

Modelos de Machine Learning para la mejora de la precisión, seguridad y eficiencia en la gestión de datos biométricos

Carlos E. Alvez, Graciela R. Etchart, Juan J. Aguirre, Ernesto M. Miranda, Silvia E. Ruiz, Juan C. Teza, Jesús F. FROLA da SILVA, Kevin Inda

Autores: Facultad de Ciencias de la Administración. Monseñor Tavella 1424. Concordia. CP (3200), Entre Ríos. Argentina.

Contacto: carlos.alvez@uner.edu.ar

ARK: <https://id.caicyt.gov.ar/ark:/s22504559/0nog529fw>

Resumen:

La gestión de acceso es un problema cada vez más recurrente para los usuarios, dado que normalmente no es sencillo recordar varias contraseñas para accesos a diferentes sistemas, especialmente para aquellos con altos niveles de seguridad. Para resolver este problema se utilizan ampliamente los sistemas biométricos. A pesar de su importancia, estos sistemas aún presentan desafíos en cuanto a precisión, seguridad y eficiencia: precisión para disminuir las tasas de falsos positivos y falsos negativos; seguridad de datos biométricos, referido a la protección ante ataques tanto a nivel de captura, transmisión como almacenamiento de las plantillas recogidas durante los procesos de enrolamiento, identificación y verificación; eficiencia en el proceso de identificación (búsqueda 1 a n), dado que las bases de datos biométricas pueden ser enormes. Por lo antes expuesto, en este trabajo, se propondrá en primer lugar, para tratar el problema de la precisión utilizando modelos basados en técnicas de ML para el reconocimiento biométrico. En segundo lugar, se analizará el diseño de una plataforma segura para la captura, transmisión y almacenamiento de información, para lo cual se desarrollarán enfoques basados en métodos de detección/identificación de ataques mediante ML y/o sistemas recomendadores. Por último, dado que las comparaciones entre plantillas no son triviales (como es el caso de datos alfanuméricos), es importante encontrar formas que reduzcan el espacio de búsqueda o métodos de indexación de dominio específicos, se analizarán métodos de indexación que permitan reducir los espacios de búsquedas en el proceso de identificación.

Palabras clave: Bases de Datos, Machine Learning, Biometría, Seguridad de Datos, Eficiencia.

Objetivos propuestos y cumplidos

Objetivo general: proponer modelos basados en Machine Learning (ML) y mecanismos de bases de datos para la mejora en la precisión, seguridad de datos y eficiencia en la gestión de datos biométricos.

Objetivos específicos propuestos:

1. Identificar los diferentes métodos de ML utilizados en las aplicaciones biométricas para tratar las problemáticas relacionadas con la precisión.
2. Analizar la aplicabilidad de técnicas de ML y sistemas recomendadores para la detección de ataques inteligentes a sistemas biométricos.
3. Analizar diferentes métodos de indexación y/o segmentación de la base de datos para mejorar el proceso de identificación.
4. Proponer modelos basados en técnicas de ML que permitan mejorar la precisión en el reconocimiento biométrico.
5. Proponer métodos de detección/identificación de ataques mediante ML y sistemas recomendadores.
6. Diseñar modelos que permitan la segmentación e indexación para mejorar la eficiencia en el proceso de identificación.

Objetivos específicos cumplidos: los objetivos propuestos han sido cumplidos. En la sección de resultados, en la Tabla 1, se detallan los artículos presentados relacionados por objetivo.

Marco teórico y metodológico

El presente proyecto es una continuidad del PID 7044 “*Gestión de Datos Biométricos en Bases de Datos Objeto-Relacionales*”, en el cual se propuso una arquitectura genérica para la gestión de datos biométricos en Bases de Datos Objeto-Relacionales. Dicha arquitectura contempla la captura y procesamiento de rasgos biométricos, interoperabilidad mediante la gestión de registros de transacciones ANSI/NIST ITL 1-2011 y el sistema de reconocimiento biométrico [1]. Además, para implementar esta arquitectura se desarrolló una extensión de un *Sistema de Gestión de Datos Biométricos en Bases de Datos Objeto-Relacionales (SGBDOR)* que permite la gestión simple y eficiente de los datos del dominio, sobre todo, las consulta SQL se pueden realizar utilizando los métodos de búsqueda definidos en el sistema de tipos específicos de la extensión, aún por usuarios que no sean expertos en el dominio [2]. Los logros obtenidos en el proyecto anterior mostraron resultados satisfactorios en lo referente a gestión de metadatos e interoperabilidad, que fueron sus principales objetivos. Sin embargo, desde el punto de vista de los sistemas de reconocimiento biométricos, se presentan otros desafíos como la precisión, seguridad de datos y eficiencia. En este sentido, las diferentes técnicas de Machine Learning (ML), como complementos de los mecanismos existentes en los sistemas de gestión de bases de datos, ofrecen diversas opciones para afrontar los mismos.

Respecto a la precisión, se hace referencia a la probabilidad de errores de verificación e identificación (falsos positivos y falsos negativos). Cuando se comparan rasgos biométricos, uno de los parámetros a definir es el umbral (diferencia) que divide los casos aceptados y rechazados. Esto lleva a que, si el umbral es muy pequeño, las tasas de falso negativos aumentarán, y si el umbral es muy amplio, aumentarán las tasas de falso positivos. Una de las alternativas que se han propuesto es la fusión, tanto mul-

