

DELFOS Project

Leticia Morales Trujillo¹, Nicolás Sánchez Gómez¹,
Maria Dolores Acuña¹, Julián Alberto García García¹ David Lizcano²,

¹Universidad de Sevilla (US)
Escuela Técnica Superior de Ingeniería Informática
Avda. Reina Mercedes s/n. 41012, Sevilla, España.

²Universidad a Distancia de Madrid (UDIMA)
Collado Villalba A-6, 15, 28400, Madrid, España.

{lmtrujillo,nicolassg,macuna,juliangg}@us.es, david.lizcano@udima.es

Abstract. *DELFOS (Decision Support System based on Machine Learning Algorithms Designed to Detect Hidden Genetic Pathologies by Big Data Analysis of Electronic Medical Records) es un sistema experto de soporte a la decisión clínica (DCSS) basado en Big Data e inteligencia artificial (IA) orientado a mejorar el asesoramiento genético mediante la integración de historias clínicas electrónicas (HCE) y datos genómicos. El proyecto, financiado por Red.es en la convocatoria 2021 de proyectos de I+D en IA (C005/21-ED) y desarrollado por G7 Innovations Solutions SL en colaboración con la Universidad de Sevilla e INEBIR¹, busca impulsar la transformación digital en clínicas de reproducción asistida, donde la demanda de asesoramiento genético supera la disponibilidad de genetistas clínicos y se prevé que aumente con el avance de la medicina personalizada. DELFOS se integra con historias clínicas electrónicas mediante el estándar FHIR (Fast Healthcare Interoperability Resources) y, en su estado actual, ha sido validado en un piloto clínico en INEBIR para dos modelos predictivos (infertilidad masculina y endometriosis). En infertilidad masculina, el piloto con 40 pacientes obtuvo un 47,5% de acierto, un 7,5% de predicciones parcialmente correctas y un 45% no coincidentes con el diagnóstico clínico, destacando la necesidad de ajustar umbrales, especialmente en teratozoospermia. En endometriosis, el piloto con 14 pacientes mostró un 100% de acierto cuando existían datos genéticos disponibles y una reducción significativa del rendimiento cuando no se disponía de dicha información.*

1. Problema a resolver y a quién beneficia

El asesoramiento genético [Abacan et al. 2019] en tratamientos de reproducción asistida requiere una evaluación integral del paciente que combina historia clínica, pruebas específicas y, cuando corresponde, información genómica [Poornima et al. 2020] [Seraji et al. 2021]. No obstante, en la práctica este proceso se ve limitado por la escasez de especialistas, la complejidad del análisis genético y el volumen de datos generado por tecnologías como la secuenciación de nueva generación (NGS) [Soini et al. 2006]. Como consecuencia, esto puede provocar retrasos en diagnósticos, en la estimación del riesgo hereditario y en la toma de decisiones

¹Instituto para el Estudio de la Biología de la Reproducción Humana

clínicas. A ello se suma la falta de interoperabilidad entre sistemas sanitarios, que dificulta el acceso conjunto a información clínica y genética y reduce la eficiencia del proceso asistencial.

Con el fin de abordar estas limitaciones, el objetivo global del proyecto es desarrollar un sistema experto basado en tecnologías Big Data e IA capaz de descubrir patologías a partir del análisis de historias clínicas electrónicas, correlacionar variables clínicas cuya relación no se evalúa adecuadamente e integrarlas con datos de secuenciación del exoma. Con ello, DELFOS pretende aportar información útil para la gestión del paciente, reforzar el cribado sensible y la detección temprana de factores de riesgo genético y apoyar la toma de decisiones clínicas en entornos de reproducción asistida, favoreciendo además su transferibilidad e integración en distintas organizaciones sanitarias. Para alcanzar estos objetivos, DELFOS se concibe como un proyecto multidisciplinar que combina conocimiento clínico y genético en reproducción asistida (INEBIR), investigación en ingeniería del software, interoperabilidad y analítica de datos sanitarios (Universidad de Sevilla), y capacidad de desarrollo e integración de soluciones en entornos reales (G7 Innovations Solutions SL).

Como resultado, DELFOS beneficia principalmente a genetistas clínicos y especialistas en reproducción asistida al proporcionar un apoyo a la decisión que integra información clínica y genómica e incorpora un enfoque de cribado y alerta temprana; de forma indirecta, también contribuye a mejorar la experiencia de los pacientes al facilitar decisiones más ágiles y basadas en datos.

