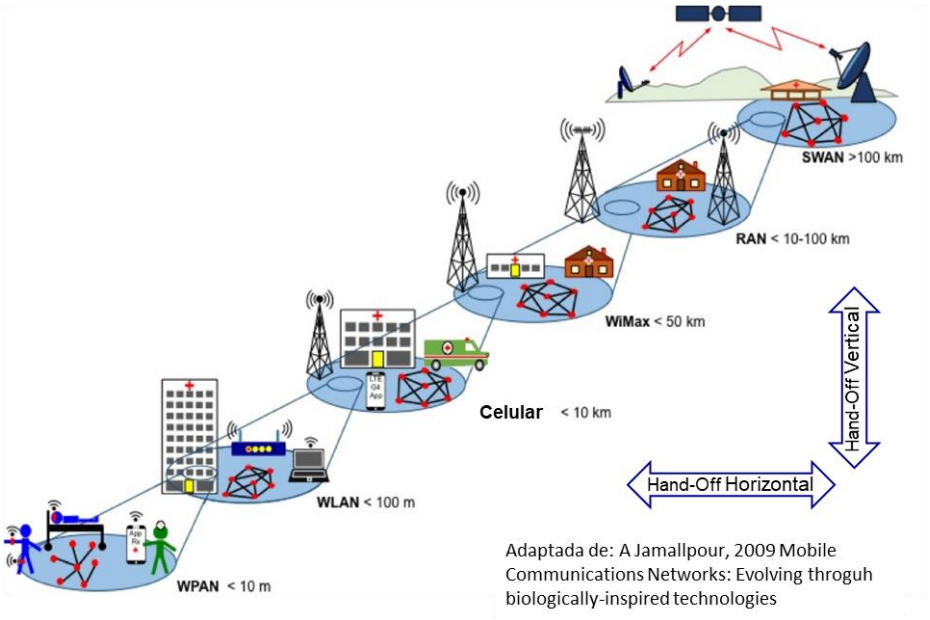


Figura 2. Cobertura ubicua de salud, lograda con tecnologías inalámbricas, de telemedicina y telesalud integradas

A continuación se presenta una breve descripción de las principales tecnologías y estándares utilizados actualmente en redes y sistemas inalámbricos de telemedicina que pueden ser parte de un sistema de cuidado de la salud ubicuo.

Redes Inalámbricas de Área Personal (Wireless Personal Area Networks – WPAN). Las WPANs son un conjunto de redes basadas en el uso del estándar IEEE 802.15, compuesto por los estándares IEEE 802.15.1 (Bluetooth) e IEEE 802.15.4 (Zigbee, 6LowPAN), enfocados en conectividad inalámbrica con dispositivos móviles, portátiles y fijos en un espacio personal (< 10 metros). El término popular de Redes de Área Corporal (Body Area Networks – BAN), o tecnología vestible (Wearable), también aplica las WPANs centradas o desplegadas cerca o sobre personas. El trabajo realizado por Casillas, Villarreal-Reyes, Gonzalez, Martínez y Pérez-Ramos [9] compara diferentes arquitecturas de redes inalámbricas de sensores (Wireless Sensor Networks – WSN) sobre WPANs, así como WPANs para aplicaciones del cuidado de la salud. En este trabajo se describen diferentes tecnologías abiertas y propietarias para salud móvil usadas en monitorización remota de diferentes variables fisiológicas de pacientes en casa.

Redes Inalámbricas de Área Local (Wireless Local Area Networks – WLAN). Las WLANs son un conjunto de redes desarrolladas alrededor del estándar IEEE 802.11 como el equivalente inalámbrico a las redes de área local (LANs por sus siglas en inglés). Desde el inicio, las WLANs fueron dise-

ñadas para operar en exteriores e interiores de casas y oficinas, incluso, en salas de reunión, restaurantes y lugares con muchas personas, como centros de conferencia, parques, estadios y aeropuertos. Aunque inicialmente fueron diseñadas principalmente para aplicaciones de casa y oficina, posteriormente su uso se extendió a Redes Privadas Virtuales (Virtual Private Networks – VPN) y Voz sobre IP (VoIP) [24]. La versión WiFi es probablemente la tecnología de red inalámbrica de banda ancha más ampliamente utilizada alrededor del mundo actualmente. Es utilizada ampliamente para aplicaciones de cuidado de la salud, redes inalámbricas para el personal administrativo, pacientes, familiares y visitantes en los hospitales e instalaciones de salud. Sin embargo, es muy probable que las tecnologías celulares 5G sean una competencia importante para las WLANs, dado que 5G también ofrece banda ancha, soporta mayor movilidad de los usuarios y ofrece una cobertura mucho más amplia.

Redes de Telefonía celular/móvil. Son las redes comunes inalámbricas o de telefonía móvil que se usan en el día a día. En estas redes, un extremo del enlace de comunicaciones es un dispositivo de radiocomunicaciones portátil (p. ej., un smartphone), mientras que en el otro es una estación base fija con una conexión de banda ancha a la red pública de telefonía o de datos. Inicialmente, las redes de telefonía celular/móvil fueron desarrolladas para servicio de telefonía analógica en los años 1980s (primera generación, 1G) y, rápidamente, evolucionaron a redes de acceso digital durante los años de 1990 (Segunda Generación, 2G) y a redes de datos compatibles con IP durante los años 2000 (Tercera Generación 3G). Para la Cuarta Generación (LTE-Advance, 4G), en los años 2010, se realizaron actualizaciones para aplicaciones de Internet de banda ancha, las cuales son ampliamente usadas actualmente. En este mismo sentido, Keikhosrokiani [31] presenta una extensa descripción de diversas tecnologías móviles para el cuidado de la salud y sus principales características técnicas, aplicadas a redes médicas y del cuidado de la salud, dispositivos y aplicaciones, así como la percepción de su desempeño. Basado en nuevos desarrollos, se espera que la siguiente generación se despliegue durante los años 2020 (Generación 5, 5G). Esta tecnología incluye comunicaciones Máquina a Máquina (M2M), aplicaciones de banda ancha en cuasi-tiempo real, múltiples estaciones base con arreglos de antena masivos, nanoceldas, etc. [21], teniendo un fuerte impacto positivo en la presencia de servicios para el cuidado de la salud basados en comunicaciones móviles e inalámbricas.

