UNIVERSIDAD
DE MÁLAGA

LENGUAJES Y
CIENCIAS DE LA
COMPUTACIÓN
UNIVERSIDAD DE MÁLAGA

TESIS DOCTORAL
Tecnologías Informáticas

Integration and analysis of biomedical data from multiple clinical trials

E.T.S.I. Informática
R.D. 99/2011

Autor

Sandro Hurtado Requena

Directores

Dr. Ismael Navas Delgado
Departamento
Lenguajes y Ciencias de la Computación
Universidad de Málaga

Dr. José Manuel García Nieto
Departamento
Lenguajes y Ciencias de la Computación
Universidad de Málaga

Noviembre 2022

UNIVERSIDAD DE MÁLAGA

AUTOR: Sandro José Hurtado Requena

https://orcid.org/0000-0003-0990-480X

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña SANDRO JOSÉ HURTADO REQUENA
Estudiante del programa de doctorado TECNOLOGÍAS INFORMÁTICAS de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: INTEGRATION AND ANALYSIS OF BIOMEDICAL DATA FROM MULTIPLE CLINICAL TRIALS

Realizada bajo la tutorización de JOSÉ FRANCISCO ALDANA MONTES y dirección de ISMAEL NAVAS DELGADO Y JOSE MANUEL GARCÍA NIETO

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 10 de NOVIEMBRE de 2022

Fdo.: Doctorando/a	Fdo.: Tutor/a
Fdo.: Director/es de tesis	

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

Contents

Acknowledgements

First, I would like to sincerely thank my director, Dr. José Francisco Aldana Montes and my supervisors Prof. Ismael Navas Delgado and Prof. José Manuel García Nieto for their guidance and support throughout these years of developing the work that is part of this Thesis. They have allowed me to do a doctoral thesis and trusted me all these years. Also, I thank them for sharing their academic/labour and personal experiences with me.

My sincere thanks to Prof. Anton Popov for the lovely stay I had in those five months at National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute". They gave me a huge opportunity for living an excellent experience and also improving my research.

I would also like to thank, my colleagues of the Khaos research group, who have shared with me so many hours of work in the laboratory. They have helped me turn the hard work hours into moments of joy, laughter and friendship.

I would also like to thank my girlfriend and friends for being with me during difficult times. They can make the sun shine every day, even on cloudy days.

Finally, I would like to thank my family for their support and for being the fundamental pillar in my life. They have taught me to never give up despite adversities and that with effort and sacrifice, we will be able to get anywhere.

Resumen

Introducción

Los rápidos avances en la tecnología de secuenciación de próxima generación (NGS del inglés Next-generation Sequencing), y el consiguiente crecimiento y disponibilidad de datos biológicos [1, 2], han llevado a las organizaciones a enfrentarse a nuevos desafíos que le permitan analizar y descubrir información relevante, más allá de los obtenidos por los métodos tradicionales. De este modo, los profesionales clínicos pueden combinar los datos biológicos con otra información clínica y específica del paciente, como las historias clínicas electrónicas, los hábitos, la ascendencia y los factores ambientales, lo que les permite analizar y encontrar información pertinente más allá de lo que puede obtenerse mediante enfoques convencionales.

El sector sanitario genera mucha información sobre evaluaciones médicas, declaraciones de pacientes, tratamientos y prescripciones [3]. No obstante, el diagnóstico de las enfermedades es una tarea difícil en la medicina moderna. Comprender el diagnóstico preciso de los pacientes mediante el examen y la evaluación médica es la responsabilidad más crítica de estos profesionales.

Los datos clínicos proceden de numerosas fuentes de información, como los datos obtenidos mediante diversas técnicas masivas de secuenciación paralela del ADN, datos fisiológicos como el electrocardiograma, el encefalograma [4], repositorios biomédicos e incluso redes sociales públicas y científicas [5]. Entre estas fuentes también se encuentran las imágenes médicas, que comprenden la mayor parte de los datos de los pacientes (sobre todo en el caso de los pacientes oncológicos), los factores de riesgo de la enfermedad, los datos multiómicos, los regímenes/procedimientos terapéuticos y los datos de seguimiento.

Asimismo, se utilizan técnicas de recogida de datos a través de diversos sensores (tecnologías del internet de las cosas (IoT del inglés Internet of Things)). Entre otros, la recogida de datos a través de sensores de acelerometría, que permiten monitorizar la actividad de los pacientes [6, 7] e incluso detectar una posible caída de los mismos [8].

Este rápido aumento de los datos clínicos ha puesto de manifiesto la necesidad de desarrollar nuevas y sofisticadas herramientas para la gestión y el análisis de datos en la investigación clínica y la medicina personalizada. Los sistemas de gestión de ensayos clínicos (CTMS del inglés Clinical Trial Management System) se utilizan para recuperar datos significativos de los ensayos clínicos, obtener una visibilidad temprana de las enfermedades y encontrar terapias alternativas. Con la ayuda de estas herramientas informáticas de apoyo a la investigación clínica, es posible adquirir nuevos conocimientos relacionados con la salud y descubrir nuevos medicamentos [9]. Estas aportaciones también pueden consolidarse en bases de datos para contribuir a otros estudios farmacológicos de carácter investigativo, para realizar minería e ingeniería de datos y para apoyar a los expertos en el diagnóstico. En consecuencia, los CTMS se han convertido en una herramienta de apoyo esencial para la investigación clínica [10]. Estas herramientas traen consigo la posibilidad de recopilar información clínica relevante y realizar diversas técnicas de análisis que ayudan a comprender el mecanismo molecular y las posibles terapias de las enfermedades humanas e incluso

proporcionan información biológica y médica que permite una atención médica individualizada.

Cabe destacar que la gestión de los datos clínicos que intervienen en los estudios de NGS es una tarea difícil, dados los continuos obstáculos que se encuentran en el mantenimiento del sistema durante la inscripción de los pacientes, el proceso de adquisición de muestras del estudio clínico y los diferentes pasos para la preparación de los flujos de trabajo de procesamiento de datos clínicos. La mayoría de estas dificultades se deben a la naturaleza dinámica y heterogénea de los datos clínicos. Igualmente, existe una necesidad constante por parte de los investigadores en este campo de ampliar las funcionalidades de estos CTMS e integrar los datos de las operaciones clínicas en múltiples sistemas e incluso ser capaces de trasladarlos a diferentes tipos de muestras. Diversas organizaciones que participan en ensayos clínicos, especialmente en centros médicos académicos, tienen una gestión de datos y un control de calidad complejos [11, 12].

No obstante, la mayoría de las herramientas clínicas carecen de estas características, ya que son exclusivamente competentes para cubrir casos específicos y están limitadas para adaptarse a los continuos cambios en la práctica de los ensayos clínicos. A este respecto, el principal problema de los CTMS de código abierto es que es difícil personalizar un flujo de trabajo que no sea propio de una única solución [13, 14]. Por lo general, estos sistemas no permiten la integración de diferentes fuentes heterogéneas de datos clínicos. Por lo tanto, es difícil que un CTMS de este tipo gestione una variedad de ensayos clínicos, lo que suele ser necesario en la mayoría de centros médicos. Resulta una tarea difícil comprender un fenómeno como una enfermedad con un solo tipo de datos. Es por ello que la mayoría de estos sistemas tienen limitaciones de diseño debido a la naturaleza heterogénea y dinámica de los datos clínicos, ya que actualmente no pueden satisfacer las necesidades de adaptación a los continuos cambios en la práctica de los ensayos. En este sentido, es un hecho que se necesita más investigación para mejorar el acceso y la integración de datos sanitarios heterogéneos con el fin de mejorar las necesidades no cubiertas en la atención médica de pacientes.

Por otro lado, los modelos de Inteligencia Artificial (IA) son cada vez más predominantes en la investigación biomédica y la práctica clínica [15]. Estos modelos se han mostrado prometedores en muchos campos, como la estratificación y el modelado de riesgos, la detección personalizada, el diagnóstico de enfermedades moleculares, el pronóstico y la predicción de la respuesta a la terapia [16, 17]. La IA puede ser impulsada por la convergencia de conjuntos de datos clínicos anotados a gran escala, los avances en el aprendizaje automático (ML del inglés Machine Learning), las herramientas de software de código abierto, el rápido aumento en la potencia de procesamiento, y el almacenamiento en la nube. A tal efecto, la integración de los sistemas de IA en los CTMS es esencial, ya que la IA puede lograr un éxito especializado en algunas tareas sanitarias en las que puede ayudar a los médicos a determinar el pronóstico de enfermedades y los procedimientos quirúrgicos. Por ello, los CTMS deben realizar análisis y predicciones exhaustivas de los datos para obtener indicadores de calidad. La combinación de IA y CTMS se propone como un gran avance en la práctica de la medicina en un futuro próximo.

A pesar de los tremendos avances llevados a cabo en la tecnología NGS y las herramientas de software bioinformático, existen limitaciones para hacer frente a enfermedades complicadas y genéticamente heterogéneas. En lo que a nosotros concierne, actualmente muy pocos sistemas permiten analizar los datos clínicos de los pacientes para diagnosticar enfermedades, por consiguiente, estos sistemas no son capaces de cubrir el *"gap"* existente entre bioinformáticos, genetistas moleculares y clínicos. En este respecto, la integración de múltiples flujos de datos clínicos podrían tener un impacto en la práctica de los ensayos. Los algoritmos de ML pueden utilizar enfoques de integración de datos de varios conjuntos de datos biomédicos para obtener resultados más precisos y mejorar la comprensión de los sistemas biomédicos. Aun así, la mayoría de los CTMS no tienen en cuenta la importancia de la integración de datos en biomedicina y sólo consideran la variación de un único tipo de datos. En este sentido, pueden perderse muchos patrones esenciales que sólo pueden observarse al considerar múltiples niveles de datos biomédicos.

Por lo tanto, es necesario integrar distintas fuentes de datos para que los algoritmos puedan hacer detección de enfermedades y predicciones aún mas precisas [18]. Todavía hay margen de mejora de la tendencia actual hacia la medicina de precisión mediante el diseño de métodos personalizados con efectos significativos en las vías diagnósticas y terapéuticas [19].

Motivación

De acuerdo con el *"Hype Cycle"* de Gartner [20], hay un reto en el campo de la gestión y administración de datos con respecto a la integración de los datos en las herramientas software actuales. Gartner subraya la importancia de la integración de datos en las herramientas: *"Las organizaciones necesitan herramientas de integración de datos para apoyar la gestión de datos distribuidos y proporcionar datos a través de diversos casos de uso. Esto incluye la integración de datos para apoyar la analítica, la ciencia de datos, la integración de aplicaciones, la preparación de datos de autoservicio, etc."*. En este sentido, ofrece flexibilidad, escalabilidad y extensibilidad en la infraestructura para garantizar que los datos se puedan consumir en múltiples casos de uso en las instalaciones, en varias nubes o en cualquier forma de híbrido. Asimismo, facilitará la integración de estas herramientas con otros sistemas y habilitará la interconexión de algoritmos. En esta Tesis hemos abordado este reto creando FIMED, una herramienta software de soporte al experto clínico que permite la recolección de datos clínico de sistemas y fuentes de origen dispares para convertirlos en información significativa y valiosa. Esta herramienta ayuda al investigador clínico al proceso de integración de datos, contribuyendo al avance en la gestión de datos de ensayos clínicos. Asimismo, incluye algoritmos de análisis para mejorar la investigación, aprovechando al máximo la información clínica recopilada.

Como objetivo transversal de esta Tesis, se han investigado e intregado nuevas propuestas algorítmicas que utilicen este tipo de datos, enfocadas al ML y optimización, por ser de especial interés en los estudios de investigación clínica. En este sentido, el objetivo es simular la fisiología de los sistemas vivos en los sistemas biológicos como un conjunto de componentes que interactúan, más que como un conjunto de componentes físicos individuales. Este enfoque tiene la ventaja práctica de proporcionar información sobre la regulación u optimización de elementos específicos del sistema, teniendo en cuenta su impacto en el conjunto del mismo. Además, es esencial modelar las interacciones entre los numerosos componentes que conforman un sistema biológico de este tipo para comprender mejor el complicado comportamiento global observado y los procesos físicos subyacentes. El empleo de enfoques computacionales modernos que puedan llevar a cabo un estudio integrado de fuentes de datos tan dispares es crucial y desafiante al mismo tiempo para aprender los respectivos modelos a gran escala.

Todas estas razones nos llevan a definir la principal hipótesis de esta Tesis, *"La integración automática de datos clínicos heterogéneos procedentes de múltiples fuentes dará lugar a análisis avanzados y a la generación de nuevas propuestas algorítmicas que utilicen este tipo de datos. Los datos biomédicos se combinarán como variables predictoras para permitir un modelo completo y adquirir resultados más relevantes"*

Tradicionalmente, la medicina general debe adaptar los tratamientos en función de las características clínicas y biológicas del paciente para ofrecer una atención óptima. En los últimos años, muchos estudios han puesto sus esfuerzos en el camino hacia la medicina de precisión [21, 22, 23, 24, 25]. Sin embargo, siguen existiendo muchas lagunas de investigación cuando se intenta abarcar los problemas de un ecosistema de datos biomédicos. Varios cuellos de botella frenan la transición de la medicina convencional a la personalizada. En este sentido, la investigación presentada en esta Tesis se llevó a cabo para responder a las siguientes preguntas de investigación en la gestión y el análisis de datos biomédicos hacia la medicina de precisión.

Cuestiones de Investigación

- **Q1**: ¿Es suficiente una única fuente de información clínica para explicar enfermedades complejas como el cáncer?
- **Q2**: ¿Cómo se puede mejorar la inferencia e interpretación de redes biológicas complejas para extraer información útil que ayude al diagnóstico de enfermedades o al descubrimiento de nuevos biomarcadores?
- **Q3**: ¿Es posible mejorar la calidad de los métodos de IA en el ecosistema de datos biomédicos aprovechando grandes cantidades de datos sin etiquetar recogidos en entornos Big Data?
- **Q4**: ¿Se puede conceder al experto clínico la capacidad de interpretar los resultados de modelos complejos de IA?

Asimismo, estas principales cuestiones de investigación se asocian a diversos retos que se llevan a cabo en problemas del mundo real en un entorno de datos biomédicos.

Retos de Investigación

- **Ch1**: Disponibilidad y análisis de datos biomédicos. Es necesario diseñar nuevos métodos y herramientas de software dedicados a mejorar la recopilación de datos, la gestión y el análisis avanzado de datos biomédicos utilizando diferentes estrategias. Aprovechar las nuevas herramientas de integración de datos clínicos y las estrategias de análisis puede impulsar potencialmente un cambio real en las terapias personalizadas [24].
- **Ch2**: Se necesitan nuevos métodos computacionales y experimentales para inferir y explicar las redes biológicas. Es un reto reconstruir las redes biológicas e interpretarlas para los investigadores clínicos debido a la complejidad que presenta la alta dimensión de los datos biomédicos [26].
- **Ch3**: Se necesitan nuevas estrategias de IA que sean capaces de proporcionar resultados confiables mediante el entrenamiento de modelos con pequeños conjuntos de datos clínicos etiquetados e incluso con conjunto de datos sin etiquetar. Un reto común es la escasez de datos etiquetados. Es necesario diseñar nuevas estrategias que consideren casos de uso con escasez de datos etiquetados [27].
- **Ch4**: La explicabilidad de la toma de decisiones automatizada en la medicina de precisión. Las sofisticadas técnicas de aprendizaje automático han logrado recientemente un gran éxito predictivo para muchas aplicaciones biomédicas. Sin embargo, es un reto explicar el resultado clínico de este modelo de caja negra [28]. Esta explicabilidad es esencial para el ámbito clínico, donde las decisiones afectarán a la vida de los pacientes.

Por lo tanto, la principal motivación de esta Tesis es indagar en los principales retos definidos anteriormente con el objetivo de ayudar a mejorar el ecosistema de datos biomédicos existente hacia una medicina de precisión.

Fases y objetivos

Esta Tesis se centra en contribuir con soluciones realistas a los sistemas de investigación clínica a través del diseño y desarrollo de una herramienta de software de gestión de ensayos clínicos junto con la aplicación de técnicas de IA y optimización para el análisis de datos. Por lo tanto, el objetivo principal de este trabajo es hacer frente a los problemas del mundo real en un ecosistema

de datos biomédicos. En el desarrollo de la presente Tesis se han abarcado problemas reales, entre ellos: la detección de enfermedades y búsqueda de nuevos biomarcadores para predecir la eficacia de ciertos tratamientos de cáncer, monitorización de pacientes mediante sensores y análisis de imágenes cancerígenas. En este sentido, se ha trabajado con diferentes tipos de datos clínicos, como pueden ser: datos de expresión génica, imágenes biomédicas, información clínica de pacientes, datos provenientes de sensores, datos de series temporales, etc. Asimismo, se han desarrollado nuevas herramientas y propuestas algorítmicas a modo de componentes software que permiten al experto clínico realizar la integración y el análisis exhaustivo de los datos como una línea de productos de software con elementos combinables, mediante estratégicas de IA y de optimización. Los diferentes componentes pueden interactuar, facilitando la interconexión, reutilización y trazabilidad de la cadena de valor de los datos. Finalmente, la XAI se lleva a cabo para ofrecer interpretabilidad de los resultados al investigador clínico. En este sentido, se focaliza el estudio en la explicabilidad de los algoritmos de aprendizaje profundo (DL del inglés Deep Learning), ya que estos algoritmos se perciben como *"cajas negras"* complejas y poco interpretables.

En concreto, los principales objetivos de esta Tesis se detallan a continuación:

1. **Objetivo 1: Diseñar y desarrollar una nueva solución de software para consolidar grandes cantidades de datos clínicos heterogéneos e integrar los análisis en múltiples ensayos.**

   (a) Diseñar la estructura de la base de datos de la herramienta con un motor de base de datos NoSQL (MongoDB).

   (b) Desarrollar la infraestructura software para crear flujos de trabajo de análisis de datos biomédicos que puedan aprovechar la integración de datos de múltiples fuentes.

   (c) Proporcionar herramientas para el análisis y la visualización de datos clínicos.

   (d) Desarrollo de una interfaz de usuario (GUI) que permita al investigador utilizar las funcionalidades de consolidación de datos, análisis y visualización de forma fácil e intuitiva.

2. **Objetivo 2: Desarrollo de soluciones de optimización meta-heurísticas para la inferencia de redes de regulación génica a partir de datos de expresión genética de pacientes.**

   (a) Adaptar algoritmos de optimización multiobjetivo bien conocidos para hacer frente a la inferencia de GRNs.

   (b) Proponer nuevos *emsemble* de algoritmos de optimización para la inferencia de redes de regulación génica.

   (c) Proporcionar a los biólogos de sistemas técnicas de optimización para inferir GRNs consistentes.

   (d) Proporcionar a los biólogos de sistemas un conjunto de herramientas de visualización para explorar la red construida.

3. **Objetivo 3: Desarrollo de soluciones de ML para el análisis de datos clínicos procedentes de sensores de monitorización de pacientes.**

   (a) Investigar el diseño de nuevas estrategias algorítmicas de ML adaptadas a casos de uso específicos con datos de sensores.

   (b) Investigar nuevos enfoques de integración de datos que combinen datos etiquetados y no etiquetados procedentes de sensores.

   (c) Proporcionar nuevos análisis de datos y flujos de trabajo para un seguimiento eficiente de los pacientes en tiempo real.

4. **Objetivo 4: Explicabilidad de los algoritmos de ML en casos de uso biomédico.**
   (a) Estudio de la eficacia de los algoritmos actuales de explicabilidad en casos de uso biomédicos.
   (b) Proponer la explicabilidad de los algoritmos de ML a los casos de uso biomédicos.
   (c) Estudio de nuevas métricas de evaluación para algoritmos de explicabilidad en casos de uso biomédicos.

5. **Objetivo 5: Abordar problemas del mundo real y académico en el contexto del ecosistema de datos biomédicos.**
   (a) Análisis de datos de expresión génica en un caso de uso real para el descubrimiento de biomarcadores de eficacia de tratamiento en la enfermedad del cáncer de piel Melanoma.
   (b) Validación del uso de *multi-objective swarm optimizers* en la reconstrucción de GRNs utilizando datos reales de pacientes con cáncer de melanoma.
   (c) Análisis de los datos de un grupo de pacientes con obesidad y enfermedades cardiovasculares para el reconocimiento de la actividad humana (HAR del inglés Human Activity Recognition) en el sistema sanitario de Andalucía (España).
   (d) Aplicación de estrategias de explicabilidad para algoritmos de DL en el diagnóstico del cáncer de Melanoma.

Contribuciones científicas

Las principales cuestiones y retos presentados en la motivación de esta Tesis pueden asociarse a grandes rasgos con las contribuciones incluidas en este estudio.

- En el **Capítulo 3** se presenta FIMED (*Flexible Management of Biomedical Data*), una nueva herramienta software para dar solución al primer reto (**Ch1**) en el contexto de el ecosistema de datos biomédicos. En este sentido, FIMED hace uso de base de datos NoSQL para aliviar algunas de las limitaciones impuestas por las bases de datos relacionales. Esta solución de software ha sido diseñada para apoyar a los expertos clínicos en la integración y el análisis de datos clínicos heterogéneos procedentes de múltiples fuentes de información de una manera sencilla, incremental y dinámica, evitando las limitaciones de las herramientas actuales.

- El segundo reto (**Ch2**) se trata en los trabajos descritos en los **Capítulos 3** y **4**. Estos capítulos centran su investigación en aportar mejoras en la reconstrucción de las redes de regulación génica (GRN). Estas GRN definen las interacciones entre los productos del ADN y otras sustancias en las células, lo cual es muy relevante para la investigación clínica. En consecuencia, una mayor comprensión de las GRN permite entender mejor los mecanismos que causan diversos trastornos y enfermedades. De esta forma, la interpretación de estas GRN sirve de ayuda a los profesionales clínicos para diagnosticar las enfermedades en sus fases más tempranas, descubrir nuevos biomarcadores de la progresión de la enfermedad e incluso en el diseño de nuevos tratamientos farmacológicos.

- El **Capítulo 5** esta directamente relacionado con el tercer reto (**Ch3**) definido en esta Tesis. Tradicionalmente los modelos avanzados de IA requieren una gran cantidad de datos etiquetados para ofrecer resultados prometedores. Sin embargo, el etiquetado de datos es costoso

y difícil de realizar en situaciones del mundo real. En concreto, muchos procedimientos de recogida de datos clínicos en los sistemas sanitarios modernos se realizan a través de tecnologías IoT mediante sensores [29] colocados en el paciente. La recopilación de datos a través de sensores da lugar a entornos de Big Data con cantidades masivas de datos. Por lo tanto, es inviable el etiquetado de todos los datos en un entorno Big Data para entrenar los modelos de ML. A este respecto, en el Capítulo 5 proponemos desarrollar nuevas estrategias de ML que puedan utilizar la cantidad limitada de datos etiquetados y las enormes cantidades de datos sin etiquetar para ayudar a los algoritmos a aprender, mejorar su rendimiento y generalizar su conocimiento. Además, este capítulo también se abarca el primer reto (**Ch1**), ya que en este trabajo se diseñó una metodología para la integración datos procedentes de sensores de acelerometría de distintas fuentes de información y en diversos formatos.

- En el **Capítulo 6** se propone cubrir el cuarto reto (**Ch4**). En este trabajo se explora distintos métodos para proporcionar explicabilidad a los resultados de los algoritmos complejos de inteligencia artificial en el ámbito de la medicina. Las aplicaciones de la ciencia de los datos y los métodos novedosos de aprendizaje automático todavía se perciben como las llamadas *cajas negras*. Sin embargo, en el campo de la medicina, es crucial predecir el resultado clínico del paciente y cuantificar el impacto del fármaco y, al mismo tiempo, tener en cuenta el sexo, la edad y otras características de forma interpretable. Es esencial explicar los resultados de estos algoritmos al experto clínico, mejorando así la interpretabilidad de los resultados. En consecuencia, este trabajo se centra en la investigación y evaluación de estrategias para la explicabilidad de algoritmos complejos de IA en problemas del mundo real en los que el modelo debe tener un buen rendimiento y ser fácilmente interpretable.

Publicaciones de la Tesis

Para transmitir los resultados de la investigación, esta Tesis ha realizado tareas de divulgación científica en revistas y congresos. En concreto, se han publicado tres artículos en revistas indexadas en el Journal of Citation Report (JCR). Asimismo, dos artículos en un congreso internacional. Para destacar la relevancia de esta Tesis, se muestra a continuación la relación de publicaciones realizadas durante su desarrollo:

Artículos de revista (JCR)

1. Sandro Hurtado, José García-Nieto, Ismael Navas Delgado, José Francisco Aldana Montes, ***FIMED: Flexible management of biomedical data [30]***.

   Este artículo está dedicado a presentar FIMED, una herramienta de software que integra datos de investigación clínica procedentes de diferentes fuentes de información y formatos de forma dinámica, flexible y transparente para el experto clínico. Gracias a la integración de estos datos, es posible desarrollar técnicas integrales en el análisis de datos clínicos mediante estrategias de ML y optimización, en combinación con datos tabulares, datos de secuenciación, datos de imagen biomédica, datos procedentes de sensores, etc., para romper las barreras de acceso y aplicabilidad relacionadas con las técnicas de análisis en el ámbito de la salud. Además, la herramienta ofrece una interfaz gráfica de usuario (GUI) para facilitar la gestión de la información y los análisis de los ensayos clínicos.

2. Sandro Hurtado, José García-Nieto, Ismael Navas Delgado, Antonio J.Nebro, José Francisco Aldana Montes, ***Reconstruction of gene regulatory networks with multi-objective particle swarm optimisers [31]***.

   En este artículo, se investiga el comportamiento de un conjunto de algoritmos de optimización multiobjetivo basados en diferentes estrategias en el contexto de la inferencia de GRNs. El

objetivo principal es proporcionar a los biólogos de sistemas pruebas experimentales sobre qué técnica de optimización se comporta con mayor éxito para inferir GRNs consistentes.

3. Sandro Hurtado, José García-Nieto, Anton Popov, Ismael Navas Delgado, ***Human Activity Recognition from sensorised Patient´s Data in Healthcare: A Streaming Deep Learning-based Approach [32]***.

   Este trabajo presenta una propuesta de análisis y monitorización de datos de pacientes en tiempo real a partir de sistemas IoT. Para ello, se utiliza un enfoque de DL mediante una estrategia de combinación de datos de sensores etiquetados y no etiquetados. En este sentido, el modelo puede aprovechar una gran cantidad de datos sin etiquetar disponibles extrayendo características relevantes de estos datos, lo que aumentará el conocimiento en las capas más internas del modelo para mejorar las predicciones. Por lo tanto, el modelo entrenado puede generalizar bien cuando se utiliza en casos de uso del mundo real. Además, se lleva a cabo un proceso de streaming para clasificar patrones de movimiento de pacientes con obesidad en condiciones de tiempo real, lo cual es crucial para la monitorización diaria de pacientes a largo plazo. Este trabajo es el resultado de una colaboración internacional y una estancia de investigación de cinco meses en el grupo de investigación del profesor Anton Popov (Universidad Técnica Nacional de Ucrania "Instituto Politécnico Igor Sikorsky de Kiev").

Artículos de conferencia

1. Sandro Hurtado, José García-Nieto, Ismael Navas Delgado ***A Service for Flexible Management and Analysis of Heterogeneous Clinical Data [33]***.

   En este trabajo, para mostrar la capacidad de integración y la flexibilidad para adaptarse a las nuevas herramientas y funcionalidades que ofrece FIMED, se ha desarrollado FIMED 2.0. El objetivo es proporcionar a los usuarios nuevas funcionalidades para realizar análisis más precisos. La motivación de este trabajo surgió a partir del trabajo propuesto anteriormente: *Reconstruction of gene regulatory networks with multi-objective particle swarm optimisers* [31]. En este sentido, se investiga la inferencia de GRNs incorporando nuevos algoritmos de GRNs. Además, se propone una estrategia *emsemble* de algoritmos de GRNs basado en un sistema de votación para permitir a los usuarios clasificar las interacciones génicas más críticas consensuadas entre las salidas similares de un conjunto de algoritmos de GRNs. De esta manera, se pueden indicar los pares de genes más importantes en el proceso de regulación. Además, se añaden herramientas de visualización a esta nueva versión de FIMED para proporcionar a los usuarios una visión profunda de las redes mediante una mejor representación gráfica.

2. Sandro Hurtado, Hossein Nematzadeh, José García-Nieto, Miguel-Ángel Berciano-Guerrero, Ismael Navas Delgado, **On the Use of Explainable Artificial Intelligence for the Differential Diagnosis of Pigmented Skin Lesions [34].**

   El objetivo de este artículo es que el resultado de las aplicaciones basadas en el ML sea comprendido por los usuarios finales, principalmente cuando utilizan datos médicos y toman decisiones críticas. Este artículo representa un primer intento de investigar empírica y técnicamente la explicabilidad de los métodos modernos de XAI en un conjunto de datos de clasificación de imágenes de melanoma.

Cabe destacar que otras contribuciones científicas se desarrollaron con anterioridad al inicio de la Tesis como: *Análisis de datos de acelerometría para la detección de tipos de actividades* [35]. Se realizó un estudio de viabilidad para clasificar los datos de actividad física de pacientes con problemas cardiovasculares obtenidas mediante pulseras de acelerometría. En este estudio se propusieron redes neuronales profundas, concretamente redes neuronales recurrentes como LSTM, en un caso

de uso en el reconocimiento de la actividad humana. Nos remitimos a este artículo como un estudio previo al trabajo *Human Activity Recognition from sensorised Patient´s Data in Healthcare: A Streaming Deep Learning-based Approach.* Gracias a la estrategia semi-supervisada llevada a cabo en este artículo, que propone la combinación de datos etiquetados y no etiquetados proveniente de sensores, permite entrenar un modelo robusto para generalizar el conocimiento a más pacientes en casos de uso reales.

Conclusiones

La investigación biomédica está en constante crecimiento, ya que se han realizado muchos avances tecnológicos para ampliar las bases de datos, desde la secuenciación de próxima generación hasta otros datos biomédicos disponibles, como las imágenes biomédicas, los metadatos de anotación, los recursos de IoT, etc. Debido a la gran cantidad de datos clínicos generados a diferentes escalas y en múltiples dimensiones, se pueden realizar análisis más completos para mejorar la calidad de vida de los pacientes y prevenir o detectar enfermedades. Estos datos pueden combinarse con otros conjuntos de datos relacionados con el problema. Esta integración de datos puede ser beneficiosa, ya que el estudio de un conjunto de datos tan heterogéneo y procedente de diferentes fuentes de información puede revelar patrones interesantes o información adicional que si los datos se analizaran de forma independiente. Por consiguiente, la integración y la transmisión de datos entre organizaciones clínicas son de suma importancia. Asimismo, en las últimas décadas, el desarrollo de algoritmos de Inteligencia Artificial ha desempeñado un papel esencial en la investigación biomédica. Ha demostrado unas capacidades extraordinarias para interpretar y analizar grandes escalas de datos clínicos y desarrollar modelos predictivos.

La presente tesis aborda un reto clave en la actualidad: la importancia de integrar, combinar y analizar varias fuentes de datos clínicos heterogéneos para apoyar a los médicos en su toma de decisiones. Este enfoque se refiere a la combinación de varios de estos análisis clínicos de forma integrada. Esta integración de datos clínicos permite realizar análisis más complejos con un gran potencial para dar lugar a avances relevantes en muchos campos de las ciencias de la vida. Algunos ejemplos son el diagnóstico de enfermedades y la identificación de nuevos biomarcadores.

En concreto, esta Tesis indaga en las principales cuestiones y retos de investigación formulados en la motivación de este trabajo relacionados con la implantación de la medicina de precisión o la asistencia sanitaria asistida por IA. A este respecto, esta Tesis responde a la pregunta de lo lejos que estamos (tecnológicamente hablando) de resolver estas cuestiones pendientes y también intenta mejorar el ecosistema de datos existente hasta el punto de abordar estas cuestiones. En resumen, las principales aportaciones de esta Tesis se muestran a continuación, respondiendo a las cuestiones de investigación formuladas en la motivación de esta Tesis:

- **Capítulo 3. Contribución a la gestión y el análisis flexible de datos biomédicos heterogéneos**. En el contexto de los sistemas de gestión de datos clínicos, existen continuas limitaciones en el proceso de adquisición de datos debido a la naturaleza heterogénea de los datos clínicos. La mayoría de los sistemas que se encuentran en el estado del arte no pueden cubrir estas limitaciones. Además, pocas de estas herramientas permiten el análisis de los datos clínicos para el diagnóstico de enfermedades. Respondiendo a la pregunta **(Q1)** expuesta en la motivación de esta Tesis, la necesidad de alinear los datos clínicos con estrategias innovadoras de análisis de IA ha estimulado el desarrollo de la integración de datos, la implementación de herramientas de análisis y la representación del conocimiento en el análisis de decisiones heterogéneas y predictivas. En este sentido, hemos diseñado e implementado una nueva herramienta software para la gestión y el análisis flexible de datos biomédicos de múltiples fuentes (FIMED) para cubrir el primer reto **(Ch1)** y las limitaciones mencionadas.

FIMED implementa internamente un flujo de trabajo con varios componentes para la recogida y gestión de datos con adaptabilidad a múltiples ensayos, análisis de datos y visualización de los mismos. En este sentido, FIMED permite al investigador clínico realizar un flujo de trabajo completo en la práctica de los ensayos clínicos. Esta herramienta posibilita el diseño de formularios electrónicos personalizados y fácilmente modificables gracias a la flexibilidad que proporciona su motor de base de datos NoSQL. De esta forma, FIMED permite realizar el proceso de recogida de datos de forma incremental sin necesidad de redefinir el esquema. Además, proporciona una rápida disponibilidad de los datos y permite el almacenamiento de diferentes muestras asociadas al paciente para aportar información adicional, y de esta manera realizar análisis más exhaustivos.

FIMED incluye varias herramientas de análisis para ayudar al experto clínico a detectar enfermedades o buscar nuevos biomarcadores. Entre ellas se encuentran los algoritmos de análisis de la expresión génica, algoritmos para la reconstrucción de redes biológicas y la visualización de datos para anotar la funcionalidad de los genes e identificar los genes involucrados en una enfermedad. La implementación de estas herramientas de análisis fue nuestro primer intento de responder a la pregunta de investigación **(Q2)** y cumplir con el segundo reto **(Ch2)** propuesto en esta Tesis. La motivación era mejorar la inferencia e interpretación de redes biológicas complejas para extraer información relevante para el diagnóstico de enfermedades. Además, FIMED ha sido validado en un caso de uso real con datos de expresión génica de pacientes de Melanoma metastásico. Cabe destacar que FIMED ha sido actualizado en varias ocasiones y está en continuo mantenimiento. En su última versión (FIMED 2.0), se añadieron nuevas herramientas de análisis y de visualización enfocadas al problema de la reconstrucción de redes de regulación génica con el fin de facilitar el diagnóstico de la enfermedad al experto clínico, como se expone en (Ch2). Cabe destacar que FIMED ha sido diseñado con capacidad de extender sus funcionalidades software de forma sencilla, convirtiendo a FIMED en una robusta herramienta de investigación clínica para la gestión, análisis y visualización de datos en ensayos clínicos para diversas enfermedades en estudio. Además de la instancia pública proporcionada[1], el proyecto puede ser desplegado por el personal clínico en cualquier sistema de información sanitaria, garantizando una mayor protección de los datos.

- **Capítulo 4. Contribución a la reconstrucción de redes de regulación génica mediante algoritmos de optimización multiobjetivo.** En el contexto de la reconstrucción de GRNs, los expertos clínicos pueden investigar las funcionalidades de las moléculas biológicas, como los genes, las proteínas y el ARN, y su organización en los organismos vivos. Del mismo modo, los profesionales pueden observar el comportamiento y la organización de los componentes de las células vivas y ver la relación entre ellos en los procesos moleculares para descifrar los mecanismos de múltiples enfermedades genéticas como el cáncer o la diabetes. En este sentido, numerosos estudios recientes han intensificado el uso de optimizadores de enjambre de partículas (PSO) para abordar la inferencia de los GRN. Sin embargo, todavía faltan propuestas basadas en formulaciones multiobjetivo. Por ello, la motivación principal de esta contribución es cubrir el reto **(Ch3)** propuesto en la motivación de esta Tesis, en el que se requieren nuevas metodologías computacionales y experimentales para explicar las redes biológicas debido a la complejidad de la alta dimensión de los datos biomédicos. En este sentido, este trabajo propone aplicar y evaluar un conjunto representativo de optimizadores de enjambre de partículas multiobjetivo (MOPSOs), que utilizan diferentes estrategias de archivo (hipervolumen y agregación) y, en consecuencia, diferentes enfoques para la selección de líderes en el contexto de la inferencia de GRNs. Por lo tanto, este trabajo intenta

[1] https://khaos.uma.es/fimed/

obtener conclusiones imparciales sobre cuál de ellos (y otros MOPSOs relacionados) podría ser utilizado por los expertos en estudios in silico/in vivo para encontrar nuevas posibles interacciones génicas que participen en las regulaciones genéticas.

El ajuste óptimo de los parámetros en *S-Systems* se aborda actualmente con metaheurísticas de optimización continua. En este sentido, se han adaptado una serie de variantes de MOPSO, OMOPSO [36], MOPSO [37], VEPSO [38], SMPSO [39], DMOPSO [40] y MOPSOHv [41], para abordar (por primera vez) la inferencia de GRNs. Estas técnicas han sido seleccionadas por constituir un conjunto heterogéneo de optimizadores multiobjetivo, que realizan diferentes procedimientos de aprendizaje e inducen diferentes comportamientos. En este sentido, se ha realizado una exhaustiva comparación experimental sobre datos de expresión génica con redes de referencia de los retos DREAM3 y DREAM4 [42] basados en organismos reales (E.Coli y Levadura). Por último, se realizan experimentos significativos para inferir redes a partir de muestras in vivo de IRMA y de muestras de cáncer de melanoma de pacientes reales. La capacidad de reproducir el comportamiento biológico se evalúa en términos de convergencia y diversidad algorítmica y en términos de precisión de las redes inferidas respecto a los estándares de oro.

- **Capítulo 5. Contribución al análisis de series temporales en tiempo real con datos biomédicos procedentes de dispositivos IoT**. En el contexto del análisis de series temporales biomédicas, las técnicas de aprendizaje profundo se consideran herramientas potentes que permiten extraer las características más predictivas de conjuntos de datos altamente complejos. Una característica clave que diferencia el aprendizaje profundo de otras técnicas de aprendizaje automático es su capacidad para aprender representaciones directamente de las estructuras de datos sin utilizar descriptores de estructura predefinidos. Esta capacidad elimina la necesidad de procesos convencionales de selección y reducción de características. Sin embargo, los enfoques de aprendizaje profundo requieren grandes cantidades de datos y el etiquetado de estos datos es costoso y requiere mucho tiempo. El etiquetado de datos es un reto cuando se trata de problemas del mundo real en entornos no controlados y aún más cuando se trabaja en casos de uso de Big Data donde se considera una cantidad mínima de datos etiquetados y una cantidad masiva de datos sin etiquetar. En este sentido, la principal aportación de este trabajo es abordar el tercer reto de esta Tesis **(Ch3)**, en el que proponemos nuevas estrategias de IA para tratar con conjuntos de datos clínicos poco etiquetados y no etiquetados. Una forma excelente de abordar estos problemas es adoptar un enfoque semi-supervisado, que puede emplear datos no etiquetados con un pequeño número de ejemplos etiquetados. Por lo tanto, se ha propuesto una estrategia de HAR semi-supervisada para la monitorización de pacientes con sobrepeso en un caso de uso real en el sistema sanitario que implica una tarea de fusión de datos sensorizados por acelerómetro de muestras etiquetadas/no etiquetadas. En concreto, este trabajo pretendía clasificar las actividades diarias de 300 pacientes, lo que equivale a 30 TB de datos privados de movimiento en bruto. Sin embargo, no se disponía de datos etiquetados en nuestro conjunto de datos. Por este motivo, se ha recopilado e integrado un conjunto de conjuntos de datos del estado del arte en el entorno del problema HAR para utilizarlos como datos públicos etiquetados. Para la integración de los datos, se ha utilizado una metodología exhaustiva basada en técnicas de interpolación, normalización, muestreo de datos y desbalanceo de clases, ya que los datos se han recogido de distintas fuentes y dispositivos, en otros formatos y con distintas frecuencias de muestreo **(Ch1)**.

  Además, se ha realizado un estudio preliminar para observar qué conjuntos de datos tenían una distribución similar a nuestro conjunto de datos sin etiquetar. A continuación, se ha entrenado un modelo *CNN-Encoder-Decoder* semi-supervisado con datos públicos etiquetados y privados no etiquetados. El modelo tiene la capacidad de aprender las características

más relevantes de los datos no etiquetados y luego utilizarlas para clasificar las actividades. En este sentido, la extracción de conocimiento de los datos no etiquetados a través de la parte no supervisada del modelo *(Encoder-Decoder)* se almacena y se utiliza como punto de partida para el entrenamiento del modelo en la parte supervisada. Se ha llevado a cabo una experimentación exhaustiva para la selección y validación del modelo, en la que se ha evaluado esta estrategia con distintas cantidades de datos sin etiquetar. El flujo de trabajo de análisis resultante se despliega en un clúster de nodos Spark, por lo que se predice la clasificación continua de 30 TBs de datos de sensores para un grupo de pacientes. El propuesto clasifica adecuadamente los patrones de movimiento en condiciones de tiempo real, lo cual es crucial para la monitorización diaria de pacientes a largo plazo. Representa un paso adelante para cumplir con los desafíos identificados en [6], que consiste principalmente en la generación de plataformas de reconocimiento de actividad en tiempo real y el desarrollo de un modelado no supervisado más preciso para este problema. Por lo tanto, podemos concluir que nuestra estrategia de recopilación e integración de datos, junto con el enfoque en *streaming* semi-supervisado de aprendizaje profundo para la clasificación de actividades, es una solución en esta dirección.

- **Capítulo 6. Contribución a la Inteligencia Artificial eXplicable para la clasificación de imágenes biomédicas**. En el contexto de la inteligencia artificial en medicina, se han desarrollado numerosas técnicas de IA que han logrado recientemente un gran éxito predictivo para muchas aplicaciones biomédicas. Sin embargo, en muchos casos, explicar el resultado clínico de modelos muy complejos es un reto. Por ello, este trabajo propone estudiar y desarrollar técnicas adicionales que permitan clarificar los resultados de estos modelos de *caja negra*, lo cual es esencial para el ámbito clínico en el que las decisiones afectarán a la vida de los pacientes, como se expone en el cuarto reto de esta Tesis **(Ch4)**. La principal contribución es proporcionar al experto clínico la capacidad de interpretar los resultados obtenidos por los algoritmos. En este sentido, se ha desarrollado una metodología para evaluar la calidad de los algoritmos de explicabilidad mediante un conjunto de métricas en un conjunto de datos de imágenes de Melanoma. En primer lugar, un algoritmo preentrenado (Resnet) clasificó imágenes de cáncer de piel de melanoma para la detección temprana de la enfermedad. En segundo lugar, se han aplicado dos de los algoritmos de explicabilidad post-hoc más utilizados (LIME y SHAP) para explicar y validar los resultados obtenidos por el clasificador. Estos algoritmos devolvieron, como resultado, las imágenes de Melanoma con las características más críticas (super/pixeles) de la imagen para realizar la predicción. Finalmente, en el núcleo de este trabajo, se propuso evaluar experimental y técnicamente los resultados de estos algoritmos en términos de reproducibilidad y tiempo de ejecución. En ambos casos, LIME se comporta mejor que SHAP para este caso de uso.

Durante este trabajo, hemos realizado numerosas contribuciones al contexto del Ecosistema de Datos Biomédicos de varias maneras. Desde el punto de vista de la gestión de datos clínicos, hemos diseñado FIMED, que permite la recogida, integración y gestión flexible de datos biomédicos procedentes de múltiples fuentes. Además, hemos propuesto técnicas de IA y optimización y las hemos analizado para proporcionar al investigador herramientas de análisis completas para la detección de enfermedades o el desarrollo de nuevos fármacos. Además, desde el punto de vista de las aplicaciones, hemos abordado varios problemas del mundo real en áreas de la salud, mostrando la utilidad de nuestras propuestas para abordar problemas que podrían surgir en el ámbito académico y en la industria.