tibiométrica como monobiométrica. Casos de fusión multibiométrica se presentan en [3], que presenta un sistema multimodal basado en un modelo de fusión de nivel de puntuación utilizando iris, huella dactilar y palmar y en [4], donde diseñan un marco de fusión multimodal para imágenes de rostro y huellas dactilares utilizando una matriz de imágenes basada en bloques. Un caso de fusión monobiométrico se presenta en [5], donde combina diferentes dispositivos de un mismo rasgo (huella dactilar), para mejorar la precisión. Estos casos de fusión presentados, si bien muestran una mejora importante en lo que hace a la precisión, tienen algunos inconvenientes. En primer lugar, debido a que debe utilizarse más de un dispositivo de captura, puede afectar la eficiencia tanto a nivel de captura como en los procedimientos de verificación e identificación y, por otro lado, en la usabilidad debido a que los usuarios (aún cooperativos), pueden presentar alguna resistencia en cuanto a utilizar más de un dispositivo. Otros trabajos tratan de lidiar con estos problemas utilizando diferentes técnicas de ML en lugar de enfoques estadísticos (comúnmente utilizados en biometría) para mejorar la precisión. Un caso se presenta en [6], donde se realizan diferentes experimentos con rasgos de la huella dactilar, la firma en línea y fuera de línea utilizando máquinas de soporte vectorial (SVM), algoritmo k-vecinos más cercanos (k-NN), red neuronal de percepción multicapa (MLP), clasificador de árboles extra (ETC) y clasificador de bosques aleatorios (RFC). De los resultados de los experimentos se desprende que los clasificadores de aprendizaje en conjunto superaron a los métodos convencionales. El enfoque de ML también se puede combinar con fusión, como se presenta en [7], donde se utiliza Histograma de gradientes orientados (HOG), luego se aplica el modelo de mezcla gaussiano (GMM) para encontrar valores de función de densidad de probabilidad (PDF) y luego se combinan estas características usando fusión de nivel de puntuación. En este sentido se orientará este proyecto para la mejora de la precisión, basado principalmente en el enfoque de ML.

En lo que respecta a la seguridad de los datos biométricos, se hace referencia a que los sistemas de autenticación biométrica son vulnerables a una serie de ataques, especialmente a los ataques de reproducción/falsificación.

Cuando el sistema está sujeto a ataques de falsificación, es a menudo acompañada de la filtración de información biométrica del usuario, que también es una forma de vulneración de la privacidad. En los sistemas de autenticación biométrica, hay dos maneras posibles de revelar información privada:

- Divulgación de la privacidad en un entorno práctico: las personas pueden revelar su información biológica en cualquier momento de la vida real, como las huellas dactilares dejadas después de tocar algunos objetos, la firma que queda al pagar con tarjeta de crédito, la imagen del rostro e incluso imagen del iris contenida en fotos de alta definición, la voz grabada en áreas públicas, etc.
- Divulgación de la privacidad en un entorno de red: la información biométrica puede ser robada, manipulada o usada durante la captura, el almacenamiento y la transmisión.

Por lo tanto, uno de los principales desafíos en la seguridad de los datos biométricos es la protección de los mismos tanto a nivel de captura, transmisión como almacenamiento. Meng junto a otros autores [8] señalaron en un sistema genérico de autenticación biométrica para teléfonos móviles, ocho posibles puntos de ataque y estudiaron las medidas posibles para contrarrestarlos. Obviamente, la seguridad y la privacidad

de los datos biométricos es una cuestión crítica. Sin embargo, no se ha tenido plenamente en cuenta en muchos de los sistemas biométricos existentes. Según Rui & Yan en [9], muchos investigadores no han tenido en cuenta los posibles ataques a la hora de diseñar sus sistemas. En la bibliografía consultada aún no se encuentra una revisión exhaustiva sobre la seguridad de datos y preservación de la privacidad en la autenticación biométrica. Entre las propuestas existentes para tratar este problema, en [10] se ofrece una visión general de las teorías, metodologías, técnicas, normas y marcos de los sistemas biométricos. Además, se realiza un análisis de los equipos utilizados, la extracción de las características, la seguridad de los almacenamientos y realiza propuestas para garantizar la corrección de los resultados obtenidos a partir de un mecanismo de decisión inteligente. En [11], se proporciona un enfoque analítico para evaluar el rendimiento y la resistencia a ataques del BAS (Sistema de autenticación biométrico), así como su compensación, mediante el modelado de los datos biométricos y el estudio del efecto de la extracción de características y las configuraciones ML en el procesamiento de los datos.

Para la protección de los datos biométricos, se pueden utilizar diferentes mecanismos similares a la protección de datos de cualquier dominio, como en [12], que describe la arquitectura basada en la colocación del módulo de criptografía de curva elíptica dentro del software de gestión de bases de datos (DBMS) y en [13], que propusieron una mejora de dos esquemas de autenticación de contraseñas basados en el timestamp y nonce para evitar el ataque de spoofing del servidor. Sin embargo, los ataques a sistemas biométricos pueden ser efectuados utilizando diferentes mecanismos inteligentes. En este sentido, en [14] se presenta el concepto de Máquinas Adversarias, estudiando el problema de la seguridad de los sistemas biométricos cuando operan en presencia de atacantes inteligentes y adaptables que manipulan los datos para comprometer el funcionamiento normal del sistema. En [9] se analizan las amenazas de la autenticación biométrica y se propone una serie de criterios con respecto a la autenticación segura y la preservación de la privacidad.

Para fortalecer la seguridad y la privacidad de los datos biométricos, una técnica de ML y de inteligencia artificial que ha demostrado ser muy prometedora es el sistema de recomendaciones. Estos sistemas se pueden utilizar para generar listas priorizadas de acciones de defensa [15], para detectar amenazas internas [16], para supervisar la seguridad de la red [17] y para acelerar otros análisis [18].

Respecto a los problemas relacionados a la eficiencia, se presentan principalmente por la necesidad de utilizar un método de identificación (comparación de 1 a n), con n muy grande. Para esto, las soluciones pueden consistir en métodos de indexación y/o segmentación de la base de datos. Este tema ha sido poco abordado en la literatura. Un trabajo relacionado a la indexación de datos biométricos se presenta en [3], donde se utiliza k-tree para indexación de plantillas de iris. Existen otras estructuras de índice basadas en métricas que podrían utilizarse para realizar índices de dominio en área de biometría, como M-tree [19] y Metric Index [20]. Una alternativa a la indexación para mejorar la eficiencia, es utilizar técnicas de ML (clustering) para la segmentación de la base de datos biométrica. De esta manera, se puede reducir el espacio de búsqueda mejorando los tiempos en el proceso de identificación. Un caso se puede ver en [21], donde se utiliza K-means para segmentar una base de datos de huellas dactilares. Básicamente, segmenta la base de datos por similitud. Para la búsqueda de una huella

particular, comienza por el segmento con mayor similitud, luego por el siguiente, y así sucesivamente. Estas dos alternativas para mejorar la eficiencia no son incompatibles. De hecho, se puede dividir la base de datos en segmentos e indexar cada segmento como se presenta en [22], donde se utiliza k-means para segmentar la base de datos y Hashing Multi-Index (MIH) para indexar los segmentos.