Interoperabilidad Mundial para Acceso por Microondas (Worldwide Interoperability for Microwave Access - WiMAX). WiMAX es una tecnología basada en IP y en el conjunto de estándares IEEE 802.16 para comunicaciones inalámbricas. Estos fueron desarrollados para desplegar acceso de banda ancha de última milla con una cobertura de hasta 50km utilizando estaciones fijas como una alternativa inalámbrica a diferentes tecnologías como cable, la Línea de Suscriptor Digital (DSL por sus siglas en inglés) y el par de cobre para acceso de banda ancha en casa. Existen diversos casos reportados que utilizan WiMAX para aplicaciones para el cuidado de la salud. Por ejemplo, servicios multimedia para telemedicina y una propuesta relacionada con QoS para la transmisión en banda ancha de datos y video de forma inalámbrica a distancias desde 5 a 15 km, utilizando estaciones móviles, y

hasta 50 km con estaciones fijas, en la República de Macedonia [15]. Otro caso de uso es el descrito en [54], ubicado en Ghana.

Redes Inalámbricas de Área Regional (Wireless Regional Area Networks - WRAN). El estándar IEEE 802.22 para comunicaciones inalámbricas está enfocado en proveer acceso inalámbrico de banda ancha en áreas rurales, con un radio de cobertura entre 10 y 60 km a través del uso de la tecnología de radios cognitivos (Cognitive Radio, CR). El propósito de este estándar es operar en los canales disponibles en las bandas blancas de VHF y UHF para TV y, así, aprovechar el mayor rango y las mejores condiciones de propagación a frecuencias más bajas, comparadas con las características de la banda para la tecnología celular. Se basa en establecimiento de enlaces inalámbricos directos entre una estación base (Base Station - BS) y hasta 512 usuarios (Customer Premises Equipment - CPE) dentro un área de cobertura dinámica. Las BS y CPEs utilizan radios cognitivos para controlar todas sus características de RF y evitar interferencia hacia/desde otros usuarios de las bandas blancas de TV. En [35] se presenta una comparación entre las tecnologías WRAN y WiMAX para el uso de aplicaciones de telemedicina en México.

4.1 Redes Satelitales de Área Amplia (Satellite Wide Area Networks - SWAN)

Las SWAN son redes satelitales de cobertura muy amplia. El límite de su cobertura depende de la ganancia y directividad de las antenas y de la altura de la órbita y dirección alrededor de la tierra. La mayoría de los satélites en red están colocados en una órbita geoestacionaria. Comúnmente, las SWANs con aplicaciones para salud están basadas en el uso de la tecnología de red denominada Terminales de Apertura Muy Pequeña (Very Small Aperture Terminal - VSAT), la cual es compatible con IP. En las SWANs se crea un enlace de radio entre múltiples terminales remotas VSAT y una estación terrena central (Hub) conectada a la Internet. En [56] se describe una red VSAT como parte de las tecnologías utilizadas en China para servicios hospitalarios civiles y militares. Cerca de 211 hospitales son parte de la red de telemedicina descrita, incluyendo 60 estaciones militares en campos remotos de China. Como parte del proyecto TESHEALTH [32] de la Agencia Europea Espacial, se desarrolló una plataforma interactiva de telemedicina desplegada sobre una red híbrida satélite/terrestre para conectar hospitales rurales y centros de salud en diferentes países europeos. Finalmente, en [11] se describen los servicios requeridos para el cuidado de la salud, principalmente para la población en zonas rurales de la India y cómo están siendo abordados a través de la red de su agencia espacial, llamada Indian Space Research Organization (ISRO), y diversas agencias del Estado relacionadas con la salud y la familia. Se describe cómo se conectan 45 hospitales rurales y 15 hospitales de alta especialidad en la India.

Redes Virtuales Privadas (Virtual Private Networks – VPN). Debido a la proliferación de conexiones de Internet de bajo costo, es común para las compañías implementar redes virtuales privadas (VPNs), las cuales consis-

ten en túneles virtuales que se generan utilizando diferentes tecnologías de red como la telefónica, celular, satelital, etc., como enlaces físicos entre los nodos VPN. Básicamente, una VPN se utiliza para habilitar la transmisión privada de información sobre una red pública, como las mencionadas un poco antes, utilizando mecanismos de seguridad como la encriptación de información y algoritmos para restringir el acceso a la VPN de usuarios no autorizados. Las VPNs pueden, entonces, proveer enlaces seguros a través de WANs públicas, como la Internet, creando túneles basados en IPSec y protocolos como Capa de Puertos Seguros (Secure Socket Layer - SSL) y Seguridad de Capa de Transporte (Transport Layer Security - TLS). De manera adicional, utilizando el protocolo MPLS (Multiprotocol Label Switching por sus siglas en inglés) y una VPN (MPLS-VPN) se puede habilitar un backbone. Para habilitar un backbone que permita el despliegue global de servicios para el cuidado de la salud, una VPN debería poder usar cualquier tipo de red de telecomunicaciones compatible con la Internet y las tecnologías mostradas en la Fig. 2, dado que debe habilitar la conectividad entre usuarios en diferentes ubicaciones en el mundo. Un aspecto importante a considerar es que, actualmente, los servicios de VPN, operación, seguridad, confiabilidad y gestión están bajo el control de los proveedores de servicios y no del usuario, por lo que la interoperabilidad no está garantizada.