Trabajos futuros

Como líneas de investigación futuras en general, planeamos continuar esta propuesta de integración y análisis de datos clínicos de diferentes ensayos con el objetivo principal de mejorar el acceso y la integración de datos sanitarios heterogéneos. También pretendemos mejorar las técnicas de análisis existentes considerando la integración de datos de diferentes fuentes. En este sentido, queremos seguir trabajando en el diseño y desarrollo de estrategias que permitan abordar los problemas en un ecosistema de datos biomédicos y así seguir mejorando en los retos que definimos en la motivación de esta Tesis hacia la consecución de una medicina de precisión y personalizada.

Asimismo, se han identificado diferentes líneas de investigación para trabajos futuros. En este apartado se presentan algunas de las más destacadas:

- Como futura línea de trabajo pretendemos seguir mejorando las limitaciones que implican el primer reto **(Ch1)** de la Tesis. En este sentido, planeamos seguir actualizando FIMED para asegurar la compatibilidad futura con más casos de uso. Así, consideraremos la adaptabilidad a más formatos de muestras de expresión génica, otras enfermedades y la integración con otras herramientas analíticas o algoritmos (algoritmos avanzados de GRNs, nuevos enfoques de ML para el análisis de datos de sensores en tiempo real y análisis de imágenes clínicas, y nuevos algoritmos de XAI). Además, tenemos previsto seguir investigando nuevas formas de integrar los datos en la IA. La integración múltiple de datos permite realizar análisis más complejos con el potencial de lograr avances adecuados en múltiples campos biomédicos. En este sentido, planeamos desarrollar algoritmos que aborden explícitamente la diversidad de datos y los combinen infiriendo un único modelo. Mediante esta estrategia podremos integrar las fuentes de datos dentro de la construcción del modelo predictivo para combinar datos multi-ómicos, imágenes biomédicas e información clínica del paciente en un único modelo robusto. Esta línea de investigación aborda los límites planteados por el enfoque convencional de guiar el análisis ML de forma independiente, combinando un conjunto de datos diversos y extrayendo conclusiones significativas de los datos integrados.

- Enfocándonos en el segundo reto **(Ch2)**, estamos interesados en trabajar en la adaptación de diferentes optimizadores, como la Evolución Diferencial, con parámetros y operadores específicos para la reconstrucción eficiente de GRNs. Para ello, el uso de modernas técnicas de autoconfiguración ayudaría a encontrar una sintonía precisa para los GRNs. Además, el diseño de nuevas estrategias de codificación y coevolución parece ser una línea optimista para mejorar el poder predictivo de los algoritmos. En este sentido, el desarrollo de enfoques paralelos distribuidos podría mejorar el rendimiento de las redes a gran escala. Desde la perspectiva de la modelización de redes, también tenemos previsto trabajar en nuevos enfoques que requieran menos parámetros que *S-System* para ser ajustados. Además, queremos estudiar la integración de datos multiómicos heterogéneos para la inferencia de GRNs. Aunque los datos de micromatrices de ADN se emplean habitualmente para la inferencia de redes, la reconstrucción de GRNs utilizando únicamente datos de micromatrices es fundamentalmente limitada, ya que el valor informativo de dichos datos está restringido por aspectos tecnológicos y biológicos. En consecuencia, sugerimos que se investiguen técnicas más avanzadas para reconstruir con mayor precisión la estructura y la dinámica de las GRNs mediante la combinación de tipos adicionales de datos biológicos, como los datos de experimentos alternativos y bases de datos diferentes.

- Como línea de investigación futura relacionada con el tercer reto **(Ch3)**, en esta Tesis se ha propuesto un enfoque semi-supervisado para ayudar a aprovechar los datos no etiquetados de nuestro conjunto de datos junto con los datos etiquetados recogidos e integrados de la literatura. Hay que reconocer que los resultados son prometedores, ya que hemos aprovechado

el conocimiento de los datos no etiquetados para ayudar al aprendizaje del modelo. Sin embargo, tenemos previsto seguir trabajando en nuevas estrategias para mejorar la calidad de los resultados. Planeamos desarrollar una metodología más robusta basada en técnicas de aprendizaje por transferencia para integrar datos de acelerometría de diferentes fuentes y formatos. Pretendemos realizar un flujo automático que elija aquellos conjuntos de datos de la literatura con la distribución de datos más similar a nuestro conjunto de datos privado y los integre en un único conjunto de datos listo para ser utilizado **(Ch1)**. También tenemos previsto desarrollar estrategias de IA utilizando modelos de aprendizaje automático para ayudar al etiquetado automático de datos.

- Como línea de investigación futura relacionada con el cuarto reto **(Ch4)**, es necesario seguir investigando para diagnosticar las técnicas de toma de decisiones de la IA aplicando los métodos de la XAI. En este sentido, pretendemos investigar nuevas métricas para evaluar los resultados obtenidos de los algoritmos XAI. Como primer paso, proponemos que los resultados de la XAI sean reproducibles y replicables. Por lo tanto, el modelo de entrenamiento debe producir resultados consistentes, y también, el modelo debe funcionar consistentemente incluso cuando se entrena con diferentes muestras de datos. Además, planeamos diseñar nuevas técnicas de visualización de datos masivos para obtener interpretaciones precisas y comprensibles para el experto humano. Asimismo, desde el punto de vista algorítmico, un trabajo futuro de XAI es abordar la explicabilidad del DL en el conjunto de datos de melanoma mediante la mejora de LIME, así como abordar otro tipo de conjuntos de datos de imágenes médicas.

Part I
Introduction and context

Chapter 1

Introduction

In the last few years, practitioners can perform deeper analyses thanks to current advancements in next-generation sequencing (NGS) and the rapid growth and availability of biological data [1, 2]. They can combine biological data with other clinical and patient-specific information, such as electronic health records (EHR), habits, ancestry, and environmental factors, which enable them to analyze and find pertinent information beyond what can be obtained through conventional approaches.

The diagnosis of illnesses is a challenging task in modern medicine. Understanding the precise diagnosis of patients through medical examination and evaluation is the most critical responsibility of practitioners. The healthcare industry generates much information about medical evaluations, patient statements, treatments, prescriptions, and other topics [3].

Clinical data come from numerous sources of information such as data obtained through various massive parallel DNA sequencing techniques, physiological data such as Electrocardiogram (ECG), Encephalogram (EEG) [4], biomedical repositories and even public and scientific social networks [5]. These sources also include medical images, which comprise the bulk of patient data (particularly for cancer patients), disease risk factors, multi-omics, therapy regimens/procedures, and follow-up data.

Similarly, data collection techniques also include sensors (Internet of Things (IoT) technologies), among them, the collection of data through accelerometer sensors, which allow for monitoring the activity of patients [6, 7] and even detect a possible fall of them [8].

This rapid increase in clinical data has highlighted the need for developing new sophisticated tools for data management and analysis in clinical research and personalized medicine. Clinical Trial Management Systems (CTMS) are utilised to retrieve meaningful data from clinical trials, gain early visibility into problems, and find alternative therapies. With the help of these clinical research support software tools, it is possible to acquire new knowledge related to health and discover new drugs [9]. These contributions can also be consolidated in databases to contribute to other pharmacological studies with an investigative nature, to carry out data mining and engineering and to support experts in diagnosis.

CTMS have become an essential support tool for clinical research [10]. These tools bring with them the possibility of collecting relevant clinical information and performing various analysis techniques that help understand the molecular mechanism and potential therapies for human diseases and even provide biological and medical information to enable individualized medical care.

The management of clinical data involved in NGS studies is a challenging task, given the continuous obstacles encountered in system maintenance during patient enrollment, the clinical study sample acquisition process, and the different steps for preparing clinical data processing pipelines. Most of these difficulties occur due to clinical data's dynamic and heterogeneous nature.

Moreover, there is a constant need for researchers in the field to extend the functionalities of these CTMS and integrate data from clinical operations in multiple systems and even be able to deal with different types of samples. Regardless, most clinical tools lack these features, as they are exclusively competent to cover specific case studies and are limited in adapting to the continuous changes in clinical trial practice. Many organizations participating in clinical trials, especially in academic medical centres, have complex data management and quality control processes [11, 12].

The major problem with open source CTMS is that it is difficult to customize a fixed workflow [13, 14]. Therefore, it is difficult for such a CTMS to manage a variety of clinical trials, which is often necessary at academic medical centres. However, most of these systems have design limitations because the heterogeneous and dynamic nature of clinical data, as they are currently unable to meet the needs of adapting to continuous changes in trial practice.

Moreover, clinical systems behave as highly complex systems. It is challenging to understand a phenomenon such as a disease with a single type of data. Nevertheless, they do not allow the integration of different sources of heterogeneous clinical data. In this sense, it is a fact that more research is needed to improve the access and integration of heterogeneous health data to enhance medical care in areas of high unmet public health needs.

In terms of data analytics, Artificial Intelligence (AI) models are becoming more predominant in biomedical research and clinical practice [15]. These models have shown promising in many fields, including risk stratification and modelling, personalized screening, diagnosis of molecular disease, prognosis, and response prediction to therapy [16, 17]. AI can be fuelled by converging large-scale annotated clinical datasets, deep learning breakthroughs, open-source software tools, cheap and quickly rising processing power, and cloud storage. In this sense, integrating AI systems into CTMS is essential since AI can achieve specialized success in some healthcare tasks where they can support doctors in determining the prognosis of patients and surgical procedures. Hence, CTMS must carry out exhaustive analyses and predictions of the data to obtain quality indicators. The combination of AI and CTMS portends to alter the practice of medicine in the distant future.

However, to the best of our knowledge, few systems allow the analysis of patients' clinical data to diagnose clinical diseases, so they are limited in filling the gap among bioinformaticians, molecular geneticists and clinicians. Even with the tremendous advancement in NGS technology and bioinformatics software tools, more improvements are required to deal with complex and genetically heterogeneous diseases.

In that respect, innovative developments could have a therapeutic impact by integrating multiple clinical data flows from diverse sources. These sources include medical images, disease risk factors, data from multi-omic studies, therapy procedures and regimens, and follow-up information. In this sense, there is still room for improvement of the current trend toward precision medicine, producing personalized methods with significant effects on diagnostic and therapeutic pathways [19].

With this motivation, machine learning algorithms use data integration approaches from various biomedical datasets to obtain more accurate results and improve our understanding of biomedical systems. Even so, in most of these software tools, CTMS do not consider the importance of data integration in biomedicine and only considers the variation in a single type of data. In this regard, many essential patterns that can only be observed when considering multiple levels of biomedical data may be missed. Therefore, it is necessary to integrate many different types of data so that the algorithms can make an accurate predictions [18].

1.1 Motivation

To emphasize the importance of Data Integration Tools in the Hype Cycle, Gartner says, *"Organizations need data integration tools to support distributed data management and provide data across*

diverse use cases. These include data integration to support analytics, data science, application integration, self-service data preparation, etc" [20]. Data integration offers flexibility, scalability and extensibility in infrastructure to assure data are consumable across multiple use cases on-premises, multi-cloud or any form of hybrid. As observed in Figure 1.1, Data Integration Tools and non-relational databases (NoSQL), such as Wide-Column Database Management System (DBMS), are progressing until they reach the final market since they are already in the process of the plateau of productivity. For this reason, this Thesis aims to develop a proposal for a data integration tool using a NoSQL database.

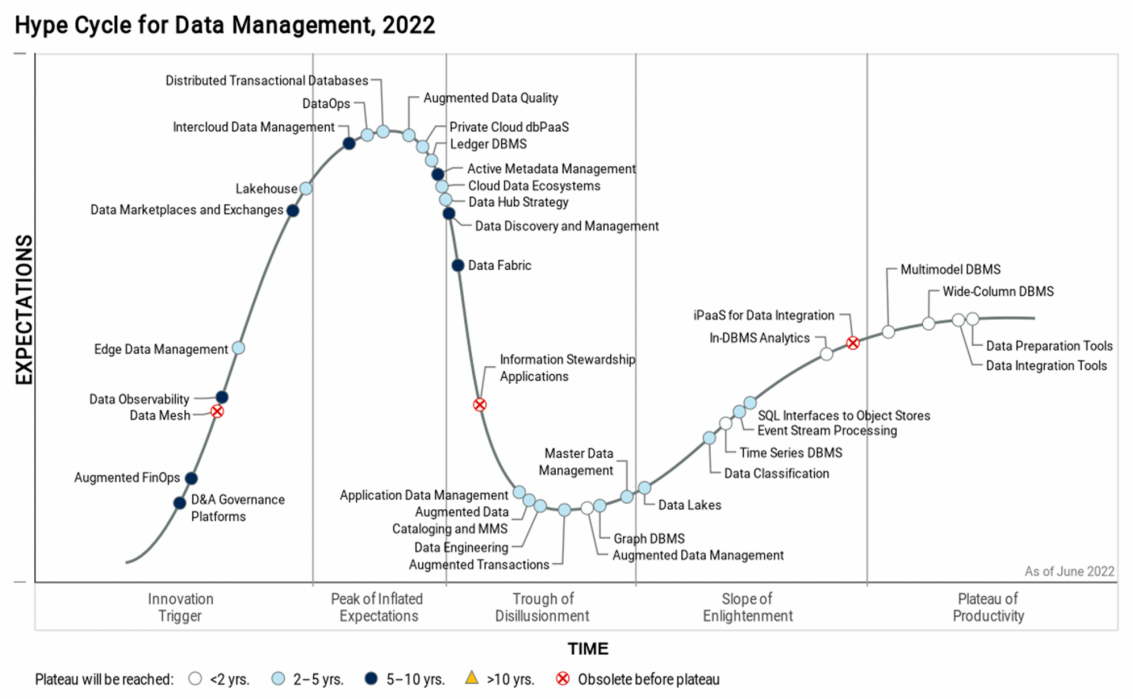


Figure 1.1: Gartner, Hype Cycle for Data Management, 2022.

As a transversal target of this Thesis, we are interested in investigating and integrating new algorithmic proposals that use this type of data, focused on Machine Learning (ML) analysis and optimization, because it is of special interest in some relevant clinical research studies. In this sense, the aim is to simulate the physiology of living systems in biological systems as a set of interacting components rather than as a set of individual physical components. This approach has the practical advantage of providing information on the regulation or optimization of specific system elements, considering the impact on the whole system. Furthermore, it is essential to model the interactions between many components that make up such a biological system to comprehend better the observed complicated global behavior and the underlying physical processes. The employment of modern computational approaches that can conduct an integrated study of such disparate data sources is crucial and challenging at the same time to learn respective large-scale models.
All these reasons lead us to define the hypothesis of this Thesis, *"The automatic integration of heterogeneous clinical data from multiple sources will lead to advanced analysis and the generation of new algorithmic proposals that use this type of data. Biomedical data will be combined as predictor variables to allow a complete model and acquire more relevant results".*

Traditionally, general medicine must tailor treatments based on the patient's clinical and biological characteristics to provide optimal care. In the last few years, many studies have put their efforts on the road to precision medicine [21, 22, 23, 24, 25]. However, many research gaps remain when attempting to encompass problems in a biomedical data ecosystem. Several bottlenecks slow the transition from conventional to personalized medicine. In this regard, the research presented in this Thesis was carried out to answer the following research questions in the management and analysis of biomedical data towards precision medicine.

Research Questions

- **Q1**: Is enough a single clinical data source of information to explain complex diseases such as cancer?
- **Q2**: How can inference and interpretation of complex biological networks be improved to extract helpful information within high-dimensional data to aid in disease diagnosis or the discovery of novel biomarkers?
- **Q3**: Is it possible to improve the quality of AI methods in the biomedical data ecosystem by leveraging large amounts of unlabeled data collected in Big Data environments?
- **Q4**: Can the clinical expert be granted the ability to interpret the results of complex AI algorithms?

In addition, these main research questions can be associated with various challenges that are carried out on real-world problems in a biomedical data environment.

Challenges

- **Ch1**: Biomedical data availability and analysis. There is a need to design new methods and software tools dedicated to improving data collection, management and advanced analysis of biomedical data using different strategies. Leveraging new clinical data integration tools and analysis strategies can potentially drive real change in personalized therapies [24].
- **Ch2**: New computational and experimental methodologies are required to explain biological networks. It is challenging to reconstruct biological networks and interpret them for clinical researchers due to the complexity of the high-dimensional nature of biomedical data [26].
- **Ch3**: AI strategies to deal with small labeled clinical datasets or no labeled data. A common challenge is the scarcity of labeled data. It is necessary to design new strategies that consider use cases with a shortage of labeled data [27].
- **Ch4**: The explainability of automated decision-making in precision medicine. Sophisticated machine learning techniques have recently achieved great predictive success for many biomedical applications. However, it is challenging to explain the clinical outcome of this black-box model [28]. This explicability is essential for the clinical domain where decisions will affect patients' lives.

Therefore, the primary motivation of this Thesis is to probe into the main challenges to the implementation of Precision Medicine defined above. Besides, this Thesis checks how far (technologically speaking) we are from solving the pending issues and improving the existing data ecosystem to the point that these issues are addressed.

1.2 Objectives and phases

This Thesis focuses on contributing with realistic solutions to clinical research systems through the design and development of a clinical trial management software tool together with the application of AI and optimization techniques for data analysis. As the main target, we focus on the application of advanced AI algorithms applied to use cases in a biomedical data ecosystem (such as gene expression data analysis, patient monitoring with wearable sensors, detection of skin cancer in its earliest stages, etc.), and optimization algorithms in the reconstruction of gene regulatory networks. Moreover, eXplainability in Artificial Intelligence (XAI) is carried out to offer interpretability of the results to the clinical researcher. In this respect, we focused our study on the explainability of Deep Learning (DL) algorithms since these algorithms are perceived as complex *"black boxes"* and not easily interpretable. In concrete, the main objectives of this Thesis are detailed as follows:

1. **Objective 1: Design and develop a new software solution to consolidate large amounts of heterogeneous clinical data and integrate analyses across multiple assays.**
   (a) Design the database structure of the tool with a NoSQL database engine (MongoDB).
   (b) Develop the software infrastructure to create biomedical data analysis workflows that can take advantage of data integration from multiple sources.
   (c) Provide tools for analysis and visualization of clinical data.
   (d) Development of a user interface (GUI) that allows the researcher to use the data consolidation, analysis and visualization functionalities easily and intuitively.
2. **Objective 2: Development of meta-heuristic optimization solutions to reconstruct gene regulatory networks from patient gene expression.**
   (a) Adapt well-known multi-objective optimization algorithms to cope with the inference of GRNs.
   (b) Propose new ensembles as a gene regulatory network inference made from prior network algorithms.
   (c) Provide systems biologists with optimization techniques to infer consistent GRNs.
   (d) Provide systems biologists with a rich set of visualization tools to explore the constructed network.
3. **Objective 3: Development of deep learning solutions for analyzing clinical data from patient monitoring sensors.**
   (a) Investigate the design of new ML algorithmic strategies adapted to specific use cases with sensor data.
   (b) Investigate new data integration approaches combining labeled and unlabeled data from sensors.
   (c) Provide new data analysis and workflows for efficient real-time patient monitoring.
4. **Objective 4: Explainability of deep learning algorithms in biomedical use cases**
   (a) Studying the effectiveness of current explainability algorithms in biomedical use cases
   (b) Propose the explainability of ML algorithms to biomedical use cases
   (c) Studying new evaluation metrics for explainability algorithms in biomedical use cases

5. **Objective 5: Tackle real-world and academic problems in the context of biomedical data ecosystem**
   (a) Analysis of gene expression data in a real-world use case for the early detection of the skin cancer disease Melanoma.
   (b) Use of Multi-objective swarm optimizers in the reconstruction of GRNs is validated using real patient data of Melanoma cancer.
   (c) Analysis of real-time series data from a group of overweight patients with cardiovascular disease is analyzed for Human Activity Recognition (HAR) in the healthcare system of Andalusia (Spain).
   (d) Application of explainability strategies for deep learning algorithms in the diagnosis of Melanoma cancer.

1.3 Thesis contributions

The main questions and challenges presented in the motivation of this Thesis can be roughly associated with the contributions included in this study.

- An approach to solve the first challenge (**Ch1**) is presented in **Chapter 3**. In this chapter, we have described FIMED (Flexible Management of Biomedical Data) [30]. FIMED uses a NoSQL database to alleviate some of the limitations imposed by relational databases. This software solution has been designed to support clinical experts in integrating and analyzing heterogeneous clinical data from multiple sources of information in a simple, incremental and dynamic way, avoiding the limitations of the current tools.

- The second challenge (**Ch2**) is covered in the works described in **Chapters 3 and 4**. These chapters investigate the Gene Regulatory Networks (GRNs) reconstruction issue. These gene regulatory networks define the interactions between DNA products and other substances in cells, which is highly relevant to clinical research. Consequently, a greater understanding of GRNs would enable a better understanding of the mechanisms that cause various disorders. To help clinical professionals create novel treatments to treat diseases at their earliest stages or discover novel biomarkers of disease progression, we seek to provide enhancements to GRNs reconstruction.

- **Chapter 5** is directly related to the third challenge (**Ch3**). Traditional ML algorithms require a large amount of labeled data to deliver promising results. However, labeling data is costly and challenging to perform in real-world situations. In particular, many clinical data collection procedures in modern healthcare systems are carried out through IoT technologies with sensors [29]. Data collection through sensors requires Big Data environments with massive amounts of data. Therefore, it is unfeasible to label the data to offer it to train ML algorithms. Hence, in Chapter 5 we propose to develop new ML strategies that can use the limited quantity of labeled data and the vast amounts of unlabeled data to help the algorithms learn, improve their performance, and generalize their knowledge. Moreover this chapter also covers the first challenge (**Ch1**) since, in this work, a methodology for integrating accelerometry data from different sources of information in various formats was designed.

- In **Chapter 6** we attempt to cover the fourth challenge (**Ch4**). We have explored different methods to provide explainability when using complex artificial intelligence algorithms in medicine. Data science and machine learning applications are still perceived as black boxes. However, in the medical field, it is crucial to predict the patient's clinical outcome and

quantify the impact of the drug and, at the same time, take sex, age, and other features into account in an interpretable way. It is essential to explain the results of these algorithms to the clinical expert, thus enhancing the interpretability of the results. Consequently, this work concentrates on investigating and evaluating strategies for the explainability of complex AI algorithms in real-world problems where the model needs to perform well and be easily interpretable.

Following the main hypothesis defined above, Figure 1.2 illustrates the conceptual structure of this Thesis. It is worth noting that the main aim of this work is to cope with real-world problems in a biomedical data ecosystem. In Chapter 3, we have developed FIMED as the core component of this Thesis. As exposed above, FIMED can integrate clinical data from different sources and analysis strategies. In this sense, the Yellow Components refer to the different types of clinical data collected, analyzed and integrated in the context of this Thesis. Likewise, it has focused on the strategic research of AI (Green Components) and optimization (Blue Components) that have been applied to improve the current trend toward precision medicine significantly, resulting in more reliable and personalized approaches with a high impact on diagnostic and therapeutic pathways. Moreover, Big Data technologies have been used for real-time data processing with the Apache Spark cluster computing system. The proposed approach implies a paradigm shift from the definition of statistical and population perspectives to individual predictions, which allows preventive actions and the planning of more effective therapies.

Figure 1.2: Conceptual block involving the components used into this Thesis.

It is worth mentioning that the components that make up this Thesis can be combined. It aims to allow the clinical expert to perform the integration and exhaustive analysis of the data as a line of software products with combinable elements (AI and optimization strategies). The different components can interact, facilitating the interconnection, reuse and traceability of the data value chain.

In this sense, this Thesis covers a diversity of heterogeneous clinical data from different sources of information and in various formats. They have been integrated in some way to later use them by one or more algorithmic proposals of Artificial Intelligence by combinable analysis products. The approach to integrating heterogeneous clinical data in this Thesis can be summarized in three main categories depending on the modelling phase in which the integration occurs. Figure 1.3(A) represents the combination of diverse heterogeneous clinical data sources into a single data set, which is later analyzed by AI strategies. In Figure 1.3(B), each data source generates an independent model during prediction integration. Model weighting might combine these models' predictions, or the outcomes could be explored independently. Moreover, Figure 1.3(C) shows how algorithms explicitly address the diversity of data and combine them by inferring a single model. This integration does not fuse the input data or produce individual models for each data source. Instead, it integrates the data sources within the construction of the predictive model.

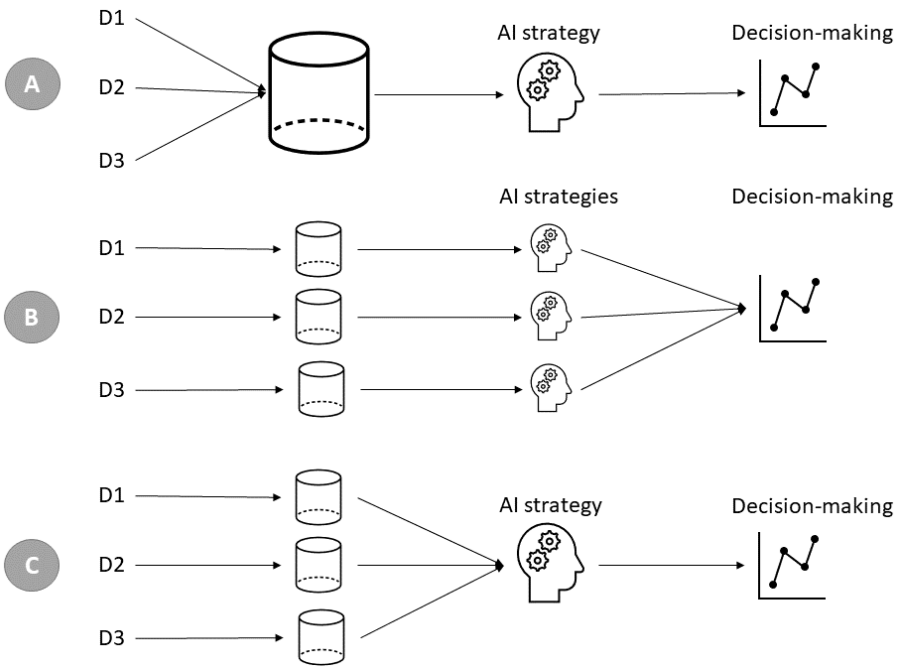


Figure 1.3: A general approach to the integration of heterogeneous clinical data depending on the modeling phase in which the integration occurs.

An overview of how we could integrate the different components defined in this Thesis is shown in Figure 1.4. A clear example is represented in a use case in detecting and diagnosing Melanoma cancer. This use case has been one of the most predominant during the development of this Thesis since we have worked from different points of view, with different types of samples and analysis strategies that can be combined to ensure the effectiveness of the results. In Chapter 3 gene expression samples from patients with melanoma cancer have been analyzed to observe how the genes involved in this process are activated or inhibited over time and to be able to diagnose or

monitor the disease or even compare samples from a set of patients. Likewise, in Chapter 4, gene regulatory network reconstruction strategies have been carried out with this same type of samples to observe how genes interact in transcription processes and discover new biomarkers involved in the disease. In addition, in Chapter 6 melanoma images have been analyzed using various Deep Learning techniques for detection at the earliest stages. Likewise, XAI techniques have been used to explain and interpret the results so that the clinical expert can trust the results of the AI predictions.

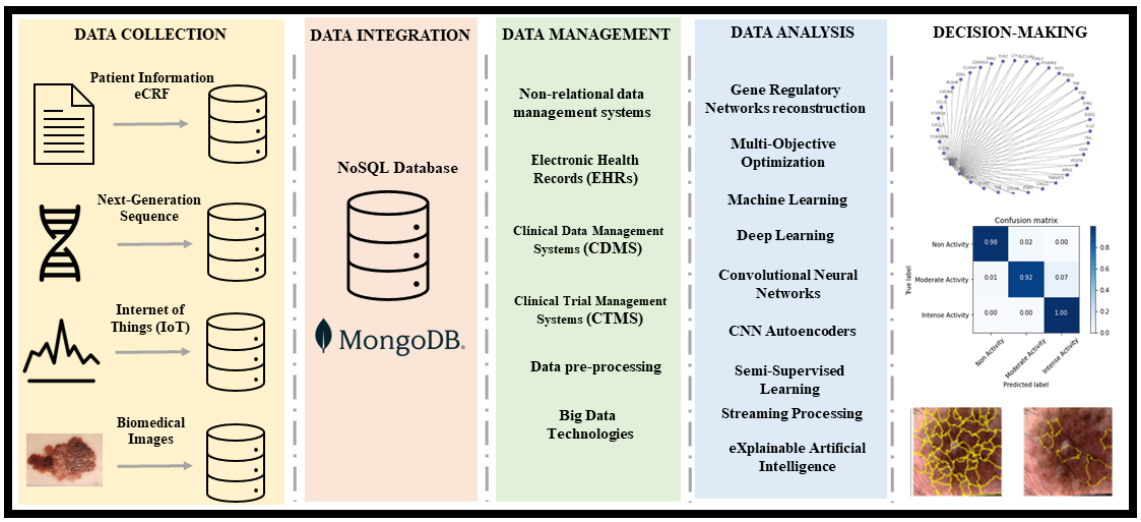


Figure 1.4: An overview of how the different components of this Thesis can combine and interact from the point of view of a precision medicine approach.

1.3.1 Thesis Publications

To transmit the research results, this Thesis has carried out scientific dissemination tasks in journals and conferences. In concrete, three articles in journal indexed in the Journal of Citation Report (JCR) have been published. Also, two articles in an international conference. To highlight the relevance of this Thesis, the list of publications made during its development is shown below.

Journal Articles (JCR)

1. Sandro Hurtado, José García-Nieto, Ismael Navas Delgado, José Francisco Aldana Montes, ***FIMED: Flexible management of biomedical data***, in Journal *Computer Methods and Programs in Biomedicine* (Q1, Category of *Computer Science, Interdisciplinary Applications*, Rank: 20/113, Impact Factor: 7.027, DOI: `https://doi.org/10.1016/j.cmpb.2021.106496`).

   This article is devoted to present FIMED, a software tool that integrates clinical research data from different information sources and formats in a dynamic, flexible and transparent way for the clinical expert. Thanks to the integration of these data, it is possible to develop comprehensive techniques in the analysis of clinical data through machine learning and optimization strategies, in combination with tabular data, sequencing data, biomedical imaging data, data from sensors, etc., to break down the barriers of access and applicability related to

analysis techniques in the field of health. Moreover, the tool provides a web Graphical User Interface (GUI) for easing the management of the clinical trial information and analyses.

2. Sandro Hurtado, José García-Nieto, Ismael Navas Delgado, Antonio J.Nebro, José Francisco Aldana Montes, ***Reconstruction of gene regulatory networks with multi-objective particle swarm optimisers*** in Journal *Applied Intelligence* (Q2, Category of *Computer Science, Artificial Intelligence*, Rank: 46/144, Impact Factor: 5.019, DOI: https://doi.org/10.1007/s10489-020-01891-1).

   In this article, we study the behavior of a set of multi-objective optimization algorithms based on different archiving and leader selection strategies in the scope of the inference of GRNs. The main objective is to provide systems biologists with experimental evidence on which optimization technique performs most successfully to infer consistent GRNs.

3. Sandro Hurtado, José García-Nieto, Anton Popov, Ismael Navas Delgado, ***Human Activity Recognition from sensorised Patient´s Data in Healthcare: A Streaming Deep Learning-based Approach*** in Journal *International Journal of Interactive Multimedia and Artificial Intelligence - IJIMAI*, (Q2, Category of *Computer Science, Interdisciplinary Applications*, Rank: 36/113, Impact Factor: 4.936, DOI: http://dx.doi.org/10.9781/ijimai.2022.05.004).

   This work presents a proposal for analysis and real-time patient monitoring data from IoT systems. For this purpose, we use a deep learning approach using a strategy of combining publicly labeled and private unlabeled sensor data. In this sense, the model can take advantage of a large amount of available unlabeled data by extracting relevant features in this data, which will increase the knowledge in the innermost layers. Therefore, the trained model can generalize well when used in real-world use cases. In addition, a streaming process is carried out of the proposed deep learning approach to classify movement patterns in real-time conditions, which is crucial for long-term daily patient monitoring. This work results from international collaboration and research stay for five months at the research group of Professor Anton Popov (National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute").

Conference Articles

- Sandro Hurtado, José García-Nieto, Ismael Navas Delgado ***A Service for Flexible Management and Analysis of Heterogeneous Clinical Data***, in Conference *International Work-Conference on Bioinformatics and Biomedical Engineering: IWBBIO 2022*, part of the Lecture Notes in Computer Science book series (LNBI,volume 13346, DOI: https://doi.org/10.1007/978-3-031-07704-3_19).

  In this work, to exhibit the integration capabilities and flexibility to adapt to new tools and functionalities offered by FIMED, we have developed FIMED2.0. Our goal is to provide users with new functionalities to perform further and more accurate analyses. The motivation for this work arose from the previous work ***Reconstruction of gene regulatory networks with multi-objective particle swarm optimisers***. In this sense, we place our interest in studying GRNs inference incorporating new GRNs algorithms for a principled comparison among GRNs gene network reconstructions. Also, an ensemble of GRNs is proposed based on a voting system to allow users to rank the most critical gene interactions (top-k genes/edges) between the similar outputs of a set of GRNs. So this can indicate the gene pairs most important in the regulatory process. Moreover, visualization tools are added to this new version of FIMED to provide users with a deep insight into the networks through better graphic plotting.

- Sandro Hurtado, Hossein Nematzadeh, José García-Nieto, Miguel-Ángel Berciano-Guerrero, Ismael Navas Delgado, **On the Use of Explainable Artificial Intelligence for the Differential Diagnosis of Pigmented Skin Lesions**, in Conference *International Work-Conference on Bioinformatics and Biomedical Engineering: IWBBIO 2022*, part of the Lecture Notes in Computer Science book series (LNBI,volume 13346, DOI: https://doi.org/10.1007/978-3-031-07704-3_26).

  The nutshell of this article is that the outcome of the machine learning-based applications should be understood by end users, mainly when using medical data and making critical decisions. This paper represents a first attempt to empirically and technically research the explainability of modern XAI methods on a Melanoma image classification data set.

It should be noted that other conference papers were developed earlier at the beginning of the Thesis such as ***Análisis de datos de acelerometría para la detección de tipos de actividades*** [35]. A feasibility study was conducted to classify physical activities obtained by accelerometry bracelets in patients with cardiovascular problems. This study proposed deep neural networks, specifically recurrent neural networks such as LSTM, in a use case of Human Activity Recognition. We refer to this article, a previous study of our paper ***Human Activity Recognition from sensorised Patient´s Data in Healthcare: A Streaming Deep Learning-based Approach***. Thanks to the semi-supervised strategy combining public labeled from other datasets and our private unlabeled sensor data, this work can train a robust model to generalize the knowledge to more patients in actual use cases.

1.4 Thesis organization

This Thesis is structured into three main parts. Part I contains the current chapter that consists of an introduction to the work done, presenting the motivation to carry it out, the objectives that have been sought, the phases that have been followed to achieve those objectives and the main contributions of the Thesis. Chapter 2 focuses on describing the principles of all the concepts covered, such as general concepts of Clinical Trial Management Systems, non-relational data management systems, IoT systems, Big Data, Artificial Intelligence, Machine Learning, Deep Learning, Semi-supervised learning, Autoencoders, gene regulatory networks, multi-objective optimization and Particle Swarm Optimization (PSO).

Part II, describes three methodologies and associated tools in terms of technology and domain to approach the Thesis goals. Chapter 3 presents FIMED as core or principal component. This software tool supports the clinical expert in integrating and analyzing heterogeneous clinical data from multiple sources. Chapter 4 presents a study based on multi-objective optimization algorithms for the reconstruction of GRNs from gene expression data. Chapter 5 introduces a new semi-supervised deep learning-based approach to the problem of human activity recognition for patient monitoring from wearable devices. It empirically shows the generalizability of the method in a real-world use case with a group of overweight patients in the healthcare system of Andalusia (Spain). Chapter 6 introduces a methodological study to evaluate explainable artificial intelligence algorithms in interpreting the results of complex machine learning models in applying medical problems.

Finally, Part III includes Chapter 7, which presents the main conclusions and the future research lines that are planned to be worked based on to the results of this PhD Thesis.

Chapter 2

Context and fundamentals

This Chapter provides an overview of artificial intelligence, optimisation and big data in the context of the acquisition, management and analysis of heterogeneous biomedical data. First, a general context of the importance of the integration of heterogeneous biomedical data is presented. Then some basic Big Data concepts in the domain of healthcare are described.

This Chapter also exposes the fundamentals of Artificial Intelligence, focusing on predictive modelling for healthcare data analysis as the main area of research. Then, it describes the main fundamentals of optimisation, focusing on Gene Regulatory Networks Multi-Objective optimisation, thus giving an insight into the kind of optimisation carried out in this Thesis.

Finally, the fundamentals of eXplainable Artificial Intelligence (XAI) are presented to contextualise the importance of explainability of AI algorithms in medicine.

2.1 Clinical data management environments

Electronic health records (EHRs) offer chances to improve patient care, incorporate performance metrics into clinical practice, and allow clinical research [43]. Several issues have been raised, including the difficulty of recruiting trial participants, the intrusive nature of data gathering, and the generalizability of the findings. There is a lot of interest in using electronic health records to buck these trends. Observational studies, embedded pragmatic or post-marketing registry-based randomised studies, or comparative efficacy studies are anticipated to use electronic health records as the primary data source [44]. EHRs may be utilised to evaluate research viability, ease patient recruitment, and speed data collection at baseline and follow-up, advancing this strategy for randomised clinical trials [45].

A clinical trial is an investigation of human subjects intended to discover or verify the clinical, pharmacological, or pharmacodynamic effects. Therefore, the conduct of clinical trials is an essential part of the development of new drugs. A Clinical Trial Management Systems (CTMS) is a comprehensive system for managing clinical trials that can help effectively provide patient oversight. This kind of systems do not need to be always integrated with the EHR of the health care system, and so it can be managed in an isolated way to fulfill all the ethical criteria of the clinical trial. A CTMS can include functions to recruit subjects, record case report forms (CRFs), control the overall project schedules, enter the results data, conduct statistical analyses, and monitor the conduct of the clinical trial [46]. Clinical Data Management Systems (CDMS) have become essential in clinical trials to handle the increasing amount of data collected and analyzed [47]. The main goal of CTMS processes is to deliver high-quality data by minimizing the number of mistakes and missing data and gathering maximum data for analysis [48].

In the context of clinical data management systems, in Chapter 3 of this Thesis, we developed a software tool (FIMED) that supports flexible and dynamic EHRs and integrates other possible biomedical data sources related to the patient's clinical trial. For example, this tool allows the clinical expert to combine EHRs with additional patient information (e.g. reports, scanner images, electrocardiograms, encephalograms, etc.). Thus, it will allow for more robust and comprehensive analyses toward a deeper understanding of the molecular mechanisms in a particular disease.

Gene expression regulation is a fundamental molecular mechanism involved in almost every aspect of life, from homeostasis to development, metabolism to behavior, reaction to stimuli to disease progression [49]. Using gene expression data in clinical trials allows biologists to broadly monitor the amount of gene expression, which comprises RNAseq and DNA microarray (Serial analysis of gene expression) [50]. In this respect, FIMED has been used to store and analyze gene expression data from Melanoma cancer patients. Furthermore, in Chapter 4, exhaustive algorithms have been developed to infer gene regulatory networks from patient gene expression data.

Biomedical Signal processing is widely used in clinical trials to solve many problems in the bioinformatics context. Biomedical signals come from electrical activity in the human body in diverse forms, such as electrocardiography (ECG), electrocorticography (ECoG), electromyography (EMG), and electrooculography (EOG) [51]. These biomedical signals have been used as raw input data for algorithms in numerous studies to provide analytical tools to diagnose human diseases [52]. Specifically, in this Thesis (Chapter 5), we have used this type of biomedical data to analyze and monitor obese patients with data collected through wearable sensors.

Biomedical Image processing is also commonly used in diverse clinical applications as an essential part of the existing healthcare system for performing non-invasive diagnostic treatments. It entails creating functional and illustrative models of the internal organs and systems of the human body for use in clinical analysis. It comes in various forms, including X-ray-based techniques including traditional X-rays, molecular imaging, Computed Tomography (CT), mammography, Magnetic Resonance Imaging (MRI), and ultrasonic imaging [53]. Clinical images are increasingly

employed in addition to these medical imaging techniques to identify a variety of disorders, particularly those that pertain to the skin [54, 55]. In this sense, Melanoma skin cancer images have been used in this Thesis (Chapter 6), as input data for training advanced algorithms that allow diagnosing the disease in its earliest stages.

2.1.1 Non-relational data management systems

One of the most commonly adopted systems worldwide for clinical data storage is the Relational Database Management System (RDBMS). The data in RDBMS is highly structured in the form of tables with relations incorporated among them, where Structured Query Language (SQL) is used to communicate with the stored data. While there is a possibility to store some of the clinical data in the structured format, due to the sporadic nature of the new sources and their scalable needs, the relational model is not practical when the requirement of fields is high. This is because it will lead to empty fields resulting in insufficient storage [56]. However, it faces a drawback of rigidity for which the data should always be in the form of tables. Since it is evident from the recent searches and experiments that clinical data is heterogeneous in nature, a shift from the traditional storage in a relational database to an advanced non-relational database format becomes a necessity. Non-relational databases show true signs of usability in clinical applications where large volumes of heterogeneous data are collected and generally unstructured [57].

Non-relational databases, also commonly known as NoSQL (Not Only SQL), are not built primarily on tables, and generally do not use SQL for data manipulation. Indeed, NoSQL databases came into existence due to the limitations of the traditional relational database systems [58]. Additionally, classic relational database management systems are unable to handle the rapid growth of the data with different (or without) structures of information. This technology cannot satisfy agile and highly iterative application development approaches required by the existing processing scenarios of Big Data [59]. For that reason, Non-relational databases are used by a health information management system [60], when working with clinical data, either structured, semi-structured or unstructured data.

In [61], an interesting table of advantages of No-SQL databases compared to RDBMS in an in-clinical use case in COVID-19 data management is shown (see table 2.1 where the main differences are highligthed).

RDBMS	**Non-RDBMS**
Table-oriented with fixed, predetermined, and restrictive schema	Document-oriented databases that are schemaless
Can be only scaled vertically which is limited by budget	Can be scaled horizontally to provide more resilience and lower costs
A very rigid schema and making regular changes is not feasible	It has no constraints and provides adaptability
Can handle data coming in low velocity	Can handle data coming in high velocity

Table 2.1: Difference between RDBMS and non-RDBMS

In [62], NoSQL databases are classified depending on how it is defined their data model, as *key-value store*, *column-oriented store*, *document-oriented store* and *graph databases*:

- ***Key-value***. These database store items as alpha-numeric identifiers (keys) and associate values in simple. The values may be various from simple text strings to more complex like

lists or sets. Examples of this type of database are: Redis [63], Tokyo Cabinet-Tokyo Tyrant [64] or Scalaris [65].

- ***Column-oriented***. The data model of this kind of database is defined as rows and columns, although columnar manner is preferred over the traditional row manner. As Column-oriented databases we could highlight the following: HBase [66], HadoopDB [67] or Apache Cassandra [68].

- ***Document Oriented Store***. These are extended key-value stores in which the value is represented as a document encoded in standard semi-structured formats such as XML, JSON, or BSON (Binary JSON). A document has a flexible schema through adding or removing its attributes at runtime. Unlike the opaque content of values in key-value stores, document stores know the format of documents and support indices and search functionalities based on their attribute names and values. Some instances of document databases are: MongoDB [69] or Apache CouchDB [70], Amazon DynamoDB [71] and Couchbase [72].

- **Graph store database.** The data model store the data in a graph structure to depict the relationship between data by warehousing data in the form of nodes, edges, and properties. Examples of this type of database are: Node4j [73], Virtuoso [74] and Stardog [75]

In Chapter 3, we propose the use of NoSQL databases as a solution for effective management of clinical trial data, since the transformation of the data into a structured format for data analysis are extremely challenging issues in electronic health records development [76]. In this sense, we use MongoDB as core in our software tool for flexible management of biomedical data (FIMED) for the following reasons:

1. **Flexibility**: MongoDB is beneficial in terms of flexibility to deal with the characteristics of clinical data. The dynamic nature of clinical data allows to organise them without being confined to a predefined structure. In this sense, MongoDB allows a schema-less database design or, if required, to tackle semi-structured or structured data. Consequently, FIMED has enormous flexibility when declaring the structure of dynamic form schemas. Database schema does not require to be defined entirely beforehand, and the data structure can change over time without needing to update previous database entries. Thus, it helps in decreasing redundancy and, in turn, improving efficiency.