Otro punto relacionado con la eficiencia, aunque fuera del alcance de este proyecto, es cómo afecta a la eficiencia el tratamiento de la seguridad de los datos biométricos. En este sentido en [11], utilizan técnicas de extracción de características y de ML para mejorar el rendimiento, pero también reducen el esfuerzo del adversario para romper el sistema utilizando datos falsificados.

Metodología utilizada para lograr los objetivos planteados:

1. Se realizó un relevamiento continuo de los trabajos científicos y técnicos relacionados con el objeto de estudio y haciendo especial hincapié en: métodos de ML utilizados en el dominio de los sistemas biométricos, tratamiento de la precisión, seguridad de datos y eficiencia en la gestión de datos biométricos, métodos de indexación y/o segmentación de la base de datos biométrica.
2. Se identificaron y compararon las diferentes técnicas de ML utilizadas en las aplicaciones biométricas para tratar problemáticas relacionadas con la precisión, para lo que se utilizaron bases de datos biométricas para el testeo de indicadores.
3. Se analizó la aplicabilidad de técnicas de ML y sistemas recomendadores a la detección de ataques inteligentes hacia sistemas biométricos. Se estudió la aplicabilidad de los servicios de razonamiento rebatible basados en preferencias en el dominio de los sistemas de reconocimiento biométrico.
4. Se estudiaron diferentes métodos de indexación y/o segmentación de la base de datos para mejorar el proceso de identificación biométrica (búsqueda 1 a n). Por un lado, se realizó un estudio comparativo de diferentes métodos de indexación de plantillas biométricas generadas en el laboratorio, teniendo como indicador el tiempo medio de recuperación y, por otro lado, se analizaron algoritmos de ML (clustering) y se midieron las capacidades de segmentación de la base de datos para reducir los espacios de búsqueda.
5. Se propuso un modelo utilizando técnicas de ML en base a los resultados obtenidos en los indicadores de precisión del punto 2. Las técnicas de ML se utilizaron principalmente para la comparación de plantillas biométricas.
6. En base a los resultados obtenidos en el punto 3, se diseñó una arquitectura segura para la gestión de datos biométricos utilizando métodos ML y sistemas recomendadores para la detección y respuesta frente ataques inteligentes.
7. Se presentó un marco metodológico para sistemas biométricos con sus respectivas fases, tareas y salida, el cual establece un vocabulario común entre los participantes del proyecto constituyendo una guía para coordinar las actividades.

Resultados

Los principales resultados se muestran a continuación. Para cada caso, se indica su relación con los artículos presentados que se detallan en la Tabla 1:

- **Una profundización del estado del arte en cuanto a los métodos ML y su uso para la identificación de personas mediante sistemas biométricos.** Se identificaron los diferentes métodos de aprendizaje automático utilizados en las aplicaciones biométricas y se trataron las problemáticas relacionadas con la precisión en el reconocimiento biométrico (Tabla 1 - artículos A, C, E y O).

Para evaluar la performance de los algoritmos ML, en general, y en particular para el aspecto que interesa en esta actividad del proyecto (la precisión), se estudiaron diferentes métricas a utilizar. Además de la precisión, se analizaron otras métricas relacionadas con la performance como son el costo computacional y los tiempos de respuestas. Por otro lado, se necesitó disponer de bases de datos con capturas de los rasgos de interés, por lo que se analizaron bases de datos disponible para el rasgo biométrico estudiado (iris en nuestro caso). Se puede ver el detalle de los pasos llevados a cabo para la evaluación de performance en (Tabla 1 - artículo M).

- **Un estudio detallado de los métodos de indexación/segmentación de la base de datos para mejorar la performance en los métodos de identificación.** Se analizaron diferentes métodos de indexación y/o segmentación de la base de datos para mejorar el proceso de identificación (utilizando el rasgo de iris). Se estudiaron diferentes alternativas que ofrecen las bases de datos para mejorar la eficiencia en el proceso de identificación.

Para mejorar la eficiencia del proceso de identificación biométrica (búsqueda 1 a n), se plantearon dos alternativas: por un lado, un estudio sobre las diferentes técnicas de indexación de plantillas biométricas teniendo como indicador el tiempo medio de recuperación y, por otro lado, algoritmos de ML (clustering) para segmentación de la base de datos para reducir los espacios de búsqueda. De ambas alternativas, se trabajó con los métodos de indexación, dado que se llegó a la conclusión de que los algoritmos de segmentación existentes, no son aplicables para dimensiones muy altas como es el caso del código del iris.

Luego del análisis de los métodos de indexación que permiten realizar índices de dominio en base al código del iris, se concluyó que el índice M-tree [23] [24] es el que mejor se adapta para para aplicar al código del iris, dado que, con ello, se logra el propósito de segmentar la base en espacios métricos y lograr una mejor performance en las búsquedas 1 a n (Tabla 1 - artículo L).

- **Implementación de una aplicación para la extracción del código del iris métodos de comparación e indexación.** Para la generación masiva de IrisCode, se desarrolló una aplicación basada en diferentes componentes y librerías open source. A esta aplicación, basta con indicarle la carpeta donde se encuentran las imágenes de iris y realiza el proceso automático de la generación del código de iris con sus correspondientes máscaras de ruido (posiciones de bits que deben ser descartados por ruido/oclusión).

- **Propuesta modelo basado en técnicas de Deep Learning para la mejora de la precisión en el reconocimiento de iris.** Como alternativa a las técnicas ML, se analizaron diferentes mecanismos de Deep Learning (DL) y Redes Neuronales Convolucionales (CNN) en búsqueda de una mejora en la precisión en la comparación (matching). Se realizó un análisis comparativo de las diferentes metodologías utilizadas en [25] [26] [27] [28] [29] [30] y los modelos de redes neuronales utilizados para la segmentación del iris.