5 Inteligencia Artificial y Big Data en Salud

El uso de la inteligencia artificial (IA) en servicios de diferentes ámbitos promete mejorar su desempeño y el acceso a estos servicios. En el caso particular de los sistemas de e-Salud, el uso de la IA, en conjunto con la existencia de una gran cantidad y variedad de datos, permitiría habilitar diferentes aplicaciones para mejorar la cobertura y acceso efectivo a servicios de salud, esto mediante la habilitación de sistemas inteligentes de e-Salud. En esta sección se discuten las bases de la IA y Big Data y su relación con servicios enfocados al cuidado de la salud.

5.1 Inteligencia Artificial

La Inteligencia Artificial (IA) y Aprendizaje Máquina (Machine Learning - ML) son tecnologías relativamente nuevas que el diccionario Oxford define como: “La capacidad de las computadoras u otras máquinas de mostrar o simular comportamiento inteligente”. El diccionario Merriam-Webster en línea las define como: “1: una rama de las ciencias computacionales que se enfoca en la simulación de comportamiento inteligente en una computadora; 2: la capacidad de una máquina para imitar el comportamiento humano inteligente”. De manera más específica, la IA es definida por un White Paper de IBM como la ciencia y conjunto de tecnologías computacionales inspiradas en la forma en que los humanos sienten, aprenden, razonan y tomas acciones a través de su cuerpo y sistema nervioso, que permite que las máquinas tengan la habilidad de llevar a cabo tareas similares a las humanas con diferentes niveles de complejidad [23].

Actualmente, utilizando IA es posible detectar patrones con mayor precisión en más tipos de datos. A través del uso de algoritmos complejos, las má-

quinas que usan IA pueden realizar muchas tareas en algunos dominios específicos, tales como tareas generales (reconocimiento visual y de voz, procesamiento y traducción del lenguaje), tareas formales (juegos que involucran aprendizaje) y tareas expertas (análisis de ingeniería y diagnóstico de enfermedades). Diversos estudios muestran resultados sobresalientes de la IA aplicados a la salud, como el estudio de la corporación Mitre en 2017, que está enfocado en procesos asistidos por IA para toma de decisión en el cuidado de la salud. En este estudio se afirma que las redes neuronales profundas pueden desempeñarse tan bien como el personal médico en algunas tareas de diagnóstico específicas, incluyendo herramientas de IA en aplicaciones orientadas a salud que pueden ser usadas en dispositivos móviles como los teléfonos inteligentes [26]. Este estudio muestra diferentes aplicaciones basadas en IA relacionadas con servicios de salud, dando diferentes ejemplos relevantes, observaciones y resultados. Pero, principalmente, en [26] se establece que, actualmente, la sociedad estaría más dispuesta a aceptar aplicaciones de salud potenciadas por IA debido a tres factores: 1) frustración por el sistema legal; 2) disponibilidad de dispositivos inteligentes conectados en la sociedad; y 3) mayor aceptación de servicios en casa como los provistos a través de Amazon y otras compañías. Algunos de sus principales hallazgos y recomendaciones incluyen: la aceptación de las aplicaciones de IA en la práctica clínica; la disponibilidad de datos de calidad para entrenamiento para construir y actualizar aplicaciones de IA; llevar a cabo campañas de recolección de datos a gran escala para obtener flujos de datos faltantes; crear competencias importantes de IA; y entender las limitaciones de la IA en el cuidado de la salud.

El Aprendizaje Máquina (ML) es un subconjunto de la IA, donde el método científico es utilizado para representar, entender y usar conjuntos de datos utilizando algoritmos. El desempeño de estos algoritmos mejora conforme son expuestos a una mayor cantidad de datos; esto, sin la necesidad de explícitamente programar nuevas reglas, dado que se aprenden de los nuevos datos. Un ejemplo de ML aplicado a dermatología es el siguiente: imágenes de las irregularidades de la piel y el diagnóstico relacionado se alimentan a las computadoras; el algoritmo compara, procesa y aprende las variaciones entre las diferentes imágenes. Cuando llegan nuevas imágenes, el sistema propondrá un diagnóstico basado en el aprendizaje de los ejemplos previos. La IA y, en particular, las técnicas de ML estudian sistemas que aprenden a realizar clasificaciones no lineales mediante entrenamiento supervisado o no supervisado, o una combinación de ambos. La clasificación supervisada necesita un conjunto de muestras ya etiquetadas para entrenar la máquina y otro conjunto para la validación. Las redes neuronales, las máquinas de vectores de soporte (SVM), Adaboost y Naive Bayes son algunos ejemplos de aprendizaje supervisado. Por otro lado, los algoritmos no supervisados no necesitan un conjunto etiquetado [14].