2. **Scalability**: MongoDB has been designed to operate using a cluster configuration, making it a great choice if scalability and computational effort are required. In general, NoSQL systems are the most suitable for query speed because their performance is efficient, and they are more scalable.

3. **Big Data compatibility**: The emergence of Big Data technologies to handle gigantic volumes of structured and unstructured data all at low cost has right suited it to take the position of data archival solution. MongoDB is designed to establish long-term storage needs, an effective and prompt search of content with the help of keywords or full texts and cost-efficient services. The challenges associated with the velocity, volume and variety of data can be tracked down in a swift, elegant, and agile manner, thereby making MongoDB a scalable back-end data archival solution to such clinical data.

4. **Security**: MongoDB provides some features to FIMED in terms of database security such as authorization and authentication. In addition, it allows data encryption while it is in transit over the network or at rest in storage and backups by the administrators

5. **High functionality and extensibility**: FIMED has been designed to include clinical data analysis tools as a decision-making support tool for clinical experts. In this sense, a semi-structured schema has been designed in the MongoDB database, so that new analytics capabilities can be added, and it can adapt to heterogeneous different data from various sources.

2.1.2 Big Data

Big Data refers to large, complex datasets that are beyond the capabilities of traditional data management systems of storing, managing, and processing in a timely and economical manner. Big Data technologies handle vast amounts of structured, semi-structured, and unstructured data. Compared with traditional datasets, Big Data typically includes masses of unstructured data that need more real-time analysis [77].

Healthcare is one of the domains of application that can significantly benefit from the increasing amounts of data and its availability [78]. Entities, including health care providers, pharmaceutical companies, research institutions, and government agencies, have begun to compile massive amounts of data from research, clinical trials, and public health and insurance programs. The consolidation of data from various sources has significant potential [79]. Physicians are beginning to diagnose and treat patients with a concept known as evidence-based medicine, which involves reviewing large amounts of data aggregated from clinical trials and other treatment pathways on a large scale and making decisions based on the best available information. With Big Data technologies, a physician can look at nationwide trends on what course of treatment would work best for the patient to prescribe the best medications. The aggregation of individual data sets that would otherwise prove meaningless provides doctors with the information needed to make better, more holistic medical decisions.

2.1.2.1 IoT in healthcare

Internet of things (IoT) has been generally defined as *"dynamic global network infrastructure with self-configuring capabilities based on standards and interoperable communication protocols; physical and virtual things in an IoT have identities and attributes and are capable of using intelligent interfaces and being integrated as an information network"* [80, 81]. However, there is not an single or universal definition. Also, IoT has been defined as things-connected network, where things are wirelessly connected via smart sensors [82]. IoT is an inter-connected world-wide network based on sensory, communication, networking, and information processing technologies.

IoT has improved the sanitary system by integrating numerous IoT devices to collect real-time physiological data from patients, including blood glucose levels, temperature monitors, and many other patient vital signs that can be obtained from wearable sensors [83]. Indeed, IoT applications in healthcare have enabled innovative management of the healthcare processes, disease management, assisted living, clinical monitoring remotely, self-caring, detection of some events such as seizure detection, fall detection to help Parkinson's gait disturbance, stroke rehabilitation, neurologic monitoring, and cardiac to reduce medication and human errors [84]. Moreover, the most common IoT applications in healthcare are in some areas such as home healthcare, mobile healthcare or e-healthcare [85].

In this sense, Chapter 5 of this Thesis follows an architecture of IoT in healthcare systems in a real-world case study of Human Activity Recognition (HAR) from sensorised patient data [32]. In this work, we proposed to discriminate basic posture change movements or activities of a group of patients with obesity and cardiovascular problems. In this sense, we can provide tools to practitioners to follow the daily routine of their patients and thus prevent a sedentary lifestyle.

This work was an ongoing collaboration project in the context of a real-world healthcare system (in Andalusia, Spain).

2.1.2.2 Apache Spark

Matei Zaharia initially developed Apache Spark at UC Berkeley's AMPLab in 2009 [86]. Apache Spark is an open-source cluster computing framework, consisting in an ultra-fast engine for storing, processing and analyzing large volumes of data. It is open source and is managed by the Apache Software Foundation [87]. It has emerged as the framework for Big Data analytics with libraries for scalable machine learning, graph analysis, streaming, and structured data processing [88]. Its processing power speeds-up the detection of patterns in data, the organized classification of information, the execution of intensive computation on data and parallel processing in clusters. Moreover, it provides advanced application programming interfaces in Java, Scala, Python, and R.

Apache Spark is mainly composed of several components including Spark core [89]:

- **Spark SQL**: allows access to data in a structured way. It also facilitates Spark integration with Hive, ODBC, JDBC and business intelligence tools [90].
- **Spark Streaming**: provides support for real-time data processing. It is done through a small batch packaging system [91].
- **MLlib - Machine Learning library**: provides a library of powerful machine learning algorithms [92].
- **GraphX**: provides a graph processing API for parallel graph computing [93].

Moreover, Spark internally uses Resilient Distributed Dataset (RDD) to create performant, scalable data pipelines and algorithms. RDDs offer fault-tolerant, parallel data structures that enable users to deliberately store data on disc or in memory and control how it is partitioned. RDD also perform a wide range of operations (e.g., map, filter, union, etc.) and actions (e.g., reduce, collect, count, etc.). Hence, RDD allows the development of different workloads efficiently [89]. Furthermore, as an alternative to RDD, Spark supports Dataframe and Dataset, which are distributed data collection and better handling, from version Spark 1.6.

In Chapter 5, the Apache Spark framework has been used in the Big Data context to classify the daily activity of 300 overweight patients, close to 30 TBs of accelerometer sensor-based data. In this Thesis, we use Spark for parallel processing and analyzing streaming data. So, the presented proposal can classify movement patterns in real-time conditions, which is crucial for long-term daily patient monitoring.

2.2 Artificial Intelligence in biomedical environments

Artificial intelligence (AI) is defined as *"a field of science and engineering concerned with the computational understanding of what is commonly called intelligent behavior, and with the creation of artifacts that exhibit such behavior"* [15]. One of the pioneers of contemporary computer science and artificial intelligence was the British mathematician Alan Turing (1950). He characterized intelligent behavior in a computer as the capacity to complete cognitive tasks at a level comparable to that of a human; this definition later became known as the "Turing test" [94]. AI served as the impetus for a great deal of contemporary research, and early studies on the functioning of the mind contributed to the development of modern logical thought. AI systems are programs that allow computers to behave in ways that give the impression that a computer is intelligent [95]. AI can also refer to circumstances in which machines can mimic human minds in learning and analysis

and thus be used to solve problems. The term Machine Learning (ML) is another name for this type of intelligence [96].

AI has advanced quickly in recent years in a vast number of areas such as the IoT [97], machine vision [98], autonomous driving [99], natural language processing [100], and robotics [101]. Especially, AI has been taking importance in biomedical fields in the last few years since modern medicine is faced with the challenge of gathering, analyzing and applying a large amount of knowledge necessary to solve complex clinical problems. Hence, researchers have been actively attempting to integrate AI into healthcare to enhance analysis and treatment outcomes and, consequently, increase the efficacy of the overall healthcare and medicine [102, 103, 104]. The development of AI systems aimed at assisting the clinician in formulating a diagnosis, making therapeutic decisions, and predicting outcomes has been linked to the development of medical AI. They are made to help healthcare professionals perform daily tasks that require data manipulation and knowledge [105, 106]. As shown in Figure 2.1, research interest in AI applications in biomedicine has gained importance in the past years, especially in the last five years, and continued growth in future can be forecast [107].

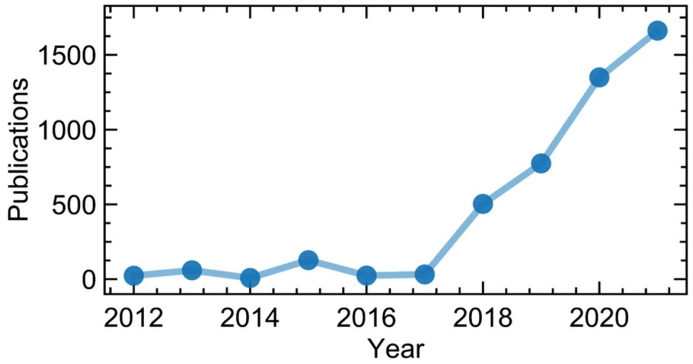


Figure 2.1: Research interest in AI applications in biomedicine [107]

2.2.1 Machine Learning

Machine Learning (ML) is a specific family of methods of artificial intelligence that gathers knowledge from training data and enables computer systems to carry out complicated tasks deftly [108]. Traditional AI systems typically rely on hard-coded rules to define each step of how to solve a problem. In contrast, a machine learning model uses many data to identify features and complete a predefined job. Then it discovers the best way to provide the desired result. Model construction is automated by ML when it comes to creating models. Machine learning is an iterative process that enables the computer to modify its strategies and results in response to new events and data [109]. There are three main categories of machine learning algorithms:

1. **Supervised learning**. In general, supervised machine learning refers to a framework that is trained on labeled data [110]. The data labels are categorized at each data point into one or more groups. The nature of these labeled data (training data) is exploited by the supervised framework, which predicts new data categories (test data).

2. **Unsupervised learning**. It refers to learning without labeling. The goal is to identify shared patterns between data points, such as cluster creation and data point allocation to these clusters [111].

3. **Reinforcement learning**. Reinforcement learning focuses on learning via experience. An agent interacts with its environment and attempts to optimize a reward function in normal reinforcement learning circumstances. During the training and learning process, the agent's goal is to comprehend the impact of its choices and identify the best methods for increasing its rewards.

Additionally, hybrid approaches are also being developed, such as semi-supervised learning (using partially labeled data) [112]. Semi-supervised learning is focused on performing specific learning tasks using both labeled and unlabeled data. It takes advantage of the substantial amounts of unlabeled data present in many use cases in addition to the more common smaller sets of labeled data [113].

It is important to note that supervised learning requires a significant amount of training data, which is both an inefficient and time-consuming operation. In contrast, unsupervised learning does not require any labeled data and instead groups the data into clusters depending on how similar the data points are using either a clustering or a maximum likelihood approach. This method's primary drawback is that it cannot accurately cluster anonymous data. Hence, other paradigms are employed to address issues where most data are unlabeled, such as semi-supervised learning. In this regard, semi-supervised is used to overcome the weaknesses of both supervised and unsupervised learning since it can learn from small amounts of training data and label unknown or test data [114]. These developments have pushed the limits of medical AI by improving already-existing technology and expanding our understanding of illnesses. In particular, these methods produce practical insights by allowing models to discover new patterns and categories rather than being constrained by preexisting classifications, as in the supervised paradigm [115].

Semi-supervised learning models can be divided into semi-supervised learning for classification and semi-supervised learning for clustering. In this PhD Thesis, we will focus on semi-supervised classification algorithms. Many different semi-supervised classification algorithms have been proposed during the past 20 years. These approaches vary concerning the semi-supervised learning presumptions they are founded on, how they employ unlabeled data, and how they interact with supervised algorithms [113]. One of the most predominant methods in semi-supervised learning is the inductive method. The goal of inductive method is to build a classifier that can produce predictions for any object in the input space [116]. Unlabeled data is used in inductive methods to train this classifier. It can be divided into three main categories:

1. *Wrapper methods*: In wrapper methods, classifiers are trained on labeled data and then make predictions on the unlabeled data to produce more labeled data. After that, classifiers can then be retrained using both the existing labeled data and this pseudo-labeled data.

2. *Intrinsically semi-supervised methods*: Unlabeled data is incorporated into the learning method's objective function or optimization process in intrinsically semi-supervised methods. Like semi-supervised support vector machines, these techniques are straightforward expansions of supervised learning techniques.

3. *Unsupervised preprocessing*: Unsupervised preprocessing methods utilize labeled and unlabeled data at two distinct stages. The unsupervised step often consists of either the automated extraction or transformation of sample features from the unlabeled data (feature extraction).

In Chapter 5 of this Thesis, we will use a hybrid approach based on an unsupervised preprocessing method. In this use case, we have a massive amount of unlabeled data, so we will use an unsupervised preprocessing method to do automatic feature extraction from the unlabeled data to improve model learning.

2.2.2 Deep Learning

Non-linear data models called Artificial Neural Networks (ANNs) are designed to mimic the cortical biological neural network in the brain. In the 1940s, fundamental pattern recognition (PR) gave rise to ANN [117]. A variety of artificial neural units make up the ANN. These networks enable computers to mimic the cognitive and logical processes of the human brain by simulating the learning process through different layers in the cortex. That allows for multiple processes like identifying objects in images, processing languages or recognizing and translating voice commands [118]. They have become especially desirable analytical tools in medicine because of their ability to learn from experience, analyze non-linear data, deal with ambiguous information, and generalize information. Thus, it allows the application of the model to other independent data.

If an artificial neural network includes more than three hidden layers, it is said to be a deep neural network. Deep Learning (DL) is considered a new learning paradigm within machine learning. DL includes a collection of learning algorithms to train complex prediction models rather than just one method [119].

Building an ML system required subject expertise and human engineering to create feature extractors that converted data into usable representations in which a learning algorithm could recognize patterns. Traditional ML models need to explore hand-crafted features. On the contrary, DL is a type of representation learning that consists of numerous layers of representations. It involves feeding data so it may create the representations required for pattern recognition. The representation of one layer (starting with the data input) is fed into the next layer and changed into a more abstract representation. These layers are often ordered consecutively and comprise many primitive, nonlinear operations. The input space iteratively warps as data travels through the system's layers until data points can be distinguished [120].

One of the essential benefits of DL is that it automates the feature extraction process, reducing the need for some manual human supervision. For example, in biomedical image classification, DL models automatically extract image features to optimize the model's performance. Hence, deep neural networks can process the data directly and enable the creation of end-to-end predictive models by carrying out all the processing steps typically associated with designing a conventional ML model, such as feature extraction and learning (Figure 2.2).

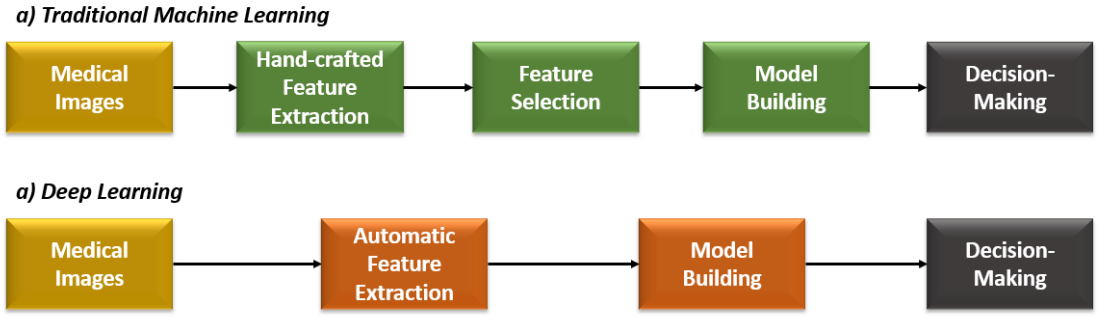


Figure 2.2: Artificial Intelligence workflow for image classification, in which traditional ML workflow includes hand-crafted human features. In contrast, the DL workflow provides a deep feature extraction to extract the most relevant features from images automatically.

Deep learning models usually outperform traditional ML methods because they scale to huge datasets and improve with additional data. Also, DL algorithms can include multiple data types as input, which is an especially relevant feature for heterogeneous healthcare data (Figure 2.3).

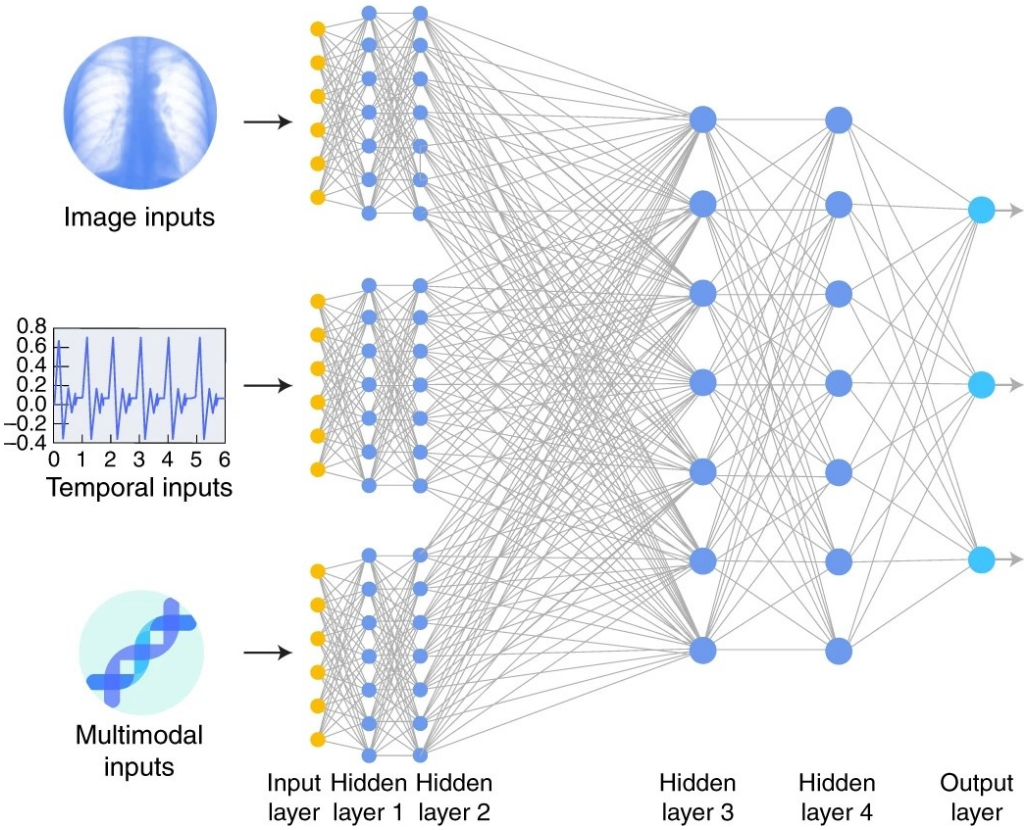


Figure 2.3: Deep neural network develops a practical specialization in its low-level interactions networks using biomedical images, raw time series, and other forms of data as input. Then, combined data is passed to higher layers. Biomedical data interactions are becoming increasingly crucial in the healthcare industry. Image modified from reference [121].

The system's ability to learn from enormous volumes of data in an unsupervised or semi-supervised fashion is one of the critical features of DL. Therefore, it may use a general-purpose learning technique to learn from data on its own without the guidance of an expert. It has become the de facto computing paradigm for ML experts, thanks to recent advances in medical disciplines [107].

2.2.2.1 Convolutional Neural Networks

At the Large Scale Visual Recognition Challenge (ILSVRC2012) [122], the term *Convolutional Neural Network* (CNN) was first introduced in 2012. For the first time, CNN outperformed conventional pattern recognition techniques by halving the error rate on the image classification task. Nowdays, CNN can be considered the most popular deep learning architecture, which has many characteristics with traditional NN. A three-dimensional arrangement of neurons in a CNN differs from a standard NN in that it connects with a portion of the previous layer rather than the full layer when given an image as input, as observed in Figure 2.4.

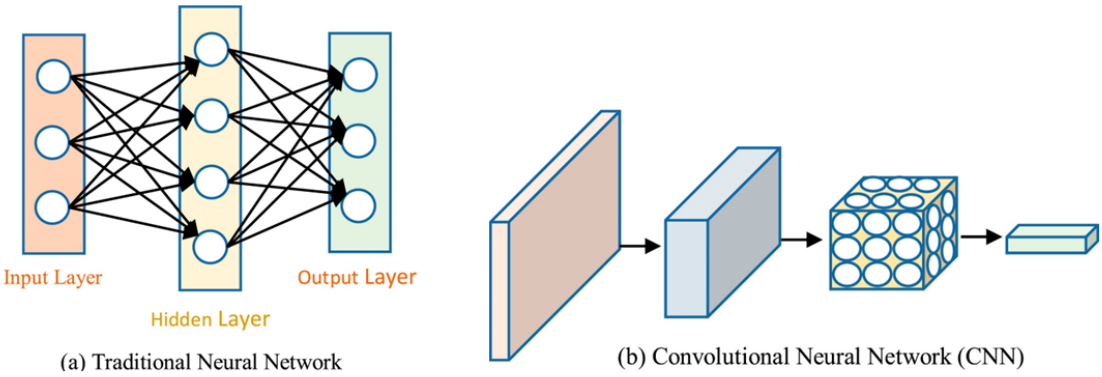


Figure 2.4: (a) Traditional full connected Artificial Neural Network (ANN), (b). Convolutional Neural Network (CNN). Image modified from [123].

The convolutional layer is the initial layer in these neural networks. Each node in the convolutional layer processes only a small portion of the visual field. After the convolutional layers, corrected layer units, or rectified linear unit layer (ReLU), are applied, allowing the CNN to handle complex input. ReLU helps to improve non-linearity and the training speed, since it is a non-linear function. The ReLU function returns 0 if very negative values are passed and remain the same if these values are positive. When the output value is 0, its derivative is also 0, which results in neurons' death (Figure 2.5).

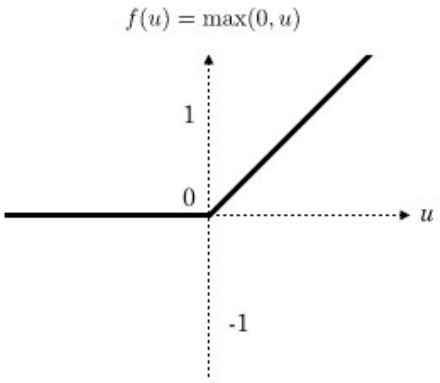


Figure 2.5: Rectified linear unit layer (ReLU). Figure taken from reference [124].

Additionally, CNN comprises others layers such as pooling layer or fully connected layer. Since computations are based on neighboring pixels, the pooling layer down-samples the input values to minimize the spatial dimensionality of the input to reduce computational cost and prevent overfitting [98]. Only relevant data should be extracted, and unnecessary information should be discarded via a perfect pooling process. Hence, down-sampling is applied to divide the input into rectangular pooling regions, as shown in Figure 2.6.

The final layer of the architecture of CNN is often a fully connected layer, similar to the hidden layers of a classic NN in that all of its neurons are coupled to those in the layer before it.

In CNN, each layer enhances the input characteristics that can be used to distinguish any

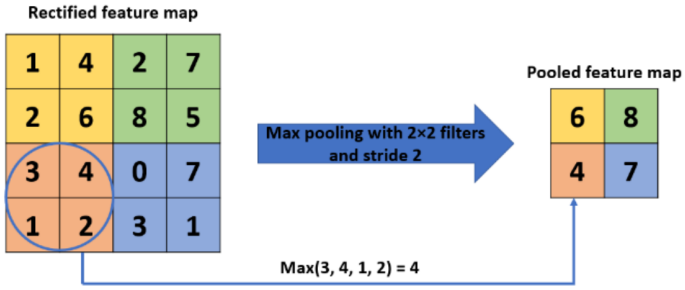


Figure 2.6: Example of Max-Pooling operation. Figure taken from reference [125]

irrelevant variation from the lower layer. For example, in the case of image processing, the first layer of representation often indicates the presence or absence of edges at specific places and orientations in the image when the input is an array of pixel values. The next layer emphasizes specific edge configurations while ignoring any minute differences within the edges to show the existence of motifs. Subsequent layers would identify elements as combinations of these pieces, as this layer can aggregate motifs into larger combinations that approximate parts of known objects. As the number of layers increases, complex functions can be learned if enough of these transformations are combined. For this reason, DL is driving AI research and is considered the industry standard for sophisticated computational models [107].

2.2.2.2 Autoencoders

As commented above, most of the work to date is limited to using a low amount of labeled data to train a supervised model. As a solution, deep neural networks can be used as semi-supervised feature extraction (Unsupervised preprocessing) method to uncover latent representations of the input data [126]. One of the most recognized feature extraction methods is the autoencoder, a neural network containing one or more hidden layers [127]. Autoencoders are trained to learn valuable representations by compressing and reconstructing unlabeled data to learn some patterns from unlabeled data. Thus, the goal of recreating its input is the most well-known example. The network is compelled to find a means to compactly represent its input data by incorporating a hidden layer with a small number of nodes, commonly referred to as the representation layer. The representation layer supplies features once the network has been trained [128]. Figure 2.7 offers a schematic illustration of a typical autoencoder.

Autoencoders seek to reduce the input space's size without significantly reducing its information content. Thus, autoencoders operate by default, presumption that lower-dimensional substructures in the input space represent the data. Additionally, they presume that two samples on the same lower-dimensional substructure have the same label when used as a preprocessing step for classification [129]. Many variants of AE have been developed, such as denoising AE [130], stack denoising AE [119], marginalized denoising AE [131] and Variational AE [132].

In Chapter 5 of this Thesis, we work on a Human Activity Recognition (HAR) use case with obese patients where we have a massive amount of unlabeled data, and the labeled data is minimal. To cope with the limitation of labeled data, we propose to use a semi-supervised autoencoder

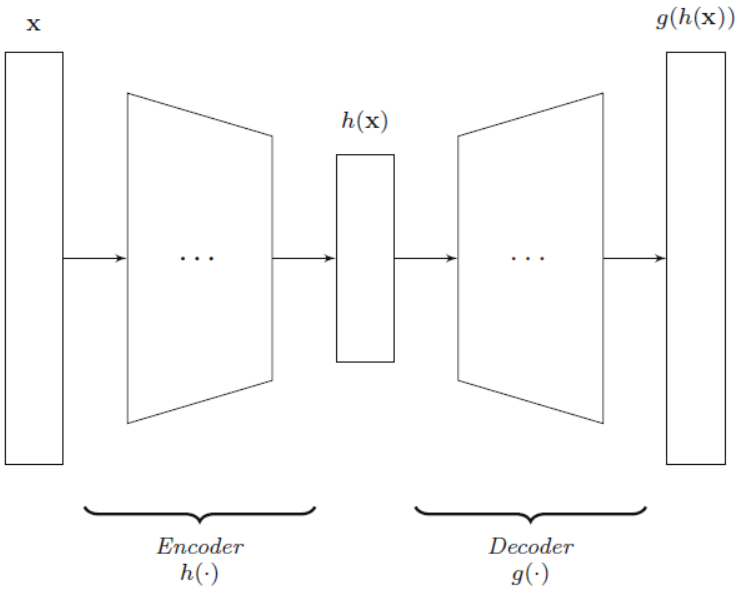


Figure 2.7: Basic architecture of Autoencoder [113]. The basic architecture of an Autoencoder is composed of three main layers: input layer x, hidden layer ($h(x)$) and output layer ($g(h(x))$). In fact, the input layer and the hidden layer compose the enconder h, which compressed an input vector x to its latent representation $h(x)$. Also, the hidden layer and the output layer compose the decoder g, which try to reconstruct the latent representation back to the original x. Hence, the output its the difference between the input x and the corresponding reconstruction $g(h(x))$.

algorithm for classification to extract relevant features from the unlabeled data that help to improve the prediction of the results and generalize the model knowledge.

2.3 Fundamentals of optimization

This section outlines the fundamentals of optimization, and essential concepts are formally defined. We focus on multi-objective algorithms and metaheuristics for dealing with the problem of gene regulatory network reconstruction.

Optimization is an intuitive concept for most of us; it is about finding the best solution to a problem or a good solution, which is essential in real-world problems. The optimization problem consists of maximizing or minimizing a function relative to a set of data, which may have a range of possibilities in a given situation [133]. In this respect, a metaheuristic consists of stochastic algorithms for solving complex optimization problems, which do not guarantee obtaining the optimal solution of a given problem, but when properly tuned allow to obtain near-optimal solutions, often the optimal one, with bounded computation effort. Some typical applications in optimization can be to minimize the costs of a product, maximize the profits of a company, organize the tasks of a company optimally or find the shortest route between two points. Traditionally, these problems could be solved without the aid of computational assistance. However, nowadays, novel optimization methods are used to deal with real-world problems. To define an optimization problem, we have to take into account some concepts: objective function or fitness function, design variables, and constraints. The objective function, $f(x)$, is the output to be maximized or minimized. For example, if we want to minimize the price of a product, the fitness function is the cost of the

product. Design variables are the inputs $(x_1, x_2, x_3, ..., x_n)$ that can be adjusted to change the fitness function, and constraints are limits placed on the value of design variables. Finally, all these variables or quantities are placed in the *solution space or search space* of the problem [134].

Without loss of generality, a formal definition of an *optimization problem*, assuming the minimization case, is defined as follow:

Definition 1 (Optimization problem)**.** *An optimization problem is formalized as a pair* (S, f)*, where* $S \neq \emptyset$ *represents the search space of the problem, and* f *represents the* fitness function*, that is defined as:*

$$f : S \to \mathbb{R} \ . \tag{2.1}$$

Thus, solving an optimization problem consists in finding a solution, $x^* \in S$*, that gratifies the following inequality:*

$$f(x^*) \leq f(x), \quad \forall \ x \in S \ . \tag{2.2}$$

A problem where the fitness function is to be maximized (instead of minimized) does not restrict the generality of the results. It can also be established with this standard problem statement since maximization of a function $f(x)$ is the same as minimizing the negative of $f(x)$ as follows [135, 136]:

$$\max\{f(x)|x \in S\} \equiv \min\{-f(x)|x \in S\} \ . \tag{2.3}$$

Depending on the domain to which S belongs, we can define *binary* $(S \subseteq \mathbb{B}^*)$, *integer* $(S \subseteq \mathbb{N}^*)$, *continuous* $(S \subseteq \mathbb{R}^*)$, or *heterogeneous* $(S \subseteq (\mathbb{B} \cup \mathbb{N} \cup \mathbb{R})^*)$ optimization problems.

2.3.1 Multi-objective optimization

So far, we have talked about mono-objective problems; however, in real-world scenarios, two or more objective functions must be optimized, all equally important for the task. To deal with these problems, we need multi-objective optimization, where our goal is to optimize several objectives simultaneously. In this sense, the formal formulation of a multi-objective optimization problem is presented as follows:

$$\begin{aligned} &\text{minimize} \quad \{f_1(x), f_2(x), ..., f_k(x)\} \\ &\text{subject to} \quad x \in S \subset \mathbb{R}^n \end{aligned} \tag{2.4}$$

with $k \geq 2$ conflicting objective functions $f_i : S \to \mathbb{R}$ and where x is a vector of decision variables from the feasible set S. We can denote an objective vector by $z = f(x) = (f_1(x), f_2(x), ..., f_k(x))^T$.

Generally, in multi-objective optimization, there exists no single solution within the search space where all the objectives reach their individual optima. It has many optimal solutions with different trade-offs. These optimal solutions are called Pareto optimal solutions, which may contain a set of optimal solutions, possibly infinite. Hence, the purpose may be to discover a representative set of Pareto optimal solutions, quantify the trade-offs in satisfying the different objectives, and find a single solution that meets the preferences of a Decision-Maker (DM), normally a human being.

With the goal of being more specific, in (2.4), a design variable vector $x' \in S$ and the corresponding objective vector z are called Pareto optimal if there does not exist another $x \in S$ such as $f_i(x) \leq f_i(x')$ for all $i = 1, ..., k$ and $f_j(x) < f_j(x')$ for at least one index j.

In multi-objective optimization the concept of dominance is used to determine if one solution is better than other solutions. As shown in Figure, a solution a dominate a solution b if a is no

worse than b in all objectives and a is better than b in at least one objective. In the same way, a is non-dominated by b. The solution of the multi-objective optimization problem is characterized by the set of non-dominated solutions, known as Pareto set.

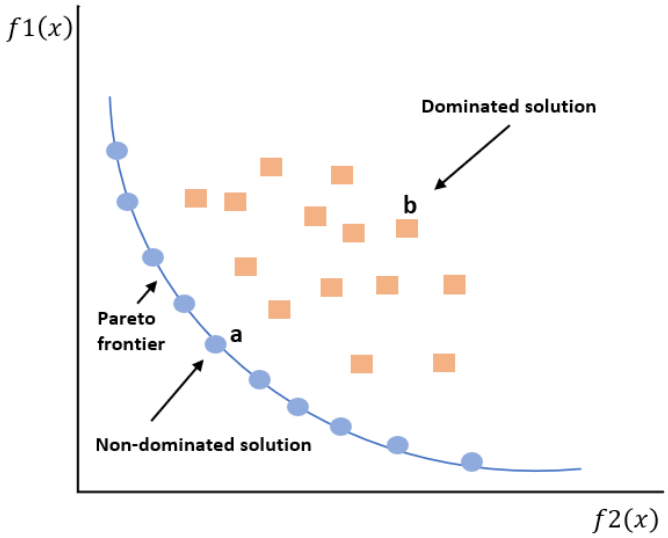


Figure 2.8: Example of a Pareto frontier for a multi-objective optimization problem with two objective functions.

2.3.1.1 Multi-Objective optimization in Gene Regulatory Networks reconstruction

The network biology approach has demonstrated great promise in systems biology research [137]. In the recent decade, network biology has been a popular tool for representing, combining, and revealing intracellular relationships and mechanisms [138, 139].

In biological networks, interactions between genes, proteins, and metabolites are coordinated as part of cellular processes. The fundamental building blocks of network topologies that encode the dynamics of physiological responses and various regulatory motifs are molecular interactions [140]. Biological networks attempt to present a complete cell or organism map, including their collective biological actions and co-expressed traits. They are undoubtedly the best tools for studying complicated diseases like cancer [141].

The biology of a living organism can be explained by transcriptional information stored in DNA, converted into RNA, and then translated into proteins to carry out diverse tasks in a cyclical process that also involves the reproduction and propagation of the information itself [142].

The process of reading information from the genome to produce the set of proteins required for the growth and operation of a living organism is known as gene expression [143]. More specifically, we might characterize this idea as a sequence of cellular processes that strictly adhere to the fundamental principle of molecular biology. DNA's nucleotide sequence eventually provides instructions on which amino acids to combine and in what order to create different biological proteins [144]. Genes are DNA segments containing nucleotide sequences responsible for storing the knowledge required to produce particular polypeptide chains or a particular set of proteins. Since DNA is divided into genes, they can be thought of as nucleotide sequences. Although all cells share a similar genome, each cell's transcriptome is a product of its biological activity and, consequently, the cellular tissue to which it belongs. Genes are selectively translated into RNA in each cell.

Gene expression regulation is the process of gene activation and deactivation that enables cell specialization. It is typically explained by interactions between gene products and transcription factors [145]. This process is referred to as the regulation of gene expression. After RNA is translated into proteins, making them unique in every cell assembly, proteins serve as the final products that further specify the cell's functionality [146].

In recent years, the concept of Gene Regulatory Networks (GRNs) has grown widespread as an influential applied biology strategy for explaining the complex and highly dynamic set of transcriptional interactions (gene to gene interactions) due to its easy-to-interpret features [147]. For the reconstruction of Gene Regulatory Networks (GRNs), gene expression data is essential, which provides information about gene expression by acting as the raw input data. In general, gene expression data is represented by a matrix where a_{ij} represents the gene expression value of the $j-th$ gene $(1 \leq j \leq n)$ in the $i-th$ experiment $(1 \leq i \leq p)$.

$$\mathbf{G} = \begin{bmatrix} a_{11} & \dots & a_{i1} & \dots & a_{p1} \\ \vdots & & \vdots & & \vdots \\ a_{1j} & \dots & a_{ij} & \dots & a_{pj} \\ \vdots & & \vdots & & \vdots \\ a_{1n} & \dots & a_{in} & \dots & a_{pn} \end{bmatrix} \tag{2.5}$$

GRN can be represented as a directed graph compose of gene-to-gene interactions, where regulators genes are connected to target genes by interaction edges based on the information provided by gene expression data [148] (Figure 2.9). Transcription Factors (TF), which may function as both activators and repressors, RNA-binding proteins, and regulatory RNAs are all regulators of gene expression. Understanding biological processes such as cell growth and division, cell differentiation, and development depends on identifying regulatory links between transcriptional regulators and their targets [149].

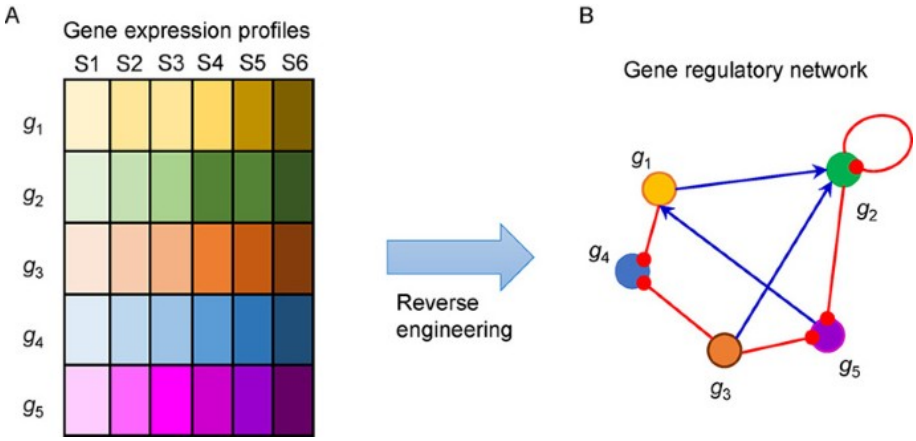


Figure 2.9: Reconstruction of GRNs (B) from gene expression profiling data (A). Image taken from reference [150].

The automatic reconstruction of GRNs is a complex problem found in computational biology [151], which consists in tuning parameters of a model that quantitatively reproduces the dynamics of a given biological system. Since, biologists still struggle to catalogue, predict, and comprehend every GRN relationship across all species and cellular contexts, a variety of computational methods has been studied that are capable of inferring the topology of gene interactions

to form networks [152]. In this regard, multiple optimization techniques such as evolutionary algorithms [153, 154, 155], and especially particle swarm optimization [156, 157, 158, 159, 160], have been applied to the inference of gene regulatory networks from gene expression time-series. However, their precisions are strongly influenced by the quality of available datasets and the characteristics of the learning model used for such predictions. In this sense, S-System [161] framework can obtain a good trade-off between biological relevance and mathematical flexibility. It internally uses an Ordinary Differential Equations (ODE) system, which is a helpful framework to fit continuous variations of genetic regulations over time. Nevertheless, ODE systems require additional computational effort to tune parameters of kinetic orders and rate constants from a usually short amount of gene expression data. Moreover, as usually observed in biological systems, a sparse topology of the network should be accurately reproduced, so the early detection of significant node connections constitutes a major challenge in this process. To cope with these issues, the inference of GRNs has been traditionally tackled as a global optimization problem [162], which has demanded the use of specialized optimization techniques [154, 157, 158] to deal with it.

With this motivation, in Chapter 4 of this Thesis, a study of different GRN reconstruction techniques has been carried out, giving information to the expert biologist as to which method provides the best results. It focuses on PSO due to the relative accurate behaviour and fast convergence. It proposes a set of multi-objective particle swarm optimisers (MOPSOs) using different archiving strategies (hypervolume and aggregation) and consequently different strategies for the selection of leaders, in the context of the inference of GRNs.

2.4 Explainable artificial intelligence in biomedical environments

Thanks to the extraordinary progress in AI research, it has been possible to apply it to many different fields and disciplines, including medicine and healthcare. With the advent of more complex and comprehensive algorithms, such as ensemble algorithms and deep learning algorithms, significant advances have been made in developing recommendation and support systems for the clinical experts. However, due to the complexity of these algorithms, which behave like "black boxes", we cannot obtain a clear interpretation of the results. Many of these algorithms' decisions are still poorly comprehended [163]. Neural networks typically consist of many layers connected via many nonlinear intertwined relations, so it is unfeasible to comprehend how the neural network came to its decision fully. The concern is mounting in various fields of application that these black boxes may be biased in some way like "When do you succeed?.", "When do you fail?.". In this regard, the medical sector requires a high level of responsibility and, therefore, transparency. The clinical expert must know why the algorithm obtains such results, as shown in Figure 2.10. Hence, failure is not an option in medical decisions [164]. In some situations, human lives depend on early detection to stop a fatal disease by taking drugs in the earliest stages. Thus, despite the current encouraging results in the performance of medical tasks shown by AI, they are far from flawless and far from usable by the scientific community in real situations [165].

All of these indicate that the interpretation of ML models must provide knowledge of the model operations and predictions or visualize the model's discrimination rules on what could cause the model to be perturbed [166].

In this sense, the term eXplainable Artificial Intelligence (XAI) was coined to cover all these limitations in the effectiveness of current AI systems [167]. Conceptually, XAI proposes more explainable models while maintaining a high level of learning performance. Thus, these models can be understandable by humans with reliable predictions.

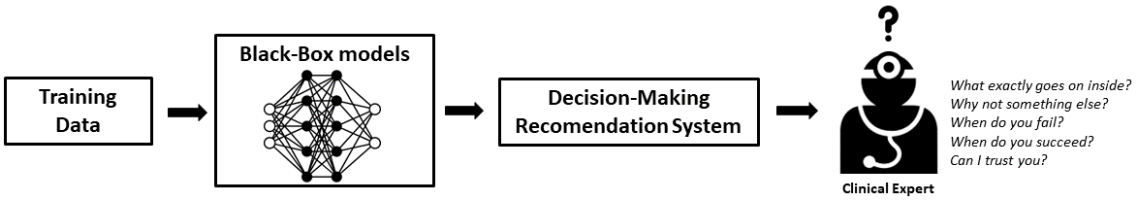


Figure 2.10: Black-box problem in decision-making when using AI solutions.

2.4.1 Explainability vs Interpretability

Several terms are essential to understanding the characteristics of an XAI system, such as explainability and interpretability.

When a model is interpretable, people may easily comprehend how it works and the reasons behind its conclusions [168]. Several ML models are inherently interpretable, such as Logistic Regression models. Humans may consult the weights and odds ratios to understand how the model works and the coefficients to understand the reason behind specific predictions [169]. Also, Decision Tree models can be considered interpretable, even though they become less interpretable as their complexity increases, as in the case of ensemble and deep learning algorithms.

Explainability is the capacity to express how an AI decision has been made to a broader range of end-users in language humans can comprehend [170]. It allows humans to essentially understand how a model works and how decisions are made without accounting for all the details of their computations [171]. In this sense, explainability refers to how predictions are interpreted in the presence of novel cases. In contrast, interpretability refers to how the model is interpreted after training on data [172]. Moreover, explainability is often applied to black-box models in which this technique generate an interpretable model to explain the behavior of the complex model. Consequently, explanations are reasonable approximations of a model's structure and behavior when interpretations directly account for those aspects [173].

2.4.2 Intrinsic vs Post-hoc methods

In general, there are two main categories of XAI methods: post hoc and intrinsic methods [174]. Besides, these methods can be classified into model-specific or model agnostic. Model-specific explanations are specific solutions for a single model or group of models. In contrast, model agnostic explanations can be used for any machine learning model. Usually, agnostic methods analyze the relationship between input and output features. In addition, these methods can be categorized into local or global methods. Local methods explain individuals' predictions while global explanations can explain the entire model behavior [175].

Intrinsic methods refer to models that are simple enough to comprehend but complex enough to accurately match a relationship between input and output [176]. Frequently, these are the classic machine learning models that are intrinsically explainable such as support vector machine, linear regression, logistic regression, k-nearest neighbor, decision trees, and others such as rule-based learners or general additive models. Intrinsic methods require a human to be able to reason about the entire decision-making process of the model internally, or they mandate the usage of a small number of features [176]. By employing this technique, we may determine what portion of the input data contributes to the categorization choice made by any classifier.

Post-hoc explanation examines a trained model (a neural network in deep learning) to gain knowledge of the acquired relationships. The main difference between intrinsic and post-hoc methods is that post-hoc explanation makes the neural network explainable. In contrast, intrinsic

explanation trains a neural network before attempting to explain its behavior [177]. As depicted in Figure 2.11, using the intrinsic XAI method allows the medical XAI application to examine medical data and provide decisions and explanations to physicians. Alternatively, if the medical application were to use post-hoc XAI, the black box methods would be applied to the medical data for decision making, followed by post-hoc XAI that would explain those of the black box methods.