En base al análisis, se decidió implementar dos arquitecturas de CNN. La primera es la una arquitectura CNN llamada VGG16 y se realizó un análisis comparativo con respecto a otra CNN llamada SNN Timing-based Working Memory Spiking Neural Network o Red neuronal de picos de memoria de trabajo basada en el tiempo (TW-SNN). El estudio se enfocó en comparar ambas arquitecturas para determinar cuál obtiene los mejores resultados según las métricas de precisión y eficiencia.

Como resultado del análisis comparativo se determinó que las redes neuronales mencionadas funcionan con una alta tasa de precisión. Sin embargo, la CNN VGG16 resulta más sencilla de implementar y bajo las métricas utilizadas tiene mejores prestaciones que la implementación y el rendimiento que la CNN TW-SNN. Como contrapartida, se debe destacar que las CNN adolecen de un problema de eficiencia energética. Comparadas con redes neuronales naturales y ML, son tremendamente ineficientes, pues requieren más computación y más energía. Los trabajos relacionados se presentan en: (Tabla 1 - artículos I, J, K, N y S).

- **En base al análisis de técnicas de ML y sistemas recomendadores para la detección de ataques inteligentes hacia sistemas biométricos,** se realizó una propuesta que consiste en una arquitectura con capacidades para detectar una acción intrusiva en el sistema biométrico y recomendar una acción de respuesta con el fin de mitigar o reducir el daño causado por la intrusión.

En relación a los procesos biométricos, la detección y respuesta frente a intrusiones puede ser utilizada como una capa de defensa contra los ataques que surgen en el canal de comunicación. Aunque numerosos sistemas detectores de intrusión han sido estudiados y efectivamente aplicados en diferentes escenarios de la vida real, la naturaleza dinámica del dominio donde se producen los ataques en ocasiones conduce a situaciones de información en conflicto, principalmente cuando se razona en presencia de información incompleta y potencialmente contradictoria. En este escenario, la argumentación representa una alternativa útil para hacer frente a estas cuestiones [31]. El mecanismo de inferencia sobre el cual se basan, permite decidir entre conclusiones contradictorias y adaptarse fácilmente a entornos cambiantes.

La argumentación constituye un área de estudio de especial interés en el ámbito de la IA, principalmente, porque permite razonar con información incompleta e incierta, y permite manejar inconsistencia en los sistemas basados en conocimiento. Este tipo de razonamiento es particularmente atractivo para sistemas recomendadores.

Una de las principales contribuciones en relación a este objetivo es la propuesta de un Sistema de Respuesta a Intrusión Basado en Argumentación para extender las capacidades de razonamiento de los sistemas biométricos incorporando argumentación al proceso de responder frente a una intrusión. Un aspecto clave del framework propuesto es la recomendación de contramedidas a desplegar basada en mecanismos de inferencia. La contramedida se modela a través de una serie de postulados básicos, que se pueden extender o relajar fácilmente. En una posible instanciación del framework propuesto, estos postulados se traducen en reglas rebatibles DeLP [32], que permiten inferir conclusiones tentativas con respecto a la contramedida frente a una intrusión detectada. El proceso de argumentación llevado a cabo por el servicio de razonamiento basado en Programación Lógica Rebatible [33][34] permite filtrar selectivamente información para sugerir acciones y disminuir los tiempos de respuesta ante la presencia de una posible intrusión en un sistema biométrico basado en red.

Como prueba de concepto para favorecer la comprensión del funcionamiento del principal componente de la arquitectura propuesta se presentan dos instanciaciones utilizando formalismos concretos para representar el conocimiento, las consultas y respuestas. Una instanciación utiliza el sistema argumentativo de programación en lógica rebatible, siendo una de sus características destacadas la posibilidad de manejar información contradictoria y/o incompleta. La otra instanciación, basada en el formalismo propuesto por Burón Brarda et al. [35][36][37], posibilita llevar a cabo el proceso de elección de la contramedida adecuada a partir de un proceso de decisión multicriterio basado en argumentación. Esto permite dar prioridad a aquellos atributos que mejor definen una recomendación relevante, en función de las políticas de seguridad que se deseen aplicar y diversas alternativas de respuesta. Una característica común a ambos formalismos y de gran relevancia, es que las conclusiones obtenidas pueden explicarse fácilmente mediante el proceso de razonamiento argumentativo, el cual permite ofrecer explicaciones como conjuntos coherentes de razones a favor o en contra de la acción de la respuesta recomendada. Los argumentos que se construyen permiten explicar los resultados del razonamiento a los responsables de la toma de decisiones humanas, de manera que se aclare la situación y se mejore la calidad de las decisiones. Estas explicaciones razonadas son útiles para generar confianza en los usuarios y favorecer la retroalimentación en el sistema.

Por otra parte, uno de los componentes de la arquitectura propuesta incluye un sistema de detección de intrusión basado en uso indebido. En este tipo de sistemas, el poblamiento de la base de firmas es una operación costosa tanto intelectualmente como en tiempo para un humano. Con el propósito de mejorar la actualización de las bases de datos de firmas del detector de intrusión, se realizó un estudio preliminar sobre distintas herramientas de ML, a fin de proponer una solución aceptable en relación a ese propósito. Como parte de la solución propuesta, se incorpora un componente opcional que utiliza una herramienta basada en aprendizaje automático para generar firmas orientadas a la detección de intrusiones personalizadas para el sistema biométrico. En esta tarea se utilizó SIRUS (Stable and Interpretable RULe Set)

[38], un nuevo algoritmo de reglas para la clasificación, que adopta la forma de una lista corta de reglas, que brinda un útil nivel de interpretabilidad y favorece la traducción del modelo de predicción a reglas IDS personalizadas para la detección de ataques dirigidos al sistema biométrico, al tiempo que se mantiene un entorno explicable e interpretable. (Tabla 1 - artículos B, D, F, P, Q y R).

- **Se propuso un marco metodológico general para la implementación de sistemas biométricos.** El marco metodológico basado en CRISP-DM [39], cubre los aspectos relacionados a la aplicación de un modelo de ML para sistemas de reconocimiento biométrico (Figura 2). Este marco de trabajo brinda una orientación que permite que los distintos proyectos y profesionales que desarrollen aplicaciones basadas en ML puedan mejorar el éxito y la eficiencia de las mismas. (Tabla 1 - artículos E, G y S).