Aprendizaje Profundo (Deep Learning - DL) es una familia de métodos automáticos de ML que ha ganado considerable atención en la comunidad científica, rompiendo récords de referencia en áreas como reconocimiento de voz y visual en el área de la salud, así como datos clínicos. Se diferencia de los métodos convencionales de ML en su capacidad para aprender la representación óptima de datos sin procesar a través de transformaciones no lineales consecutivas, logrando niveles cada vez más altos de abstracción y complejidad [3], [48]. Dada su capacidad para detectar patrones abstractos y complejos, DL se

ha aplicado en estudios como el cáncer, las enfermedades cardiovasculares, la diabetes y las enfermedades psiquiátricas. Desde la perspectiva de un sistema de análisis de datos, los sistemas basados en DL buscan ser completamente inteligentes, programables e interactivos, cuya operación comience inmediatamente desde la entrada de datos sin procesar hasta la salida final de patrones reconocidos. Recientemente, la técnica que se está utilizando es la computación cuántica basada en qubits, que van desde cincuenta hasta unos pocos cientos [38]. La aplicación de Quantum Machine Learning (QML) es una de las aplicaciones más alentadoras, siendo investigada activamente por varios grupos de investigación [4]. En general, existe un desafío para desarrollar nuevas técnicas capaces de explotar las ventajas de la computación cuántica para mejorar el aprendizaje automático.

Estas tecnologías de IA se aplican cada vez más a la atención médica con el único propósito de mejorar la calidad de vida de los pacientes. Para obtener más información relacionada con el uso de IA en el cuidado de la salud, lea [36], [34], [27], [2].

5.2 Big Data

Big Data es el resultado de grandes cantidades de datos generados por dispositivos, sistemas, empresas y organizaciones empresariales, relacionados con todos los campos del conocimiento, entre ellos, la medicina, la salud y el cuidado de la salud. Dada la gran cantidad de datos generados por sistemas, máquinas, vehículos, personas, sensores, IoT y otros dispositivos, y su creciente proliferación en todas partes, se requieren nuevas formas y técnicas para capturar, almacenar, procesar, mostrar y, cuando sea necesario, analizar y organizar dichos datos. Las arquitecturas IoT generan diferentes tipos de datos en grandes volúmenes a velocidades muy altas, generados por todo tipo de sensores y dispositivos RFID, datos que eventualmente irán a bases de datos y otros repositorios de almacenamiento y datos [43]. Dado que los datos generados en el mundo moderno son enormes y siguen creciendo exponencialmente, los datos estructurados y no estructurados (texto, imágenes y archivos de audio, web, correos electrónicos, etc.) inundan rápidamente a las organizaciones. Por otro lado, se sabe que "Big Data ha atraído mucha atención recientemente en el gobierno, las industrias, las ciencias, la ingeniería, la salud y la medicina, las finanzas y de manera destacada en las empresas", como lo menciona Ajah []1. Big Data y Business Analytics se están desarrollando e implementando para analizar estos grandes volúmenes de datos, pero cada empresa necesita una visión diferente de los crecientes volúmenes de datos transaccionales recibidos. El análisis de datos en tiempo real ayuda a las organizaciones a ver el pasado y prever el futuro "al saber lo que ocurrió (descriptivo), comprender por qué sucedió (diagnóstico), anticipar lo que podría suceder (predictivo) y, en última instancia, determinar cómo influir en el futuro su ocurrencia (prescriptivo)", como lo describe Ajah [1]. Desde otro punto de vista, existe un fuerte interés en las posibles ventajas económicas del análisis de Big Data, indicando cómo "el potencial económico de Big Data representa el mayor desafío y consiste en encontrar valor en el gran volumen de datos no estructurados en (o cerca de) tiempo real. La tendencia a utilizar esta información para la analítica empresarial se está convirtiendo en una práctica de gestión a nivel mundial" [53].

En cuanto al uso de Big Data en los servicios de salud, la Royal Society del Reino Unido publicó en 2006 un estudio con una serie de recomendaciones sobre las posibles contribuciones de las TIC a la salud y medicina, indicando que "las TIC ayudarán a que sea posible obtener datos tanto del individuo como de la comunidad más amplia. En el caso del individuo, se podrá obtener información clínica detallada en todos los niveles del organismo humano (sistema, órgano, tejido, célula, proteína y gen). Esto, junto con información médica más general dentro del registro de salud de un paciente (peso, regímenes de ejercicio, dieta y riesgo genético), permitirán un análisis de datos mucho mejor, lo que conducirá a estudios epidemiológicos mucho más precisos". También se menciona cómo la Comisión Europea ha trabajado en un registro de salud electrónico paneuropeo (EHR) que "permitiría realizar estudios epidemiológicos muy detallados con la capacidad de monitorear continuamente el estado de salud de la nación y la UE y rastrear el desarrollo de epidemias en una etapa muy temprana. Esto podría generar muchos beneficios, como un desarrollo y una distribución de vacunas más efectivos" [45]. Otro estudio diferente muestra cómo "Hoy en día, los grandes datos son un tema candente para la minería de datos y el IoT; también discutimos las nuevas características de los grandes datos y analizamos los desafíos en la extracción de datos, los algoritmos de minería de datos y el área del sistema de minería de datos. Con base en la encuesta de la investigación actual, se propone un sistema de minería de big data sugerido" [12]. Finalmente, la Guía de estrategia de Schumacher también describe cómo "Hoy en día, los datos de salud significan Big Data. Esto, a su vez, significa que replicar todos los registros de salud a cada miembro de la cadena de bloques requeriría un uso intensivo del ancho de banda, un desperdicio de recursos de red y plantearía problemas de rendimiento de datos" [47]. Para obtener más información relacionada con el uso de Big Data en el cuidado de la salud, lea [30] y [7].