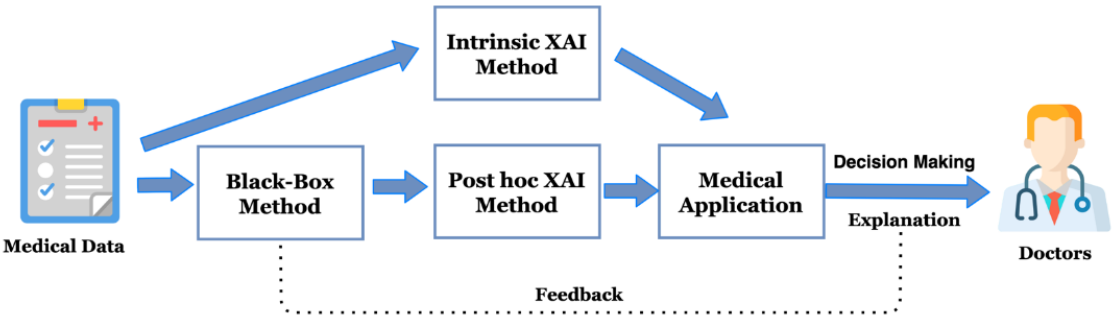


Figure 2.11: General workflow of XAI medical applications in which post-hoc or intrinsic methods provide explanations of the outcome of the black-box method to the clinicians. Image taken from reference [178].

Many risks are involved while making medical decisions since deep learning's black box aspect represents state-of-the-art in medical image processing. XAI is being used more frequently by medical imaging researchers to explain the outcomes of their algorithms. An explanation is adequate if it clarifies the reasoning behind a neural network's choice and helps it make sense. In this respect, Chapter 6 of this Thesis is focused on the use of explainability in deep learning models. A deep neural network often uses hundreds to millions of weights. This type of network is neither sparse nor well suited for a human to replicate and rationalize a model's whole decision-making process internally. Therefore, we intend to explain the performance of these neural networks in diagnostic biomedical imaging. In particular, we have focused on detecting melanoma skin cancer by image processing. Thus, we will apply explainability strategies to deep learning models in the context of biomedical imaging applying post-hoc methods.

Numerous post-hoc methods have been used for black-box explanation in diagnosing biomedical imaging. In particular, surrogate methods allow explaining the outcome of complex deep learning models by implementing an interpretable model, such as linear models, on the outputs of the deep learning model [179]. In Chapter 6, two of the most well-known methods are used: *Local interpretable model-agnostic explanations* (LIME) [180] and *SHapley Additive exPlanations* (SHAP) [181]. When applying these methods to image classification models such as our convolutional neural network for predicting deseases, they provide visual explanations, also called saliency mapping. Visual explanation methods display the essential elements of a picture that influence a choice. Although the majority of saliency mapping techniques use backpropagation-based techniques, others employ perturbation-based or multiple instances learning-based techniques.

LIME is a method that aims to explain each individual prediction by approximating any black box machine learning model with a local and interpretable model, such a linear model. LIME employ perturbation-based techniques to perturbs the original data points, which feeds them into the black box model and looks at the resulting outputs. The approach then adjusts the weights of those additional data points based on how close they are to the original point. Ultimately, it uses those sample weights to fit an interpretable model, such as a regression or decision tree model, to the dataset with variations. The newly trained explanation model can then be used to explain each original data point.

SHAP is a model-agnostic method that uses traditional Shapley values from game theory [182] to explain the output of any machine learning model. It provides local explanations by determining the marginal contributions of each feature to the model's outcome.

2.5 Computational infrastructure

Almost all the experiments conducted in this Thesis have been performed in a virtualization environment running on a private high-performance cluster computing platform. This infrastructure is located at the Ada Byron Research Center at the University of Málaga (Spain), and comprises a number of IBM hosting racks for storage, units of virtualization, server compounds and backup services. The physical platform comprises an IBM Chassis BladeCenter H type and BULL server with a high-performance VMX5300, unified CPU, memory and storage. It also contains a NovaScale Blade BL265 only for storage. The virtualization layer supports computing resources through VMWare ESXi (and VMWare vCloud), storage through IBM SVC (SAN Volume Controller), backup through Veem Backup and IBM Protectier (virtualization of tapes) and desktops are managed with VMWare Horizon and Virtual Cable UDS. The general characteristics of this virtualization installation are: a processor with 256 cores and 578.94 GHz, 2.75 TB RAM, 84.41 TB storage space, 10 Gbps LAN network, 1 Gbps internet network, iSCSI 10 Gbps and FC 8 Gbps storage network.

Moreover, in others experiments with massive data processing, a super-computing platform infrastructure is used, which hardware is managed by a Slurm middleware acting as the distributed task scheduler. It takes part in the Picasso Supercomputer (RES node) located in the Bio-Innovation Building of the University of Málaga.

Part II
Methodology, analysis and results

Chapter 3

Contribution to flexible management and analysis of heterogeneous biomedical data

In the last decade, Clinical Trial Management Systems (CTMS) have become an essential support tool for data management and analysis in clinical research. However, these clinical tools have design limitations since they cannot cover the needs of adaptation to the continuous changes in the practice of the trials due to the heterogeneous and dynamic nature of the clinical research data. These systems are usually proprietary solutions provided by vendors for specific tasks.

In this chapter, is devoted to present FIMED, a software solution for the *Flexible Management of Biomedical Data* from multiple trials, which can contribute positively by improving clinical researchers' quality and ease in clinical trials. This tool allows a dynamic and incremental design of patients' profiles in the context of clinical trials, providing a flexible user interface that hides the complexity of using databases. Clinical researchers can define personalized data schema according to their needs and clinical study specifications. Thus, FIMED allows the incorporation of separate clinical data analysis from multiple trials. A real-world use case has demonstrated the efficiency of the software for a clinical assay in Melanoma disease, which has been indeed anonymized to provide a user demonstration. Moreover, FIMED is flexible enough to be used in the context of any other illness where clinical data and assays are involved.

An instance of this tool is freely available on the web at `https://khaos.uma.es/fimedV2`. It can be accessed with a demo user account, *"researcher"*, using the password *"demo"*.

3.1 Introduction

Current advances in Next-Generation Sequencing (NGS), together with the consequent fast-growth and availability of biological data [1, 2], enable practitioners to combine these data with other clinical and personal information of patients, such as electronic health records, habits, inheritance and environmental factors; and therefore perform deeper analyses [183]. It promotes the development of sophisticated tools for data management and analysis in clinical research and personalized medicine [184, 185, 186].

Managing clinical data involved in NGS studies is a challenging task [187], given the continuous obstacles encountered in the system maintenance during patient enrollment, the acquisition process of clinical study samples and the different steps for the preparation of processing pipelines of clinical data. Most of these difficulties are indeed produced due to the dynamic and heterogeneous nature of clinical data [188]. The variability of clinical data concerning the type of data requires special attention in data management, systems since large volumes of heterogeneous data are integrated from multiple sources with different structures and data formats.

There is a myriad of research efforts implementing software applications focused on managing clinical information, which traditionally relied on the use of relational database management systems, such as MySQL, Oracle, or Microsoft SQL Server [189, 190]. Although the relational data model is the most extended established approach to data management, it introduces certain limitations when dealing with clinical data [186]. In this sense, since relational databases require the schema design to be set up before introducing data, this demands that software engineers know the structure of the data that will be stored and the characteristics they possess in advance [191]. Later modifications in the schema, once the users are introducing data through an application, are complex as they need to be done by engineers and can have consequences, such as data loss and data inconsistencies [192]. However, data collection could produce cases where new clinical variables need to be considered [193]. For this reason, using relational databases to store clinical research data would cause dispersed tables with empty fields due to schema changes. In consequence, we can identify several features that would be of interest in clinical research tools [13]:

- Dynamically storing the clinical data from multiple clinical trials;
- Allowing to expand their functionalities;
- Integrating data from different clinical operations in multiple systems;
- Transferring data to different types of samples to target different analysis;
- Being adequate to the special characteristics of clinical data;
- Using a database schema that grants sufficient adaptability to face the continuous changes in the practice of clinical trials;
- Enabling ways to secure patients' information.

As commented before, data management in clinical trials is a complex and multidimensional process without information technology. In that respect, the main challenge presented in this chapter is to give design and implementation details of FIMED, a more complete, usable, and high quality data management system for clinical trials that meets the needs of the clinical expert without programming skills and providing flexible management and analysis of clinical research information.

3.2 Related works

Different software platforms exist devoted to storing and processing large volumes of heterogeneous data from multiple data sources, which are also focused on performing computation on encrypted data in different application domains. Some prominent examples in this sense are TrajMesa [194], a holistic distributed NoSQL trajectory storage engine. TrajMesa can manage many trajectories and support plenty of query types efficiently; PERSIST [195], is a middleware architecture that externalizes the complexity of a federated cloud, storage architecture and the complex storage logic from the SaaS application to storage policies. This platform also allows tenants to enforce different storage- and privacy-related requirements at a fine-grained level and supports the dynamic (re)configurability of the underlying federated cloud storage architecture. PERSIST offers support for run-time cross-provider polyglot persistence and the confidentiality of sensitive data through encryption.

A third solution is CryptDICE [196], a distributed data protection system that provides built-in support for several different data encryption schemes, supports making appropriate trade-offs and execution of these encryption decisions at diverse levels of data granularity, and integrates a lightweight service that performs dynamic deployment of User Defined Functions (UDF), without performing any alteration directly in the database engine for heterogeneous NoSQL databases. This leads to realizing low-latency aggregate queries and also avoiding expensive data shuffling. Finally, SecureNoSQL [197] aims to cover data confidentiality and the integrity of the datasets residing on a cloud server. In this last platform, a secure proxy does the required transformations, and the cloud server is not modified. The construction applies to all NoSQL data models, especially those oriented to a document-store data model.

Similarly, in the specific case of clinical data management, there are many software packages already developed, some of them freely available to clinicians [198, 199, 200, 201, 202, 203, 204, 205]. In this regard, OpenClinica [200] is one of the most prominent tools designed to capture clinical trial data. This web-based tool allows designing electronic Case Report Forms (eCRF), firstly building them in any spreadsheet program and uploading them via the user interface. However, the forms must be uploaded again in the tool if users want to modify or update them in OpenClinica. For this reason, the users will be hindered as they will have to constantly load the forms into the tool due to the heterogeneity of the clinical data and the changes that may occur in the different trials. REDCap [202] is a research electronic data capture tool where clinical researchers declare the fields in a spreadsheet using metadata and send it to the REDCap team. The computer scientists design the tables in the database and deploy the web application for the specific case. However, additional modifications in the data structure need to be approved by the REDCap team. Thus, some changes are not allowed due to the database limitations, which limits the system's flexibility.

There are many other software tools for the management of clinical data (such as [199, 201, 203, 204]), with more or less similar functionalities as described in some surveys [206, 207, 208]. In this regard, the survey presented in [209] indicates that most clinical data management systems are web-based platforms based on the needs of a specific clinical trial in the shortest possible time. Therefore, these systems do not fully support the clinical data management process and lack flexibility and extensibility in terms of development. Similarly, as argued in [210], the systems used to collect study data are often redundant to systems used in patient care. Consequently, the data collection in studies is inefficient, and data quality may suffer from unsynchronized datasets, non-normalized database scenarios and manually executed data transfers. A solution proposed in [210] consists of OpenCampus Research, an open adoption software (OAS) that provides a familiar environment for state-of-the-art research database management. However, practically none of these tools include the possibility of analyzing the clinical data of patients in terms of disease exploration.

3.3 Proposed approach

As found in the literature, most of the clinical data management systems used in the research centres are weak in supporting the data management process and managing clinical trial's workflow [209]. Most of the systems consist in web-based platform oriented to the needs of a specific clinical trial in the shortest possible time. Moreover, most of these systems use relational databases. They do not fully support the process of clinical data management since lacked flexibility and extensibility for system development. Besides, these systems do not consider an entire clinical trial workflow, since data collection/management is typically separated from data analytics.

With the motivation of approaching all these features, we designed and developed FIMED [30], a tool for the flexible management and analysis of clinical research information. It is a "*do-it-yourself*" tool that allows users to build their forms in a simple, incremental and dynamic way to facilitate multiple source data collection. FIMED offers an advance in the functionalities offered by these tools providing users with an easy-to-use tool for the flexible design (including later modifications on it) of eCRFs according to the clinician's needs.

FIMED has been developed using MongoDB to alleviate some of the limitations imposed by relational databases. As commented in Chapter 2, MongoDB is a non-relational database and document-oriented database [211], where a schema does not require to be defined entirely beforehand, and the data structure can change over time without needing to update previous database entries. Thus, any new data entry can introduce schema changes without declaring them at the schema level. This allows a flexible design of the databases at the data load phase. FIMED provides a web user interface in which users focus on inserting the data they collect in clinical trials. As soon as they detect new fields to be added, they are just included in the latest data inserted in the database, so the model is updated. MongoDB has been designed to operate using a cluster configuration, making it a great choice when scalability is required.

FIMED also provides analysis tools for clinical trials in order to minimize bias. It also enables functionalities for gene expression data analysis in a semi-automatic and straightforward way using heat maps, cluster heat maps and gene regulatory networks inferred from inserted data. This provides practitioners with early insights into the gene expression samples of a patient (discovering changes in their gene expression levels) or sets of patients (discovering patient clusters with a possible clinical correlation).

In this section, we describe the design issues of FIMED, focusing mainly on its internal architecture, which has been changing and increasing over time, to offer an easy-to-use view of its main features and add new analytics functionalities. Among the covered features, we describe the main components that compose a general workflow in FIMED, including clinical data collection and management, data mapping to provide adaptability to multiple trials, data analysis and data visualization.

3.3.1 Architecture of FIMED

FIMED internally implements a workflow as depicted in Figure 3.1, which consists of several phases: data collection, integration, analysis and visualization. Thanks to the web interface, the user is guided through this workflow, so internal data mappings and adaptations are automatically conducted.

3.3.1.1 Data collection

We have designed a core MongoDB schema to integrate clinical information and other related information, such as gene expression data. This schema can be observed in the JSON Code Snipped 3.1 as a single collection of users, each corresponding to a MongoDB document in the

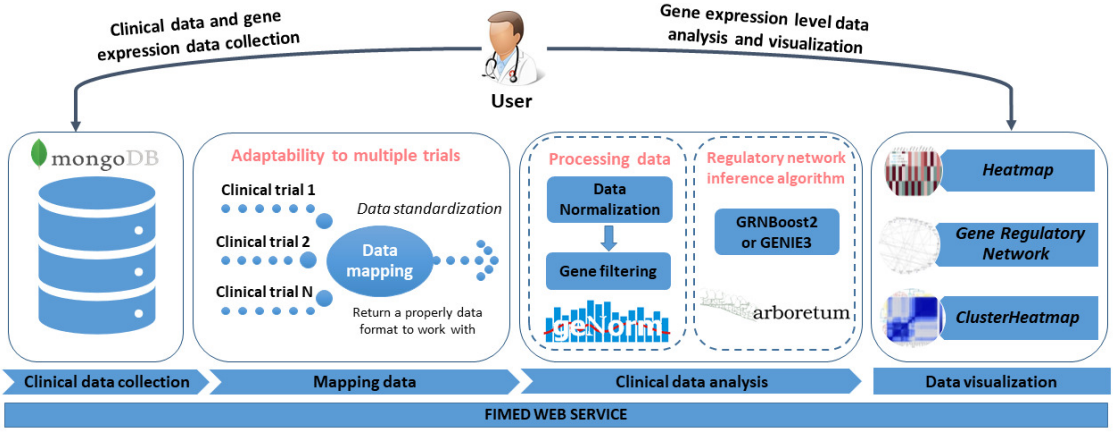


Figure 3.1: General clinical trial workflow in FIMED. This workflow contains several components such as data collection, mapping data, data analysis and data visualization.

database. Each entry (user) in this collection contains the list of patients who have undergone clinical trials with this user (clinician). The user can store clinical trial information of each patient (e.g. name, gender, date of birth, medical records, medication, diagnosis date, disease's progress, etc.) and associated files obtained from, for example, gene expression assays in different formats. Moreover, the user could attach files associated with the patient as additional information (e.g. reports in PDF, scanner images, signed informed consent, etc.).

Code Snippet 3.1: Core JSON Schema. It constitutes the initial document structure from which the database is incrementally adding new elements and updating existing ones.

```
{   "_id": <ObjectId()>,
    "Name": <String>,
    "Surname": <String>,
    "Password": <String>,
    "Patients": [{
            "_id": <ObjectId()>,
            "_patientInformation": <Object>,
            "_files": [
            {
                "filename": <String>,
                "metadata": <Object>,
                "gridFS": <Object>
            }],
            "_clinicalSamples": [
            {
                "sample_name": <String>,
                "metadata": <Object>,
                "gridFS": <Object>
            }]
        }],
    "Form": <Object>,
    "Analysis_results": [
        {
            "name_analysis": <String>,
            "results": <String>
        }]
}
```

This database is supported by a web-based front-end to facilitate the data collection, which is dynamically adapting user needs along with the database scheme. The data insertion process starts with an initial form with fields according to the core JSON schema. The user can add new fields dynamically, just indicating the field name and inserting data for this field in the current patient. The fields used in previous records are automatically shown when adding information for a new patient. The system also learns from the fields introduced by any user, so when a new field is created, its use is available to other users, which can fill it from now on (when required). The system processes the data in real-time, so for each field in which the user introduces information, it is automatically linked with the corresponding clinician records when applicable.

Patients' information is secured in the whole database. For this proposal, FIMED uses the Advanced Encryption Standard (AES). AES is an encryption algorithm [212], which employs the same secret key to encrypt and decrypt the data. FIMED uses the 256-bit keys, the longest AES allowed and recommended to achieve robust data security. Hence, the encryption key (secret key) is a random combination of a suitable length of 256 bits that is generated on the server side during the registration process each time a user is registered in FIMED.

This encryption key is used in the first instance to encrypt the user password. It is also used as a secret key to encrypt patients' information in the database and decrypt the data when retrieving it by querying it. Thus, we avoid that the clinicians registered in the application can access information that does not correspond to them. Hence, FIMED protects the sensitive information of users and their patients.

The proper management of cryptographic keys is essential to the effective use of encryption products. Loss or corruption of these keys can lead to loss of access to systems and data, as well as making a system completely unusable unless it is reformatted and reinstalled. For this reason, FIMED saved the cryptographic keys to a MinIO [1] cluster. This cluster is an internal network only accessible from the server where the APIs are allocated. This enables a fine grain track on the access to these keys, replication of the information to avoid losing keys and an additional security level provided by MinIO.

3.3.1.2 Gene expression data mapping

A series of mapping processes have been developed to translate gene expression data into a suitable format for processing and analysis. This enables the tool to import gene expression data in different NGS file formats from different providers (Nanostring [2], Affymetrix [3], etc.). In this sense, we first perform a gene expression data parsing process since different brands of machines for gene expression profiling will produce different formats. The parsing process aims to extract the gene names, class names and gene expression values to obtain a gene expression matrix. These data provide us with the level of expression of each gene in the patient's temporary samples to be later pre-processed and analyzed by the tools offered by FIMED, as observed in Figure 3.2.

FIMED supports gene expression samples in RCC (*Reporter Code Count*) format. It has been tested with RCC examples and real samples from the Immune Profiling Panel Nanostring^{TM}. Each RCC file contains the count for each target mRNA molecule in a sample. From each RCC file, we can extract Code Class, Gene Name, Accession and Count (see Figure 3.2 A), essential lane attributes to carry out the normalization process and obtain the gene expression matrix. At the end of this step, the system has uniform gene expression data, and additional transformations are done to ensure high-quality results.

[1] https://min.io/
[2] https://nanostring.com/
[3] https://www.affymetrix.com/index.affx

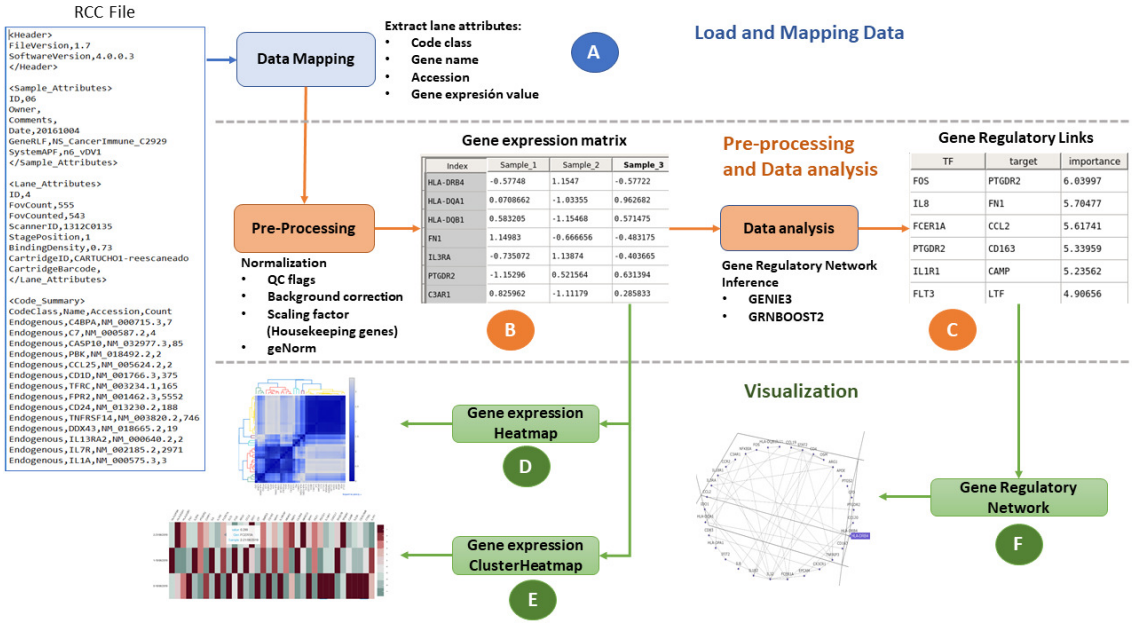


Figure 3.2: FIMED provides functions for (A) loading, mapping, (B) pre-processing, (C) and analyzing data from different profiling panels. It also offers data visualization through (D) gene expression heatmap, (E) gene expression cluster heatmap, and (F) reconstruction of gene regulatory network visualizations.

3.3.1.3 Gene expression pre-processing

In this step of the workflow, the gene expression files previously generated are pre-processed since the variation of gene expression data is the aggregation of biological variations that could include possible bias or noise produced during the gene sequencing process. FIMED focuses on normalization in this stage.

A standard normalization process is carried out to reduce technical variations from experiments within the different files so that the remaining variance can be attributed to the underlying biology of the system under study. It is worth noting that the most common variations originated in the sample or the platform. Thus, the normalization of this variability is essential since the precision and accuracy of the analysis techniques in gene expression assays depend on it [213]. In this sense, normalization allows users to compare gene expression samples directly.

Samples include quality control flags: positive and negative control genes. The positive control linearity ensures that the samples maintain a certain linear relationship. Background correction is achieved with the use of negative control samples. A certain threshold is calculated as two standard deviations of the negative control values over the geometric mean of reference genes. A filtering process is carried out to filter less expressed genes. Those most stable reference (*housekeeping*) genes will be identified, using the algorithm *geNorm* [4] [213] .These genes will be used to calculate the scale factors for the rest of the sample. This way, we can calculate the specific normalization factors for each sample. As a result of this normalization and filtering process, the gene expression matrix is obtained (as illustrated Figure 3.2 B), which will be used in the data analysis and visualization processes.

[4] https://genorm.cmgg.be/

3.3.1.4 Gene expression data analysis

After the pre-processing step, the use of a series of gene expression data enables the possibility of exploring how different genes are connected (through gene interactions). Analyzing these interactions helps produce networks of interactions focused on Transcription Factors (TFs). In order to infer possible interactions between them, two distinguished algorithms are provided in the *arboretum* Python package [5], which are integrated into FIMED:

- **GENIE3** is a generic and straightforward algorithm based on feature selection with tree-based ensemble methods. It breaks down the prediction of a regulatory network involving p genes into p separate regression problems. The expression pattern of one of the genes (target gene) is predicted from the expression patterns of all the other genes (input genes) in each regression problem. An input gene's importance in predicting a target gene's expression pattern is interpreted as a possible regulatory connection. The network is then recreated by aggregating putative regulatory relationships across all genes to generate a ranking of interactions [214, 215].
- **GRNBoost2** uses Gradient Boosting Machine regression with early-stopping regularization to estimate regulatory networks. A tree-based regression model is trained for each gene in the dataset to predict its expression profile using the expression values of a collection of putative transcription factors (TFs). This algorithm is based on the GENIE3 architecture [216].

The gene inference analysis algorithms result in a collection of regulatory interactions between transcription factors and their target genes. An example can be observed in Figure 3.2 C, where the importance value is the strength of the interaction. The resulting links are then used to create and visualize the gene regulatory network, which can be plotted as network graphs with different layouts (see Figure 3.2 F).

3.3.1.5 Visualization

After the pre-processing, the most variable genes are used to perform a variety of visualizations to provide users with a rich set of tools for validating targets through comparing different patient samples or patients in the same disease stage. Hence, a set of analytic functionalities are used to discover patterns in the change in the gene expression levels. In concrete, FIMED provides three main visualizations, as represented in Figure 3.2 (D,E,F):

- Heatmaps. A heatmap is a graphical representation of a data matrix. The cell values are represented with different colours depending on their values. This is useful to discover relationships between elements visually.
- Cluster heatmaps. They follow the same principles as heatmaps but re-ordering the data matrix to aggregate those sub-matrices with similar values.
- Gene interaction network. The interactions between different genes can be presented as a graph where nodes represent genes and arc their interactions. The arcs can represent the strength of the interaction employing the arc shape or length.

3.3.2 Performance evaluation

Additionally, we evaluated the performance of FIMED via locust [6], an open source Python-based user load testing tool. We have used the locust tool to perform different data loads and operations,

[5] https://arboretum.readthedocs.io/en/latest/
[6] https://locust.io/

stressing the API of FIMED with a series of queries. The performance of the time response of FIMED has been studied for data insertion, deletion and retrieval. For this purpose, we have configured the locust API to simulate up to 5000 users interacting with the FIMED API with a spawn rate of 50 users (users spawned/second = 50). We have defined the behavior of the users in the locust API with Python code as follows: each user will make a POST request to create his/her form, and then (s)he will fill in that form with his/her patient's information, then (s)he will search for the patient and finally (s)he will delete it.

FIMED performs appropriately as the number of simultaneous users increases, with an average time response of 74 (ms), 81 (ms), 42 (ms) and 50 (ms), respectively, for each of the requests presented before. When we reach approximately 4500 users, the FIMED API performs moderately in some requests, and the time response grows speedily. Probably, this could be explained by the fact that FIMED is served through a standard Tomcat 9 Web application service, and more simultaneous requests are received than can be handled by the currently available request processing threads.

3.4 Use cases

In this section, a practical set of use cases are conducted with authentic expression data from metastatic Melanoma patients used in VIGLA-M [217]. This use case replicates this work from data acquisition to the integration and analysis based on advanced visualizations. In order to enable users to explore FIMED functionalities, an instance has been deployed on servers of the super-computing infrastructure of the Ada Byron Research building (University of Málaga). In this instance, users can freely manage their patient data or test it with sample data using the demo user provided[7] that contains anonymised patient data. This sample data enables new users to explore an example of how their databases could be developed. However, the user should create a new free account to use the tool. Thus, users will have independent workspaces, where each one can only access its own patients' data. After logging in, users can use different options on the main page (Figure 3.3).

Option **Form design** (Figure 3.3 A) allows the user to define new data fields to include any patient's clinical information. Thanks to the flexibility provided by MongoDB, the initial database schema can then be increased in a personalized way. Users can create dynamically new fields of any simple type *String, Number, Date, Boolean*. It is also possible to define compound fields with nested sub-fields, constituting a hierarchical organization. Once the user has designed the form (for inserting the patient's data), the new fields will be stored in the database as *Attributes* (Keys in JSON and MongoDB terminology), so the database scheme is incrementally designed. It is worth noting that users could adapt the database schema to any case of study in handling data in clinical trials.

At this point, the process of inserting a patient's information into the application is performed through option **Add patient(s)** (Figure 3.3 B), which enables a different kind of data to be stored in the database. First, patients' clinical information is introduced in the data fields previously declared on the form. The user interface dynamically extracts the data insertion schema from the database. This way, whenever a new field is added, it will automatically appear in the user interface. In addition, new fields are recommended to other users for future forms. Second, files containing gene expression assays associated with the patient can be loaded using the browse functionality. Accordingly, new meta-data fields could be added to the gene expression files to provide additional information to the samples. Depending on the file type, it could be used in different analyses. Third, it is also possible to insert additional files to guarantee the complete

[7] Demo user grants: username *"researcher"* and password *"demo"*

Figure 3.3: Main panel of the FIMED web application. The first main option is the (A) form design, which enables the user to create its own fields with corresponding attributes in the database. Then, (B) Add patient(s) option enables to store multiple sources of clinical data in the database. (C) Search patient(s) helps users to enhance the search process. Finally, (D) Gene expression level analysis option enables three principal analyses with gene expression levels: heatmaps, cluster heatmaps and gene regulatory networks.

patient information (e.g. doctor's reports, test results, scanned images, signed informed consent, etc.).

Similarly, FIMED provides the user with a search engine to help clinicians to retrieve their patient's information. The option **Search patient(s)** (Figure 3.3 C) offers a dynamic interface to facilitate this functionality, as well as filter options according to existing data fields to enhance the search process. The search tool provides access not only to view the data but also to modify them. In this sense, the patient's information can be updated anytime. In this operation, new fields could also be created when required.

A last main option comprises the data analysis and visualization, which is offered by clicking on **Gene expression level analysis** (Figure 3.3 D). As commented before, FIMED currently enables three principal analyses with gene expression levels: heatmaps, cluster heatmaps and gene regulatory networks. Depending on data availability, these analyses could be performed for one single patient and for aggregated data from several patients, hence allowing comparisons among different individuals, some of them acting as control samples.

Intending to show the potential of using FIMED, we have tested the tool in use cases conducted with actual sequence data from metastatic Melanoma patients [217]. Thus, we have validated the management and analytical functionalities generating indicative analysis and visualization in cancer research.

In these use cases, we have inserted clinical information of three Melanoma patients through the FIMED web service. Firstly, we designed the form for this clinical assay. Then we inserted the clinical information of the patients into the tool. This clinical case has a set of 5 simple fields and 1 composed field, as seen in Code Snippet 3.2.

Code Snippet 3.2: Data Schema in Melanoma use cases

```
{   "Form":
    {
        "_id":<ObjectId()>,
        "Attributes":
        {
            "Patient Code": <Number>,
            "Sex": <String>,
            "Birth Date": <Date>,
            "Blood pressure": <String>,
            "Observations": <String>,
            "Hospital admission":
            {
                "Hospital name": <String>,
                "Hospital address": <String>
            }
        }
    }
}
```

Moreover, we have used gene expression data using the Immune Profiling Panel Nanostring^{TM} (770 genes). This panel has been specifically designed for cancer projects studying the immune aspects of the disease. The panel includes 24 different immune cell types, common checkpoint inhibitors, CT antigens, and genes covering both the adaptive and innate immune responses. For this case, the analysis component works with RCC files[8], starting from the data normalization with a housekeeping based method. This platform can analyze 12 samples in each cartridge, providing 12 RCC files with the gene counts for each gene panel. These files are stored in FIMED associated with the patient's code, the sample collection date and the experiment date. FIMED has been

[8]See https://khaos.uma.es/fimedRCC for examples copied from https://github.com/hbc/sen-Nanostring and so licensed under MIT License.

shown to ease the process of flexible and dynamic collecting of patients' clinical information without needing a database re-engineering process.

3.4.1 Use case 1: Heatmap clustering

At this point, users can perform a first main analysis based on generating heatmaps and hierarchical clusters with dendrograms. As illustrated in Figure 3.4, users can select gene expression samples from one or more patients to constitute the gene expression dataset to be analyzed. In this process, a sliding element is provided to set a parameter for extracting only those most variable gene expression levels as a percentage of the total number of genes in the panel. Therefore, a series of different analyses can be generated in a given session according to this parameter. Thus, resulting clusters and heatmaps can be visually compared and inspected by a simple click-and-drag feature to zoom in and a click-once feature to zoom out.

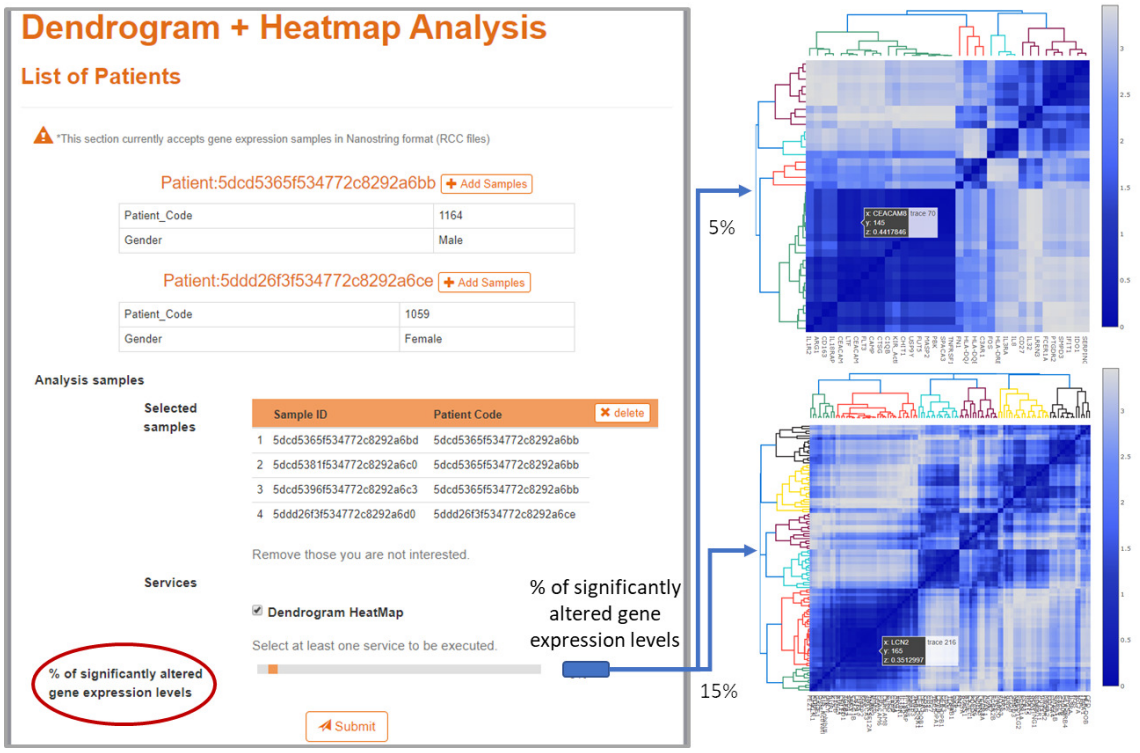


Figure 3.4: Selection panel of gene expression files and visualization of resulting Cluster heatmaps according to different percentages of significantly altered gene expression levels. In this example is observed the results for four samples (three from first patient and one from the second one) with two filtering percentages. Thus, the result on the right-up side shows a case with only the 5% of the most representative genes.

The long-term goal would be to identify unexpected relationships between genes expressed in a similar way that would help identify new drug targets or new biomarkers of the patient's expected evolution in their treatment. This is an ongoing work in collaboration with the regional hospital through the Biomedical Institute of Málaga (*Hospitales Univesitarios Regional y Virgen de la Victoria de Málaga, Instituto de Investigaciones Biomédicas - IBIMA*) using FIMED.

3.4.2 Use case 2: Reconstruction of gene regulatory networks

Another exciting analysis comprises the inference of gene regulatory networks, which can now be extracted from the gene expression levels previously stored in FIMED. A gene regulatory network consists of a set of genes (acting as transcription factors) that regulate (activate or inhibit) each other's expression. The nodes are the genes themselves, and their connections represent the regulatory mechanisms of their genetic expression. Two genes are connected if one regulates, positively or negatively, the expression of the other.

Figure 3.5 shows the selection panel offered in FIMED to generate and visualize gene regulatory networks. Similarly to the previous functionality, a sliding parameter is used to extract only those most variable gene expression levels as a percentage of the total number of genes in the panel. In addition, a statistical cutoff parameter is provided to limit the maximum number of links in the network, which is useful to enhance visualization, as it just centres on the most important genes and their relationships. Nonetheless, this interactive graph functionality allows the user to manually move the network and explore the connectivity between the nodes and hence, clearly inspect the network's topology. In this regard, users can select different layouts for network representation: Force-directed layout or Circular layout.

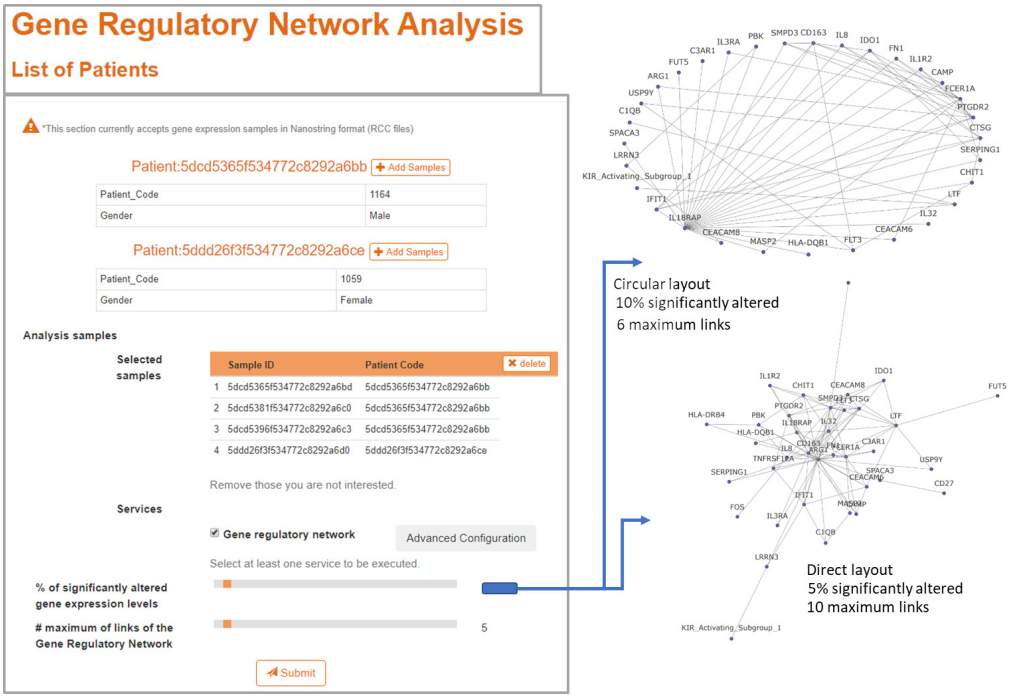


Figure 3.5: Selection panel of gene expression files and visualization of resulting gene regulatory networks according to different percentages of of significantly altered gene expression levels.

An interesting experiment consists in inferring a set of different networks, which are obtained using different random seeds, although using the same parameters of the percentage of the total number of genes in the panel to 5%, and the maximum number of allowed links to 10. This way, it is possible to discover those genes that, with a high frequency, are attractors of multiple links (interactions) with other genes. These attractors are then considered as hubs in the transcriptional regulatory network, which are usually identified to be used as diagnostic and prognostic markers and possibly for targeted therapy. In the case of the sample Melanoma data stored in FIMED, inferred

networks are frequently generated with hubs in genes: ARG1, IL18RAP, CD163, FCER1A, HLA-DQA1 and IL3RA. These genes are usually identified to have an adaptive resistance to immune regulatory factors in pathology [218], so this could support the clinician with new useful information for adjusting the treatment process.

3.5 FIMED 2.0

To exhibit the integration capabilities and flexibility to adapt to new tools and functionalities offered by FIMED, we have developed FIMED2.0 [33]. Our motivation for developing FIMED 2.0 stems from our experience with FIMED [30]. Our goal is to provide users with new functionalities to perform further and more accurate analyses. In this new version, we are interested in studying GRNs inference incorporating new algorithms for a principled comparison among gene network reconstructions. Also, an ensemble of GRNs inference techniques has been proposed based on a voting system to allow users to rank the most critical gene interactions (top-k genes/edges) between the similar outputs of a set of networks, so it can indicate the gene pairs most important in the regulatory process. Moreover, visualization tools have been added to this new version of FIMED to provide users with a deep insight into the networks through better graphic plotting. As a result, the primary goal would be to establish links between genes that are expressed similarly, which would lead to the discovery of novel therapeutic targets or biomarkers for the patient's expected treatment progression.

3.5.1 Architecture of FIMED 2.0

As commented above, FIMED 2.0 is an extension of FIMED. Hence, FIMED 2.0 implements the same workflow as FIMED (Figure 3.1), which consists of several phases: data collection, integration, clinical data analysis and visualization. However, FIMED 2.0 includes further gene regulatory network analysis and data visualization to annotate gene functionality and identify hub genes. This version allows the practitioner to use four different network construction methods: data assimilation, linear interpolation, tree-based ensemble or gradient boosting machine regression. Figure 3.6 summarizes this tool's architecture, emphasizing the new elements included in this extension. Originally FIMED integrated two distinguished algorithms (GENIE3 and GRNBoost2) as shown in 3.3.1.4, provided in the *arboretum* Python package [9]. However, this version of the tool includes two new gene inference algorithms that will provide the user with a broader comparison of the results to improve their analysis capacity. Moreover, FIMED 2.0 provides an ensemble gene regularity inference functionality that enables users to examine which algorithms produced similar reconstructions.

3.5.1.1 Gene regulatory network inference analysis

To summarize, FIMED 2.0 includes the following new functionalities for GRNs reconstruction:

- **PANDA** (Passing Attributes between Networks for Data Assimilation) is a message-passing model that integrates protein-protein interaction, gene expression, and sequence motif data to reconstruct genome-wide, condition-specific regulatory networks as a model. In this regard, the generated networks are more accurate than those constructed using individual data sets. Gene regulatory network generating with PANDA can also capture information about specific biological mechanisms and pathways that other methodologies had ignored [219].

[9] https://arboretum.readthedocs.io/en/latest/

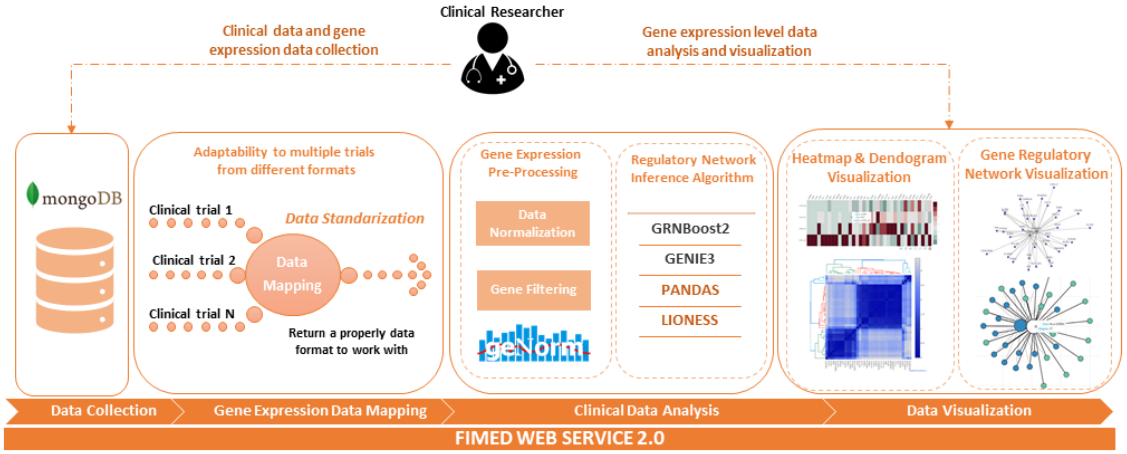


Figure 3.6: General workflow of FIMED 2.0. In this new version of FIMED, algorithmic functionalities for GRN reconstruction and new visualization tools have been added.