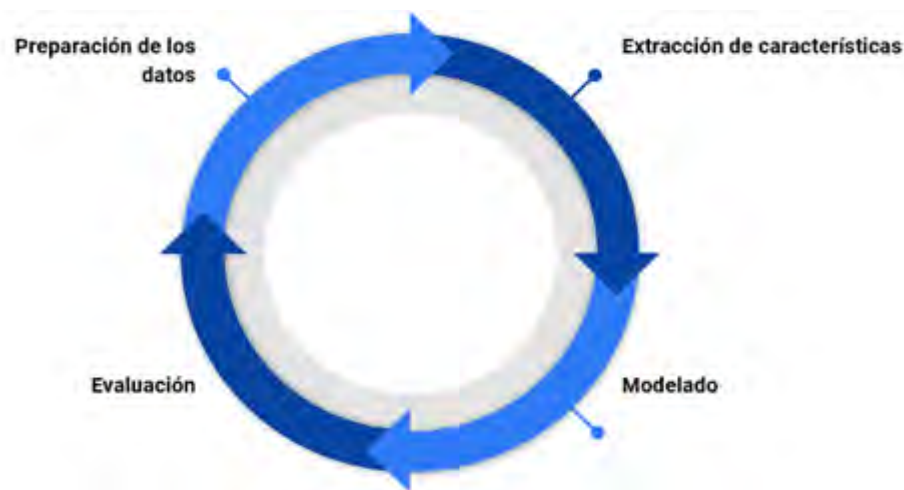


Figura 2: Fases del marco metodológico propuesto. Fuente: Ruiz S., et al. "Hacia la definición de un marco metodológico para el desarrollo de un sistema de reconocimiento biométrico mediante técnicas de Machine Learning". WICC. 2022.

El marco metodológico propuesto presenta un enfoque más flexible e iterativo entre las distintas fases, donde se coloca en el centro el análisis y exploración de los datos permitiendo la redefinición de los objetivos. Esto implica una nueva perspectiva, ya que la metodología CRISP-DM se diseñó pensando en las tareas asociadas para el cumplimiento de los objetivos definidos a priori, siguiendo una trayectoria secuencial.

Para verificar las bondades del marco metodológico, el mismo se aplicó a un caso de estudio concreto de un sistema de reconocimiento de iris. En cada fase, se detallaron los métodos y técnicas elegidos para realizar los procesos de localización, normalización y extracción de características de la imagen del iris. En la fase de modelado, las tareas se centraron en entrenar un modelo de ML para así poder realizar inferencias en base al conjunto de datos de entrenamiento.

Los resultados antes mencionados, se presentaron en diferentes artículos de congresos nacionales e internacionales, así como también, se hicieron un par de presentaciones en una revista para su evaluación y publicación. Los mismos se encuentran detallados por objetivos en la Tabla 1.

Tabla 1: Artículos relacionados por objetivos.

ID	Artículos publicados	Objetivo
A	Frola, F., Chesñevar, C. I., Alvez, C. E., Etchart, G., Miranda, E., Ruiz, S., Teze, J. C. (2019, Junio). Framework SDF Machine Learning en transacciones financieras y detección temprana de fraudes. In XXI Workshop de Investigadores en Ciencias de la Computación (WICC 2019, Universidad Nacional de San Juan)	1
B	Etchart, G., Teze J. C., Alvez C., Simari G. R. Implementación de herramientas computacionales para el manejo de preferencias en servicios de razonamiento rebatible. XXI Workshop de Investigadores en Ciencias de la Computación – WICC 2019, Facultad de Ciencias Exactas, Físicas y Naturales - Universidad Nacional de San Juan, San Juan	5
C	Frola, F., Alvez, C., Chesñevar, C. I. (2020, Junio). Un primer acercamiento a un modelo predictivo ajustable por umbrales para detección de fraudes financieros. ASAI, Simposio Argentino de Inteligencia Artificial (49 JAIIO, Buenos Aires (CABA), Argentina)	1
D	Etchart G., Teze J. C., Alvez C., Martínez M. V., Simari G. I. Hacia la Detección de Intrusión en Sistemas Biométricos Utilizando Argumentación Rebatible. 8° CoNaISI 2020, Universidad Tecnológica Nacional - Facultad Regional San Francisco	5
E	Ruiz S., Carlos Alvez, Graciela Etchart, Ernesto Miranda, Juan Aguirre, Fabian Frola, Kevin Inda. Machine Learning en biometría: adaptación de la metodología CRISP DM en un sistema de reconocimiento biométrico. 8° CoNaISI 2020, Universidad Tecnológica Nacional - Facultad Regional San Francisco	1, 4, 6
F	Etchart Graciela R., Teze Juan C.L., Alvez Carlos E., Martínez M. Vanina, Simari Gerardo I. Utilizando Argumentación Rebatible en la Detección de Intrusión en Sistemas Biométricos. XXIII Workshop de Investigadores en Ciencias de la Computación – WICC 2021, Universidad Nacional de Chilecito, La Rioja	5
G	Ruiz, S., Alvez, C., Etchart, G., Miranda E. (2022, Abril). Hacia la definición de un marco metodológico para el desarrollo de un sistema de reconocimiento biométrico mediante técnicas de Machine Learning. XXIV Workshop de Investigadores en Ciencias de la Computación (WICC 2022, Universidad de Champagnat)	6
H	Ruiz S., Alvez C., Marco Metodológico para el Desarrollo de un Sistema de Reconocimiento Biométrico Mediante Técnicas de Machine Learning. I Congreso Argentino de Ciencias de la Computación (CACIC).2022. Universidad Nacional de La Rioja	6
I	Inda Kevin “Mecanismos de Deep Learning y Redes Neuronales para el Reconocimiento Efectivo del Iris” Jornadas de Jóvenes Investigadores. AUGM. Ciencia, tecnología e innovación. Asociación de Universidades Grupo Montevideo. Facultad San Francisco Xavier. Bolivia	2, 4
J	Kevin Inda. Análisis, Implementación y Comparación de las Redes Neuronales Vgg16 y Tw-Snn para el Reconocimiento del Iris: Un Enfoque Comparativo. Trigésima Edición de las Jornadas de Jóvenes Investigadores de la Asociación de Universidades Grupo Montevideo, República del Paraguay, del 11 al 13 de octubre de 2023	2, 4
K	Kevin Marcelo Inda, Carlos Alvez. Análisis e Implementación de una CNN Basada en la Arquitectura VGG16 para el Reconocimiento del Iris, XXV Workshop de Investigadores en Ciencias de la Computación (WICC 2023), realizado los días 13 y 14 de Abril de 2023 en Junín, Bs. As	2, 4