2. Arquitectura tecnológica y tecnologías empleadas

DELFOFOS se implementa mediante una arquitectura modular (ver Figura 1) diseñada para integrar información clínica y genómica y soportar el ciclo completo de captura, procesamiento, análisis y validación.

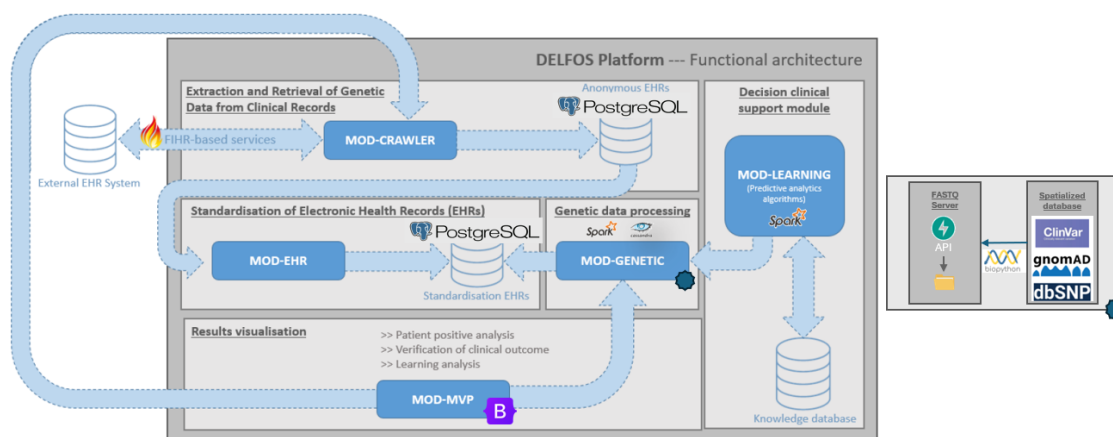


Figura 1. Diagrama de arquitectura tecnológica con tecnologías empleadas

El sistema inicia con la **extracción y estructuración de los datos clínicos** desde **sistemas EHR externos**, asegurando la compatibilidad e interoperabilidad mediante el estándar **FHIR**, lo que permite estructurar e intercambiar información procedente de historias clínicas electrónicas. A partir de ahí, el módulo **MOD-CRAWLER** realiza la captura y anonimización de los registros y, posteriormente, **MOD-EHR** ejecuta la norma-

lización clínica antes de persistir la información en **PostgreSQL**, facilitando una base consistente para el análisis.

En paralelo, DELFOS incorpora la gestión de información genómica a partir de ficheros de secuenciación en formato FASTQ (uno de los formatos estándar para el almacenamiento de secuencias genómicas), para lo cual dispone de una API implementada con FastAPI que permite la consulta y gestión. Con el objetivo de enriquecer el análisis, el sistema correlaciona la información clínica con variantes genéticas utilizando bases de datos especializadas como **ClinVar**, **dbSNP** y **gnomAD**, a las que accede mediante **Biopython**. El módulo **MOD-GENETIC** se encarga de comparar la secuencia del genoma del paciente con la de referencia del genoma humano para detectar variantes relevantes, apoyándose en **Apache Spark** para el procesamiento distribuido y en **Cassandra** para el almacenamiento escalable de los datos resultantes. Sobre esta infraestructura, el módulo **MOD-LEARNING** entrena y ejecuta modelos de **aprendizaje automático** utilizando **MLlib (Machine Learning Library) de Spark**, ajustándolos iterativamente con validaciones médicas. Los resultados se presentan mediante el módulo **MOD-MVP**, optimizado para la visualización y validación clínica con **Bootstrap**.

3. Resultados y trabajo futuro

En el estado actual del proyecto, DELFOS ha sido pilotado en INEBIR para evaluar su desempeño en la identificación de patologías, centrándose en infertilidad masculina y endometriosis. INEBIR ha colaborado activamente en el desarrollo y pilotaje, con participación clínica en el diseño y validación del programa de IA para ambos modelos. Como base para el desarrollo de los modelos, en endometriosis se partió de una cohorte de 5.143 pacientes, con 270 casos positivos (aproximadamente 5,25%), utilizando una partición 80/20 para entrenamiento y prueba (4.150/993). En infertilidad masculina, el modelo se desarrolló con una cohorte retrospectiva de 15.264 pacientes, empleando un conjunto final de 27 variables derivadas del seminograma.