- **LIONESS** is a linear framework to relate a set of networks, each representing a different biological sample. The average of individual component networks reflecting the contributions of each member in the input sample set can be thought of as an "aggregate" network predicted from a collection of N samples [220].

- **Gene regulatory network ensemble**. For further analysis, an ensemble approach has been developed in this proposal as a gene regulatory network inference made from the four prior networks (GENIE3, GRNBoost2, PANDA, LIONESS). The ensemble approach has been designed since network inference algorithms are naturally noisy. It remains a challenge to identify whether these changes represent actual cellular responses or whether they emerged by random coincidence. In this sense, the ensemble internally develops a voting system to rank the top-k edges composed of similar outputs of a set of GRNs. Thus users can examine the top-k edges produced by similar reconstructions of the GRN algorithms.

3.5.1.2 Gene regulatory network inference visualization

One of the advantages of FIMED 2.0 is its power related to visualization features thanks to the availability of better graphic plotting, where users can interactively explore the constructed network. Many interactive visualizations allow users to actively move the network and examine the connections between nodes, allowing users to see the network's structure in detail. In this sense, users can choose from different network representation layouts: Circular layout or Force-directed layout, as shown in Figure 3.7. It is worth noting that FIMED 2.0 offers a new graph visualization in which clicking on a given gene will highlight this gene and its related neighbors and the information associated with them (Figure 3.7 C).

These rich visualization tools allow users to observe the most critical nodes representing genes and arcs representing interactions between them. By changing the arc form or length, the arcs can express the strength of the interaction. Users can compare different patient samples or patients at the same sickness stage.

Additionally, as mentioned before, an ensemble powerful visualization tool has been developed combining various GRNs models, where users can examine the top-k edges between gene interactions. Each similar reconstruction is represented in different edge colours. In this way, users can

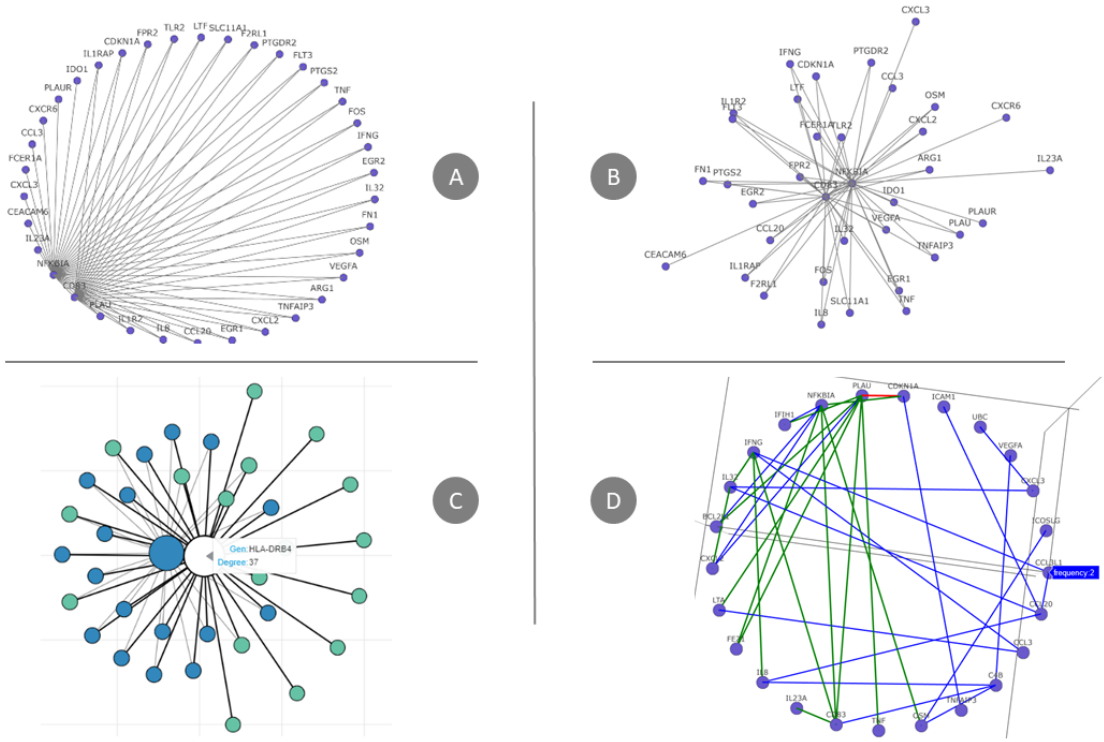


Figure 3.7: Gene regulatory network representations with different layouts: Circular layout or Force-directed, and dynamic plotting.

observe edges frequency in different colours to see the most important gene interaction in the ensemble voting system as depicted in Figure 3.7 D, where grey edge colour represents the frequency of 1, blue edge frequency of 2, green edge colour frequency of 3 and red edge colour frequency of 4.

3.5.2 Use case: Reconstruction of gene regulatory network

To show the potential of using the new functionalities of FIMED 2.0, the tool has been tested with real-world scenarios involving patients with advanced Melanoma [217], as in the previous version of FIMED. Thus, we have validated the new analytical functionalities and visualization techniques producing appropriate analysis and visualization in cancer research. For this proposal, we have used the FIMED 2.0 online interface to enter the clinical information of two Melanoma patients. In this sense, for this clinical trial, we have used a customized eCRF already designed for Melanoma use cases (Code Snippet 3.2). Then, the clinical information of the patients was entered into the tool. Thanks to MongoDB's flexibility, the primary database structure can be increased in a customized manner.

In addition, other files providing gene expression assays related to the patient have been loaded in FIMED 2.0. As a result, new meta-data fields in the gene expression files have been introduced to offer more information to the samples.

FIMED 2.0 has been deployed on our servers to enable users to explore the new functionalities, where users can manage their patient data or test it using sample data given by the demo user

provided [10]. This demo user includes anonymised patient data to allow new users to see an example of how their databases may be built. Users can also establish a new free account in which each user will have an independent workspace to design a particular database schema for their clinical trial.

As exposed above, new functionalities in terms of GRNs algorithms have been added to the tool, as well as new features in the visualization part that improves the ability of users to discover important gene-to-gene interactions and to inspect the topology of the network thanks to the availability of better graphic plotting.

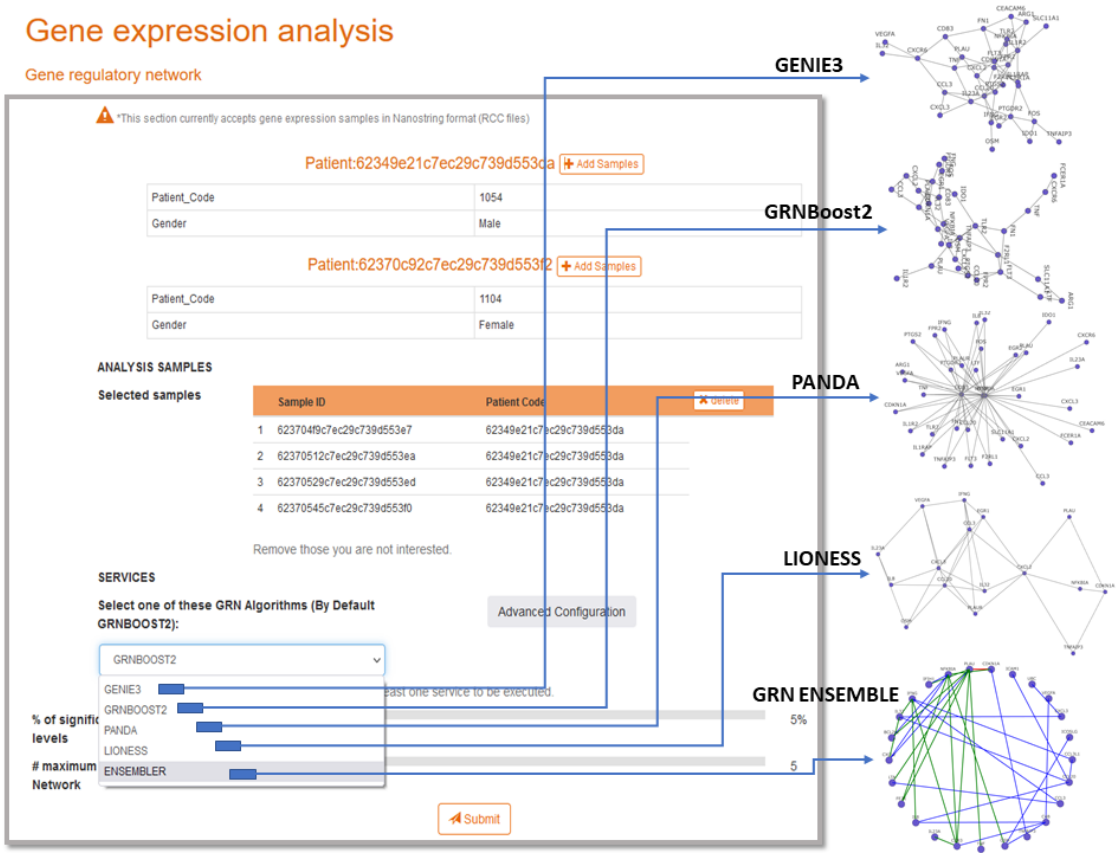


Figure 3.8: Selection panel of FIMED 2.0 that allow users to perform gene regulatory network analysis and visualizations from gene expression data.

In Figure 3.8, the selection panel of FIMED 2.0 allows users to perform gene regulatory network analysis and visualizations. Five GRNs visualizations have been performed, corresponding to each of the GRNs algorithms provided in FIMED 2.0. An experiment has been carried out, which consists in inferring a set of different networks (GENIE3, GRNBoost2, PANDA, LIONESS) and comparing the results of each of the networks. Besides, the ensemble algorithm based on a voting system has been executed to provide users with an entirely deep insight into the most important gene interactions coming from similar reconstructions of the GRNs algorithms.

Users will then be able to distinguish the frequency that each interaction between genes is repeated through similar GRNs outputs since each frequency is represented with a different edge

[10] Demo user grants: username *"iwbbio"* and password *"demo"*

colour (frequency 1: grey, frequency 2: blue, frequency 3: green, frequency 4: red). Therefore, the most important gene interactions (highest frequency) are represented with a red edge colour.

Furthermore, only the most variable gene expression levels as a fraction of the total number of genes in the panel can be extracted using a sliding parameter. A statistical cutoff parameter is also supplied to limit the maximum number of linkages in the network, which improves visualization because it focuses only on the most relevant genes and their interactions. It is worth mentioning that these features might provide clinicians with new information for improving treatment outcomes by allowing users to find genes and gene interactions that could be utilised as diagnostic and prognostic indicators and focused therapy.

3.5.3 Current status and implementation details

FIMED has been developed in *JAVA, JSP*, and *JavaScript* languages and follows a Model-View-Controller (MVC) software design pattern to manage the MongoDB database. The user interface is served through a standard Tomcat 9 Web application service. FIMED provides a user-friendly web application with all major browsers supported. The web interface has been designed to guide the user in the tasks of clinical data collection and database organisation transparently and straightforwardly. In addition, this tool provides gene expression data analysis by means of the visualization of clusters, anomalies, changes in patterns, etc., with open source libraries (*Plotly*[11] and *Bokeh*[12]). FIMED also provides an Open Source version for being installed on clinicians' servers to secure patients' information. This also enables the extension of FIMED by external developers providing new functionalities. Additionally, the Model-View-Controller (MVC) software design pattern has been implemented, employing an API connecting the web user interface with MongoDB. FIMED has an MIT license. Consequently, anyone can contribute. In its current status, it is entirely usable. However, it is certain that new features will be added in the near features, and some changes in the architecture will be foreseeable from the experiences we gain when using it and from the feedback of interested users.

FIMED is an active project that is in continuous development. In this sense, FIMED is now in its second version FIMED 2.0. It is released freely on the web for the community at `https://khaos.uma.es/fimedV2/`. The current version improves the previous version of FIMED regarding Gene Regulatory analysis inference. New distinguished GRNs algorithms (PANDA and LIONESS) have been integrated into the tool to provide users with better analysis capacities to increase their understanding of GRNs. Besides, an ensemble has been designed based on a voting system of a similar reconstructions network from a set of four GRNs algorithms (GENIE3, GRNBoost2, PANDA and LIONESS). Moreover, new visualization features have been added, guaranteeing users new ways of exploring gene networks to inspect the network's topology, thanks to better graphic plotting. All these new functionalities of the tool have also been tested in a use case conducted with real-world gene expression data from Melanoma cancer to find new biomarkers for predicting the patient evolution during the treatment, which results will be relevant for the medical community. These data have been stored in FIMED 2.0 so that users can explore the tool with a demo user *"iwbbio"* and password *"demo"*.

3.6 Discussion

Despite the availability of many eCRF tools designed to capture clinical trial data, most lack a flexible integration of clinical information since the clinicians cannot design and modify the forms according to their needs. Moreover, we have observed that most of these tools required a significant

[11] `https://plot.ly/`
[12] `https://bokeh.pydata.org/en/latest/`

time investment to create CRFs and a thorough study, making their use complicated for small-scale investigators.

Table 3.1: FIMED in comparison with other related systems according essential features shared in almost all systems encountered in the literature.

	OpenClinica [200]	REDCap [202]	TrialDB [199]	Phoenix [203]	Progmatic [201]	Dados-P. [204]	openCDMS [198]	PhOsCo [205]	FIMED [30]
1	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	✓	✓	✓	✓	✓	✓	✓	✓	✓
5	✓	✓	✓	✓	✓	✓	✓	✓	✓
6	✓	✓	-	✓	✓	-	✓	✓	✓
7	✓	✓	✓	✓	✓	✓	✓	✓	✓
8	✗	✗	-	✓	-	✓	✓	✓	✓
9	✗	✗	-	-	✗	✗	✗	-	✓
10	✓	-	-	-	✓	✓	✗	-	✓
11	✓	✓	-	✓	-	-	✓	-	✗
12	✗	✗	✗	✗	✗	✗	✗	✗	✓
13	✓	✗	✗	✗	✗	✗	-	✗	✓
14	✗	✗	✗	✗	✗	✗	✗	✗	✓
15	✗	✗	✗	✗	✗	✗	✗	✗	✓
16	✗	✗	✗	✗	✗	✗	✗	✗	✓

Before developing FIMED, we extensively studied general features in this kind of system. We outlined several of these essential features shared in almost all systems. In order to alleviate some of the limitations encountered in the literature, we present other features in the proposed tool that we have not observed in the systems found in the scientific literature. However, they are crucial for the collection, management and analysis of the clinical information of the study subjects. These features are presented below:

1. Enabling ways to secure patient's information.
2. Ensuring that retrieved data regarding each subject is only attributable to that subject.
3. Creating eCRF.
4. Providing support for several types of fields (such as dates, text, numerical values) and in various formats/ support for all basic field data types).
5. Supporting Web-based interfaces.
6. Providing software to be hosted locally to protect sensitive data.
7. Being Open source.
8. Providing user-friendly interfaces so that users can create CRFs and enter data directly on the interface.
9. CRFs should be easy to modify once created.
10. Should be able to contain non-traditional fields such clinical images, samples, etc.
11. Exporting to formats such as Excel, Pdf, Xml, Html, and CSV.
12. Dynamically storing the clinical data from multiple clinical trials.
13. Allowing to extend their functionalities.
14. Transferring data to different types of samples to target different analysis.

15. Using a database schema that grants enough adaptability to face the continuous changes in the practice of clinical trials.

16. Integration of analysis tools in order to examine the data to understand a disease.

Table 3.1 shows a comparison between FIMED and a set of related tools found in the literature, according to the list criteria above. As can be observed, desirable features related to dynamism in the integration phase, adaptability, scalability and advanced analytics are covered by FIMED, which represents an advantage with regard to these compared tools.

3.7 Conclusions

In this Chapter, FIMED has been detailed as a software tool for clinical data collections allowing clinicians without programming skills to manage clinical research information. It provides many functionalities in order to facilitate data management by clinicians, such as (I) personalized form design ("do-it-yourself") dynamically adapting to each of the patient's entries in the application; (II) browse functionality to store gene expression assays associated to the patient with metadata to grant additional information to the samples; (III) the modification and the update of the data over the time; and (IV) a search tool to provide direct access to the data with different filter options. Additionally, FIMED integrates analysis tools for clinical trials to allow clinicians to perform different types of analysis towards a deeper comprehension of the molecular mechanisms in a particular disease through interpreting results. Moreover, FIMED offers mechanisms to extend the software with new components to expand its functionalities.

FIMED incorporates gene expression analysis algorithms and offers visualization tools for exploring these data: Heatmaps, Cluster Heatmaps and Gene Regulatory Networks. FIMED has taken the experience acquired with the development of VIGLA-M [217] in the analysis of gene expression data and has been tested with use cases conducted with actual sequence data from metastatic Melanoma patients. This previous work also provided relevant needs from the clinical assay data management from the clinician point of view, as clinicians found limitations in improving or extending the data collected during the process. Its usability in this real scenario has been validated since we obtained our first real clinical insights. In this sense, it has been evident how this tool can be easily integrated into different use cases, making FIMED a robust clinical research tool for data management, analysis and visualization in clinical assays in different studied diseases. Apart from the public instance provided, the project can be deployed by IT administrators in any health information system, ensuring higher protection of the clinical data.

Chapter 4

Contribution to the reconstruction of gene regulatory networks with multi-objective optimization

In this chapter, sophisticated search methods are studied in the computational reconstruction of gene regulatory networks from gene expression data. Among these techniques, particle swarm optimization-based algorithms stand out as prominent techniques with fast convergence and accurate network inferences. A multi-objective approach for the inference of GRNs consists of optimizing a given network's topology while tuning the kinetic order parameters in an S-System, thus preventing unnecessary penalty weights and enabling the adoption of Pareto optimality based algorithms. In this study, we empirically assess the behavior of multi-objective particle swarm optimisers based on different archiving and leader selection strategies in the scope of the inference of GRNs. The main goal is to provide system biologists with experimental evidence about which optimization technique performs with higher success for the inference of consistent GRNs. The experiments involve time-series datasets of gene expression taken from the DREAM3/4 standard benchmarks and in vivo datasets from IRMA and Melanoma cancer samples. Our study shows that multi-objective particle swarm optimizer OMOPSO obtains the best overall performance. Inferred networks offer biological consistency under in vivo studies in the literature.

4.1 Introduction

In the last years, multiple optimization techniques such as evolutionary algorithms [153, 154, 155], and especially particle swarm optimization [156, 157, 158, 159, 160], have been applied to the inference of gene regulatory networks (defined in Chapter 2 of fundamentals) from gene expression time-series. The automatic inference GRNs is a complex problem found in computational biology [151], which consists in tuning parameters of a model that quantitatively reproduces the dynamics of a given biological system. In this regard, advanced computational models exist that are capable of inferring the topology of gene interactions to form networks. However, their precisions are strongly influenced by the quality of available datasets and the characteristics of the learning model used for such predictions. In this sense, S-System [161] framework can obtain a good trade-off between biological relevance and mathematical flexibility. It internally uses an Ordinary Differential Equations (ODE) system, which is a helpful framework to fit continuous variations of genetic regulations over time. Nevertheless, ODE systems require additional computational effort to tune parameters of kinetic orders and rate constants from a usually short amount of gene expression data. Moreover, as usually observed in biological systems, a sparse topology of the network should be accurately reproduced, so the early detection of significant node connections constitutes a major challenge in this process. To cope with these issues, the inference of GRNs has been traditionally tackled as a global optimization problem [162], which has demanded the use of specialized optimization techniques [154, 157, 158] to deal with it. In this problem, for the evaluation of the quality of solutions, a common strategy is to use aggregative fitness functions based on Mean Squared Error (MSE) [221, 222, 159] between inferred and observed (from data) gene expression values, which incorporate additional penalty terms based on sums of the magnitude of kinetic orders. A different approach to avoid weighting penalty values is to use a multi-objective formulation, which enables minimizing the MSE using S-System and a Topology Regularization (TR) value at the same time [153]. This way, optimizing kinetic order and rate constant parameters is possible while the topology of a given network is obtained.

Although some recent studies have intensified the use of particle swarm optimisers to deal with the inference of GRNs [157, 158], there is still a lack of proposals based on multi-objective formulations. Therefore, the primary motivation in this work is to apply and evaluate a representative set of multi-objective particle swarm optimisers (MOPSOs), which use different archiving strategies (hypervolume and aggregation) and, consequently, different strategies for the selection of leaders in the context of the inference of GRNs. The main goal is to obtain unbiased conclusions concerning which of them (and other related MOPSOs) could be used by experts in studies in silico/in vivo to find new possible gene interactions taking part in genetic regulations.

The rest of this chapter is organized in the following sections: Next section reviews related proposals in the literature. Mathematical models and methods are described in Section 4.3, where the multi-objective approach for the inference of GRNs is also detailed. Section 4.4 describes the algorithmic variants selected for evaluation. Section 4.5 reports the experimentation methodology and Section 4.6 analyses the obtained results. Finally, Section 4.7 reports conclusions and future lines of research.

4.2 Related works

The inference (or reconstruction) of GRNs has been traditionally dealt with different techniques, from the basic Boolean networks [223], to the continuous and stochastic models [224]. Recent studies have focused on adapted machine learning strategies such as non-parametric models based on decision trees [225] and recurrent neural networks [226].

Among continuous models for GRN representation, S-System [161] based on coupled ODE

provides mathematical flexibility enough to describe the reaction kinetics of the constituent parts and capture the biological dynamics system. However, it requires a large set of parameters to be tuned, which implies a certain limitation of this model when applied to large-size networks [154]. Even though S-System shows a remarkable ability to predict correct regulations, it finds difficulties when inferring the topology of the networks. To cope with this issue, similar studies [221, 222] have proposed using decoupled ODE frameworks to disaggregate internal equations that model genes without strong explicit interactions.

The optimal tuning of parameters in S-Systems is currently dealing with metaheuristics for continuous optimization such as: Genetic Algorithms (GAs) [162, 227, 155], Differential Evolution (DE) [228] and especially Particle Swarm Optimization (PSO) [221, 222, 229], due to the relative accurate behavior and fast convergence usually experimented by this last technique. In terms of multi-objective formulations, there exist similar approaches in the literature that tackle the inference of GRNs [227], [230], [231] and [232]. However, most of these proposals used single-objective aggregation functions for solution evaluation [230, 231], and they did not consider S-System models with mechanisms for topology regularization. Recently, in [153], a multi-objective cellular genetic algorithm was proposed to optimize parameters in S-System at the same time the topology of the network is regularized. This approach showed accurate performance for DREAM3 and small in silico networks. However, extensive comparisons and deeper experimentation with in vivo samples are still pending for testing such a multi-objective approach.

In this regard, many representational MOPSO variants have been recently applied with success to many different problems, such as: cost-based feature selection in biological classification [233]; variable-size cooperative co-evolutionary PSO for feature selection on high-dimensional data [234]; and bare-bones MOPSO environmental/economic dispatch [235], among others. However, the application of multi-objective particle swarm optimisers to the inference of GRNs is still an open issue. To the best of our knowledge, it has been partially approached in a past study [236] without using the S-System model and for just one synthetic dataset. In the present chapter, a thorough experimentation is conducted on a representative set of multi-objective PSO variants to evaluate their actual performance and usefulness in the context of standard in silico benchmarks, as well as for in vivo datasets.

4.3 Reconstruction of gene regulatory networks

The task of computationally reconstructing GRNs is aimed at offering mechanisms to capture the dynamics of biological systems from gene expression time-series datasets. A practical mathematical framework for modelling the dynamics of a network is the S-System [237], which consists of a set of differential equations, as modeled next:

$$f(t,X) = \begin{pmatrix} \alpha_1 \prod_{j=1}^{n+m} X_j^{g_{1j}} - \beta_1 \prod_{j=1}^{n+m} X_j^{h_{1j}} \\ \vdots \\ \alpha_n \prod_{j=1}^{n+m} X_j^{g_{nj}} - \beta_n \prod_{j=1}^{n+m} X_j^{h_{nj}} \end{pmatrix}, X(0) = X_0 \tag{4.1}$$

with X being an n-dimensional vector of elements, so the m-dimensional independent variables are expressed as $X_{n+j}, j = 1, \cdots, m$. In this biological model, X_i is the expression level of the i^{th} gene, parameters $\alpha_i, \beta_i \in \mathbb{R}_+^N$ are rate constants ($N = n + m$), and $g_{ij}, h_{ij} \in \mathbb{R}^{N \times N}$ are kinetic orders that regulate the synthesis and degradation of gene X_i influenced by X_j (usually called transcription factor).

In this problem, S-System models are commonly evaluated with Runge-Kutta numerical methods [222], since they successfully fit the model with the time series obtained from the gene expression values, which come from different experiments in separate periods. Nevertheless, using these methods entails updating $n+m$ variables, which usually implies a high computational cost. Moreover, optimization processes iteratively evaluate a set of candidate solutions, i. e., S-System parameters, in a given population and computing the Runge-Kutta method requires to solve recursively a set of equations in each solution evaluation. Therefore, in this approach, the computational cost increases along with the number of genes implied in the network.

A strategy to alleviate this extra effort is proposed in [238], which uses decoupling ODE systems based on data collocation. This way, it is possible to compute equations independently that refer to genes without featured interactions, reducing the computational time spent at each evaluation. In this method, dynamic variables X modeling genes (Equation 4.1) are calculated by a set of functions $X(t) = \sum_{j=0}^{N} x(j)\phi_j(t)$, where $x(j)$ is an expansion coefficient of $X(t)$ and $\phi_j(t)$ represents a set of polynomial shape functions. A linear Lagrange polynomial method is trained with time-series data of each gene of the target network. In the resulting S-System, each new iteration is computed as follows:

$$x_{n+1} = x_n + 0.5\eta(f[x_{n+1,exp}, \theta] + f[x_n, \theta]) \tag{4.2}$$

with $x_{n+1,exp}(t)$ indicating numerical values from (in vivo/in silico) experiments in gene expressions dataset at time t; $f[x_n, \theta]$ computes Equation 4.1 for x_n, and θ is the vector of tuning parameters $\{g_{ij}, h_{ij}, \alpha_i, \beta_i | i, j = 1 \cdots N\}$ in the S-System. In this equation, η is a smoothness rate that controls the approximation overshot.

Bi-objective problem formulation

In the proposed approach, each candidate solution comprises a vector of real-value variables corresponding to the parameters to be tuned: kinetic orders (g_{ij}, h_{ij}) and rate constants (α_i, β_i), in the S-System model. Figure 4.1 represents the structure of tuning parameters encoded within a solution vector.

For the evaluation of solutions, a common practice is to calculate the difference between the gene expression levels predicted with the S-System and the times-series of samples from the original dataset, i. e., the Mean Squared Error (MSE). This error measure was standardized by [239] to be used as fitness function for the optimal inference of GRNs.

$$f^{MSE} = \sum_{i=1}^{N} \sum_{k=1}^{M} \sum_{t=1}^{T} \left(\frac{X_{k,i}^{cal}(t) - X_{k,i}^{exp}(t)}{X_{k,i}^{exp}(t)} \right)^2 \tag{4.3}$$

Equation 4.3 calculates the MSE, in which $X_{k,i}^{cal}(t)$ and $X_{k,i}^{exp}(t)$ are the expression levels of gene i in the k^{th} set of time-series at time t in the *calculated* and *experimental* data, respectively. M is the number of time-series taken into account, and T is the number of samples (gene expression values) given in the experimental data (original dataset). The goal is to optimally tune parameters θ to minimize the error function f^{MSE}.

Although MSE has been shown to be a proper mechanism to evaluate the accuracy of S-System model with fitting time-series of gene expressions, it presents certain limitations to obtain the topology of those gene interactions that model the dynamics observed. It is worth noting that parameters in S-System model present a high degree of freedom, which implies the existence of numerous local minima in the solution search space that fit the time courses of gene expressions with low MSEs. This was experimentally observed in [162], where an efficient method showed overfitting when reproducing the time dynamics, although getting trapped on local minima and

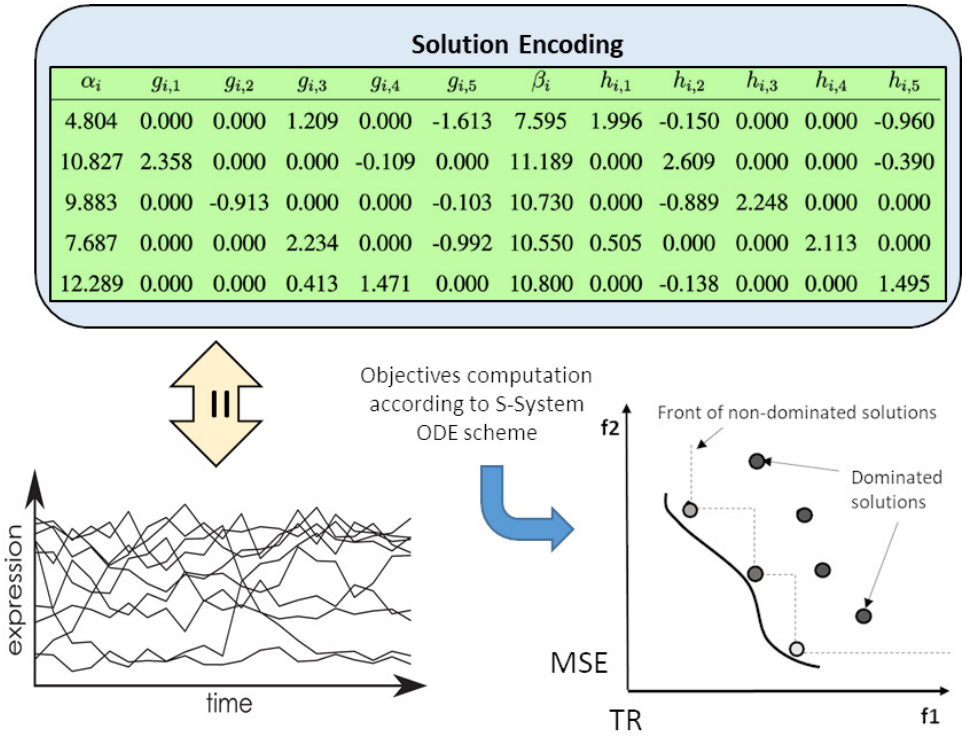


Figure 4.1: Graphical representation of solutions encoding/decoding withing the S-System setting process and the computation of the two objectives. The Pareto front approximation is constituted by using those non-dominated solutions obtained during the optimization process of MOPSOs.

failing to discover the true topology of the network. To deal with this issue, an additional pruning term for the fitness function was proposed in [228], which aims at capturing the topology of the network.

$$f_i = f_i^{MSE} + c \sum_{j=1}^{2N-I} (|K_{i,j}|) \tag{4.4}$$

Equation 4.4 aggregates function f_i^{MSE} (of Equation 4.3 for gene i) with the pruning term that comprises kinetic orders $K_{i,j}$ obtained from g_{ij} and h_{ij}. These values are joined and arranged in ascending order of their absolute values ($|K_{i,1}| \leq |K_{i,2}| \leq \cdots \leq |K_{i,2N}|$), so I is the maximum allowed cardinality degree of the network (i. e., maximum number of input/output edges of nodes). This second term includes a penalty constant c for weighting the formula. This way, an excessive cardinality in the network is penalized, which will provoke most of the gene interactions are removed when their corresponding kinetic values are low. This pruning ideally decreases false connections (false positives), while strengthening true interactions (true positives), hence promoting correct topologies from an early stage of the inference process. This strategy has been adopted in past studies where kinetic regulations are computed separately [240], although penalty term used in the present work (in Equation 4.4) was shown to obtain close to correct topologies with a higher success [228].

Nonetheless, it is worth noting that this aggregative approach still requires the use of weighting constants to provide a good balance between terms [222, 241]. This can be avoided with the use of a bi-objective Pareto dominance strategy, which enables optimization algorithms to deal with the two terms separately, even though guiding the search towards non-dominated solutions. These two terms are now adapted to be used as optimization objectives:

- (Obj. 1) f^{MSE}, to measure the prediction error of curve fitting according to kinetic and order parameters in the S-System (Equation 4.3);
- (Obj 2.) $f^{Topology} = \sum_{i=1}^{N} \sum_{j=1}^{2N-I} (|K_{i,j}|)$, to early detect the core topology commonly observed in biological networks.

An additional advantage of this strategy is to enable the use of multi-objective metaheuristics for the optimal reconstruction of GRNs, with the implicit benefits of having results in form of sets of non-dominated solutions, e.g., allowing decision makers to select solutions according to different values of time course estimations and network topologies, and promoting preference articulation for guiding the search. The use of multi-objective metaheuristics, such as evolutionary algorithms and particle swarm optimization, does not guarantee to find the optimal Pareto front, but an approximation to it. To perform accurate learning models, these techniques are developed with additional mechanisms and archiving strategies aimed at producing an accurate Pareto front approximation [242].

4.4 Evaluated multi-objective particle swarm optimization variants

The canonical Particle Swarm optimization [243] works by iteratively generating new particles positions located in a given problem search space. Each one of these new particles positions are calculated using the particle current position (solution), the particle previous velocity, and two main informant terms: the particle best previous location, and the best previous location of any of its neighbors.

Formally, in canonical PSO each particle's position vector $\mathbf{x}_i$ is updated each time step t by means of the Equation 4.5.

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \mathbf{v}_i^{t+1} \tag{4.5}$$

where $\mathbf{v}_i^{t+1}$ is the velocity vector of the particle given by

$$\mathbf{v}_i^{t+1} = \omega \mathbf{v}_i^t + U^t[0, \varphi_1] \cdot (\mathbf{p}_i^t - \mathbf{x}_i^t) + U^t[0, \varphi_2] \cdot (\mathbf{b}_i^t - \mathbf{x}_i^t) \tag{4.6}$$

In this formula, $\mathbf{p}_i^t$ is the personal best position the particle i has ever stored, $\mathbf{b}_i^t$ is the position found by the member of its neighborhood that has had the best performance so far. Acceleration coefficients φ_1 and φ_2 control the relative effect of the personal and social best particles, and U^t is a diagonal matrix with elements distributed in the interval $[0, \varphi_i]$, uniformly at random. Finally, $\omega \in (0, 1)$ is called the inertia weight and influences the tradeoff between exploitation and exploration.

To apply a PSO algorithm in multi-objective optimization, the previous scheme has to be modified to cope with the fact that the solution of a problem with multiple objectives is not a single one, but a set of non-dominated solutions. A pseudo-code of a general MOPSO is as shown in Algorithm 4.1. After initializing the swarm (Line 1), the typical approach is to use an external

archive to store the leaders, which are taken from the non-dominated particles in the swarm. After initializing the leaders archive (Line 2), some quality measure has to be calculated (Line 3) for all the leaders to select usually one leader for each particle of the swarm. In the main loop of the algorithm, the movement of each particle is performed after a leader has been selected (Lines 7-8) and, optionally, a *perturbation* operator can be applied (Line 9); then, the particle is evaluated and its corresponding personal best **p** is updated (Lines 10-11). After each iteration, the set of leaders is updated and the quality measure is calculated again (Lines 13-14). After the termination condition, the archive is returned to form the resulting Pareto front approximation.

Algorithm 4.1 Pseudocode of a general MOPSO algorithm.

```
 1: initialiseSwarm()
 2: initialiseLeadersArchive()
 3: determineLeadersQuality()
 4: generation = 0
 5: while generation < maxGenerations do
 6:    for each particle do
 7:       selectLeader()
 8:       updatePosition() // flight (Equations 4.5 and 4.6)
 9:       perturbation()
10:       evaluation()
11:       updateLocalBest()
12:    end for
13:    updateLeadersArchive()
14:    determineLeadersQuality()
15:    generation ++
16: end while
17: returnArchive()
```

Nevertheless, a number of issues have to be considered as discussed in [244] to design a MOPSO variant, which are related to: the existence of an external archive of non-dominated solutions, the selection strategy of non-dominated solutions as leaders for guiding the swarm, the velocity calculation, the neighborhood topology of particles, and the usual existence or not of a mutation (perturbation) operator. The way of dealing with these and other mechanisms lead into new different algorithmic variants, some of the most representative ones are used in this study to evaluate their performance on the inference of GRNs. A description of them is as follows:

- MOPSO [37]. It is one of the earliest multi-objective PSO algorithms that uses a secondary (i.e., external) repository of particles, which is later used by other particles to guide their own movements. The main aim of the external repository (or archive) in MOPSO is to keep a historical record of the non-dominated solutions found along the search process. This external repository consists of two main parts: the archive controller and the grid. The function of the archive controller is to decide whether a certain solution should be added to the archive or not, in accordance with an elitist decision procedure. The basic idea of the grid-based archiving method is to store all the solutions that are non-dominated with respect to the contents of the archive; the proposed strategy is that the objective function space is divided into regions, so that if the individual inserted into the external population lies outside the current bounds of the grid, then the grid has to be recalculated and each individual within it has to be relocated. The adaptive grid is really a space formed by hypercubes, each one with as many components as objective functions. Hypercubes can be interpreted as geographical regions that contain a number of individuals. When the archive is full, a solution from the most populated hypercube is removed, therefore promoting diversity in the Pareto front

approximation. In addition, MOPSO incorporates a special mutation operator that enriches its exploratory capabilities.

- OMOPSO [36]. It is an optimized version of the previous MOPSO that includes the use of the crowding distance of NSGA-II [245] for discarding leader solutions from a bounded external archive and two mutation mechanisms to promote a good convergence. The original OMOPSO also uses an $\epsilon-dominance$ strategy for bounding the amount of particles generated at each iteration. Nevertheless, the OMOPSO variant considered in this study allocates solutions dropped by the $\epsilon - dominance$ in an external archive of leaders. To compute a new particle, a leader is selected from this archive to be used as the best particle term in the velocity calculation. Binary tournament selection is performed in this operation based on the crowding values of the members included in this archive of leaders, which is bounded to the size of the swarm. At each iteration, this archive is updated, so those leader particles with larger crowding distances are removed from the archive when it is full.

- SMPSO [39]. This algorithm is inspired by OMOPSO, and is characterized by two main features: a velocity constraint mechanism and (similarly to MOPSO and OMOPSO) an external bounded archive to store the non-dominated solutions found during the search. A perturbation mechanism, implemented as a mutation operator, is also incorporated. The archive contains the current Pareto front approximation found by the algorithm, and it applies the crowding distance density estimator [245] to select particles to remove when it exceeds the maximum allowable size. For velocity calculation, a leader is selected from the archive by following binary tournament or random strategies. The solution located in the less crowded region of the Pareto front composed by all archiving solutions is then selected. The local best particle, to be used as the second term in the velocity calculation, is selected from the entire swarm by employing a dominance test, so the current best particle is replaced if it is dominated by the new one.

- MOPSOHv. In [41], a study of different leader selection mechanisms on SMPSO was conducted. In that work, the most salient variant based on the degree of contributions of particles in the hypervolume indicator I_{HV} [246] to organize the external archive of leaders, instead of using the crowding distance. The underpinning idea is to select a leader from those particles that contribute with high hypervolume to the Pareto front approximation. Therefore, the external archive of leaders contains solutions with higher values of I_{HV} indicator. The archive is managed as done in OMOPSO, although following an update strategy centered on low hypervolume values, instead of small crowding distance. In the velocity calculation phase, two solutions are randomly selected from this archive, and the one contributing with higher hypervolume is selected as the leader.

- DMOPSO [40]. It is an archive-less approach inspired by the MOEA/D [247] aggregative model, according to which, a multi-objective optimization problem is decomposed into a set of single-objective ones that are optimized at the same time. In this sense, it is worth noting that this variant partially follows the general algorithmic scheme of the pseudocode in Algorithm 4.1 (operations involving leaders' archive are avoided), but it just performs the core iteration scheme of particles movement. Therefore, in DMOPSO a set of vectors uniformly distributed $\lambda_1, \lambda_2, \ldots, \lambda_N$ are defined, with N being the swarm size. In this scheme, each particle $\mathbf{x}_i$ is associated with a vector λ_i and a neighborhood, which is defined as a set of its nearest weight vectors in $\lambda_1, \lambda_2, \ldots, \lambda_N$. Then, following the Tchebycheff scheme a scalarizing strategy is applied with the ideal reference point $z^* = (z_1^*, \ldots, z_k^*)^T$, where $z_i^* = min_{x \in S} f_i(x)$ for $i = 1, \ldots, k$ as Minimize $f^{Tch}(\mathbf{x} \mid \lambda, z^*) = \max_{i=1,\cdots,N}\{\lambda_i \cdot \mid z_i^* -$

$f_i(\mathbf{x})$ |}. Each element of the reference point (z^*) is specified by the minimum[1] value of each objective $f_i(\mathbf{x})$ among the examined solutions throughout the optimization process of DMOPSO. For velocity calculation, the local best particle term is obtained by following a similar strategy to MOEA/D when updating a neighborhood. To update the leader, the best solution in the neighborhood is selected by taking into account scalar values with regards to their corresponding weight vectors.

- VEPSO [38]. Vector Evaluated Particle Swarm optimization is a multi–swarm variant of PSO, which is inspired by the Vector Evaluated Genetic Algorithm (VEGA) [248]. VEPSO uses two or more swarms in an island-based distributed topology to probe the search space and information is exchanged among them. Each swarm is evaluated using only one of the objective functions of the problem under consideration, and the information it possesses for this objective function is communicated to the other swarms through the exchange of their best experience. The best position attained by each particle (the particle's memory) separately, as well as the best among these positions are the main guidance mechanisms of the swarm. Exchanging this information among swarms lead to Pareto approximation points. We have used a variant of VEPSO which includes, as SMPSO and OMOPSO, an external archive to store the non-dominated solutions found during the search.

In terms of computational complexity, archive-based variants (MOPSO, OMOPSO, SMPSO, MOPSOHv, and VEPSO) follow similar algorithmic scheme to NSGA-II and canonical MOPSO, which are shown to have $\mathcal{O}(M(N+n)^2)$ in the worst case, as explained in [249]. According to this work, M is the number of functions to be optimized, while N and n are the size of the archive and swarm, respectively. In the case of archive-less variants, DMOPSO follows similar scheme to MOEA/D, which shows complexity of $\mathcal{O}(MnT)$, with T (the neighbour size) representing the number of solutions that compute the scalarizing strategy each iteration.

In summary, from the initial MOPSO with external archive, different variants have been appearing with improvements consisting in: crowding based leader selection (OMOPSO), velocity constraint (SMPSO), hypervolume contribution leader organization (MOPSOHv) and archive-less based on an aggregative decomposition scheme (DMOPSO). Contemporary to MOPSO, the VEPSO variant is also considered as it implements a totally different scheme based on distributed swarm-islands and a common archive. For the integration of these versions, we have adapted the implementation provided in the jMetal 5.1 [250] framework[2] to deal with the inference of GRNs.

4.5 Experimentation

In order to assess the performance of the selected MOPSOs on the inference of GRNs, we have followed a standard procedure that comprises both, in silico and in vivo time-series from gene expression data of different organisms. In particular, we focus on the DREAM3 and DREAM4 challenges[3] as the in silico data, and we use a cell cycle regulatory subnetwork in Saccharomyces cerevisiae (IRMA) and Melanoma patiens' samples as the in vivo biological datasets. Gene expression data from Melanoma patients can be found in FIMED 2.0[4] (Chapter 3), where GRNs have also been performed.

DREAM3 in silico challenge [42] is nowadays a standard benchmark for GRNs reconstruction, which consists of gene expression datasets from two organisms: *E.coli* (*Escherichia coli*) and *Yeast*

[1] Without loss of generality, we assume minimization for algorithmic definitions.
[2] Online Available at URL `http://jmetal.sourceforge.net/`
[3] Online Available at URL `http://dreamchallenges.org`
[4] Online Available at URL `https://khaos.uma.es/fimedRCC` for examples copied from `https://github.com/hbc/sen-Nanostring` and so licensed under MIT License.

(*Saccharomyces cerevisiae*), with two dimensions according to the number of genes taken into account: 10 and 100, for each of them. 10 size networks involve 4 time-series with 21 samples per gene, while 100 size networks comprise 46 time-series of 21 samples. The true topology of the networks are obtained from in vivo GRNs of *Escherichia coli* and *Saccharomyces cerevisiae*, which present different patterns of density and topology. Target graphs in DREAM3 Challenge are directed and they do not distinguish between inhibitors or inductors. Similarly, DREAM4 in silico challenge consists of networks of sizes 10 and 100. The size 10 network data consists of 5 simulated networks, each of which contains 21 time points and 5 replicates. DREAM4 also provides 5 networks with size 100 genes with 21 time points and 10 replicates each one. Melanoma samples have been taken from the Immune Profiling Panel Nanostring^{TM} (770 genes), which have been specifically designed for cancer projects studying immune aspects of the disease. The panel includes 24 different immune cell types, common checkpoint inhibitors, CT antigens, and genes covering both, the adaptive and innate immune response. Moreover, this platform can analyze 12 samples in each cartridge, so it provides 12 RCC files with the gene counts for each of the gene panels.