L	Juan J. Aguirre, Carlos E. Alvez, Graciela E. Etchart. Índices de Dominio para Identificación Eficiente Mediante Código de Iris. XV Workshop de Investigadores en Ciencias de la Computación (WICC 2023), realizado los días 13 y 14 de Abril de 2023 en Junín, Bs. As	3,6
M	Carlos Eduardo Alvez, Silvia Ester Ruiz, Ernesto Martín Miranda, Juan José Aguirre. Evaluación de la performance de métodos de clasificación para el reconocimiento del iris. 11° Congreso Nacional de Ingeniería Informática / Sistemas de Información- CoNaISI 2023, 2 y 3 de noviembre de 2023 San Miguel de Tucumán	2,6
N	Kevin Inda, Deep Learning en Biometría: Análisis e Implementación de una CNN Basada en la Arquitectura VGG16 para el reconocimiento del Iris. XX Jornadas Nacionales de Administración e Informática, Concordia, 22 de noviembre de 2022	2,4
O	Fabián Frola, Carlos Chesñear, Carlos Alvez. Hacia un abordaje probabilístico para detección de fraude financiero en tarjetas de crédito y débito utilizando Machine Learning. XX Jornadas Nacionales de Administración e Informática, Concordia, 22 de noviembre de 2022	1
P	Eduardo Roberto Jacobo, Javier Echaiz, Carlos Eduardo Alvez. Análisis comparativo entre VPN y enlace rentado. XX Jornadas Nacionales de Administración e Informática, Concordia, 22 de noviembre de 2022	5
Q	Graciela R. Etchart, Juan C.L. Teze, Carlos E. Alvez, María V. Martinez, Gerardo I. Simari. Utilizando Argumentación Rebatible en la Detección y Respuesta ante Intrusión en Sistemas Biométricos. XXIX Congreso Argentino de Ciencias de la Computación - CACiC 2023, realizado los días 9 y 12 de octubre en Luján, Bs. As.	5
ID	Artículos presentados para evaluación al momento del informe	Objetivo
R	Graciela Etchart, Juan Carlos Teze, Carlos Eduardo Alvez (en proceso de publicación). Marco de aplicación basado en argumentación para la respuesta a intrusiones en sistemas biométricos. Revista Ingeniare.	5
S	Silvia Ester Ruiz, Carlos Eduardo Alvez, Ernesto Martín Miranda (en proceso de publicación). Caso de estudio: implementación de un marco metodológico en un sistema de reconocimiento de iris mediante técnicas de Machine Learning. Revista Ingeniare.	4,6

Conclusiones

El objetivo de este trabajo fue proponer modelos basados en Machine Learning (ML) y mecanismos de bases de datos para la mejora en la precisión, seguridad de datos y eficiencia en la gestión de datos biométricos, para lo cual se desarrollaron diferentes actividades para los logros de estos objetivos, tomando como rasgo biométrico, el iris humano.

Respecto a la precisión, se realizaron experimentos para analizar la performance entre un método ampliamente utilizado para comparación de código de iris (distancia de Hamming) con el método de Análisis Discriminante Lineal (LDA) que es un método de ML que permite hacer clasificaciones con clases de múltiples características o dimensiones. De la comparación de los resultados obtenidos, se puede observar que el método de ML (LDA) tuvo una mejor performance en todas las métricas obtenidas a partir de las matrices de confusión de cada caso. Por otro lado, se hicieron con métodos de DL, los cuales mostraron mejor performance que el método ML propuesto. Sin embargo, se debe destacar que las CNN adolecen de un problema de eficiencia energética. Comparadas con redes neuronales naturales y ML, son tremendamente ineficientes, pues requieren más computación y más energía.

En cuanto a la seguridad de datos se diseñó una arquitectura con capacidades de razonamiento en los sistemas biométricos, incorporando argumentación al proceso de reacción frente a un ataque inteligente con el fin de mitigar o reducir el daño causado por la intrusión. El proceso de argumentación llevado a cabo por el servicio de razonamiento basado en Programación Lógica Rebatible permite filtrar selectivamente información para sugerir acciones y disminuir los tiempos de respuesta ante la presencia de una posible intrusión en un sistema biométrico basado en red. Por otra parte, el proceso de razonamiento argumentativo permite que las conclusiones obtenidas puedan explicarse fácilmente, ofreciendo explicaciones como conjuntos coherentes de razones a favor o en contra de la acción de la respuesta recomendada. Los argumentos que se construyen permiten explicar los resultados del razonamiento a los responsables de la toma de decisiones humanas, de manera que se aclare la situación y se mejore la calidad de las decisiones. Estas explicaciones razonadas son útiles para generar confianza en los usuarios y favorecer la retroalimentación en el sistema. Se considera que esta propuesta supone un avance para conseguir un proceso de reconocimiento más seguro y una toma de decisiones sobre seguridad más informada; sin embargo, la investigación debe seguir persiguiendo garantizar la confidencialidad, la integridad y la disponibilidad de un sistema biométrico.

Respecto a la eficiencia, se analizaron dos alternativas: por un lado, un estudio sobre las diferentes técnicas de indexación de plantillas biométricas teniendo como indicador el tiempo medio de recuperación y, por otro lado, algoritmos de ML (clustering) para segmentación de la base de datos para reducir los espacios de búsqueda. Dado que los algoritmos de segmentación existentes, no son aplicables para dimensiones muy altas como es el caso del código del iris, se trabajó con métodos de indexación. Luego del análisis de los métodos de indexación que permiten realizar índices de dominio en base al código del iris, se concluyó que el índice M-tree es el que mejor se adapta para aplicar al código del iris, dado que, con ello, se logra el propósito de segmentar la base en espacios métricos, y lograr una mejor performance en las búsquedas 1 a n. Para la implementación del mismo, se desarrolló una aplicación que permite la generación masiva de IrisCode y los métodos de comparación tradicionales.