En la fase piloto de endometriosis se trabajó con 14 pacientes, diferenciando grupos según disponibilidad de datos genéticos. Cuando existía información genética, el modelo alcanzó un 100% de precisión en el piloto, mientras que sin datos genéticos la tasa de acierto se redujo significativamente, lo que refuerza el valor de integrar información genómica en el proceso predictivo y sugiere que la disponibilidad de datos genéticos influye de forma relevante en el rendimiento del modelo en este caso de uso.

Para infertilidad masculina, el piloto se realizó con 40 pacientes, usando parámetros del seminograma como cantidad de espermatozoides, motilidad y morfología, siguiendo criterios de la Organización Mundial de la Salud. El modelo emitió diagnósticos correctos en el 47,5% de los casos, predicciones parcialmente correctas en el 7,5% y predicciones no coincidentes con el diagnóstico clínico en el 45%, lo que sugiere la necesidad de optimizar los umbrales de clasificación, especialmente en la detección de teratozoospermia con el objetivo de mejorar la precisión diagnóstica en el piloto.

Además de los resultados predictivos, el proyecto ha generado dos artículos científicos asociados a ambos modelos. En paralelo, DELFOS ha mostrado alta interoperabilidad con los sistemas sanitarios mediante el estándar FHIR, facilitando el acceso a la información y el apoyo a la toma de decisiones clínicas.

A partir del estado actual, el trabajo futuro se orienta a ampliar el conjunto de da-

tos, ajustar los umbrales de clasificación y mejorar el rendimiento predictivo mediante nuevas técnicas de aprendizaje automático, manteniendo el enfoque de integración clínica-genómica. Asimismo, se prevé reforzar la capacidad de alertas y cribado temprano y facilitar su transferencia a otros entornos clínicos de reproducción asistida. Dado que DELFOS se plantea como herramienta de apoyo a la decisión y no como sustituto del especialista, será necesario reforzar la explicabilidad de las predicciones, con supervisión clínica humana y trazabilidad de las variables más influyentes [World Health Organization 2021]. También deberá evaluarse el posible sesgo derivado de la representatividad de las cohortes y monitorizarse el rendimiento por subgrupos clínicos [Collins et al. 2024]. Finalmente, en un escenario de despliegue real como software con finalidad médica, el sistema deberá alinearse con la regulación aplicable a sistemas de IA de alto riesgo en el contexto europeo [European Parliament and Council of the European Union 2024], así como extenderse progresivamente a otras patologías reproductivas.

Agradecimientos

Esta investigación ha sido respaldada por: (i) el proyecto EQUAVEL PID2022-137646OB-C31, financiado por MICIU/AEI/10.13039/501100011033 y por ERDF, EU.; (ii) el proyecto DELFOS (021/C005/00151010) de la organización española red.es, dentro de su convocatoria de 2021 para financiar proyectos de I+D relacionados con la inteligencia artificial y otras tecnologías digitales integradas en cadenas de valor (C005/21-ED); (iii) la ayuda PREP2022-000332, financiada por MICIU/AEI/10.13039/501100011033 y por ESF+; (iv) Proyecto HyperIA financiado por la Consejería de Universidades, Investigación e Innovación de la Junta de Andalucía (Andalucía, España).

Referências

- Abacan, M., Alsubaie, L., Barlow-Stewart, K., Caanen, B., Cordier, C., Courtney, E., others, and Wicklund, C. (2019). The global state of the genetic counseling profession. *European Journal of Human Genetics*, 27(2):183–197.
- Collins, G. S., Moons, K. G. M., Dhiman, P., et al. (2024). Tripod+ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378.
- European Parliament and Council of the European Union (2024). Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, 13 June 2024.
- Poornima, S., Daram, S., Devaki, R. K., and Qurratulain, H. (2020). Chromosomal abnormalities in couples with primary and secondary infertility: Genetic counseling for assisted reproductive techniques (art). *Journal of Reproduction and Infertility*, 21(4):269.
- Seraji, K., Simonson, M., and Chou, C. (2021). Exploring genetic counselors' use of pedigree symbols to represent assisted reproductive technology. *Journal of Genetic Counseling*.
- Soini, S., Ibarreta, D., Anastasiadou, V., et al. (2006). The interface between assisted reproductive technologies and genetics: technical, social, ethical and legal issues. *European Journal of Human Genetics*, 14:588–645.
- World Health Organization (2021). *Ethics and Governance of Artificial Intelligence for Health*. World Health Organization, Geneva.