4.5.1 Methodology

The methodology followed in experiments consists in running each combination of algorithm and GRN instance 25 independent times. To measure the performance of algorithms, we have considered the standard quality indicator Inverted Generational Distance Plus (I_{IGD+}) [251], since it measures convergence and diversity degrees of the resulting front approximations. This way, given a set of non-dominated solutions R, used as a reference front, and another set of non-dominated solutions A, the inverted generational distance of A, $IGD(A)$, is the average Euclidean distance form each point of R to the nearest solution in A. The modified IGD, IGD+, replaces the Euclidean distance by Equation 4.7, where $r = (r_1, r_2, ...r_m)$ is a reference point, $a = (a_1, a_2, ...am)$ is a solution, and m is the number of objectives.

$$I_{IGD+}(a, r) = \sqrt{\sum_{i=1}^{m}(max(r_i - a_i, 0)^2)} \tag{4.7}$$

Therefore, for all the distributions of results, we compute the median and interquartile range of the I_{IGD+} values. In this sense, taking into account that the optimal reconstruction GRNs is a real-world optimization problem for which true Pareto fronts (required to calculate the I_{IGD+}) are not available, a reference front is calculated for each instance. This reference front is computed by joining all the non-dominated solutions obtained by all the MOPSO variants, thorough all their executions.

To deploy all the experiments, a super-computing platform summing up 63 TFLOP/s is used, which hardware is managed by a Slurm middleware acting as the distributed task scheduler. This infrastructure takes part in the Picasso Supercomputer (RES node) located in the Bio-Innovation Building of the University of Málaga, as mentioned in Chapter 2.5.

Table 4.1 contains the algorithmic configurations used in experiments, which comprise a similar setting for common parameters. The size of the swarm is 100 and the stopping condition is reached when 100,000 function evaluations are performed. The archive size, when applicable, is set to 100. MOPSO, MOPSOHv and SMPSO use the polynomial mutation with distribution index $\eta_m = 20$, which is applied with probability $1/L$, where L is the number of problem variables. For OMOPSO and DMOPSO, the acceleration coefficients C_1 and C_2 are randomly (uniformly) set in a range of (1.5, 2.0) and the inertia weight is also randomly set in a range of (0.1, 0.5). In the case of SMPSO, acceleration coefficients C_1 and C_2 are randomly (uniformly) set in a range of (1.5, 2.5). VEPSO

algorithm use constriction factor χ of Clerc [252] instead of inertia weight, which was analytically set to 0.729.

It is worth noting that a systematic parameter tuning of algorithms is needed before performing empirical comparisons, although it requires time extra effort and appropriate background knowledge of the problem to be properly conducted. There exists a variety of methods developed for this purpose, which offer significant performance [253, 254, 255, 256]. In the present study, a partial grid-search tuning of the specific algorithm parameters have been conducted to this end on smaller DREAM 3/4 instances; for the common parameters, they have been set to common values to make a fair comparison as shown in Table 4.1.

Table 4.1: Parameter settings.

Common parameters	
Swarm size	100 Particles
Maximum number of evaluations	100,000
MOPSO [37] & MOPSOHv	
Archive Size	100
C_1, C_2	1.5
ω	0.4
Mutation	*Polynomial*
Mutation probability	$1/L$
Mutation distribution index η_m	20
Selection method	Rounds
Archive selection ratio for **g**	0.2
Archive selection ratio for **p**	0.98
OMOPSO [36]	
Archive size	100
C_1, C_2	$rand(1.5, 2.0)$
ω	$rand(0.1, 0.5)$
Mutation	uniform+non-uniform
Mutation probability	1/3 of the swarm

Common parameters	
Swarm size	100 Particles
Maximum number of evaluations	100,000
SMPSO [257]	
Archive Size	100
C_1, C_2	$rand(1.5, 2.5)$
ω	0.1
Mutation	*Polynomial mutation*
Mutation probability	$1/L$
Mutation distribution index η_m	20
Selection method	Rounds
DMOPSO [40]	
Scalarizing function	Tchebycheff
C_1, C_2	$rand(1.5, 2.0)$
ω	$rand(0.1, 0.5)$
VEPSO [38]	
Archive Size	100
Number of swarm-islands	2
C_1, C_2	2.05
Constriction factor χ	0.729

4.6 Results and analysis

In this section, results and analysis are presented from three different perspectives: first, a performance comparison of the evaluated algorithms is conducted; second, the capacity of the algorithms to obtain high quality solutions in terms of the inferred in silico and in vivo networks is analyzed, with regards to current results in the specialized literature; finally, an analysis of results with a real-world gene expression dataset of Melanoma cancer is carried out, for which significant networks are inferred and assessed in terms of biological validation.

4.6.1 Algorithmic performance

As aforementioned, the Inverted Generational Distance plus (I_{IDG+}) is used as indicator to compare the performance of the multi-objective algorithms we have selected. This metric measures the dominance-based distance from each reference point to its nearest solution in the objective space, so low values of I_{IDG+} mean good performance in terms of diversity and convergence.

The first set of results are shown in Table 4.2 that contains the median and interquartile range of the distributions of I_{IDG+} values (out of 25 independent runs), for the DREAM3 and DREAM4 instances with size 10 and the six compared algorithms. As we can observe, SMPSO obtains the best median values (with dark grey background) for five network instances and the second-best median for 4 instances (with light grey background). OMOPSO shows the best median values for

three instances, followed by MOPSO with two best median values. These three algorithms use similar multi-objective strategies with external archives, although with different selection mechanisms of non-dominated solutions to act as leader particles. This last probably determines the different performances of these algorithms from DREAM3/4 instances with 10 genes network size.

Table 4.2: Median and interquartile range of I_{IDG+} for each algorithm and instance with 10 genes size. Best and second best median results have dark and light gray backgrounds, respectively.

	MOPSO	MOPSOHv	SMPSO	OMOPSO	DMOPSO	VEPSO
10-Ecoli1	$6.08e-02_{5.6e-02}$	$1.65e-01_{8.4e+01}$	$7.59e-02_{6.4e-02}$	$7.13e-02_{6.3e-02}$	$2.04e-01_{4.4e-02}$	$2.39e+01_{2.6e+01}$
10-Ecoli2	$1.61e-01_{1.7e-01}$	$3.65e-01_{4.1e-01}$	$1.53e-01_{1.6e-01}$	$1.54e-01_{1.4e-01}$	$4.95e-01_{1.1e-01}$	$3.18e+00_{2.4e+00}$
10-Yeast1	$2.44e-01_{1.7e-01}$	$3.20e-01_{3.4e-01}$	$2.39e-01_{1.4e-01}$	$2.05e-01_{1.4e-01}$	$4.53e-01_{1.2e-01}$	$2.98e+01_{4.1e+01}$
10-Yeast2	$4.83e-01_{2.8e-01}$	$9.66e-01_{2.7e+02}$	$4.42e-01_{4.3e-01}$	$5.33e-01_{3.2e-01}$	$1.13e+00_{7.3e-01}$	$1.66e+02_{1.6e+02}$
10-Yeast3	$2.62e-01_{8.1e-01}$	$6.94e-01_{5.0e+01}$	$2.56e-01_{2.8e-01}$	$2.21e-01_{1.7e-01}$	$2.12e+00_{1.2e+00}$	$7.79e+01_{1.4e+02}$
10-Net-1	$3.71e-01_{4.9e-01}$	$7.15e-01_{1.0e+00}$	$5.67e-01_{3.8e-01}$	$5.73e-01_{5.3e-01}$	$1.75e+01_{4.6e+01}$	$2.21e+02_{1.8e+02}$
10-Net-2	$4.62e-01_{1.4e+00}$	$1.37e+00_{1.8e+01}$	$3.02e-01_{2.4e-01}$	$5.88e-01_{7.4e-01}$	$1.73e+01_{3.0e+01}$	$5.00e+02_{4.9e+02}$
10-Net-3	$2.16e-02_{1.2e-02}$	$2.98e-02_{2.8e-02}$	$2.04e-02_{6.1e-03}$	$2.08e-02_{2.0e-02}$	$2.40e-01_{8.3e-02}$	$5.94e+00_{8.7e+00}$
10-Net-4	$2.25e-02_{5.7e-02}$	$2.03e-01_{6.7e-01}$	$2.22e-02_{2.7e-02}$	$4.33e-02_{7.0e-02}$	$1.85e+00_{3.0e+00}$	$5.47e+01_{3.8e+01}$
10-Net-5	$5.36e-01_{4.5e-01}$	$7.54e-01_{1.2e+00}$	$4.04e-01_{3.9e-01}$	$3.12e-01_{4.0e-01}$	$6.23e-01_{1.1e-01}$	$3.90e+00_{1.3e+01}$

Table 4.3 shows the results of compared algorithms in terms of I_{IDG+} for large-size networks with 100 genes of DREAM3 and DREAM4 benchmarks. In this case, a different behavior is observed so that OMOPSO obtains the best median values for almost all the instances, while DMOPSO shows the best results in just one network of DREAM4 (Net-5) and presents the higher number of second-best results. The remaining variants show moderate performance.

Table 4.3: Median and interquartile range of I_{IDG+} for each algorithm and instance of DREAM3 and DREAM4 with 100 genes size. Best and second best median results have dark and light gray backgrounds, respectively.

	MOPSO	MOPSOHv	SMPSO	OMOPSO	DMOPSO	VEPSO
100-Ecoli1	$3.17e+03_{4.5e+02}$	$7.48e+02_{1.1e+03}$	$1.03e+03_{7.2e+02}$	$8.84e-01_{2.3e+00}$	$3.11e+00_{1.6e+00}$	$5.94e+00_{8.0e+00}$
100-Ecoli2	$2.65e+01_{4.6e+00}$	$4.59e+00_{2.5e+00}$	$8.34e+00_{4.9e+00}$	$2.57e-03_{6.6e-03}$	$1.05e-02_{3.6e-03}$	$2.06e-02_{3.3e-02}$
100-Yeast1	$4.86e+02_{6.5e+01}$	$1.27e+02_{7.0e+01}$	$1.91e+02_{1.4e+02}$	$1.21e-01_{4.3e-01}$	$8.06e-01_{9.7e-01}$	$1.56e+00_{4.0e+00}$
100-Yeast2	$1.31e+03_{2.1e+02}$	$2.29e+02_{1.6e+02}$	$5.34e+02_{3.9e+02}$	$1.14e+00_{3.0e+00}$	$5.92e+00_{3.8e+00}$	$1.70e+02_{1.7e+02}$
100-Yeast3	$4.69e+02_{4.4e+01}$	$9.60e+01_{5.4e+01}$	$1.65e+02_{9.0e+01}$	$3.21e-02_{9.4e-02}$	$1.06e-01_{2.0e-02}$	$8.54e-01_{9.6e-01}$
100-Net-1	$2.77e+02_{1.7e+02}$	$1.80e+01_{8.4e+00}$	$4.14e+01_{3.0e+01}$	$1.40e-01_{2.9e-01}$	$4.24e-01_{3.2e-01}$	$1.39e+01_{1.7e+01}$
100-Net-2	$3.69e+03_{1.2e+03}$	$4.17e+01_{3.2e+01}$	$1.45e+03_{1.2e+03}$	$3.04e-01_{4.9e-01}$	$5.19e+00_{2.1e+00}$	$6.29e+01_{1.4e+03}$
100-Net-3	$2.77e+02_{1.7e+02}$	$1.80e+01_{8.4e+00}$	$4.14e+01_{3.0e+01}$	$1.40e-01_{2.9e-01}$	$4.24e-01_{3.2e-01}$	$1.39e+01_{1.7e+01}$
100-Net-4	$2.39e+03_{2.1e+03}$	$4.67e+01_{6.3e+02}$	$8.70e+02_{3.8e+02}$	$3.15e-01_{4.5e-01}$	$7.91e-01_{6.3e-01}$	$2.07e+02_{8.3e+02}$
100-Net-5	$1.80e+02_{2.9e+01}$	$4.11e+01_{1.6e+01}$	$8.63e+01_{4.1e+01}$	$2.16e-02_{1.3e-01}$	$6.99e-03_{1.6e-02}$	$4.31e+01_{5.6e+01}$

In order to provide these results with statistical confidence (in this study p-value = 0.05), we have assessed the entire distributions of the indicator used in this in study with non-parametric statistical tests, because in several cases the distributions of results did not follow the conditions of normality and homoscedasticity [258] required for parametric tests. In particular, Friedman's ranking and Holm's post-hoc tests have been applied to distinguish those algorithms statistically worse than the control one (the best-ranked according to Friedman).

This way, as shown in Table 4.4 and focusing on for small size instances with 10 genes, SMPSO is the best-ranked variant according to Friedman test and it is followed by OMOPSO and MOPSO. Therefore, SMPSO is established as the control algorithm in the post-hoc Holm tests, which is compared with the rest of algorithms. The adjusted p-values (indicated as $Holm's_{Adj-p}$ in Table 4.4) resulting from these comparisons are, for algorithms OMOPSO and MOPSO, higher than the confidence level (0.05), so this means that no statistical difference can be observed with regards to SMPSO. Conversely, for the remaining variants MOPSOHv and VEPSO and DMOPSO, the adjusted p-values are lower than the confidence level, meaning that SMPSO performs statistically better than these algorithms in the context of DREAM3/4 with 10 genes size. OMOPSO and MOPSO obtained similar overall performances, although showing OMOPSO better ranking than MOPSO in terms of I_{IDG+}.

I_{IDG+} Size 10 genes		
Algorithm	$Friedman's_{Rank}$	$Holm's_{Adj-p}$
***SMPSO**	**1.60**	-
OMOPSO	2.10	8.06e-01
MOPSO	2.30	8.05e-01
MOPSOHv	4.10	8.40e-03
DMOPSO	4.91	3.20e-04
VEPSO	5.98	7.24e-07

I_{IDG+} Size 100 genes		
Algorithm	$Friedman's_{Rank}$	$Holm's_{Adj-p}$
***OMOPSO**	**1.09**	-
DMOPSO	1.90	3.38e-01
VEPSO	3.30	1.71e-02
MOPSOHv	3.69	5.65e-03
SMPSO	5.00	1.19e-05
MOPSO	5.98	2.24e-08

Table 4.4: Average Friedman's rankings with Holm's Adjusted p-values (0.05) of compared algorithms for the test set of DREAM3 and DREAM4 instances with 10 and 100 genes size. Symbol * indicates the control algorithm and column at right contains the overall ranking of positions with regards to I_{IDG+}.

When focusing on large-size instances with 100 genes in Table 4.4 (right), the $Friedman's_{Rank}$ values indicate that OMOPSO is the best-ranked algorithm (control variant), but without statistical differences with regards to DMOPSO according to the Holm's adjusted p-values. However, VEPSO, MOPSOHv, MOPSO and SMPSO perform statistically worse than the control variant (OMOPSO), as they obtained adjusted p-values lower than the confidence level (0.05).

In this regard, an interesting observation can be pointed out in terms of problem scalability, since for two algorithmic variants, SMPSO and MOPSO, they show prominent behavior for small size networks, although with poor ranking values when facing large-size networks with 100 genes. Conversely, DMOPSO performs properly in 100 genes size instances, but with moderate results on size 10. The use of a search strategy based on problem decomposition without external archive (of non-dominated solutions) could delay the convergence in DMOPSO in comparison with the other variants when facing small networks. Nevertheless, this seems to be in turn beneficial in the context of large scale solution vectors, which leads this algorithm to obtain a high ranking in terms of I_{IDG+}. In the case of OMOPSO, it shows accurate results not only on small size networks but also for large-size ones. This prominent behavior has also been experimented in a previous study [259], where several different multi-objective techniques were assessed in terms of scalability on benchmarking problems. Apart from using a similar archiving strategy to SMPSO and MOPSO, the use of restriction factors and grid selection mechanism in these to last variants seem to be responsible for a good performance in small size networks but limiting their search procedure in the context of large-size ones. This is avoided when working with OMOPSO, which leads us to suggest the use of this variant in the context of essays where the number of genes can variate from tens to hundreds.

From a graphical point of view, Figures 4.3 and 4.2 plot the reference fronts computed from all executions (in a continuous line) according to the contribution of each technique, for 10 genes size networks of DREAM3 and DREAM4 challenges, respectively. In general, SMPSO contributes with non-dominated solutions to the reference front in all the network instances, although its contribution is especially high in the specific case of Yeast3 network of DREAM3 and Net1, Net3 and Net4 of DREAM4. Other reference fronts in DREAM3 are better covered with solutions of MOPSO and OMOPSO, which ideally behave especially adapted to these specific instances. Probably, the archiving and replacement mechanism in SMPSO and OMOPSO of non-dominated solutions based on crowding distance density estimator makes these variants keep solutions in these reference fronts, whereas other methods like DMOPSO and MOPSOHv discard them prematurely.

Another interesting observation in these fronts (Figures 4.3 and 4.2) lies in the number of non-dominated solutions with different values of topology regularization terms (TR), which also show low mean errors (MSEs) with regards to the gene expression time-series. This would support human

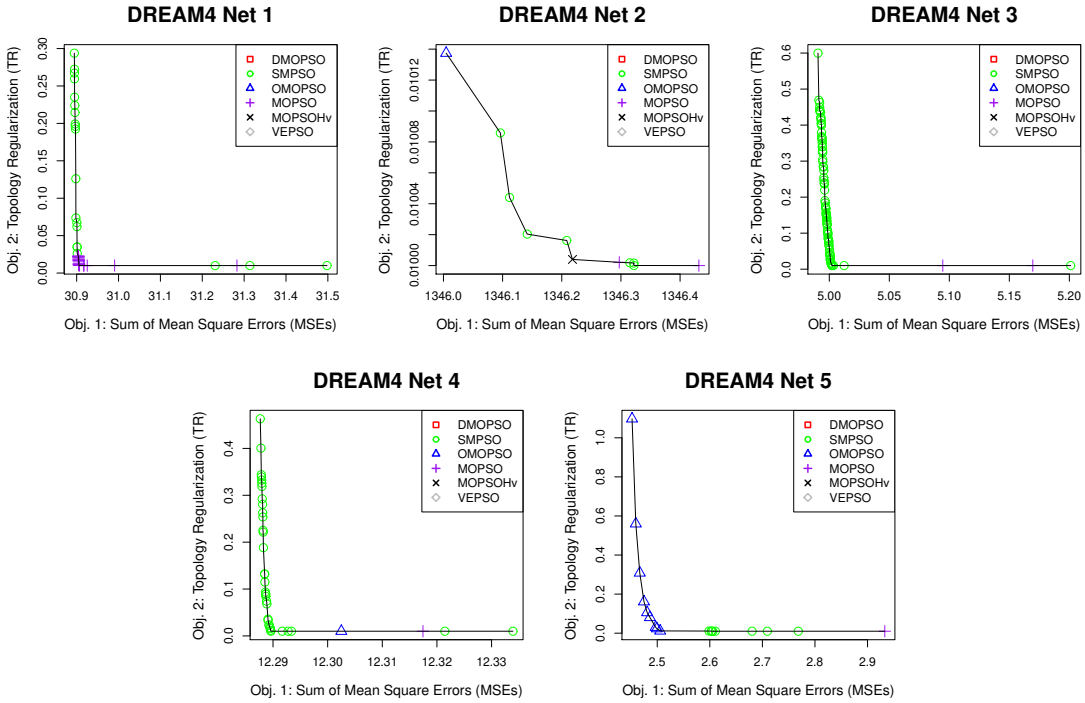


Figure 4.2: Reference Fronts with best I_{HV} values on DREAM 4 datasets.

experts in the decision-making process to select alternative networks with different topologies, but low errors, i.e., with a high ability to reproduce gene expression data.

In this sense, it can be checked the particular ability of SMPSO to obtain non-dominated solutions in the region of the reference front with low error values in Ecoli2, Yeast3, Net1, Net3, and Net4. In contrast with the other compared algorithms, OMOPSO can cover this area for Ecoli1, Yeast1 and Net5, while MOPSO performs successfully for Ecoli1 and Yeast2.

In summary, OMOPSO shows the overall best behavior for DREAM3/4 challenges. SMPSO obtains good results on small size networks, whereas DMOPSO performs accurately on large-size ones. These results are reported in terms of multi-objective standard indicator I_{IDG+} with regards to pre-computed reference fronts. Nevertheless, a more in-depth analysis is also required from the point of view of the quality of the inferred networks concerning gold-standard solutions and specialized literature. A thorough analysis in this sense is conducted next.

4.6.2 Quality of inferred (in silico) networks

To measure the quality of the networks inferred from the resulting solutions, we have followed two standard metrics as suggested in DREAM3 and DREAM4 challenges: the area under the Receiver Operating Characteristic (ROC) curve (AUROC), as well as the area under the precision-recall curve (AUPR). These values are computed with regards to gold-standard networks from Ecoli and Yeast [42]. According to this, to compute the ROC curve, the true positive rate is plotted against the false positive rate, TPR and FPR in Equations 4.8 and 4.9, respectively.

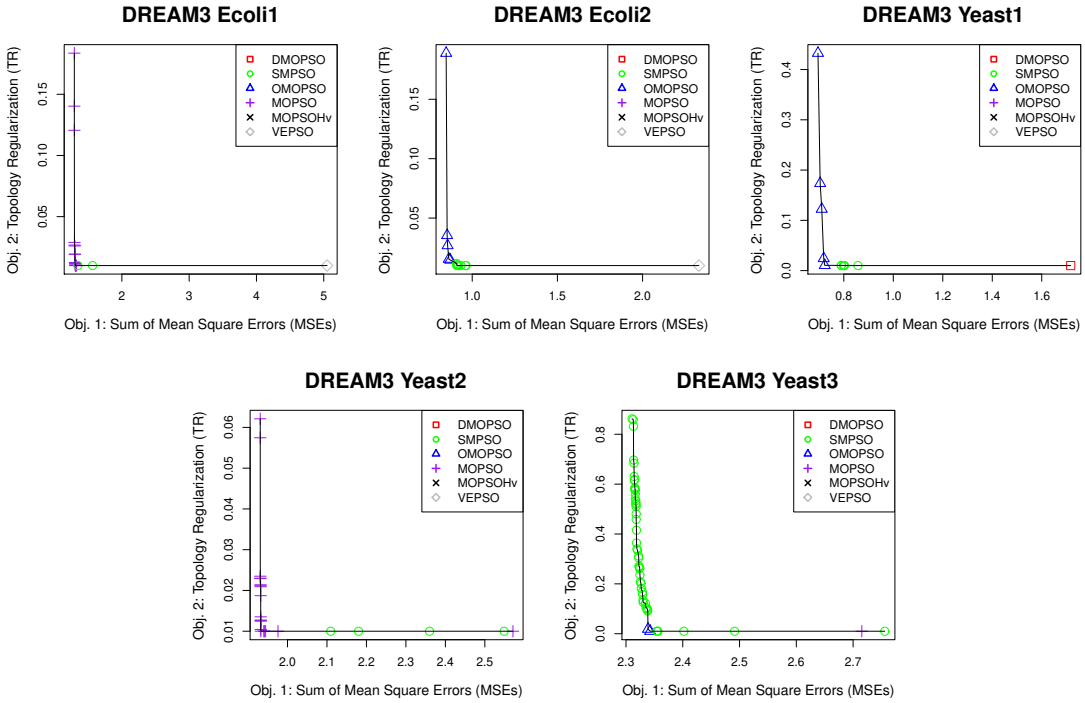


Figure 4.3: Reference Fronts with best I_{HV} values on DREAM 3 Challenge datasets.

$$TPR = \frac{TP}{TP + FN} \tag{4.8}$$

$$FPR = \frac{FP}{FP + TN} \tag{4.9}$$

$$PPV = \frac{TP}{TP + FP} \tag{4.10}$$

In these equations, TP are True Positives, FN False Negatives, FP False Positives, and TN True Negatives. The predictive performance of a network predictor is then quantified in terms of the area under the ROC curve (AUROC), which is in the range $[0, 1]$. In the case of precision-recall curve (AUPR), it plots the positive predictive value (PPV) against the true positive rate (TPR), which are computed with Equations 4.10 and 4.8, respectively. Therefore, an ideal network predictor will obtain an AUPR of 1.

The results obtained from inferred networks by all the MOPSO variants for DREAM3 and DREAM4 with size 10 are given in Tables 4.5 and 4.6, respectively. Similarly, the results for networks size 100 genes are computed in Table 4.7 (DREAM3) and Table 4.8 (DREAM4). The corresponding solutions are given from those resulting fronts with the best hypervolume, for each algorithm and instance.

In these tables, a base-line linear least squares regression method (LASSO) and those teams in DREAM3 Challenge that used the same datasets, Teams 236 and 190, are also incorporated for comparisons. This same challenge was also used to evaluate MONET [153], which performs a

Table 4.5: AUROC and AUPR for LASSO, Team 236, Team 190 (DREAM3 challenge) and MOPSO variants on DREAM3 size-10 networks.

	Ecoli1		Ecoli2		Yeast1		Yeast2		Yeast3	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
LASSO	0.500	0.119	0.547	0.531	0.528	0.244	0.527	0.305	0.582	0.255
Team 236	0.621	0.197	0.650	0.378	0.646	0.194	0.438	0.236	0.488	0.239
Team 190	0.573	0.152	0.515	0.181	0.631	0.167	0.577	0.371	0.603	0.373
MONET	0.647	0.182	0.513	0.200	0.801	0.469	0.522	0.354	0.612	0.321
MOPSO	0.582	0.252	0.523	0.212	0.811	0.471	0.512	0.332	0.623	0.324
MOPSOHv	0.578	0.345	0.604	0.277	0.633	0.177	0.454	0.241	0.554	0.289
SMPSO	0.668	0.181	0.552	0.207	0.686	0.227	0.591	0.375	0.541	0.373
OMOPSO	0.525	0.150	0.572	0.275	0.605	0.313	0.513	0.320	0.524	0.304
DMOPSO	0.623	0.337	0.456	0.204	0.613	0.223	0.603	0.418	0.558	0.321
VEPSO	0.453	0.134	0.530	0.236	0.426	0.094	0.599	0.438	0.403	0.208

Table 4.6: AUROC and AUPR for LASSO, (DREAM4 challenge) and MOPSO variants on DREAM4 size-10 networks.

	Net1		Net2		Net3		Net4		Net5	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
LASSO	0.601	0.260	0.543	0.331	0.553	0.274	0.569	0.285	0.522	0.215
MOPSO	0.645	0.262	0.506	0.175	0.518	0.184	0.627	0.275	0.541	0.149
MOPSOHv	0.471	0.153	0.596	0.298	0.588	0.220	0.642	0.206	0.581	0.160
SMPSO	0.609	0.376	0.460	0.155	0.455	0.147	0.525	0.236	0.535	0.230
OMOPSO	0.735	0.505	0.538	0.213	0.536	0.276	0.595	0.271	0.513	0.157
DMOPSO	0.537	0.208	0.554	0.367	0.547	0.353	0.653	0.281	0.448	0.116
VEPSO	0.712	0.489	0.614	0.281	0.617	0.220	0.521	0.143	0.578	0.175

similar multi-objective strategy as used in this study, so their available results are also incorporated in Tables 4.5 and 4.7.

Table 4.7: AUROC and AUPR for LASSO, Team 236 (DREAM3 challenge) and MOPSO variants on DREAM3 size-100 networks.

	Ecoli1		Ecoli2		Yeast1		Yeast2		Yeast3	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
LASSO	0.519	0.016	0.512	0.057	0.507	0.016	0.530	0.044	0.506	0.044
Team 236	0.527	0.019	0.546	0.042	0.532	0.035	0.508	0.046	0.508	0.065
MONET	0.525	0.014	0.533	0.012	0.522	0.018	0.485	0.038	0.517	0.058
MOPSO	0.502	0.018	0.501	0.015	0.525	0.018	0.492	0.038	0.516	0.061
MOPSOHv	0.486	0.018	0.487	0.014	0.534	0.019	0.499	0.041	0.486	0.059
SMPSO	0.529	0.015	0.495	0.019	0.496	0.017	0.513	0.049	0.488	0.058
OMOPSO	0.504	0.051	0.512	0.014	0.518	0.017	0.506	0.054	0.519	0.070
DMOPSO	0.485	0.012	0.513	0.017	0.514	0.017	0.510	0.040	0.503	0.055
VEPSO	0.452	0.013	0.487	0.017	0.497	0.016	0.546	0.051	0.502	0.061

The AUROC and AUPR metrics in Tables 4.5 and 4.6 indicate that MOPSO variants are in general competitive in comparison with the other techniques for 10 genes size networks. In particular, the evaluated algorithms obtained solutions with higher precision-recall (AUPR) for all the instances except for Ecoli2 and Yeast3, although with similar results to the best ones. It is worthy to note that current proposals in the literature use ad hoc operators based on a priori knowledge about the structure of the target network, then performing efficiently for these specific instances. An example of this strategy can be adopted for Ecoli2 network, which induces a star topological structure with central nodes acting as hubs for many linking edges coming from regulated genes [42]. This feature can be used to set thresholds and clusters in algorithmic operators

Table 4.8: AUROC and AUPR for LASSO (DREAM4 challenge) and MOPSO variants on DREAM4 size-100 networks.

	Net1		Net2		Net3		Net4		Net5	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
LASSO	0.510	0.016	0.482	0.027	0.507	0.016	0.430	0.024	0.486	0.023
MOPSO	0.518	0.018	0.478	0.022	0.508	0.021	0.500	0.022	0.478	0.024
MOPSOHv	0.480	0.017	0.512	0.025	0.514	0.021	0.526	0.023	0.511	0.023
SMPSO	0.510	0.023	0.526	0.029	0.508	0.021	0.517	0.021	0.488	0.019
OMOPSO	0.499	0.018	0.493	0.024	0.502	0.023	0.510	0.022	0.492	0.018
DMOPSO	0.490	0.018	0.503	0.025	0.536	0.022	0.503	0.021	0.493	0.018
VEPSO	0.520	0.020	0.507	0.025	0.526	0.021	0.476	0.020	0.523	0.022

to improve the inference power of such techniques for the specific target network, although with poor behavior when facing a different one.

In the case of MOPSO variants, network topology and strength of interaction edges are obtained from solutions encoding parameters of S-System, so implicit knowledge about GRN reconstruction is directly modeled in objectives as black-box functions. This allows researchers to use these kinds of multi-objective optimisers with generic operators, which is an advantage, since actual in vivo experiments are featured with no a priori information about the structure of the network. In this sense, there exist other related approaches in the literature like [155], which proposed a genetic algorithm hybridized with random forest using fuzzy cognitive maps for the reconstruction of GRN. This hybrid proposal was evaluated on DREAM3 instances obtaining AUROC of 0.509 and AUPR of 0.352 for Yeast2 10-size, which are in the range of MOPSO results and they are overtaken by those of SMPSO, DMOPSO and VEPSO.

Another interesting observation can be made on some variants with moderate results in terms of I_{IDG+} for 10 size networks, e. g., VEPSO and DMOSO, but with high quality solutions for some instances, especially from DREAM4 (Table 4.6). This can be explained by the fact that AUROC and AUPR are calculated from the best solutions in each experiment, although indicators are computed from the entire distribution of results and the median values are highlighted in these cases. In the specific case of VEPSO, a high value of AUROC is observed for some instances, but with low AUPR. This can be due by the existence of numerous false positives in the networks inferred by VEPSO, so the final score in this variant is limited. This is avoided in the other MOPSO variants, which show in general a good trade-off between AUROC and AUPR.

In large instances with 100 genes size, results are in general similar to those computed for size 10 and practically all the MOPSO variants obtain competitive solutions with regards to the base-line methods. Concretely, Table 4.7 shows that certain variants such as OMOPSO and SMPSO obtain outperforming networks in comparison with those of Team 236 (Team 190 was not able to submit predictions for DREAM3 size 100). In this table, for all the algorithms, AUROC results still show high precision for size-100 networks, although AUPR values deteriorate with recall in general lower than 7% (see OMOPSO in Yeast3). This last issue may be due by the low density usually observed on 100 size networks (DREAM3), for which the amount of disconnected nodes (genes with no interactions) is higher than the number of inferred edges. Additional comparisons involve the hybrid technique proposed in [155] which obtained AUROC of 0.508 and AUPR of 0.044 for specific instance Yeas2 with 100 genes (only these results were reported). For this network, multi-objective particle swarm optimisers SMPSO, DMOPSO and VEPSO obtained solutions with better AUROC and AUPR values than this compared approach. In the light of all these and previous results, we can suggest that MOPSO variants evaluated in this study show competitive behavior with regards to base-line solutions of DREAM3/4, as well as to current related algorithmic proposals in the state of the art.

4.6.3 Results on the IRMA (in vivo) network

To further test MOPSO variants, we have also used gene expression datasets from the *In vivo Reverse Engineering and Modeling Assessment* (IRMA) network [260]. In this network, the gene expression levels were measured using quantitative RT-PCR at different time points with the yeast Saccharomyces cerevisiae. IRMA instance comprises 5 genes (CBF1, GAL4, SWI5, GAL80, and ASH1) and 6 regulatory interactions among them. In addition, it can be switched on/off by culturing cells in galactose or glucose, respectively. This network is broadly used in many studies [153, 151, 155, 261] and constitutes a current gold-standard on experiments oriented to the reconstruction of GRNs.

Table 4.9 shows the AUPR values of the GRNs inferred by a set of prominent algorithmic proposals that were experimentally assessed with the IRMA network (AUROC values are not available). Among these algorithms, results of BGRMI and Jump3 are given from [151], and results of KFLR, CMI2NI, TIGRESS and GENIRF are given from [261]. In this comparison, BGRMI reports the best AUPR value for the Switch-On instance, yet with a close recall to that of SMPSO. Conversely, in the case of Switch-Off dataset, OMOPSO variant obtains the best AUPR values, followed by DMOPSO, SMPSO, MOPSOHv and LASSO. These results suggest MOPSO variants are highly competitive, not only for in silico datasets, but also for in vivo sample networks.

Table 4.9: AUPR performances on IRMA network.

Algorithm	Switch-On	Switch-Off
LASSO	0.520	0.734
Jump3	0.685	0.682
BGRMI	0.904	0.574
KFLR	0.896	0.721
CMI2NI	0.721	0.456
TIGRESS	0.714	0.452
GENIRF	0.672	0.327
MONET	0.827	0.734
MOPSO	0.502	0.533
MOPSOHv	0.702	0.746
SMPSO	0.838	0.756
OMOPSO	0.702	0.953
DMOPSO	0.502	0.783
VEPSO	0.625	0.688

4.6.4 Biological validation

Finally, a series of experiments are conducted to evaluate the MOPSO variants studied here in terms of biomedical validation, with regards to real-world gene expression datasets. In concrete, we focus on a dataset previously collected in FIMED [30] from actual clinical information of Melanoma cancer patients, including gene expression levels obtained from the NanoString[5] platform comprising the Immune Profiling Panel. This panel was curated and subjected to specific filtering techniques as commented in Chapter 3 (Gene Expression Pre-processing 3.3.1.3), so that a subset of 35 least stable genes, i.e., the ones that display most variation, was isolated to be worked in further analyses.

In a first analysis, the MOPSO variants have been run (using the parameter setting of Table 4.1) using the instance dataset of Melanoma cancer, so a series of different networks are inferred (for the

[5]Online Available in URL `https://www.nanostring.com/`

first time) that curiously share a similar structure. In this sense, we executed a Heatmap Clustering analysis in FIMED (Figure 4.4) to illustrates the frequency of repeated edges between interacting genes in the resulting networks. In this graphic, transcription factors (origins) are located in rows and target genes (destinations) in columns. As can be observed, there is a number of interaction edges with frequencies from 3 to 17 repetitions in overlapping networks. Therefore, using these edges as recurrent patterns in most of the inferences from the MOPSO variants it is possible to construct a reference network, which comprises strong interactions of genes. This reference network is plotted in Figure 4.5, which shows certain edges in green to represent frequencies higher than 10 and edges in red to represent more than 17 repetitions.

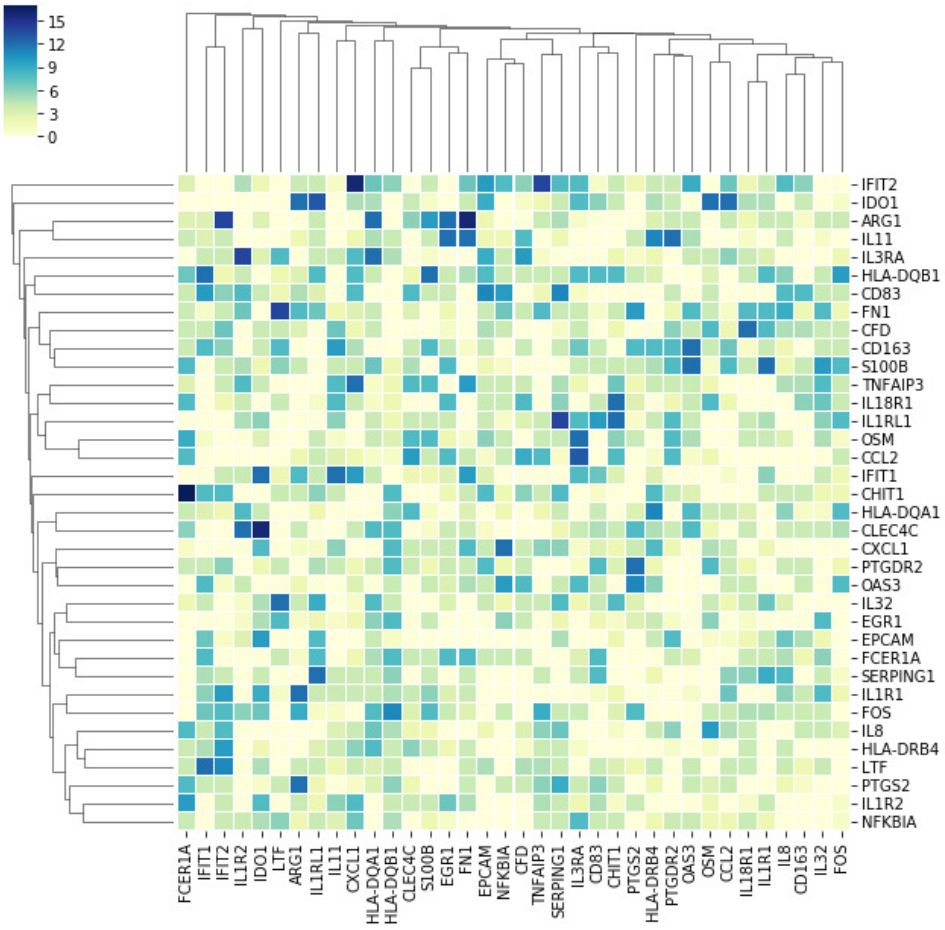


Figure 4.4: Melanoma heatmap histogram of overlapping edges in networks inferred by MOPSO variants. Target factors are located in rows and target genes in columns.

It is worth noting that practically all of the edges in this reference network are also obtained by the different algorithmic proposals used in Chapter 3 in our FIMED [30] tool for the inference of GRN, namely: GRNBoost2 and GENIE3. This is an indicative result for the validation of the inferred network, as different techniques in the literature implementing heterogeneous learning models obtain overlapping networks, but with simple variations in edges. Concretely, edges "HLA-DQA1 → IL3RA", "IDO1 → IFT1" and "ARG1 → IFIT2" are repeatedly obtained by all the

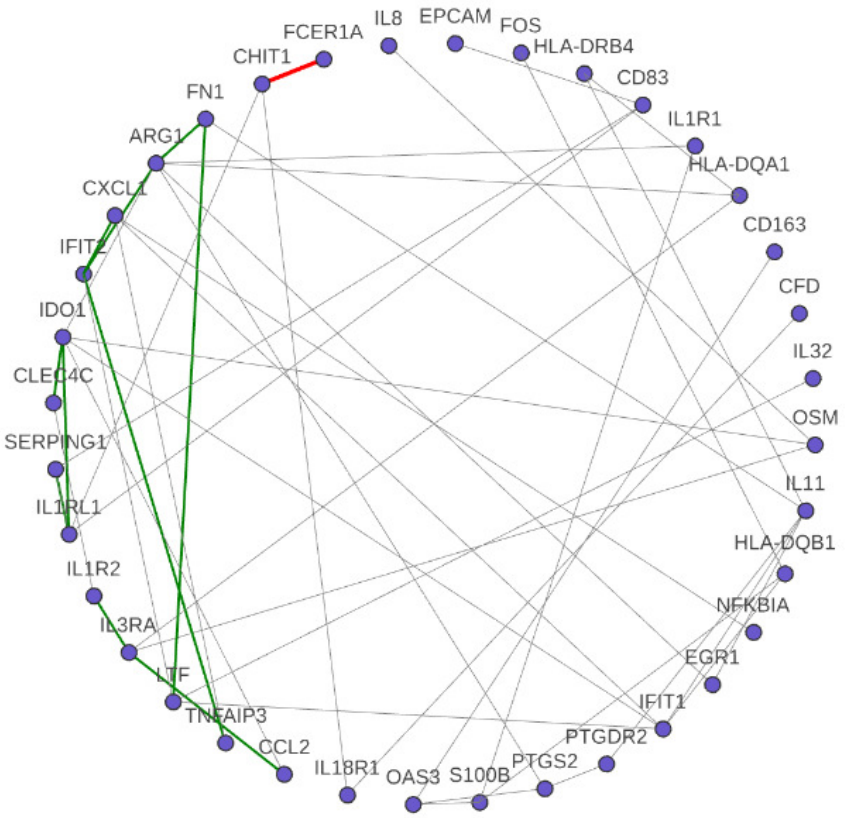


Figure 4.5: Reference network inferred by MOPSO variants studied here. Edges in green represent those with frequencies higher than 10 and edges in red are used for frequencies higher than 17.

MOPSO variants, GRNBoost2 and GENIE3, which lead us to suggest these edges would correspond to significant transcriptions of genes in the actual interaction network of Melanoma cancer.

Moreover, when checking these genes in GO (the Gene Ontology[6]) and specialized literature we are able to collect the biological terms that are related to the genes of our sample in the search of biological pathways in common. For this, Gene Set Enrichment Analysis (GSEA) for interpreting gene expression data in Melanoma cancer will be carried out. It is a validation technique based on a previous biological knowledge [262]. To bring about the analysis, GOOATOOLS (a python library for Gene Ontology analysis [263]) has been used. To observe significant pathway terms involved in the Gene Ontology for our gene samples, Fisher's exact test, as well as multiple corrections has been used as a statistical method. We select Bonferroni, Sidak, Holm and Fdr methods, since deliver rigorous results indicated in the literature [264], and p-value$\leq$ 0.05, to find the enriched terms in the Gene Ontology.

As a result, there have been 26 GO terms in total that found significant p-value$\leq$ 0.05 (enriched) in this analysis. Most of the GO terms related in biological processes are associated with antigen processing and presentation of peptide or polysaccharide antigen via MHC class II (GO:0002504), antigen processing and presentation of peptide antigen (GO:0048002), antigen processing and presentation of peptide antigen via MHC class II (GO:0002495), etc. These biological

[6]Online Available in URL http://geneontology.org/

pathways provide us a lot of information, since we know that the constitutive expression of major histocompatibility complex (MHC) class II molecules frequently occurs in Melanoma disease [265]. This biological process is connected largely with HLA-DQB1, HLA-DQA1, HLA-DRB4 genes that are directly related to melanoma as discussed in [266].