En esta sección se muestra claramente que, para que las aplicaciones de IA enfocadas a salud se puedan desplegar en los sistemas de e-Salud, es necesario recolectar, almacenar y procesar grandes cantidades de datos que permitan el entrenamiento de los algoritmos de IA. Esto no es una tarea sencilla, el volumen de datos, las diferentes ubicaciones geográficas de las personas que generan los datos, la variedad de datos y sus formatos de adquisición y almacenamiento, son retos científicos tecnológicos que se deben abordar para impulsar el desarrollo y despliegue de los sistemas inteligentes de e-Salud.

6 Conclusiones

En este capítulo se presentaron los potenciales beneficios de los sistemas inteligentes de e-Salud para contribuir a mejorar los servicios de salud y el acceso efectivo de la población a estos, así como los principales componentes de los sistemas inteligentes de e-Salud para que puedan ser desplegados. De la discusión en este capítulo se concluye que la inteligencia artificial (IA) es una herramienta que puede contribuir a mejorar el acceso efectivo a los servicios de salud. Esto, a través de habilitar aplicaciones en los sistemas de e-Salud que, por ejemplo, asistan al personal de la salud con análisis de imágenes médicas, extracción de conocimiento de notas médicas o análisis continuo de signos vitales para la detección temprana de enfermedades crónicas o agudas.

Para que las aplicaciones de salud basadas en IA sean una realidad, es importante recolectar información que permita entrenar a los modelos de IA. En este sentido, es de especial relevancia tener repositorios con datos representativos de la población a la que la aplicación será dirigida y, además, crear y tener acceso a arquitecturas de procesamiento de datos robustas y flexibles.

Debido a lo sensible de la información de salud y a las normativas que se deben observar, preservar la privacidad y seguridad de los datos desde su recolección hasta su análisis es de vital importancia. Los algoritmos para el análisis de la información podrían usar de manera automatizada la información en los sistemas de registro electrónico. Sin embargo, los algoritmos no siempre residen en la misma red que en la que se genera la información. Además, se debe cuidar también la privacidad y seguridad de los datos que se analizan y el de los resultados de los análisis. Por lo tanto, los sistemas inteligentes de e-Salud también deben observar las normativas para la protección de datos de salud de la población.

Un componente clave para los sistemas inteligentes de e-Salud son las redes de telemedicina y telesalud, las cuales permitirán recolectar información desde las distintas ubicaciones de un paciente, desplegar servicios como teleconsulta y transmitir los resultados de los análisis realizados por los algoritmos inteligentes. Sin embargo, debido al gran número de escenarios en los que se deben recolectar/entregar datos, las redes requeridas para dar soporte a los sistemas de e-Salud deben ser de diversas tecnologías para poder cumplir con los requerimientos de batería, capacidad, cobertura, movilidad, etc. Por esta razón, el desarrollo de redes heterogéneas e interoperables, que, además, ofrezcan mecanismos de seguridad y calidad de servicio es muy relevante.

Como se puede leer en la discusión de este capítulo y en los párrafos previos, los sistemas inteligentes de e-Salud presentan diversos retos científicos y tecnológicos; en el ámbito de los algoritmos de inteligencia artificial, que extraigan conocimiento de los datos de salud. Pero, para poder desplegar estos sistemas es, fundamental abordar los retos científicos y tecnológicos en otros componentes claves, particularmente en las redes heterogéneas de telemedicina y telesalud interoperables y la preservación de la privacidad y seguridad de los datos en su recolección, almacenamiento y posterior tratamiento.

Referencias

[1] Ajah IA, Nweke HF (2019). Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications. Big Data Cogn. Comput. 2019, 3, 32; doi:10.3390/bdcc3020032

[2] AoMRC (2019). Artificial Intelligence in Healthcare. Academy of Medical Royal Colleges. January 2019. https://www.aomrc.org.uk/wp-content/uploads/2019/01/Artificial_intelligence_in_healthcare_0119.pdf [Accedido en octubre 30, 2022]

[3] Arel I, Rose DC & Karnowski TP (2010). Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier], IEEE Computational Intelligence Magazine, November 2010, pp. 13-18. Digital Object Identifier 10.1109/MCI.2010.938364

[4] Arrazola JM, Bromley TR, Izaac J, Myers CR, Bradler K & Killoran N (2019). Machine learning method for state preparation and gate synthesis on photonic

quantum computers, Quantum Sci. Technol. 4 (2019) 024004. https://doi.org/10.1088/2058-9565/aaf59e

[5] Atayero AA, Adegoke E, Alatishe AS, Orya MK (2012). Heterogeneous Wireless Networks: A Survey of Interworking Architectures. International Journal of Engineering and Technology ,Volume 2 No. 1, January, 2012

[6] Barbabella F, Melchiorre MG, Quatrinni S, Papa R (2017). How can eHealth improve care for people with multimorbidity in Europe? ICARE4UE project, Policy Brief 25, ISSN 1997–8073 http://www.icare4eu.org/pdf/PB_25.pdf

[7] Bodas-Sagi DJ, Labeaga JM (2017). Big Data and Health Economics: Opportunities, Challenges and Risks. International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 4, No7, (47-52)

[8] Bhunia P (2017). The JurongHealth IT Journey – Integrating IT from the ground-up into a new digital hospital, OpenGov Asia, 28 October, 2017. https://www.opengovasia.com/exclusive-the-juronghealth-it-journey-integrating-it-from-the-ground-up-into-a-new-digital-hospital/ [Accedido en octubre 30, 2022]

[9] Casillas M, Villarreal-Reyes S, Gonzalez AL, Martinez E, Perez-Ramos A (2015). "Chapter 18: Design Guidelines for Wireless Sensor Network Architectures in mHealth Mobile Patient Monitoring Scenarios." Mobile Health: A Technology Road Map, pp. 401-428. Springer International Publishing, 2015.