Por último, se presentó un marco metodológico basado en CRISP-DM, que cubre los aspectos relacionados a la aplicación de un modelo de ML para sistemas de reconocimiento biométrico. A la vez, se implementó un caso de uso para mostrar las bondades del mismo. Para verificar las bondades del marco metodológico, el mismo se aplicó a un caso de estudio concreto de un sistema de reconocimiento de iris. En cada fase, se detallaron los métodos y técnicas elegidos para realizar los procesos de localización, normalización y extracción de características de la imagen del iris. En la fase de modelado, las tareas se centraron en entrenar un modelo de ML para así poder realizar inferencias en base al conjunto de datos de entrenamiento. En este sentido, el marco metodológico propuesto, presenta un enfoque más flexible e iterativo entre las distintas fases, donde se coloca en el centro el análisis y exploración de los datos permitiendo la redefinición de los objetivos sin seguir una trayectoria secuencial.

Como trabajo futuro, se pretende avanzar en actividades relacionadas al aseguramiento de la calidad en cada una de las tareas propuestas en el marco metodológico, que permitan afrontar los desafíos de las aplicaciones de ML identificados en forma de riesgos. Otra línea de trabajo a seguir es en la implementación del marco propuesto en un sistema de reconocimiento biométrico basado en otros rasgos de interés (como por ejemplo la huella dactilar) con el fin de comparar los resultados.

Indicadores de producción

En el marco del proyecto se dirigieron 6 tesis de posgrado, cuatro de estas fueron aprobadas durante el transcurso del proyecto, y dos se encuentran en un estado avanzado de elaboración. Además, se dirigió a un becario de iniciación a la investigación. Se fortaleció la formación de 4 docentes investigadores.

Se realizaron 17 publicaciones con referato (con presentación en congresos y eventos científicos nacionales e internacionales), se presentaron 2 artículos en revista de difusión científica (en proceso de publicación).

Se dictará 1 curso de créditos académicos como consecuencia de la investigación realizada en el 2024 y se asistió a 7 cursos como parte de la formación específica sobre temas del proyecto.

Por otro lado, se implementó una aplicación para la extracción masiva del código del iris y métodos de comparación e indexación.

Bibliografía

1. Ruiz S., Etchart G., Alvez C., Miranda E., Benedetto M., y Aguirre J. (2015). Iris Information Management in Object-Relational Databases, XXI Congreso Argentino de Ciencias de la Computación. Universidad Nacional del Noroeste de la Provincia de Buenos Aires, Junín - Argentina.
2. Alvez C., Etchart G., Ruiz S., Miranda E., y Aguirre J. (2017). Extensión de una base de datos Objeto-Relacional para el soporte de datos de iris. XXIII Congreso Argentino de Ciencias de la Computación. Universidad Nacional de La Plata – Argentina.
3. Rajesh, A., W. Singh, Gurjit, Shivam R., Tapwal A. and Anjana G. (2019). A Secure Multimodal Biometric System based on Diffused Graphs and Optimal Score Fusion. IET Biometrics.
4. Chen Y., Yang J., Wang C, Liu N. (2016). Multimodal biometrics recognition based on local fusion visual features and variational Bayesian extreme learning machine, Expert Systems with Applications, Volume 64, Pages 93-103.
5. Nanni L., Lumini A., Ferrara M., Cappelli M. (2015). Combining biometric matchers by means of machine learning and statistical approaches, NEUROCOMPUTING, 149, pp. 526 - 535
6. Leghari M., Memon S., Sahito F., Chandio A. and M. Leghari (2018). Biometric Verification Enhancement with Ensemble Learning Classifiers. 5th International Multi-Topic ICT Conference (IMTIC), Jamshoro, pp. 1-6.
7. Balaka R. N. (2018). P.V.G.D Prasad Reddy. A Multimodal Biometric System using Fingerprint and Voice. Universal Review. Volume 7, Issue XII. Page No: 1614 - 1625.
8. Meng W., Wong D. S., Furnell S., Zhou, J. (2015). Surveying the development of biometric user authentication on mobile phones, IEEE Commun. Surveys Tuts., vol. 17, pp. 1268-1293, 3rd Quart.
9. Rui Z. and Yan Z. (2019). A Survey on Biometric Authentication: Towards Secure and Privacy-Preserving Identification. IEEE Access, 7, 5994 - 6009.
10. Arslan B., Yorulmaz E., Akca B. and Sagioglu S. (2016). Security Perspective of Biometric Recognition and Machine Learning Techniques. 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, pp. 492-497.