4.7 Conclusions

In this Chapter, an empirical evaluation is conducted on a set of representative multi-objective particle swarm optimisers, based on different movement, leader selection, and archiving strategies, when optimizing the inference of gene regulatory networks. To this end, we use a multi-objective problem formulation by optimizing the topology of a given network, while tuning the kinetic orders of an S-System model, yet avoiding the weight of additional penalty terms. The experiments conducted involve time-series datasets of gene expression taken from the DREAM3/4 standard benchmarks, as well as in vivo datasets from IRMA and Melanoma cancer samples. After thorough experimentation, a series of conclusions are obtained as follows:

- The study reveals that OMOPSO shows in general the best performance for in silico and in vivo instances. SMPSO behaves properly in small size networks, so the velocity restriction mechanism implemented in this variant seems to be responsible of a low I_{IGD+} for a number of instances. Conversely, DMOPSO is more adapted to large instances. Probably, the archive-less decomposition strategy in this last variant enhances the time-series fitting for large scale parameter vectors in solutions. All this leads us to suggest the use of OMOPSO in the context of essays where the number of genes can variate from tens to hundreds.

- MOPSO, MOPSOHv and VEPSO show limited behavior in terms of algorithmic performance. Nevertheless, they usually obtain high quality inferred networks, so for several instances: Ecoli1, Yeast1, Yeast3, Net1, Net4 and Net5, these algorithms obtained accurate AUROC and AUPR values.

- When facing the in vivo network IRMA, SMPSO is able to perform accurate predictions for the Switch-On dataset, and close to BGRMI. In the case of Switch-Off, OMOPSO generates networks with the best AUPR, followed by DMOPSO, SMPSO, MOPSOHv and LASSO.

- As a matter of biological validation, it is worth mentioning that practically all of the edges in a reference network computed by the MOPSO variants are also obtained by other popular algorithmic proposals used in [267], namely: GRNBoost2 and GENIE3. This is a clear insight in terms of validation for the inferred network, as different techniques in the literature implementing heterogeneous learning models obtain overlapping networks, but with simple variations in edges. This reference network also show consistent features with regards to standard procedures in the Gene Ontology, as well as in the specialized literature.

Chapter 5

Contribution to time series streaming data analysis with biomedical data from sensors devices

This chapter focuses on the complex annotation process of a large amount of sensor-based data that can be challenging in some situations. Specifically, we focus on the problem of Human Activity Recognition (HAR), where HAR systems rely on large amounts of labelled training sensor data. However, annotating data can be challenging in some situations, especially when the granularity of the activities is great, or the user is unable or unwilling to help with the gathering process. In this sense, we propose a semi-supervised deep-learning approach in which these unlabelled data can still be used to train a recognition model. Likewise, a streaming classification process is proposed, since it is crucial for human activity recognition because getting the results in real-time is a compulsion in many situations. The proposed approach has been validated in a real-world use case with a group of overweight patients in the healthcare system of Andalusia (Spain) by classifying movement patterns in real-time conditions, which is crucial for long-term daily obese patient monitoring.

5.1 Introduction

Physical inactivity is one of the main risk factors for chronic diseases such as cardiovascular, cancer and diabetes [268, 269]. Knowing the habits and types of activity carried out by people and their relationship with these diseases is a crucial task in designing treatment strategies and prevention recommendations. Numerous advances have been crucial to deepening high-level knowledge about people's daily life [6]. One of the main objectives of HAR is to provide long-term monitoring of people's daily activities to allow medical doctors to get additional information about their patients to design care plans that may prevent or help against chronic diseases.

HAR has gained much attention in healthcare due to its wide range of applications, such as monitoring of geriatric patients, especially focused on fall detection [8, 270, 271], as well as many other studies related to chronic diseases such as Parkinson's, obesity, cardiovascular and neurodegenerative diseases [272, 273, 274, 275]. These research activities have shown that HAR can effectively improve the quality of health care for some groups of people suffering from some pathologies or chronic diseases.

HAR mainly focus on two types of methods: video-based and sensor-based. Video-based methods provide a dense feature space to allow fine-grained analysis in HAR. However, it is exposed to a highly complex background of images since an environment with very strict conditions, such as well-positioned cameras and individuals, is required for the data collection process with a high cost at the level of computing resources and energy consumption and price. Therefore, video-based methods remain limited in epidemiological studies where the evaluation of daily physical activity requires a reliable, accurate, and low-cost methodology. Sensor-based methods are widely used in scientific physical activity studies since they provide better adaptability in variable environments, high recognition accuracy and low power consumption. Furthermore, in [6] the use of accelerometers is exposed as the most used sensor in the literature since most wearable devices are equipped with them and have easy access. Additionally, an accelerometer is considered a reasonably intelligent sensor for recognizing many types of activities since most are simple body movements.

The work presented in this chapter is motivated by an ongoing collaboration project in a real-world healthcare system (in Andalusia, Spain). We focus on a sensor-based approach, with the primary purpose of discriminating basic posture change movements or activities of a group of patients with obesity and cardiovascular problems. The project aims to provide tools to practitioners to follow the daily routine of their patients and thus prevent a sedentary lifestyle. In this sense, many related studies in the literature have reported high classification accuracy [276, 277, 278, 279]. However, most of them have been tested in academic datasets on young, healthy subjects that can hardly resemble the conditions of a real patient's environment. Besides, most of these experiments have been carried out under controlled environments and restricted activity conditions.

However, as observed in actual healthcare scenarios, a series of critical issues arise related to the limited amount of available labeled data to build a classification model regarding the total volume and velocity of sensorised data. In addition, the discrimination ability of features is often difficult to capture for different classes since the variety of movement patterns in a certain group of patients, e.g. obesity and/or geriatric patients, is bounded and maintained over time. Another issue is the usual class imbalance of data registered in this kind of sensor data stream. Samples representing specific constant postures, such as sleeping, sitting, active, inactive, etc., are perceptually abundant compared to others (running, upstairs, etc.). Therefore, these challenges demand the design and development of hybrid data-driven approaches, where semi-supervised models can act at the core of data processing workflows, usually involving modern Big Data technologies.

In this chapter, a streaming classification model for HAR in healthcare systems is proposed for patient monitoring in real-time. This proposal is based on a combination strategy of public labeled/private unlabeled raw data integration, semi-supervised classification with Convolutional

Neural Networks (CNNs) and Spark streaming processing.

Guided by practical requirements, accelerometer sensor-based data have been considered in this work since low power consumption and use of resources are mandatory through long-term daily patient monitoring in uncontrolled environments. In this sense, as sensorised samples are mostly unlabeled, a data fusion task is conducted with commonly used datasets in the literature (WISDM [280], PAMAP2 [281], HUGADB [282] and USC-HAD [283]). These datasets have been previously labeled according to systematic procedures and share common attributes. This way, labeled and unlabeled samples are integrated for feeding the semi-supervised models to classify new incoming data flows, through the Spark streaming processing engine, by following a sliding window strategy.

In this approach, semi-supervised models are generated with Encoder-Decoder CNNs [284], which allow data augmentation by considering unlabeled samples and statistical features, hence embracing the global properties of the accelerometer time series. For testing purposes, a real-world case study is conducted with a group of more than 300 overweight patients in the healthcare system of Andalusia (Spain), classifying close to 30 TBs of accelerometer sensor-based data.

The remainder of this Chapter is structured as follows. Section 5.2 presents a review of related studies in the current state of the art. In Section 5.3, the proposed approach is described. The experimental procedure is explained, and the results are analyzed in Section 5.4. Finally, Section 5.6 contains concluding remarks and future work.

5.2 Related works

The discovery of patterns of human activity has led to several studies on analyzing the data collected through activity bracelets, smartwatches and smartphones [285]. Many classification methods have been used in previous studies, especially conventional approaches using Machine Learning algorithms [286] such as Extra Trees, AdaBoost, Random Forest (RF), Naive Bayes, k-nearest Neighbours (kNN), or Support Vector Machines (SVM). To name some representative studies of them, in [287] SVM was used to carry out the classification problem of HAR, collecting inertial sensor data through a smartphone mounted on the waist of the individuals. C4.5 Decision Tree and Naive Bayes classifiers were used to recognise 20 daily activities in [288]. In [289] kNN was declared the best classifier compared with C4.5 (J48) Decision Tree, Multilayer Perceptron Neural Network, Naive Bayes, logistic regression, and ensembles based on boosting and bagging. However, they still showed classification failures in similar activities.

Even when conventional approaches have obtained promising results with high-level classification accuracies in different controlled environments, these methods rely on feature-based classification guided by human domain knowledge, which supposes a heavily effort in the pre-processing data stage. Besides, the discrimination of very similar activities for these methods is still a difficult task. Deep Learning (DL) algorithms seem to be a good solution to overcome these problems since they conduct layer-by-layer structural modelling for specific feature extraction and allow the classification process after the segment pre-processing raw data. One of the first approaches can be found in [290], where HAR classification is carried out with CNNs by extracting features without domain-specific knowledge about raw data. Also, in [276], CNN is proposed to perform efficient and effective HAR using smartphone sensors by exploiting the inherent characteristics of activities and 1D time series signals, at the same time providing a way to automatically and adaptively extract robust features from raw data. Various state-of-the-art classification techniques under different scenarios are compared in [277], showing how deep neural networks perform with the best accuracy when the training data volume is drastically reduced.

Many other HAR studies have been implemented with deep learning methods, such as convolutional and recurrent approaches [274, 278, 279, 291]. In this sense, a thorough survey is reported

in [6] where new challenges and trends are identified for this area. In concrete, two of these main challenges are related to the online/streaming processing or sensorised data and the requirement of dealing with unlabeled data. These are, in fact, the direct consequence of working in real-world environments, requiring the management of high volumes of continuously sensorised data. Recent proposals [284, 292] are based on suitable semi-supervised frameworks to cope with these issues. However, they are still limited when tackling scalable data processing.

The proposed approach is conceived to cope with these limitations by combining semi-supervised Encoder-Decoder CNN dynamic models with Spark streaming processing in real-world healthcare environments.

5.3 Proposed approach

The basic methodology in the human activity recognition process consists of four phases [293] as shown in Figure 5.1. These phases are: i) selection and deployment of sensors, ii) collection of data from these sensors, iii) pre-processing and feature selection from the data and iv) use of machine learning algorithm to infer or recognize activities.

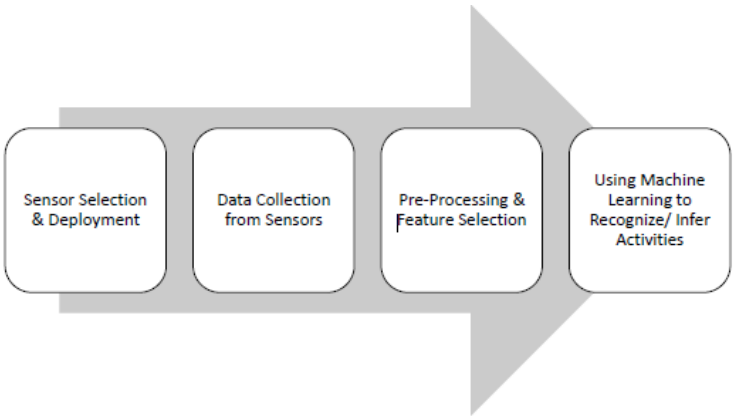


Figure 5.1: General process of human activity recognition. Figure taken from reference [294].

Our proposed approach follows the methodology presented in Figure 5.1. Additionally, it partially follows the basic methodology presented in the so-called activity recognition chain (ARC), extensively studied in [295] as a general-purpose framework for processing time series sensorised data, training and evaluating HAR workflows. A general overview of the proposed workflow in our study is illustrated in Figure 5.2, where all the elements are organized, from data acquisition to model evaluation and human activity prediction.

The main purpose of this strategy is to generate an enriched dataset that, after a feature engineering process for data fusion, is suitable for feeding semi-supervised models, avoiding bias and overfitting problems, as much as possible.

5.3.1 Sensor selection & deployment

In the sensor selection phase, a sensor-based method has been considered since they provide better adaptability in variable environments, high recognition accuracy and low power consumption [6]. Specifically, an accelerometer sensor is used for measuring acceleration. It can sense acceleration in multiple directions. In this sense, the multi-axis accelerometer can measure acceleration in x, y,

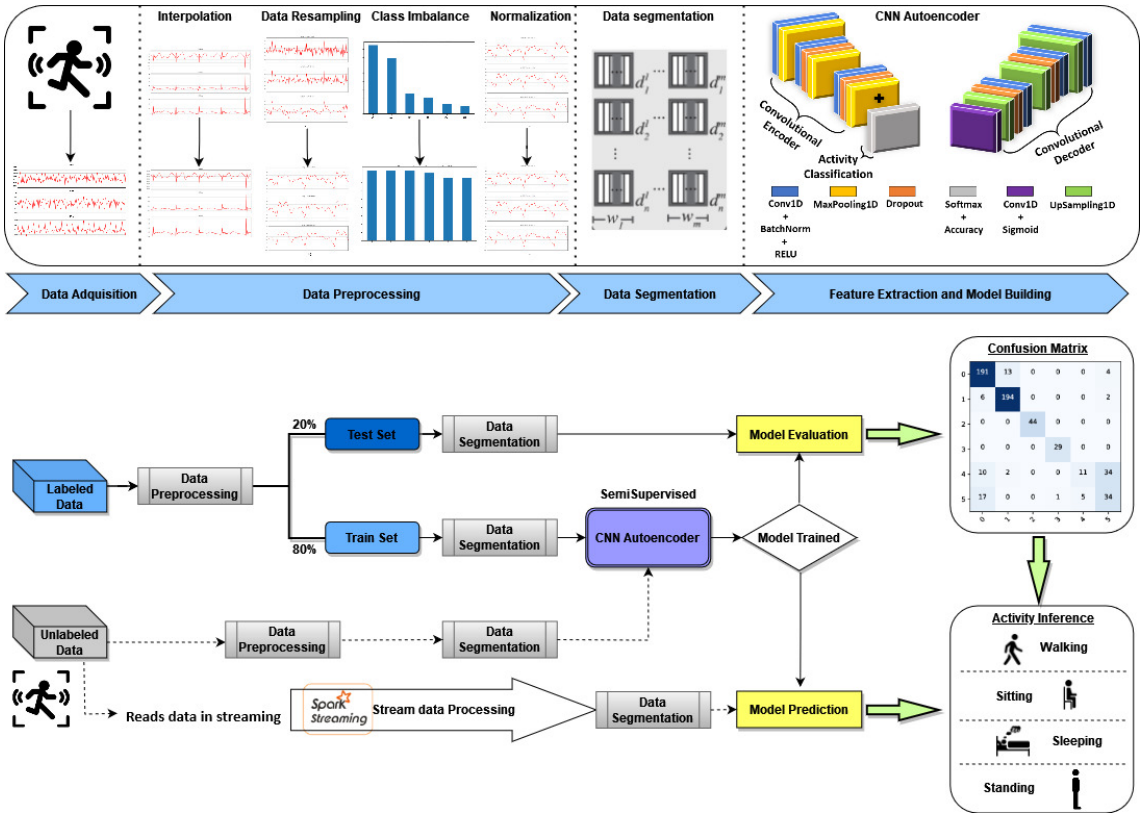


Figure 5.2: General overview of the proposed approach that is presented as a HAR workflow. This workflow is composed of several steps: (1) **Data acquisition**: the data is acquire combining unlabeled data sensors (private dataset) and from public datasets. (2) **Data pre-processing**: these data is pre-process, which involves interpolation for missing data imputation, re-sampling, class imbalance processing and normalization. Also labeled dataset is then split into two subsets with 80% of selected samples for training and 20% of remaining ones for testing. (3) **Data segmentation**: a temporal sliding window with size of 400, corresponding to roughly 4 seconds of physical activity data, and overlap of 100 (1 second) is performed to labeled and unlabeled data. (4) **Feature extraction and model training**: a CNN Encoder-Decoder model is trained with labeled and unlabeled, capturing the most relevant characteristics of the training data in order to provide activity inference of the 30TB of unlabeled data. (5) **Model evaluation**: the model is evaluated with the test sets where confusion matrix and deviated metrics are obtained (Precision, Recall, F1-score) (6) **Streaming processing and activity recognition**: once the model is evaluated and provide us promising results an Spark Streaming classification process is carry out. The whole process is repeated with a certain frequency to rebuild models with updated data. Therefore, the framework to monitor patient's movements will consider new individuals in a transparent way to the learning model, since new sensor data will be in the same Spark streaming source.

and z directions at the same time. The accelerometer is widely used in solutions for activities of daily living in the literature.

5.3.2 Data collection from sensors

In this phase, data is collected from a wearable device (GENEActiv[1]), which incorporates a MEMS triaxial accelerometer placed on the non-dominant wrist of the study subjects (Figure 5.3). Each measurement of this bracelet contains three real values on each of the sensor axes (x-y-z) with a sampling rate at 100Hz, range of +/- 8g and resolution of 12 bits. This way, after a weekly observation period, 30 TBs of raw movement data is collected from 300 patients' daily activities. This final time series dataset is a set of observations $X = (x_t^1, x_t^2 ... x_t^L)$ where each one is recorded at a specific time T and L as a length of time-step.

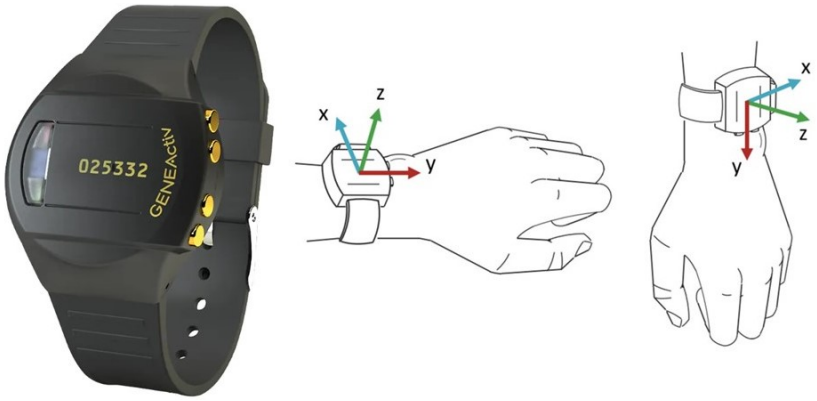


Figure 5.3: GENEActiv is a lightweight raw data accelerometer that allows for objectively continuous physical activity monitoring within clinical trials. Image taken and modified from [296].

Nevertheless, this huge amount of sensorised data still lacks class labeled features, which are required for model training. Therefore, we propose to follow a combined data acquisition strategy that merges our self-data collection from sensors (unlabeled data) with academic datasets (labeled data). The former source comprises data streams of unlabeled attributes (patients' movements) that must be classified. The latter considers a series of labeled datasets from related human activity recognition time series studies in the literature. Therefore, the proposed approach has considered a series of widely used datasets in the literature, each contributing with labeled samples for different, sometimes overlapping, activities. These datasets are: WISDM (Actitracker) [297], PAMAP2 [298], USC-HAD [299] and HuGaDB [300]. These datasets were previously labeled according to systematic procedures and shared common attributes. The time series recorded in these datasets have been collected from heterogeneous devices (smartphones and bracelets) located in different parts of the body, considering a different number of individuals and with a different sampling frequency (e.g. WISDM at 20Hz, HUGADB at 50Hz, USC-HAD and PAMAP2 at 100Hz) in the study. Moreover, they have been modeled to consider different sets of daily activities, which are recorded through different time intervals. In this respect, these activities are sometimes far from the habits observed in our patients (with obesity), so a preliminary exploration phase has been conducted to select that public dataset containing distributions more similar to our self-collected (private) data.

[1] https://www.activinsights.com/products/geneactiv/

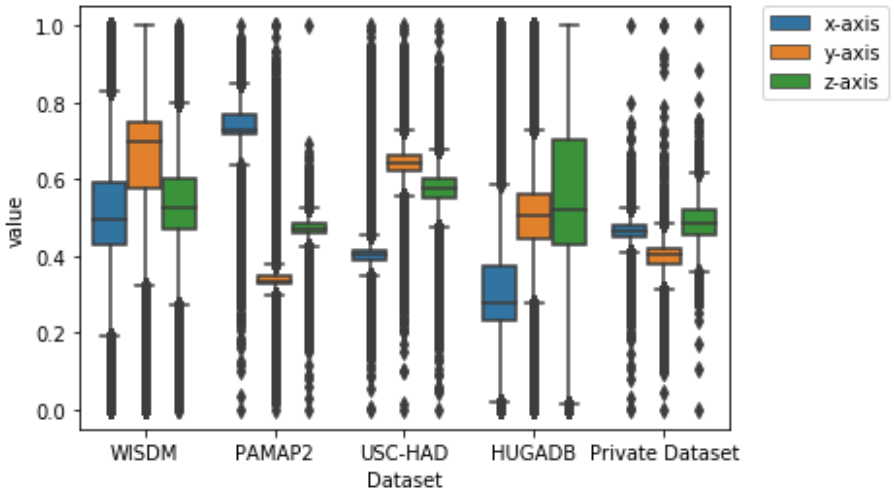


Figure 5.4: Boxplot distributions of the three accelerometer axis corresponding to WISDM, PAMAP2, USC-HAD and HUGADB, taking into account the 6 activities which have in common these datasets (walking, running, sitting, standing, downstairs, upstairs). Also our private dataset was included in the bloxplot distribution.

In this regard, Figure 5.4 shows the boxplot distributions of the three accelerometer axis (x,y,z) for each of the four considered public datasets (WISDM, PAMAP2, USC-HAD and HuGaDB), taking into account six activities which have in common these datasets (walking, running, sitting, standing, downstairs, upstairs), as well as for our private data. After this process, the WISDM dataset is selected to provide our model with labeled samples since it contains the closest axis distributions to the sensorised data of our patients. Therefore, we prevent the model from underfitting with excessive data variation. When the instances are augmented using the WISDM dataset, the model becomes more stable with a lower standard deviation. On the contrary, using all the datasets to train the model adds additional variation and significantly deteriorates the model. In concrete, WISDM (Actitracker) dataset considers six activities registered in a controlled environment: jogging, walking, ascending stairs, descending stairs, sitting and standing. A number of 36 individuals have taken part in these measures.

Public datasets were produced using different devices and under various human circumstances. Differences in the feature-space representation can be thought of in terms of the sensor modalities and sampling rates. Also, discrepancies in the marginal probability distribution can be considered in terms of distinct people performing the same activity or having the movement performed in other physical spaces. Hence a complete pre-processing procedure is carried out to homogenise all of these data sources, including those commonly detected activities among all the individuals in observation. In concrete, these shared activities are: running, walking, sitting, standing, upstairs and downstairs, which are used as labeled categories for the semi-supervised models in this proposal.

5.3.3 Data pre-processing

The success of any inference technique is highly dependent on the quality of the data fed into the model. Real-world data is often dirty and noisy and contains outliers, irrelevant or unnecessary features, and null or non-standardised values. When erroneous or raw data is often used, the resulting model tends to be biased or not perform correctly. For this reason, the transformation of

the original data in the preprocessing phase is essential. In this sense, data processing is performed on labeled and unlabeled data, which involves interpolation for missing data imputation, Data re-sampling, class imbalance processing, data normalization and data segmentation (Figure 5.2).

5.3.3.1 Data normalization

Applying data normalization techniques is critical in this study because the devices used for data collection are different in each of the selected datasets. Therefore, the input signal amplitude will vary significantly from one to another.

Raw data have been normalized through Z-score normalization. Feature standardization makes the values of each feature in the raw data have zero-mean and unit variance. This normalization is formulated in Equation 5.1, where x is the original feature vector, $x^{'}$ is the normalized value, $\tilde{x} = average(x)$ is the mean of that feature vector, and σ is its standard deviation.

$$x^{'} = \frac{x - \tilde{x}}{\sigma} \tag{5.1}$$

5.3.3.2 Missing value imputation in sensor raw data

Data collection processes are performed in a real-world scenario where data dropout may occur due to wireless sensors or possible hardware problems. It will generate raw data containing noise or missing values. Therefore, linear interpolation is conducted to tackle missing values and to fill gaps in raw data time series. This method is commonly used for time series missing value imputation. It helps estimate the missing data point using the two surrounding known data points. It searches for a straight line that passes through the endpoints x_A and x_B, as formulated in Equation 5.2, where xi are observed data, Xi are the interpolated value(s) of missing data, and α is the interpolation factor that varies from 0 to 1.

$$X_i = (1 - \alpha)x_B + \alpha x_A \tag{5.2}$$

5.3.3.3 Data re-sampling

Data re-sampling has been carried out to homogenize the frequency of the datasets since each one is arranged at a different frequency. For this purpose, down-sampling and up-sampling techniques have been applied. It is worth noting that when dealing with "waves" in time series, it is observed that low sampling frequencies tend to lose information in specific movements, where a high frequency is required to identify them correctly. For this reason, we must determine the wave frequency according to the type of recognition faced. Figure 5.5 shows an example of raw data of a patient's activities ("walking" and "cycling") collected by an accelerometer sensor on a wrist. After re-sampling, data are transformed for each activity at frequencies of 100Hz (top), 50Hz (middle) and 20Hz (bottom). The effect of data re-sampling is illustrated, and it is possible to identify some losses in the data information as long as the frequency decreases. It can be observed in Figure 5.5 a), where different wave peaks "disappear", provoking inconsistent data representations at different sampling frequencies. Therefore, a high re-sampling (100Hz) is performed to keep the informative level in samples while making data homogeneous for all the sources.

For this reason, we resampled all datasets to put them at the same frequency as our private dataset (100Hz) to keep data information. Hence, we up-sampled the WISDM dataset from 20Hz to 100Hz and HUGADB from 50Hz to 100Hz, since USC-HAD and PAMAP2 are already at 100Hz.

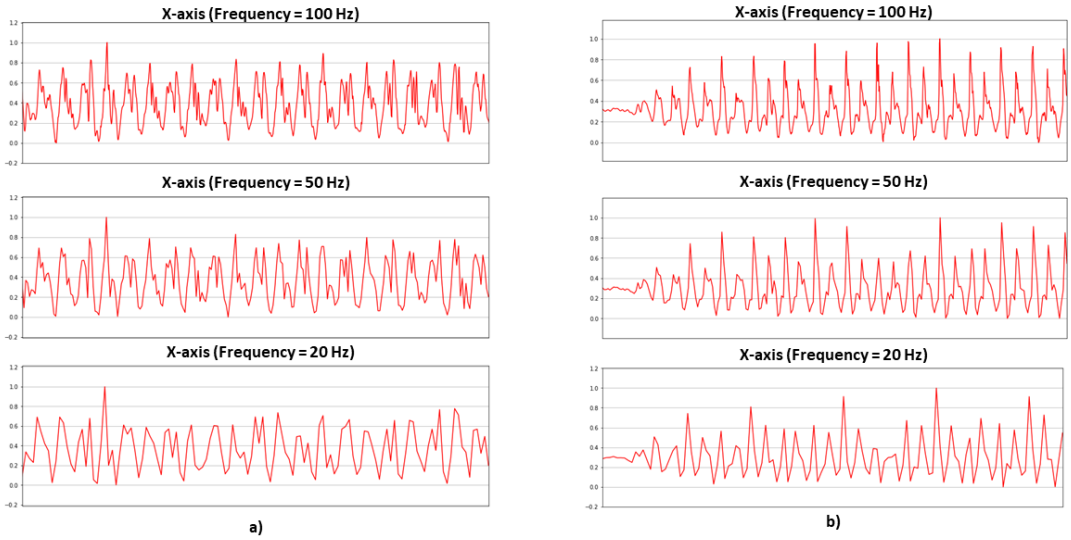


Figure 5.5: Raw data from accelerometer sensor of different activities: Walking (**a**) and cycling (**b**) at 100 Hz (top) and re-sampled data at 50 Hz (middle) and 20 Hz (bottom). It can be noticed that as the sampling rate decreases, aspects at high frequency are removed from the wave.

5.3.3.4 Class imbalance

Another quite common yet important issue registered in HAR datasets is the class imbalance. Even more in real-world sensor data from the particular case of obese patients, where the balance between classes is not guaranteed and biased to sedentary activities. For example, the "sitting" activity is more frequent in the case of overweight patients than the "running" activity, producing an important class imbalance that could lead learning models to behave with a bias towards the majority classes. As a consequence, algorithms will fail in the classification of the underrepresented minority classes, which provokes a severe decrease in the overall accuracy of the results [301].

In order to cope with class imbalance, several approaches have been used, such as oversampling and under-sampling methods at the data level [302, 303] and many other solutions at the algorithmic level trying to trade-off the class imbalance in modelling time [304, 305]. In the context of HAR, Synthetic Minority Oversampling Technique (SMOTE) [302] is a standard over-sampling method used to generate new synthetic data of the minority classes. It has shown great success in several applications where SMOTE helps to enhance the classification accuracy for imbalanced datasets. For example, in [306] data balancing was used through SMOTE oversampling approach, leading the worked model to reach high accuracy results.

By default, SMOTE re-samples all classes except the majority class; that is, the minority classes are increased to reach the total number of the majority class. However, the study in [302] suggested combining SMOTE with random under-sampling of the majority class since a high over-sampling could provoke model over-fitting. For this purpose, our methodology addresses class imbalance at the training stage by balancing classes in two separate steps: firstly, SMOTE oversampling technique is used to over-sample those minority classes to have 50% of the number of examples of the majority class. Then, under-sampling using random elimination is performed on the majority classes to have 20% more than the minority class. Then a difference of 20% between classes of samples is obtained, which helps the model to avoid problematic class imbalance, preventing the generation of synthetic data in a high percentage.

5.3.4 Data segmentation

As the last step of the data processing phase, we will segment the data into time windows. At this step, data samples are still structured in the time domain since all the axis points are collected from sensors at a certain time instant. Therefore, a segmentation stage is required to transform these input data into the frequency domain, more suitable for training deep learning models as signal processing prediction tasks. The data segmentation stage, also known as activity detection, determines which sections of the preprocessed data streams are most likely to contain information about activities. A report on activity segments is essential for activity recognition.

In our approach to data segmentation, a window is dragged over the time series data using a sliding windows approach to extract a data segment for use in later stages of the workflow. The window size has a direct impact on the recognition system's delay. Additionally, the ideal window size is not apparent a priori and can affect recognition performance [307]. In our case of patients with obesity, various time window sizes were tested. Finally, a temporal sliding window with a length of 400, corresponding to roughly 4 seconds of physical activity data, and an overlap of 100 (1 second), is performed for each axis attribute in the dataset. This overlapping among windows guarantees high numerosity of training and testing samples to train the model.

To fit the input shape of the CNN-Encoder-Decoder, it is necessary to reshape the sample obtained in the previous step. Therefore, each window comes in the shape of a matrix of values of shape N x 400 x 3, where N is the number of samples resulting from the segmentation, 400 is the time window, and 3 is the number of features to train the model (x-axis, y-axis, z-axis). In this segmentation, sliding windows are checked to contain samples from just one human activity.

5.3.5 Feature extraction and model building

One of the main challenges arising in this study is the possibility of taking advantage of dealing with labeled and unlabeled data. In this sense, using semi-supervised learning techniques constitutes a suitable option for performing predictive analysis since they allow to train models with labeled and unlabeled samples, which mainly improve generalization and avoid overfitting [284]. So, this step entails the semi-supervised learning task, which merges the labeled segments in training set with those unlabeled from sensors.

In particular, the use of CNN-based approaches has been shown to perform successfully for HAR since they provide hidden data representations and identify patterns in activity time series [290, 292]. Therefore, considering a dataset with N pairs $(x_1, t_1), (x_2, t_2), \cdots, (x_N, t_N)$, being x_i a sliding window input with length T and t_i the label representing a given activity, we adopt a semi-supervised strategy CNN Encoder-Decoder in our approach. In this, labeled samples $\{(x_i, t_i)|1 \leq i \leq N\}$ are used together with unlabeled ones $\{x_i|N+1 \leq i \leq N+M\}$ in training, to fit the model with both data sources (sensorised and academic). As argued in [292], using this semi-supervised CNN Encoder-Decoder, it is possible to learn the network and features simultaneously from the data.

In general, the encoder network maps a given input signal $x \in X \subset \mathbb{R}^{d_0}$ to a feature space $z \in Z \subset \mathbb{R}^{d_k}$, whereas the decoder takes this feature map as an input, process it and produce an output $y \in Y \subset \mathbb{R}^{d_L}$

The rationale behind the CNN Encoder-Decoder for semi-supervised classification is to include noise in all the layers of the network, so it works to minimize the distance between the clean input and the reconstructed decoder. In this way, the learning procedure can be summarized in the following steps:

1. Labeled and unlabeled data are processed by the clean encoder to compute hidden variables in the middle layers z_i^k;

2. Both labeled and unlabeled data are corrupted with Gaussian noise and transformed to an abstract representation $\tilde{z}_i^k$, by the noisy encoder;

3. Labeled data $(\tilde{\mathrm{x}}_i, 1 \leq i \leq N)$ are used to perform the prediction task on a softmax based on cross entropy cost. The predicted classes are denoted with $\tilde{y}_i$;

4. The decoder works to reconstruct unlabeled samples $(\tilde{\mathrm{x}}_i, N+1 \leq i \leq N+M)$ which are denoted with $\hat{\mathrm{x}}_i$, so they should be as close as possible to the corresponding input (x_i). To measure this similarity, square error is computed.

The cost function is formulated in Equation 5.3 as an aggregation of the supervised cross entropy of the noisy output $\tilde{y}_i$ predicting the class activity t_i for the input x_i (first term in this equation), whereas the unsupervised cost (second term in this equation) is the denoising square error between clean input x_i and their noisy reconstruction output $\hat{x}_i$.

$$Cost = -\frac{1}{N}\sum_{i=1}^{N} \log P(\tilde{y}_i = t_i | x_i) + \frac{\lambda}{N}\sum_{i=N+1}^{N+M} ||\hat{x}_i - x_i||_2^2 \tag{5.3}$$

The full structure of our CNN-Encoder-Decoder model is shown in Figure 5.6.

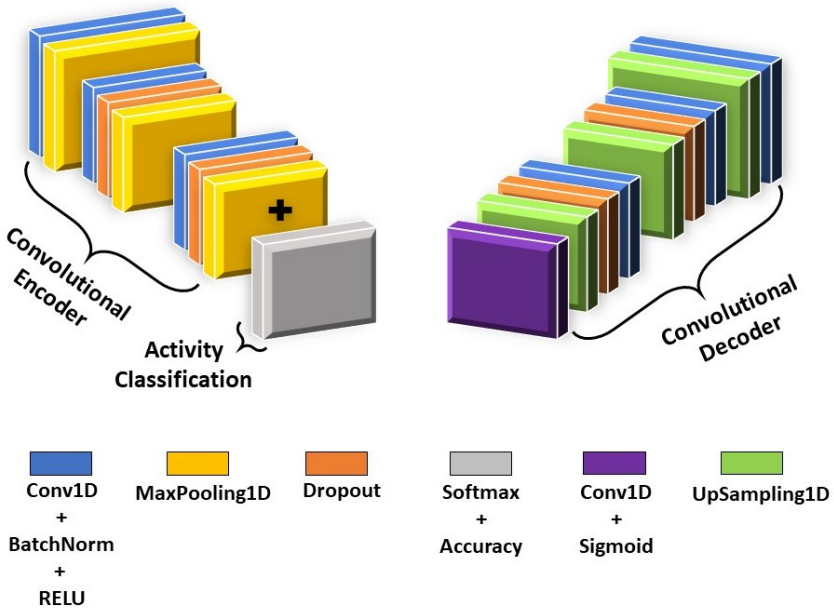


Figure 5.6: The proposed model contains an encoder part composed of three down-sampling blocks in the following structure [Conv1D + BatchNorm + MaxPooling1D + Dropout]. Moreover, each encoder layer has a corresponding decoder layer of three up-sampling blocks [Conv1D + BatchNorm + UpSampling1D + Dropout]. Finally, the Softmax output layer is added for multi-class classification.

- **Encoder**: The encoder network consists of three down-sampling blocks. Each down sampling block is composed of 1D convolutional layers with kernel size of 3, followed by a max-pooling

layer. Additionally, for each block a batch normalization is added to reduce internal co-variate shift [308], accelerating the training process of the model, and a dropout layer was added to improve generalization performance and avoid over fitting. It then follows an structure [Conv1D + BatchNorm + MaxPooling1D + Dropout]

- **Decoder**: Each encoder layer has a corresponding decoder layer. Thus, the decoder network consists of three up-sampling blocks composed of 1D convolutional layers with a kernel size of 3, followed by an up-sampling layer. As for the encoder, for each up-sampling block, batch normalization and dropout layers were added, with a structure [Conv1D + BatchNorm + UpSampling1D + Dropout].
- **Softmax**: The model is turned to a classifier by adding a Softmax output layer for multi-class classification.

Therefore, the semi-supervised CNN Encoder-decoder allows unlabeled samples from sensor streaming sources to take part in the learning model during training time, so it will avoid bias in certain classes and promote generality.

5.3.5.1 Model evaluation

Once the model is built, an evaluation step is carried out regarding the test set, where confusion matrix and deviated metrics are obtained (Precision, Recall, F1-score, etc.). It is worth noting that this test set is wholly obtained from the public dataset, in this case, WISDM. However, the model has been trained with public and private data, so final predictions are expected to show certain model generalization with moderate accuracy. The final goal is to get a prediction model suitable for a dynamic data flow environment but not for a specific dataset in a certain period.

5.3.6 Streaming processing and activity recognition

Finally, a streaming processing task is deployed through an Apache Spark environment, in which new sensorised data are pre-processed to be predicted according to the model previously built. An internal segmentation step is carried out with streaming data using a similar sliding window size used in the model training phase. This is then a continuous process of human activity label assignation of new samples regarding patient's movements, which can now be monitored by practitioners.

5.4 Experimental results and analysis

In this section, we investigate the effects of training a semi-supervised CNN Encoder-Decoder using labeled data from one public dataset (WISDM) and unlabeled data from our private dataset.

The goal is to be able to classify the 30 TB of unlabeled data. The Convolutional Encoder will compress the input signal x into a space of latent variables ($h = f(x)$), then learn how to reconstruct the data back from the reduced encoded representation. Meanwhile, the Convolutional Decoder works to reconstruct the input signal based on the information previously collected ($r = g(h)$), as observed in Figure 5.7. Therefore, the latent variable space h will capture the most relevant characteristics of the training data.

In this regard, the algorithm learns how to reconstruct the input using the Adam optimizer [309] and using the mean square error as a loss function. Therefore, the model can extract more significant characteristics from the unlabeled data that will help us make predictions.

Bayesian optimization has been used for efficient hyper-parameter tuning [310]. The hyper-parameters were tuned by performing 10-fold Stratified Shuffle Split cross-validation on the training

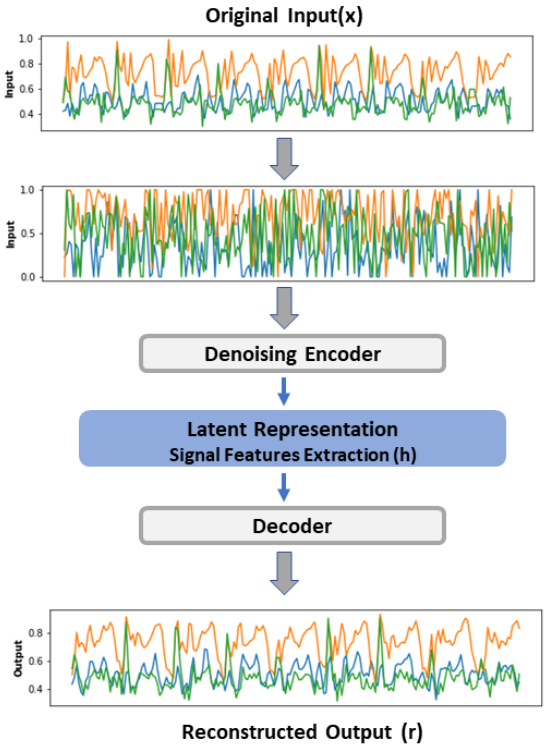


Figure 5.7: CNN Encoder-Decoder model. It contains a clean convolutional Encoder, noisy convolutional encoder, and a convolutional decoder. labeled and unlabeled data are processed by clean convolutional encoder and then corrupted with Gaussian noise. Then the convolutional decoder works to reconstruct the clean input(x) from high-level representation $r = g(h)$.

set using Bayesian optimization, obtaining a filter size of 64 for each of the 1D convolutional layers, which is activated by the ReLU function. Moreover, each of the max-pooling and up-sampling layers contains a pooling size of 2 and the dropout was set to 0.1 for each one. The Bayesian optimization was executed with a batch size of 50, 500 and 1000, obtaining the best results with 50.

In order to assess the performance of our classification methodology system, we split the available dataset into 80% train data and 20% test data. This was done based on the subjects rather than on the segmented windows. In this regard, train data contain subjects 1 to 32 of WISDM dataset and test data include the rest of the subjects (32 to 36). Thus, four subjects out of 36 are always kept isolated for each experiment to evaluate the model. This prevents over-fitting of the subjects and helps to achieve better generalization results.

To comprehensively evaluate the model, we used several evaluation metrics to evaluate the classification results: accuracy, precision, recall, F1-score, loss function, receiver operating characteristic (ROC) and normalized discounted cumulative gain (NDCG), as shown in Table 5.1. It should be noted that we opted to estimate the mean F1-score (Fm-score), that is, the mean F1-score across all the classes. It is shown in Equation 5.4 and Equation 5.5, where TP is the number of true positives in prediction, FP is the false positives, and FN is the number of false negatives.

$$precision = \frac{TP}{TP + FP} \quad recall = \frac{TP}{TP + FN} \tag{5.4}$$

$$Fm - score = \frac{2 * precision * recall}{precision + recall} \tag{5.5}$$

The CNN Encoder-Decoder has been implemented in TensorFlow using Keras. The experiments to evaluate the model have been executed on a machine with 16 CPUs (Intel(R) Xeon ® Gold 6130 CPU 2.10GHz). After each training epoch, we evaluate the model's performance on the validation set. Each model is trained for at least 50 epochs. The training stop condition is configured if there is no increase in validation performance for 10 subsequent epochs. We select the epoch that showed the best validation-set performance and applies the corresponding model to the test set.

5.4.1 Sensitivity to unlabeled sample size

In this section, we study the performance of our semi-supervised CNN Encoder-Decoder model trained with varying amounts of unlabeled data. The amount of the unlabeled data will be proportional to the percentage of samples of the labeled data used for training. Therefore, we evaluate the metrics of our model trained using unlabeled data of 10%, 20%, 30%, 50%, 80%, 100%, 150% proportion of labeled data used for training, as shown in Table 5.1. The number of unlabeled samples varies from 97,814 (10% of train labeled data) to 1,467,222 (150% of train labeled data).

Table 5.1: Metrics obtained with varying number of unlabeled examples in training set. The amount of unlabeled data is taken as a percentage of the training set of the labeled data (WISDM Dataset). The number of unlabeled samples varies from 97,814 (10% of train data) to 1,467,222 (150% of train data).

Metrics: Public data (labeled) + Private data (Unlabeled)						
%	acc	loss	recall	Fm-score	roc	ndcg
0	0.981	0.069	0.981	0.981	0.998	0.998
10	0.976	0.075	0.977	0.967	0.995	0.997
20	0.971	0.076	0.949	0.949	0.992	0.993
30	0.951	0.148	0.940	0.938	0.991	0.990
50	0.947	0.151	0.925	0.926	0.990	0.988
80	0.905	0.292	0.905	0.903	0.987	0.985
100	0.875	0.319	0.872	0.871	0.983	0.984
150	0.685	0.601	0.685	0.655	0.941	0.981

Figure 5.8 shows how the Fm-score evolves when varying the number of unlabeled examples in the experimental results. Fm-score generally decreases when there are more unlabeled samples as expected. This is explained by unlabeled data coming from a different dataset than including variation. However, it can be observed in Figure 5.8 that for percentages of unlabeled data less than 100%, we obtain a high Fm-score in the result.

Thus, our approach can potentially learn the network and features simultaneously from the data using unlabeled data in our CNN Encoder-Decoder model. Therefore, it is possible to use this model as the core predictor. To do so, we have chosen the amount of 80% of unlabeled data to classify the 30 TB from sensors since at this point, the model is still getting good results ($Fm - score = 0.90$).