[10] CDC (2022). Health Insurance Portability and Accountability Act of 1996 (HIPAA). Health Insurance Portability and Accountability Act of 1996 (HIPAA). https://www.cdc.gov/phlp/publications/topic/hipaa.html [Accedido en octubre 30, 2022]

[11] Chellaiyan VG, Nirupama AY, Taneja N. (2019). Telemedicine in India: Where do we stand?. Journal of Family Medicine and Primary Care. Volume 8, Issue 6, June 2019. pp. 1872-1876.

[12] Chen F, Deng P, Wan J, Zhang D, Vasilakos AV, Rong X (2015). Data Mining for the Internet of Things: Literature Review and Challenges. International Journal of Distributed Sensor Networks, Volume 2015, Article ID 431047, 14 pages http://dx.doi.org/10.1155/2015/431047.

[13] Chikul M, Maw HY, Soong YK (2017). Technology in healthcare: A case study of healthcare supply chain management models in a general hospital in Singapore. Journal of Hospital Administration, Dec. 2017, Vol. 6, No. 6. pp. 63-70. DOI: 10.5430/jha.v6n6p63. jha.sciedupress.com

[14] Chollet, Francois (2018). Deep Learning with Python, Manning Publications Co. ISBN 9781617294433 http://faculty.neu.edu.cn/yury/AAI/Textbook/Deep%20Learning%20with%20Python.pdf [Accedido en octubre 30, 2022]

[15] Chorbev I, Mihajlov M (2009). Building a Wireless Telemedicine Network within a WiMax based Networking Infrastructure, IEEE International Workshop on Multimedia Signal Processing, MMSP '09, Rio de Janeiro, Brazil, October 5-7, 2009. IEEE 2009.

[16] Eren H & Webster JG (Editors) (2011). Telemedicine and Electronic Medicine. 651 p, CRC Press.

[17] Evans, RS (2016). Electronic Health Records: Then, Now, and in the Future, IMIA Yearbook of Medical Informatics 2016, Yearb Med Inform 2016;Suppl1:S48-61 http://dx.doi.org/10.15265/IYS-2016-s006 Published online May 20, 2016

[18] Fong B, Fong ACM, Li CK (2011). Telemedicine Technologies: Information Technologies in Medicine and Telehealth. 2011 John Wiley and Sons, ISBN 978-0-470-74569-4

[19] Gackowski A, Czekierda L, Chrustowicz A, Cała J, Nowak M, Sadowski J, Podolec P, Pasowicz M, Zieliński K (2011). Development, Implementation, and Multicenter Clinical Validation of the TeleDICOM—Advanced, Interactive Teleconsultation System, J Digit Imaging. 2011 Jun; 24(3): 541–551. Published online 2010 May 22. doi: 10.1007/s10278-010-9303-8.

[20] Gupta M (2019). Digital tranformation in health care: 5 areas of immediate growth. Vision Critical, April 27, 2019. Vancouver, Canada. https://www.visioncritical.com/blog/digital-transformation-health-care

[21] Gupta A, Jha RK (2015). A Survey of 5G Network: Architecture and Emerging Technologies. IEEE Access, Volume 3, 2015, pp. 1206-1232. Digital Object Identifier 10.1109/ACCESS.2015.2461602

[22] Hanover, Judy, (2009). Mobile Yet Connected: The Essence of 21st-Century Healthcare Delivery, Health Industry Insights & Cisco, White Paper, July 2009, Health Industry Insights #HI217032, www.healthindustry-insights.com

[23] IBM (2018). Beyond the hype: A guide to understanding and successfully implementing artificial intelligence within your business. International Business Machines White Paper, October 2018 https://www.ibm.com/downloads/cas/8ZDXNKQ4 [Accedido en octubre 30, 2022]

[24] IEEE 802.11 (2018). IEEE 802.11 Wireless Local Area Networks, http://www.ieee802.org/11/

[25] ITU (2005). The Internet of Things, ITU Internet Reports, International Telecommunications Union, November 2005. https://www.itu.int/net/wsis/tunis/newsroom/stats/The-Internet-of-Things-2005.pdf [Accedido en octubre 30, 2022]

[26] JASON (2017). Artificial Intelligence for Health and Healthcare. JASON. The Mitre Corporation, December 2017. https://www.healthit.gov/sites/default/files/jsr-17-task-002_aiforhealthandhealthcare12122017.pdf [Accedido en octubre 30, 2022]

[27] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y (2017). Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology. 2017;0:e000101. doi:10.1136/svn-2017-000101

[28] Jurong (2011). Jurong Health Services: Transforming Care. Bringing Health to Every Home. APBN Asia-Pacific Biotech News. Vol 15, No 06, June 2011 - Healthcare, pp. 24-27. https://www.asiabiotech.com/15/1506/0024_0027.pdf [Accedido en octubre 30, 2022]

[29] Jurong (2017). Ng Kian Swan. Ng Ten Fong General Hospital, Designing for Well-being: Realizing Benefits for Patients, Visitors and Staff through Best Practice Hospital Design. Jurong Health, 12 May 2017

[30] Kaur J, Kaur K (2017). Internet of Things: A Review on Technologies, Architecture, Challenges, Applications, Future Trends. International Journal of Computer Network and Information Security (IJCNIS), Vol. 9, No. 4, pp. 57-70, 2017. DOI: 10.5815/ijcnis.2017.04.07

[31] Keikhosrokiani P, Zakaria N, Mustaffa N, Wan TC, Sarwar MI, Azimi K (2015), Chapter 30: Wireless Networks in Mobile Healthcare. Mobile Health A Technology Roadmap, pp. 687-726, Springer Series in Bio-/Neuroinformatics,

vol 5. Springer International Publishing, DOI: 10.1007/978-3-319-12817-7, January 2015.