11. Sadeghi K., Banerjee A., Javad S., Gupta K.S. (2017). Performance and Security Strength Trade-Off in Machine Learning Based Biometric Authentication Systems. 1045-1048. 10.1109/ICMLA.2017.00-12.
12. Mahajan S., Katti J., Walunj A. and Mahalunkar K. (2015). Designing a database encryption technique for database security solution with cache. IEEE International Advance Computing Conference (IACC), Bangalore, pp. 357-360.
13. Khan M. K. (2007). Cryptanalysis and Security Enhancement of Two Password Authentication Schemes with Smart Cards. IEEE International Multitopic Conference, Lahore, pp. 1-4.
14. Biggio B., Fumera G., Russu P., Didaci L. and Roli F. (2015). Adversarial Biometric Recognition: A review on biometric system security from the adversarial machine-learning perspective. in IEEE Signal Processing Magazine, vol. 32, no. 5, pp. 31-41.
15. Lyons K.B. (2014). A Recommender System in the Cyber Defense Domain. Master's thesis no. AFIT-ENG-14-M-49, Air Force Institute of Technology Graduate School of Engineering and Management, Wright-Patterson Air Force Base.
16. Thompson P. (2004). Weak Models for Insider Threat Detection. Proceedings of SPIE, vol. 5403: Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense, pp. 40-48.
17. Lewis T.A. (2014). An Artificial Neural Network-Based Decision Support System for Integrated Network Security. Master's thesis no. AFIT-ENG-T-14-S-09, Air Force Institute of Technology Graduate School of Engineering and Management, Wright-Patterson Air Force Base.
18. Wood C.J. (2014). What Friends are For Collaborative Intelligence Analysis and Search, Master's thesis, Naval Postgraduate School.
19. Zezula P., Amato G., Dohnal V., and Batko M. (2006). Similarity Search: The Metric Space Approach. Springer
20. Novak D., Batko M., Zezula P. (2011). Metric Index: An efficient and scalable solution for precise and approximate similarity search, Information Systems, Volume 36, Issue 4, Pages 721-733.
21. Manhua Liu, Xudong Jiang, Alex Chichung Kot (2007). Efficient fingerprint search based on database clustering, Pattern Recognition, Volume 40, Issue 6, Pages 1793-1803.
22. Bai C., Wang W., Zhao T., Li M. (2018). Fast exact fingerprint indexing based on Compact Binary Minutia Cylinder Codes, Neurocomputing, Volume 275, Pages 1711-1724.
23. I. DONKÓ, J. M. SZALAI-GINDL, G. GOMBOS, and A. KISS. An implementation of the m-tree index structure for postgresql using gist. In 2019 IEEE 15th International Scientific Conference on Informatics, page 6, Poprad, Slovakia, nov 2019. IEEE.
24. G. GOMBOS, J. M. SZALAI-GINDL, I. DONKÓ and A. KISS: "Towards on experimental comparison of the M-tree index structure with BK-tree and VP-tree", 2020, Acta Electrotechnica et Informatica. Vol. 20, No. 2, 2020, 19-26, DOI: 10.15546/aeei-2020-0009. <https://github.com/ggombos/mtree>
25. Jalilian, E. y Uhl, A. (2017). Iris Segmentation Using Fully Convolutional Encoder--Decoder Networks. In Deep Learning for Biometrics (pp. 133-155). (Advances in Computer Vision and Pattern Recognition). Springer Verlag. https://doi.org/10.1007/978-3-319-61657-5_6.

26. O. Oyedotun y A. Khashman. (2017). Iris nevus diagnosis: convolutional neural network and deep belief network. *Turquia: Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 25, p. 1106-1115.
27. S. Minaee, A. Abdolrashidiy y Y. Wang. (2016). An experimental study of deep convolutional features for iris recognition New York: Computer Science and Engineering Department, University of California at Riverside.
28. S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun y D. Zhang. (2019). Biometric Recognition Using Deep Learning: A Survey. *Estados Unidos: New York University*.
29. S. Tammina. (2019). Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images. *International Journal of Scientific and Research Publications*, Volume 9. DOI: 10.29322/IJSRP.9.10.2019.p9420
30. Sian Jin, Sheng Di, Xin Liang, Jiannan Tian, Dingwen Tao y Franck Cappello. (2019). DeepSZ: A Novel Framework to Compress Deep Neural Networks by Using Error-Bounded Lossy Compression. DOI: 10.1145/3307681.3326608.
31. Leiva MA, García AJ, Shakarian P, Simari GI. Argumentation-Based Query Answering under Uncertainty with Application to Cybersecurity. *Big Data and Cognitive Computing*. 2022; 6(3):91. <https://doi.org/10.3390/bdcc6030091>.
32. García, A. and Simari, G.R., "Defeasible Logic Programming: An Argumentative Approach". *Theory and Practice of Logic Programming* 4(1). 2004, pp. 95-138.
33. Tucac, M. (2011). Grupos de Servicios de Razonamiento para el Procesamiento de Consultas Contextuales en Paralelo. PhD thesis, Universidad Nacional del Sur, Bahía Blanca, Argentina.
34. Garcia, A. y Simari, G. (2014). Defeasible logic programming: DeLP-servers, contextual queries, and explanations for answers. *Argument and Computation*, 5, 63-88.
35. Burón Brarda, M.E., Tamargo, L.H., y García, A.J. (2019). An approach to enhance argument-based multi-criteria decision systems with conditional preferences and explainable answers. *Expert Systems with Applications*, 126, 171-186.
36. Buron Brarda, M. E., Tamargo, L. H., and García, A. J. (2020). Using argumentation to obtain and explain results in a decision support system. *IEEE Intelligent Systems*.
37. Buron Brarda, Martín E. and Tamargo, Luciano H. and Garcia, Alejandro Javier, Multi-Agent Multi-Criteria Decision Framework with Interactive Explanation Tools. Available at SSRN: <https://ssrn.com/abstract=4292652> or <http://dx.doi.org/10.2139/ssrn.4292652>.
38. Bénard, C., Biau, G., Da Veiga, S., y Scornet, E. (2021). SIRUS: Stable and Interpretable Rule Set for classification. *Electronic Journal of Statistics*, 15, 427-505.
39. C. Shearer, «The CRISP-DM model: the new blueprint for data mining», *J Data Wareh.*, vol. 5, pp. 13-22, ene. 2000.

PID 7058

Denominación del Proyecto

Modelos de Machine Learning para la mejora de la precisión, seguridad y eficiencia en la gestión de datos biométricos

Director

Alvez, Carlos Eduardo

Codirectora

Etchart Graciela Raquel

Unidad de Ejecución

Universidad Nacional de Entre Ríos

Dependencia

Facultad de Ciencias de la Administración

Contacto

carlos.alvez@uner.edu.ar

Cátedra/s, área o disciplina científica

Bases de Datos. Lenguajes de Programación. Sistemas Operativos. Metodologías de Sistemas I

Integrantes del proyecto

Docentes Uner: Juan José Aguirre. Ernesto Martín Miranda. Silvia Estela Ruiz. Juan Carlos Teze. Jesús F. Frola Da Silva. Becario Pid: Kevin Inda

Fechas de iniciación y de finalización efectivas

15/08/2019 y 07/01/2024

Aprobación del Informe Final por Resolución C.S. N°145/24 (24-05-2024)