[32] Kocian A, De Sanctis M, Rossi T, Ruggieri M, Del Re E, Jayousi S, Ronga LS, Suffritti R (2011). Hybrid Satellite/Terrestrial Telemedicine Services: Network Requirements and Architecture, 2011 IEEE Aerospace Conference Proceedings, March 2011, DOI: 10.1109/AERO.2011.5747335.

[33] Leiner BM, Cerf VG, Clark DD, Kahn RE, Kleinrock L, Lynch DC, Postel J, Roberts LG, Wolff S (1997). Brief History of the Internet. ISOC Internet Society, 1997. https://www.internetsociety.org/wp-content/uploads/2017/09/ISOC-History-of-the-Internet_1997.pdf [Accedido en octubre 30, 2022].

[34] McKinsey (2017). Artificial Intelligence in health insurance. McKinsey & Company. Healthcare. September 2017. https://healthcare.mckinsey.com/sites/default/files/Artificial%20intelligence%20in%20Health%20Insurance.pdf [Accedido en octubre 30, 2022]

[35] Magana-Rodriguez R, Villarreal-Reyes S, Galaviz-Mosqueda A, Rivera-Rodriguez R, Conte-Galvan R (2015). Telemedicine Services over Rural Broadband Wireless Access Technologies: IEEE 802.22/WRAN and IEEE 802.16 WiMAX. Mobile Health: A Technology Road Map, Springer Series in Bio-/Neuroinformatics, vol. 5, pp. 743-769. Springer International Publishing. 2015. doi: 10.1007/978-3-319-12817-7_32 ISBN: 978-3-319-12817-7

[36] NAM (2019). Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. National Academy of Medicine, Matheny M, Thadaney Israni S, Ahmed M, and D. Whicher D, Editors. NAM Special Publication. Washington, DC. https://nam.edu/wp-content/uploads/2019/12/AI-in-Health-Care-PREPUB-FINAL.pdf [Accedido en octubre 30, 2022]

[37] Ng, J K-Y (2012). Ubiquitous Healthcare: Healthcare Systems and Applications enabled by Mobile and Wireless Technologies, Journal of Convergence, June 2012, Vol. 3, N. 2, pp. 15-20

[38] Nielsen MA, Chuang IL (2011). Quantum Computation and Quantum Information: 10th Anniversary Edition 10th ed (New York, NY, USA: Cambridge University Press) ISBN 1107002176, 9781107002173

[39] Novillo-Ortiz, D. (2016). La eSalud en la Región de las Américas: derribando las barreras a la implementación. Resultados de la Tercera Encuesta Global de eSalud de la Organización Mundial de la Salud. Washington, DC: Organización Panamericana de la Salud.

[40] Nussbaum GM, Ault SP (1999). Using Intranets in Healthcare. Journal of Healthcare Information Management, vol. 13, no. 1, pp. 41-55, Spring 1999, Healthcare Information and Management Systems Society and Jossey-Bass Inc., Publishers

[41] OECD (2008). The Looming Crisis in the Health Workforce: How can OECD Countries Respond? 95 p. Organization for Economic Co-operation and Development, OECD Health Policy Studies, OECD PUBLICATIONS, Paris, France. 978-92-64-05043-3 http://www.oecd.org/els/health-systems/41509461.pdf

[42] panaméricaine de la Santé, O., & mondiale de la Santé, O. (2016). La eSalud en la Región de las Américas; derribando las barreras a la implementación. Resultados de la Tercera Encuesta Global de eSalud de la Organización Mundial de la Salud [Internet]. Washington, DC: OPS.

[43] Ray PP (2018). A survey on Internet of Things architectures. Journal of King Saud University. Computer and Information Sciences (2018) 30, 291-319. http://dx.doi.org/10.1016/j.jksuci.2016.10.003

[44] Roper RA, Anderson KM, Marsh CA, Flemming AC (2013). Health IT-Enabled Quality Measurement: Perspectives, Pathways, and Practical Guidance. (Prepared by Booz Allen Hamilton, under Contract No. HHSA290200900024I.) AHRQ Publication No. 13-0059-EF. Rockville, MD: Agency for Healthcare Research and Quality. September 2013.

[45] Royal Society (2006). Digital healthcare: the impact of information and communication technologies on health and healthcare. The Royal Society. Digital Healthcare. December 2006. https://royalsociety.org/-/media/Royal_Society_Content/policy/publications/2006/8218.pdf [Accedido en octubre 30, 2022]

[46] Sandhya Rani BK, Bhat S, Mukhopadhyay A (2017). A Survey of Wireless Technologies and Vertical Handoff Techniques from the Perspective of Telemedicine Scenarios, Conference: 2017 International Conference on Communication and Signal Processing (ICCSP), April 2017, pp. 1246-1251, DOI: 10.1109/ICCSP.2017.8286580.

[47] Schumacher A (2017). Blockchain & Healthcare - 2017 Strategy Guide. Method. June 2017 https://www.researchgate.net/profile/Axel_Schumacher/publication/317936859/inline/jsViewer/595244e7458515a207f7d497 [Accedido en octubre 30, 2022]

[48] Schork NJ (2019). Artificial Intelligence and Personalized Medicine. In: Von Hoff D., Han H. (eds) Precision Medicine in Cancer Therapy. Cancer Treatment and Research, vol 178. Springer, Cham. pp. 265-283, DOI https://doi.org/10.1007/978-3-030-16391-4_11

[49] Shahin, Jamal (2010). The International Telecommunications Union. Jaargang 34, nr. 154, 2010/2 pp. 11-16

[50] Scott RE, Saeed A (2008). Global eHealth - Measuring Outcomes: Why, What, and How. /making:the eHealth> connection*. Bellagio, Italy. July 13-August 8, 2008, 28p.

[51] SSA (2012), "NOM-024-SSA3-2012 Sistemas de Información de registro electrónico para la salud. Intercambio de información en salud," Diario Oficial de la Federación, pp. 79-96, 2012. https://dof.gob.mx/nota_detalle.php?codigo=5280847&fecha=30/11/2012#gsc.tab=0 [Accedido en octubre 30, 2022]

[52] SSA (2012b) "Certificación NOM-024-SSA3-2012" Acciones y Programas, https://www.gob.mx/salud/acciones-y-programas/certificacion-nom-024-ssa3-2012#:~:text=La%20NOM%2D024%2DSSA3%2D,registren%2C%20intercambien%20y%20consoliden%20informaci%C3%B3n. [Accedido en octubre 30, 2022]

[53] Suciu G, Suciu V, Martian A, Craciunescu R, Vulpe A, Marcu I, Halunga S, Fratu O (2015). Big Data, Internet of Things and Cloud Convergence – An Architecture for Secure E-Health Applications. J Med Syst (2015) 39: 141DOI 10.1007/s10916-015-0327-y

[54] Tchao ET, Diawuo K, Ofosu WK (2017). Mobile Telemedicine Implementation with WiMAX Technology: A Case Study of Ghana. J Med Syst. 2017 Jan; 41(1):17. https://doi.org/10.1007/s10916-016-0661-8

[55] Wager KA, FW Lee, Glaser JP (2017). Health Care Information Systems: A Practical Approach for Health Care Management (4th Ed.), Mar 27, 2017. ISBN-13: 978-1119337188. ISBN-10: 1119337186

[56] Wang Z, Gu H (2009). A Review of Telemedicine in China. Online Journal of Space Communication. Issue No. 14: Satellites and Health. Winter 2009. ISSN: 1542-0639A http://spacejournal.ohio.edu/issue14/research_china.html

[57] WHO (2011). World Health Organization. Safety and security on the Internet: challenges and advances in Member States: based on the findings of the second global survey on eHealth. (Global Observatory for eHealth Series, v. 4). World Health Organization 2011, WHO Library Cataloguing-in-Publication Data. ISBN 978 92 4 156439 7. https://www.who.int/goe/publications/goe_security_web.pdf

[58] WHO (2014). World Health Statistics 2014. WHO Press, Geneva, Switzerland.

[59] WHO (2016a). Global diffusion of eHealth: Making universal health coverage achievable. Report of the third global survey on eHealth, Global Observatory for eHealth, World Health Organization 2016. ISBN 978-92-4-151178-0. http://africahealthforum.afro.who.int/first-edition/IMG/pdf/global_diffusion_of_ehealth_-_making_universal_health_coverage_achievable.pdf [Accedido en octubre 30, 2022]

[60] WHO (2018). mHealth - Use of appropriate digital technologies for public health, Seventy-first World Health Assembly, WHA71/A71/20, Geneva, Switzerland, 21-26 May 2018, http://apps.who.int/gb/ebwha/pdf_files/WHA71/A71_20-en.pdf

[61] Yew HT, Supriyanto E, Satria MH, and Hau YW (2016). A Vertical Handover Management for Mobile Telemedicine System using Heterogeneous Wireless Networks. International Journal of Advanced Computer Science and Applications, Vol. 7, No. 7, 2016.

[62] Usman, Mohammed (2015). Intranet and its Significance in an Organization. June 2015. doi:10.13140/RG.2.1.3612.4326

El cuidado de la obra estuvo a
cargo de Montiel & Soriano
Editores S.A. de C. V. El tamaño
del archivo es de 51.7 MB

La digitalización avanza cada vez a mayor velocidad. Esto aplica y lo vemos cada vez más en distintos ámbitos, como el educativo, económico, social y, particularmente, el de salud. Con la digitalización, avanza también el uso de dispositivos de cómputo que, cada vez más, generan grandes cúmulos de datos que demandan desafíos científicos - tecnológicos para poder almacenarlos, accederlos, procesarlos y obtener información útil para la toma de decisiones. Resolver estos desafíos es un reto y una oportunidad para la creación de mejores productos y servicios.

El área de estudio que aborda los problemas y retos relacionados con la producción de datos a una gran velocidad, de gran variedad y en un creciente volumen, es conocida como Big Data. En este libro se presentan algunas tecnologías emergentes, métodos, algoritmos, aplicaciones y contribuciones de la comunidad científico-académica que trabaja en el desarrollo del Big Data en Salud. Estas contribuciones tienen el objetivo de coadyuvar a la creación de sistemas de información útiles y eficientes que analicen el gran volumen de datos en salud de los que actualmente se disponen, y que esto permita avanzar hacia un servicio de salud más eficiente, con un impacto en el ahorro de costos y una atención médica oportuna y más personalizada.

Este libro es producido en el marco de los Programas Nacionales Estratégicos - PRONACES - en Salud, en el Proyecto Nacional de Investigación e Incidencia - PRONAII - Ciencia de datos en Salud, del proyecto específico No. 41756 "Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud".

9 786078 857234

CONACYT
Consejo Nacional de Ciencia y Tecnología

Cinvestav

BUAP

Montiel & Soriano
EDITORES