More in-depth, Figure 5.9 shows the resulting confusion matrices when varying the amount of unlabeled data with 10%, 50% and 80% in the model training. It can be observed that the model achieves promising predictions for activities walking, running, sitting, standing and upstairs, even when increasing the number of unlabeled samples. In contrast, the model start to show limited predictions in detecting downstairs since, if we see the patterns between walking and downstairs,

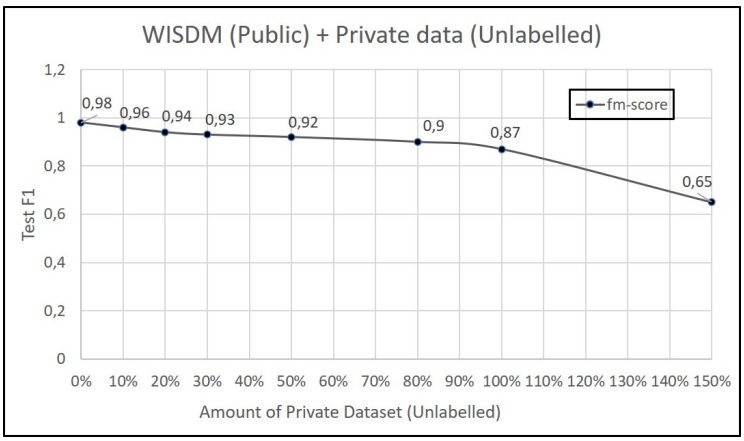
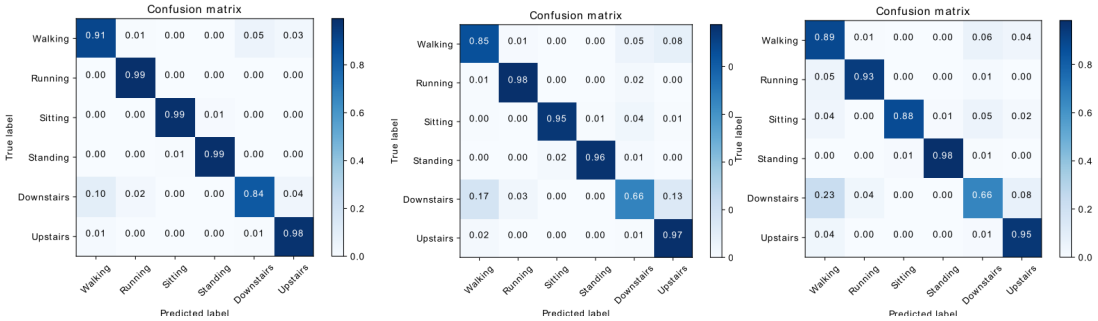


Figure 5.8: Fm-scores obtained with varying number of unlabeled examples in training set.



(a) 10% of unlabeled data (b) 50% of unlabeled data (c) 80% of unlabeled data

Figure 5.9: Illustration of confusion matrices showing the sensitivity of the networks for each individual class when varying 10%, 50% and 80% of unlabeled data when training the semi-supervised CNN-Encoder-Decoder.

they are characterized with very close signal shapes in movements, as mentioned in [280]. This is generally an acceptable precision since even for 80% unlabeled data, it still gets good predictions for all classes.

As we know, it is hard to assess performance in unlabeled data, but we still need to know if it passes *"the eye test"*. For this purpose, we classify a randomly chosen sample of unlabeled data to demonstrate that the distributions of the predictions are reasonable. It is shown in Figure 5.10 (format date is a month-day hour) how the main activity is resting (sitting and standing) as we expected. It is normal since this unlabeled data correspond to one of the 300 overweight patients in the healthcare system of Andalusia. In the same way, during the night (from 00:00 to 08:30 approximately), the patient is resting (sitting). Later, the patient is standing and starts to be more active. Then around 12:00, the patient starts to do moderate physical activity (running and upstairs). It can be seen that on both days, at 12:00 (06-05 12:00 and 06-06 12:00), the patient is physically active. This could be explained by the fact that patients follow the doctors' instructions doing daily exercise to avoid sedentary life. Afterwards, the patient does some short movements, and finally, after 00:00, resting is the main activity.

It should be noted that the classification has been carried out according to the labels that

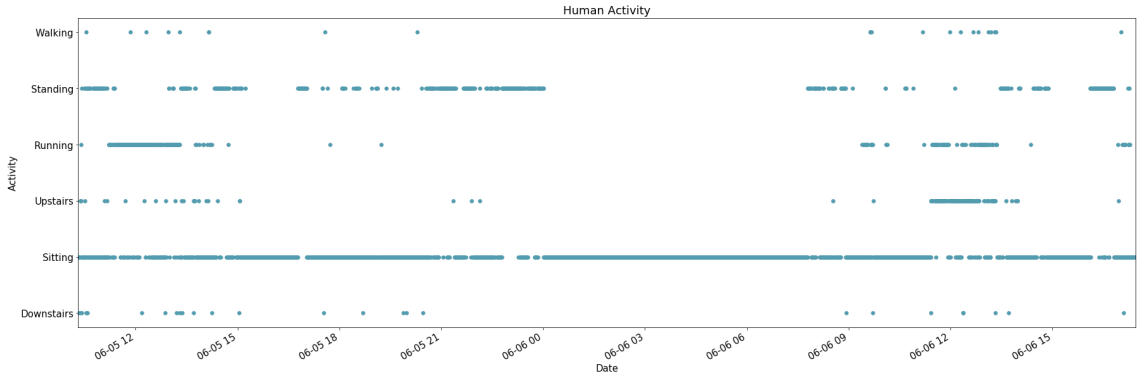


Figure 5.10: Snapshot of the Human Activity Recognition for a randomly anonymous patient. It is shown how during the night sitting (resting) is the main activity, later around 8:30, the patient starts to be more active and does short movements. Then, at 12:00 the patient seems to start some moderate activity and finally, after 00:00 resting is the main activity.

we have from the WISDM dataset. However, our private dataset provides us with long-term monitoring of patients' daily activities where we can find more activities and transitions between activities. Even so, the results obtained in Figure 5.10 seem quite reasonable to us for this first approach in which we try to address the problem of HAR in a real-world case without previously labeled activities in our dataset.

5.4.2 Additional experiments

Additional experiments have been implemented to demonstrate the feasibility of the proposed semi-supervised methodology. A first experiment was conducted to see whether the model could pass *"the eye test"* without taking into account the semi-supervised approach. Consequently, the model was trained only with raw data from the WISDM dataset. After that, a classification task was performed from a randomly chosen sample from our 30TB private unlabeled dataset. As expected, the model did not pass *"the eye test"* without using unlabeled private data in the training phase (Figure 5.11)

Moreover, the proposed methodology has been synthetically evaluated using another public dataset to simulate the unsupervised portion. In this sense, the HUGADB dataset has been considered as *"unlabeled dataset"* and WISDM as a labeled dataset. HUGADB dataset was classified with and without considering our proposed semi-supervised methodology. Finally, the model was evaluated if it could predict the activities in the HUGADB dataset. In this experiment, we concluded that using the semi-supervised approach gives us better predictions, as observed in Table 5.2. The same experiment was carried out with PAMAP2 as *"unlabeled dataset"*.

5.4.2.1 First experiment: without semi-supervision

In this first experiment, we wanted to see whether the model could pass *"the eye test"* without taking into account the semi-supervised approach. For this proposal, the model was trained only with labeled data from the WISDM dataset without considering our private unlabeled data in the training phase. Afterwards, the prediction of a randomly chosen sample (five days prediction) from our 30TB private unlabeled data set was performed, as shown in Figure 5.11. It can be observed that the model predicts running and walking downstairs as the main activities of the patient even during the nights and rarely predicts the activities of standing and sitting, even though these are

the most prevalent behaviors among obese patients. Overall, it may be said that the model cannot make reasonable predictions if the unsupervised task is not used in the training regime.

Figure 5.11: Activity classification of a randomly chosen sample (five days prediction) from our 30TB private unlabeled data set. For these predictions, the model has been trained only with labeled data from WISDM dataset without considering our semi-supervised strategy with private unlabeled data in the training phase.

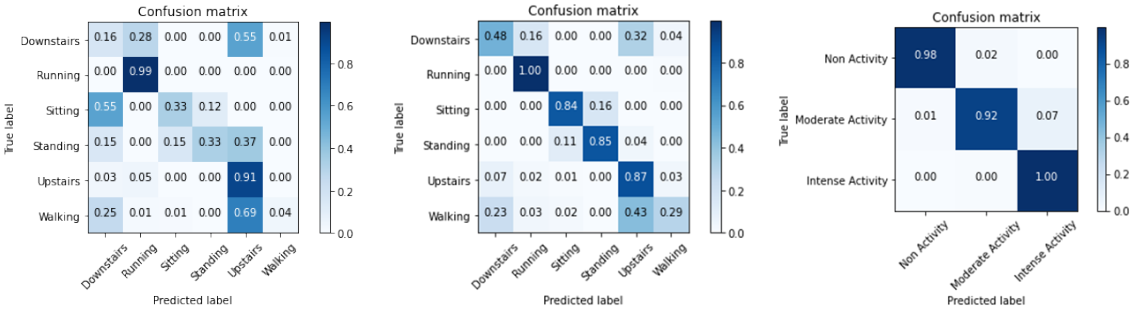
5.4.2.2 Second experiment: with semi-supervision

In a second experiment, the proposed semi-supervised methodology was synthetically evaluated using another public dataset as a simulation of the unsupervised portion. In this sense, the HUGADB dataset has been considered as *"unlabeled dataset"* since it contains in overall the closest axis distributions to the sensorised data of the WISDM dataset and the lowest standard deviation in the data as shown in Figure 5.4. Hence, we study the performance of our semi-supervised CNN Encoder-Decoder model trained with a combination of WISDM as public annotated data WISDM and 70% of the HUGADB dataset as a simulation of the unsupervised portion to classify the activities in HUGADB, as observed in Figure 5.12. First, the model has been trained only with labeled data from WISDM without considering unlabeled data in the training phase. Afterwards, the model was validated in the remaining 30% of the HUGADB dataset, as shown in Figure 5.12a. Subsequently, to demonstrate the feasibility of our semi-supervised approach, the model has been trained again, but this time 70% of HUGADB has been taken into account as a simulation of the unsupervised portion in the training phase. As previously mentioned, the model has been validated in the remaining 30% of the HUGADB dataset, as shown in Figure 5.12b. It can be appreciated that our semi-supervised approach improves the prediction results from 0.414 to 0.704 in terms of Fm-score, as shown in Table 5.2.

This second experiment has been repeated with another public dataset as a simulation of the unsupervised portion to verify the quality of the semi-supervised approach. PAMAP2 has been selected in this case since it contains different axis distributions to the sensorised data of the WISDM dataset and the highest standard deviation in the data, as shown in Figure 5.4. It is shown in Table 5.2 how the semi-supervised methodology increases the prediction results from 0.129 to 0.667 in terms of Fm-score. Also, in Figure 5.13 the semi-supervised strategy increases the accuracy in all the classes.

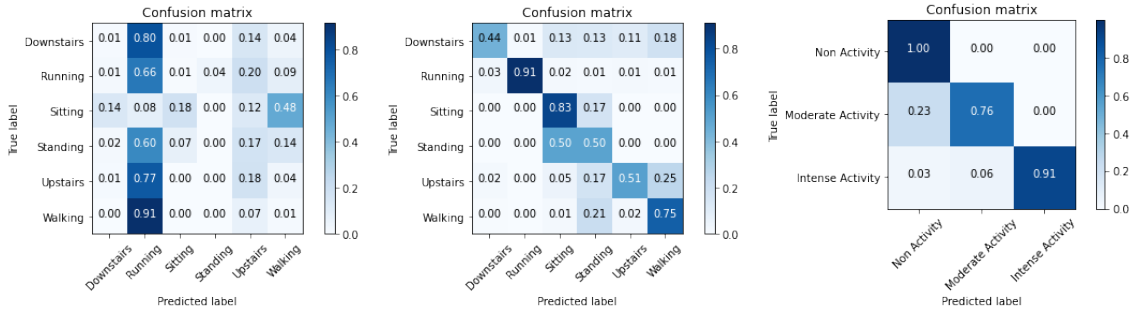
Table 5.2: Metrics evaluation with varying number of unlabeled examples in training set. HUGADB and PAMAP2 datasets have been taken as a simulation of the unsupervised portion to synthetically evaluate the proposed semi-supervised methodology.

Metrics: Public data (labeled) + Public data (Unlabeled)				
labeled/Unlabeled	%	acc	recall	Fm-score
WISDM/HUGADB	0%	0.461	0.461	0.414
WISDM/HUGADB	70%	**0.722**	**0.722**	**0.704**
WISDM/PAMAP2	0%	0.173	0.173	0.129
WISDM/PAMAP2	70%	**0.667**	**0.667**	**0.667**



(a) 0% of unlabeled data (b) 70% of unlabeled data (c) 70% of unlabeled data

Figure 5.12: Illustration of confusion matrices showing the sensitivity of the networks for each individual when varying the percentage of unlabeled data in the training regime from 0% to 70% (HUGADB as a simulation of the unsupervised portion). In addition, the dimensionality of HAR classification problem has been reduced into three basic classes in Figure(c): Non Activity (sitting and standing), Moderate Activity (walking, walking downstairs and walking upstairs) and Intense Activity (running).



(a) 0% of unlabeled data (b) 70% of unlabeled data (c) 70% of unlabeled data

Figure 5.13: Illustration of confusion matrices showing the sensitivity of the networks for each individual when varying the percentage of unlabeled data in the training regime from 0% to 70% (PAMAP2 as a simulation of the unsupervised portion). In addition, the dimensionality of HAR classification problem has been reduced into three basic classes in Figure (c): Non Activity (sitting and standing), Moderate Activity (walking, walking downstairs and walking upstairs) and Intense Activity (running).

Despite improving the quality of results with our semi-supervised approach, the model shows limited predictions in detecting some activities. For example, for the model, it is challenging to predict downstairs and to walk since, if we see the patterns between walking and downstairs, they are characterised by very close signal shapes in movements, as commented before in the paper. Furthermore, static activities can be recognised more easily than regular activities (running, walking, etc.). However, similar postures (sitting and standing) create significant complexities in case of separation due to considerable overlapping in feature space, as observed in Figure 5.13b. In general, the dimensionality of HAR classification problem can be reduced by classifying it into three basic types: Non Activity (sitting and standing), Moderate Activity (walking, walking downstairs and walking upstairs) and Intense Activity (running) as shown in Figure 5.12c and Figure 5.13c. We can obtain promising results that will allow us to provide patient activity information to doctors, which is essential to preventing obesity.

In conclusion, it can be said that the semi-supervised approach improves results when trying to predict activities from a dataset that the model has never seen before. With the semi-supervised strategy, the model can extract important features from the unlabeled data that help us to make better predictions.

5.4.3 Computational performance

To carry out the streaming classification process, the complete approach has been deployed on a virtualization environment operating on an on-premise high-performance cluster computing platform, as mentioned in Chapter 2 Section 2.5. It comprises several virtualization units that allow visualizing the cluster's performance. Concretely, this platform has 10 virtual machines, each with 16 cores (CPU 16 x 2.10 GHz), 128 GB RAM and 1 TB of virtual storage (up to 176 cores, 1408 GBs of memory and 10 TB HD storage). These virtual machines have been used with the role of Worker node (Apache Spark) to make the activity predictions. The Master node, which runs the Keras CNN Encoder-Decoder, is hosted in a different machine with 16 cores at 2.10 GHz, 128 GB RAM and 5,000 TB of virtual storage space. All these nodes use Linux 4.15.0-118-generic 64-bit distribution. The whole cluster uses Spark 3.0.1.

Additionally, an NFS distributed file system has been configured to be able to access the sensorised data from all the machines. The Master node will physically store the data (server). In contrast, the Worker nodes will behave as clients to access the data remotely. In this way, it is possible to perform the activity prediction in parallel from the different machines connected to the same network to access remote files as if they were local ones.

The classification of activities accessing a directory at the NFS distributed system for the parallelisation of Spark streaming processes. The data is passed in streaming from the repository. Each CSV file included in the directory will behave as a Spark streaming batch that will go through a segmentation process by time windows (400 rows corresponding to 4 seconds of monitoring activity) as observed in Figure 5.2. Finally, the CNN Encoder-Decoder model trained will predict the activity of each batch in streaming. The results are saved in text files using the same name as the original CSV files (See Code Snippet 5.1).

[H]

Code Snippet 5.1: Spark streaming segmentation and classification by batch

```
//Read csv in Streaming with Spark from directory
df = spark.readStream(directory)
//Load the CNN-Encoder-Decoder model
model = keras.load(model)

classify(batch, batch_id, model):
    // we set time window to 400 (4 seconds of activity)
```

```
    time_window = 400
    // raw data segmentation by time Window
    batch.map(lambda x,y: [raw_data],time_window)
    // group by time_window
    batch.reducebyKey(lambda x,y: x+y)
    // activity prediction of raw data
    batch.map(lambda r: model.predict(r))
    // save the result
    batch.saveAsTextFile(batch_id+ .txt)

// Streaming classification for each batch
df.foreachBatch(classify(batch, batch_id, model))
```

The performance of the proposed streaming solution has been evaluated through a series of experiments to measure the performance in terms of *Speedup (SN)* and the *Efficiency (EN)*. Thus we analyse the computational effort and the data management process. The standard formula of the *Speedup* calculates the ratio of *T1* over *TN*, where *T1* is the running time of the analysed algorithm in 1 processor and *TN* is the running time of the parallelised algorithm on *N* processing units (processors or cores), while the *Efficiency (EN)* is calculated as shown in Equation 5.6.

$$SN = \frac{T1}{TN} \quad EN = \frac{SN}{TN} * 100 \tag{5.6}$$

Table 5.3 shows the running time in seconds used by the Spark streaming classification approach running on 40, 80 and 160 cores with different batch sizes of raw data. This way, we have centred on file sizes of 64 MB, 128 MB, 256 MB, 512 MB and 1 GB since they are the average size of CSV files in the 30 TB data. In this sense, we measure the computational influence of using a different number of cores with different batch sizes. This table also contains the corresponding Speedup and Efficiency values to the resulting times. As mentioned, the running time is reduced concerning the increase in the number of cores used in the parallel model. The highest reduction in time is obtained when our approach is configured with 40 cores in parallel, for which the running time is reduced from 28.10 s to 6.29 s in the case of the smallest batch size (64 MB), and from 462.75 s to 8.18 s with the biggest batch size (1 GB) used in the experiments. Also, in terms of efficiency, the highest percentage, 141.48%, is reached with 40 cores with a batch size of 1 GB, reaching the best efficiency. In contrast, it decreases as the number of resources gets larger. This behaviour was somewhat expected as the particular cluster configuration involves computing overheads due to virtualisation and network communications, so a trade-off setting is reached with fewer nodes but stabilising from 80 nodes in advance. Considering the results, it is worth mentioning that both cluster configurations (80 and 160 cores) yield similar speedup and efficiency values, which indicates that the bottleneck is due to the parallel infrastructure, so increasing the number of cores does not compensate for the synchronisation and communication costs.

Table 5.3: Experimental results Spark Streaming computational performance.

	Running Time (seconds)				Speedup			Efficiency		
Batch Size	T_1	T_{40}	T_{80}	T_{160}	S_{40}	S_{80}	S_{160}	E_{40}	E_{80}	E_{160}
64 MB	28.10	6.29	7.15	7.08	4.46	3.93	3.96	11.16%	4.91%	2.47%
128 MB	69.17	4.71	4.03	4.22	14.68	17.16	16.39	36.71%	21.45%	10.24%
256 MB	124.65	5.74	10.44	10.94	21.72	11.92	11.39	54.29%	14.92%	7.12%
512 MB	244.28	5.85	34.34	34.05	41.76	7.11	7.17	104.39%	8.89%	4.48%
1 GB	462.75	8.18	124.56	115.21	56.57	3.72	4.02	141.48%	4.64%	2.51%

Therefore, according to the results, the best configuration to obtain the maximum performance in the streaming classification process with Spark is observed when using the cluster resources with 40 cores and a batch size of 1 GB (Figure 5.14). In this regard, we can consider our Spark streaming classification methodology as a real-time classification since we can classify 1 GB in 8.18s, that is approximately 12,000,000 samples rows, which is equivalent to almost one week of daily patient activities monitoring (30 TBs in 2 days and 8 hours).

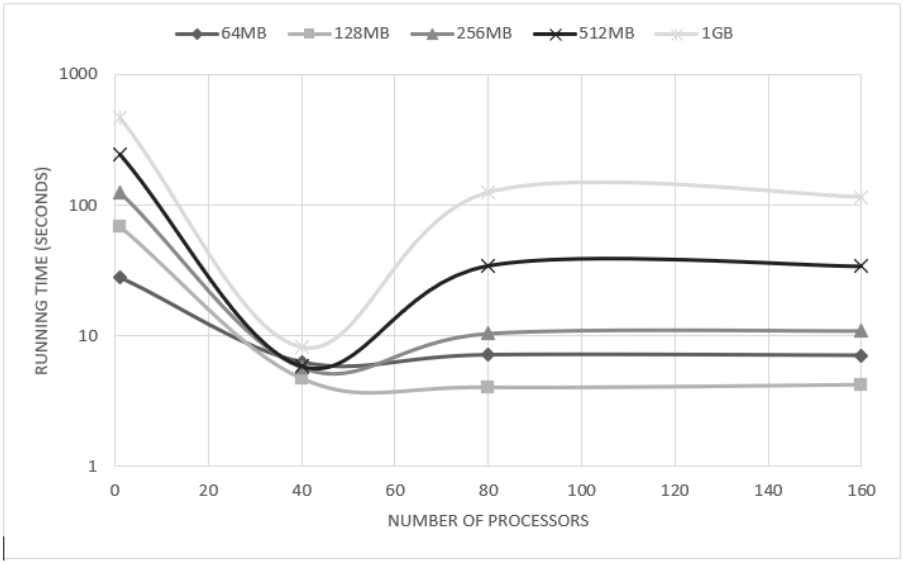


Figure 5.14: Running time in seconds (logarithmic scale) of the Spark Streaming process classification executed on 40, 80 and 160 cores in the cluster computing platform.

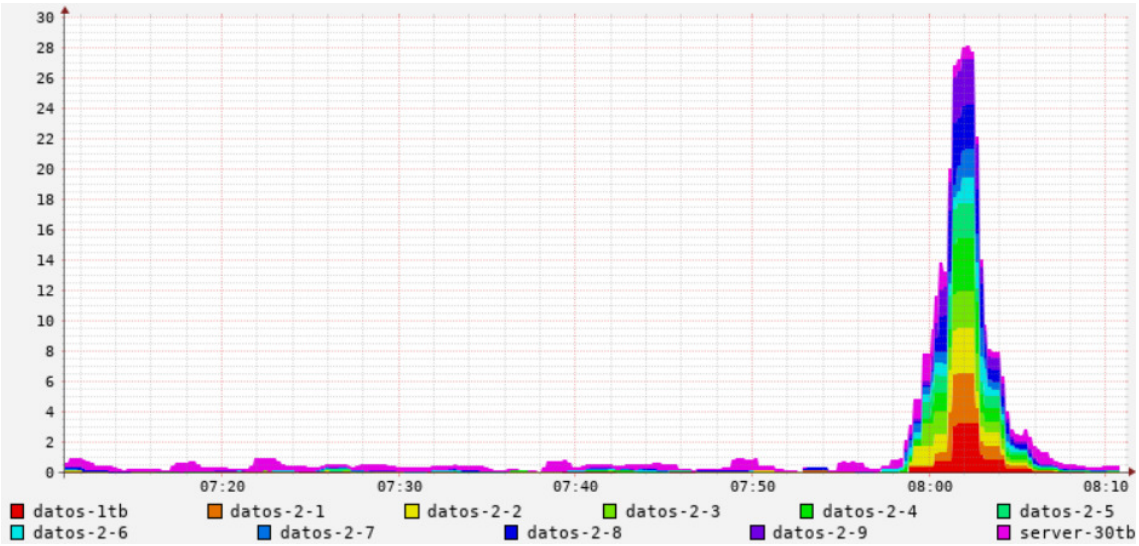


Figure 5.15: Load_one. Number of threads per node (40 cores configuration). Plot captured from Ganglia cluster monitoring system for the master node (server-30tb) and 10 worker nodes.

In terms of computational effort, we have plotted the *Load one* measure of the entire cluster while running experiments with 40 and 160 cores with a batch size of 1 GB in Figure 5.15 and Figure 5.16 respectively, to check the overall CPU load. In particular, the *Load one* computes the number of threads at kernel level that is running and being queued while waiting for CPU resources, averaged over the last minute. We could interpret this number with the number of hardware threads available on the machine and the time it takes to drain the run queue. Figure 5.15 captures a short time (close to minute 8:00) in which the Master node (Spark driver) delivers tasks to the worker nodes. They start to undertake data processing jobs when we run the experiment with 40 cores and 1 GB of batch size. The *Load one* measure in Figure 5.16 shows an increasing activity in minute 9:20 approximately, even more than in the previous experiment when increasing the number of cores to 160.

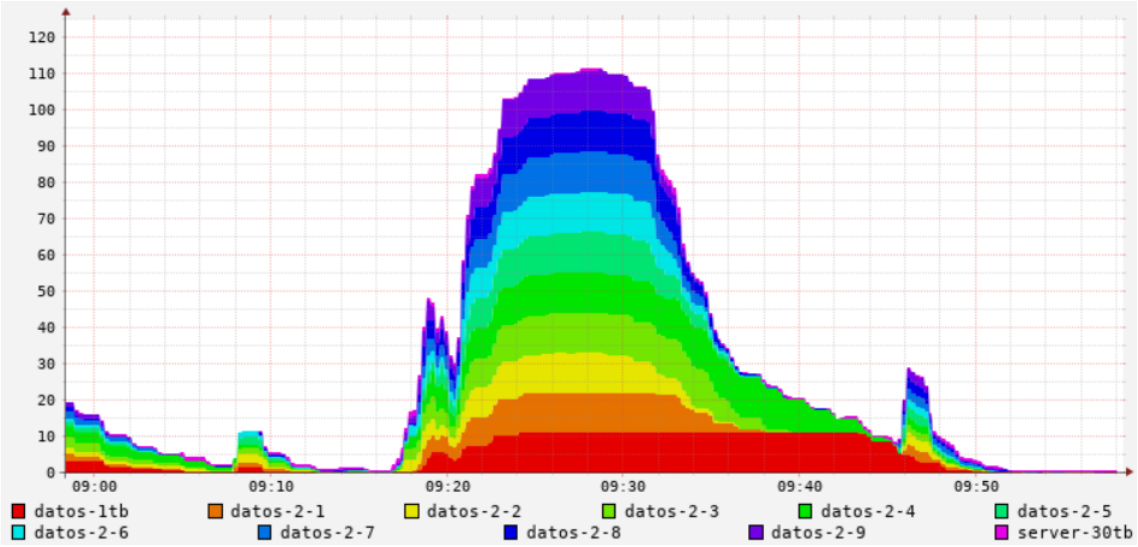


Figure 5.16: Load_one. Number of threads per node (160 cores configuration). Plot captured from Ganglia cluster monitoring system for the master node (server-30tb) and 10 worker nodes.

5.5 Discussion

In this section, to alleviate some of the drawbacks encountered in the literature, we have made an exhaustive study of general features in the existing methods, as exposed in [6, 285, 311, 312, 294, 313]. We have distinguished four main challenges in human activity recognition. These features are presented below:

- *Design issues*:

  1. *Cost*: Cost is a key factor for any technique. If accuracy of a solution is good but cost is too high, then it is of no practical use. Accelerometers are inexpensive, require relatively low power, and are embedded in most of today's cellular phones [314].
  2. *Obtrusiveness*: To be successful in practice, HAR systems should not require the user to wear many sensors nor interact too often with the application. There are systems which require the user to wear four or more accelerometers or carry a heavy rucksack with

recording devices. These configurations may be uncomfortable, invasive, expensive, and hence not suitable for activity recognition.

3. *Energy consumption*: extending the battery life is a desirable feature, especially for medical applications that are compelled to deliver critical information (Long term monitoring).
4. *Sampling rate (frequency)*: low sampling frequencies tend to lose information in specific movements.

- *Data collection protocol drawbacks:*

5. *Real-world environments (No controlled environment)*: The procedure followed by the individuals while collecting data is critical in any HAR. In [315] demonstrated 95.6% of accuracy for ambulation activities in a controlled data collection experiment, but in a natural environment (i.e., outside of the laboratory), the accuracy dropped to 66%!
6. *Large volume of data*: A comprehensive study should consider a large number of individuals.
7. *Long term patient monitoring*: most studies do not offer patient monitoring over time, which is essential to improve the problem of HAR.
8. *Data collection Flexibility*: people perform activities in a different manner which means that an acceptable number of subjects is needed for the study, so that the trained model is flexible enough to work with other subjects.

- *Model selection drawbacks:*

9. *Semi-supervised learning*: Typically, HAR systems rely on large amount of labeled training data. However, annotating data can be challenging in some situations, especially when the granularity of the activities is great or the user is unwilling to help with the gathering process. Using semi-supervised learning, these unlabeled data can still be used to train a recognition model.
10. *Deep learning*: Deep learning algorithms attempt to learn high-level features from data in an incremental manner. Nevertheless, in classical machine learning, domain experts must extract features from raw sensor data in order to make the patterns more visible for the learning algorithm.

- *Model evaluation drawbacks:*

11. *Model generalisation*: People certainly perform activities in a different manner due to particular physical characteristics. We have proposed to evaluate activity recognition systems based on the subjects rather than of the segmented windows. This prevents over-fitting on the subjects and helps to achieve better generalisation results.
12. *Latency*: Latency is a critical factor. If a solution is accurate but takes long time to provide the results, it is not practical.
13. *Real time classification/real-time decision making*: This is important for human activity recognition because getting the results in real time is a compulsion in many situations.

Table 5.4 shows a comparison between our approach and a set of related works found in the literature of HAR in this section, according to the list criteria exposed above. Desirable features related to real-world environments such as real-time processing of the sensorised data, dealing with unlabeled data and managing high volumes of continuously sensorised data are covered by our approach, which represents an advantage with regards to these compared works.

Table 5.4: Comparison of related works found in the literature on Human activity recognition. The comparison has been made according to four main challenges encountered in state of the art on human activity recognition. Additionally, our *Streaming Semi-Supervised Deep-Learning Approach* is presented in this table as *Proposal*. It is worth noting that our approach represents an advantage regarding these compared works in terms of real-time classification in real-world environments.

Features/HAR refs	[287]	[288]	[289]	[290]	[276]	[274]	[278]	[279]	[291]	[284]	[292]	*Proposal*
1. *Cost*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2. *Obtrusiveness*	✓	✗	✓	✗	✓	✓	✓	✓	✓	✗	✓	✓
3. *Energy*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4. *Sampling-rate*	✗	✗	-	✓	✗	✗	✗	✗	✓	✓	≈✓	✓
5. *Real-environment*	✗	≈✓	✗	✓	-	≈✓	✗	✓	✗	✗	✓	✓
6. *Large data-volume*	✗	✗	✗	✓	✗	✗	✗	-	✗	✗	✗	✓
7. *Long-monitoring*	✗	✗	✗	≈✓	✗	✗	✗	✗	✗	✗	✗	✓
8. *Data-flexibility*	✗	✗	✗	✓	✗	✗	✗	≈✓	✗	-	✗	✓
9. *Semi-supervised*	✗	✗	✗	✗	✗	≈✓	✗	✗	✗	✓	✓	✓
10. *Deep-Learning*	✗	✗	✗	✓	✓	✓	✗	✓	✓	✓	✓	✓
11. *Model-generalization*	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✓
12. *Latency*	-	✓	-	-	✗	≈✓	✗	✗	✗	-	✗	✓
13. *Real-time classify*	-	✓	✗	≈✗	✗	≈✓	✗	✗	✗	✗	✗	✓

5.6 Conclusions

This chapter presents a novel approach for obesity patient monitoring in healthcare systems. It comprises a combination of public (labeled) and private (unlabeled) raw data integration, semi-supervised classification with CNN Encoder-Decoder and Spark streaming processing with a sliding window to allow continuous activity recognition. This work has been validated in the context of a real-world case study with a group of 300 overweight patients in the healthcare system of Andalusia (Spain), classifying close to 30 TBs of accelerometer sensor-based data in real-time conditions, which is crucial for long-term daily patient monitoring.

The experimental results demonstrate that our proposed method can achieve significant Fm-scores training the model even with 100% of unlabeled data (proportion of data labeled used for the train) since, from this point, the results decrease below to 0.8 of Fm-score. Finally, we choose the amount of 80% of unlabeled data since, at this percentage, the model reaches a trade-off result (Fm-score = 0.90) between Fm-score and the amount of unlabeled data added to the model. Moreover, to demonstrate our model's performance, we observe that the distributions of the predictions in unlabeled data are reasonable, as shown in Figure 5.10.

In addition, we implement a Spark streaming process for the activity classification in a cluster computing platform to be able to classify the raw data sensor in real-time. For this proposal, we found the best configuration to minimize the running computation time of the streaming classification, using the cluster with 40 cores and predicting with streaming batch size of 1 GB, being able to classify one week of daily patient monitoring in approximately 8 seconds.

Chapter 6

Contribution to explainable artificial intelligence for biomedical image classification

In the last few years, eXplainable Artificial Intelligence (XAI) has attracted attention in data analytics. It shows excellent potential in interpreting the results of complex machine learning models in applying medical problems. In a nutshell, the outcome of the machine learning-based applications should be understood by end users, especially in the medical data context, where decisions must be carefully taken. Many efforts have been carried out to explain the outcome of a deep learning complex model in processes where image recognition and classification are involved, as in the case of skin lesions and Melanoma cancer. This chapter represents a first attempt (to the best of our knowledge) to experimentally and technically investigate the explainability of modern XAI methods, such as, Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive exPlanations (SHAP), in terms of reproducibility of results and execution time on a Melanoma image classification dataset. This chapter shows that XAI methods provide advantages in model result interpretation in Melanoma image classification. Concretely, LIME performs better than SHAP gradient explainer regarding reproducibility and execution time.

6.1 Introduction

Explainable AI is an artificial intelligence approach oriented to explain the results of complex machine learning algorithms [316]. Generally, it is believed that as the complexity of a machine learning algorithm increases, the understandability of the results become harder [317]. Previously, the robustness of a classification algorithm was evaluated using well-known criteria such as accuracy, precision, recall, F-score, and etc. However, in real-world scenarios, human experts usually prefer the use of understandable algorithms, even though they usually have moderate, sometimes limited, performance that other complex black-box techniques, such as deep learners. The search for explainability can lead to a loss of algorithm performance. Therefore, explainability aims to extract information from models that may have millions of parameters and present it in a form that is understandable to the human mind [318].

In fact, explainability besides accuracy are two important factors to assess the output of any machine learning algorithms [319]. One of the main categories of explainers are post-hoc model-agnostic. Post-hoc refers to those methods that are applied after training the model and not at the middle of the model training process. Model-agnostic refers to the group of explainers that are not specifically designed for a certain machine learning algorithm. XAI specifically well-adapted to provide explanation ability to deep learning output on medical datasets [320], where Melanoma cancer is not an exception.

Melanoma is the most aggressive skin tumour, with a 5-year survival rate of 93% if diagnosed in early stages, but only 27% if diagnosed at an advanced stage with the presence of metastatic disease[1]. In Spain, 6,108 cases of melanoma were estimated in 2021 (2,480 men and 3,678 women), being the fifth most frequent cancer in men and women [2]. Diagnosis in the early stages allows for better survival rates. However, it entails the difficulty of differentiating it from other pigmented skin lesions (nevus and seborrheic keratosis, mainly), which are followed up. Including artificial intelligence in the diagnosis would allow a more accurate diagnosis. In concrete, there are many efforts to melanoma diagnosis using deep learning [321, 322]. To realize trustworthy AI, XAI can be used as a technical method to ensure transparency of deep learning by helping better understand the neural network's underlying mechanisms and explaining system behaviors to users (in our case, clinicians).

This chapter is intended to be, to the best of our knowledge, a first attempt to evaluate two well-known post-hoc model-agnostic methods in XAI, namely: Local Interpretable Model-Agnostic Explanations (LIME) [323] and SHapley Additive ex-Planations (SHAP) [324], on explaining the deep learning prediction on skin lesion and Melanoma image dataset technically. Reproducibility and execution time are two major criteria for comparing LIME and SHAP.

This chapter finally concludes on which of the aforementioned method is most suitable for the explanation of Melanoma detection from an engineering point of view. The rest of this chapter is organized as follows: Section 6.3 provides related information for LIME and SHAP. Section 6.4 demonstrates the methodology and the results achieved. Finally, Section 6.5 concludes the chapter by summarizing the findings.

6.2 Related works

In the recent past, clinical researchers are increasingly using XAI methods for medical image classification to explain the output of their black-box models. A deep learning-based approach based on CNN is often used for medical image analysis. Thus, the main objective is to give a good

[1] Melanoma Cancer statistics approved by the Cancer.Net Editorial Board, 01/2021 https://www.cancer.net/cancer-types/melanoma/statistics

[2] https://seom.org/images/Cifras_del_cancer_en_Espnaha_2021.pdf

explanation of how the model came to its decision and/or can make the decision understandable. In this regard, decision-making systems should provide straightforward and explainable decisions to foster transparency and trust to the clinicians to make the in making the correct diagnosis [325].

Many XAI methods have been used in previous studies, specifically post-hoc approaches to improve the comprehensibility of deep learning models decisions. In this sense, [326] implements two post-hoc methods (LIME and SHAP), and other alternative XAI method called Contextual Importance and Utility (CIU) [327] to provide explainable decision support for in vivo gastral images. LIME is used in [328] to explain specific predictions of heart disease. Moreover, this paper uses Grad-CAM to apply local explanations to image classification models, such as CNN for predicting retinopathy. Also, XAI has been applied in the context of medicine for time series analysis, as mentioned in the survey [329], where numerous post-hoc methods aim to explain CNNs.

In [330], the authors present many XAI techniques used in medical image analysis [331, 331, 332, 333, 334, 335], paying particular attention to visual explanation, also called saliency mapping, as it is the most common form of XAI in medical image analysis. Moreover, in [336], a systematic review was conducted to investigate XAI in skin cancer screening, in which 37 publications were found. However, most of these studies were limited to applying current XAI methods to their classifier to interpret their decision-making. In concrete, more than half of the articles only used XAI algorithms superficially to depict that the models concentrated on relevant image areas. Moreover, fourteen articles evaluated the outcomes of XAI methods with the help of human supervision. For example, in [337, 338], non-medical graduate students were trained with a short tutorial on skin conditions in order to evaluate the XAI methods decision. In [339, 340], the method was evaluated by pathologists, while in [341] was evaluated by many dermatologists, and in [342] by the authors.

Therefore, these articles investigate the effect on diagnostic accuracy and dermatologists' acceptance by evaluating the results of the algorithms with human supervision. However, these research studies do not lend themselves to statistical evaluation, even though new statistical methods or validation metrics are required for XAI methods. In this sense, since classification tasks and study objectives are very heterogeneous, we propose to numerically evaluate the decision-making of the XAI methods in terms of reproducibility and execution time and compare the performance of the algorithms.

6.3 Preliminaries

This research focuses on the model-agnostic AI explainers, which provide post-hoc interpretability i.e. why the prediction model predicted its output through providing after-the-fact evidence for the outputs. These explainers are probably the most popular ones in the current literature, which consist in Local Interpretable Model-Agnnostic Explanations (LIME) [323] and SHapley Additive exPlanations (SHAP), both comprising a group of techniques that help humans visualize what an already-trained model thinks is important.

LIME uses Equation 6.1 to minimize $\xi(x)$ so that f is the prediction model which is assumed as black box, g is a model in G as a class of potentially interpretable models that tries to approximate f, Π_x is used to define the locality around the sample to be explained (perturbations from x), and $\Omega(g)$ represents the complexity of explanation that should be minimized as well as $L(f, g, \Pi_x)$.

$$\xi(x) = argmin_{g \in G}\ L(f, g, \Pi_x) + \Omega(g) \tag{6.1}$$

SHAP values are concepts coming from game theory [324]. Shapely quantifies to what extent each player (features) contributes to the game (output of prediction model). Shapely creates a power set of features firstly. The cardinality of power set is 2^n where n is the total number of features. Likewise, SHAP also requires to train 2^n models with different set of features according

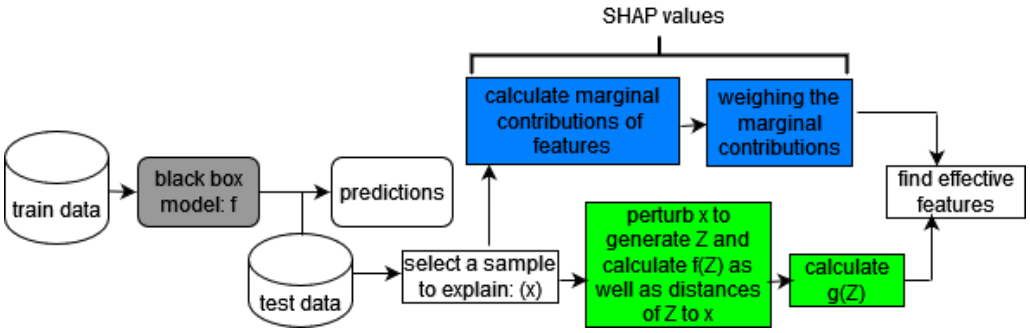


Figure 6.1: General workflow of SHAP and LIME.

to the power set. It is obvious that as the number of features is higher the number of models to be trained increases exponentially, which is treated by Lundberg et al.[324] through some approximations and samplings. Basically, calculating SHAP values has two steps, namely calculating marginal contributions of each feature and weighing the marginal contributions which can be shown in general in Equation 6.2, so that F is the entire number of (f) features and $set = 1, ..., F$.

$$SHAP_f(x) = \sum_{f \in set} [|set| \times \binom{F}{|set|}]^{-}1[Predict_{set}(x) - Predict_{set/f}(x)] \tag{6.2}$$

Figure 6.1 illustrates the difference between SHAP and LIME in general. According to this figure, LIME initially perturbs the sample to explain x to create the set $Z = z_1, z_2, ..., z_m$. Next, it selects an interpretable model (such as linear regression) to calculate the importance of features (calculating the coefficients related to each feature) via $g(Z)$. LIME finally selects the most effective features (through sorting coefficients if g is linear regression). However, SHAP builds SHAP values by calculating the marginal contribution of features and weighing them. Effective features are those with greater SHAP values. Moreover, summing the SHAP values gives exactly the difference between the output of full model and null model, which shows the additive explanations of SHAP.

While SHAP explainers are model agnostic, there exists two variations that could be used for deep learning, namely deep explainer and gradient explainer. Deep explainer approximates the conditional expectations of SHAP values using a selection of background samples, while gradient explainer explains a model using expected gradients which reformulates the integral as an expectation and combines that expectation with sampling reference values from the background dataset.

6.4 Proposed methodology

The methodology of this chapter is illustrated using a pipeline in Figure 6.2. The image dataset is online available in Kaggle Skin Lesion Images for Melanoma Classification (ISIC2019) repository[3]. It comprises more than 25,000 images with imbalanced classes (the majority of training data is nevus) which could cause an erroneous accuracy and incorrect predictions. There are many methods to balance training data including undersampling the majority class, oversampling the minority classes, applying SMOTE, and etc depending on the dataset. However, our experiments reveal that the best technique for image datasets like Melanoma is the combination of random oversampling the minority classes following by applying data augmentation.

[3]In URL https://www.kaggle.com/andrewmvd/isic-2019

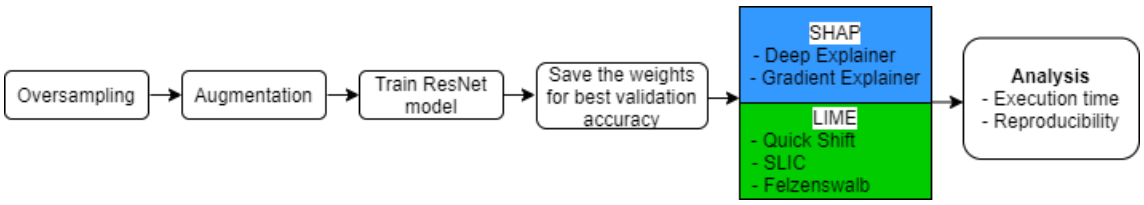


Figure 6.2: In the proposed methodology, as a first step, we propose to deal with the data imbalance problem through an oversampling strategy in the minority classes. After that, we apply data augmentation creating variations of the images that can improve the ability of the fit models to generalize what they have learned to new images. We have used a pre-trained mode (ResNet) to perform the image classification. Once the model is trained, its weights are saved for later classifications in new images. At this point, we will apply explainability algorithms (LIME and SHAP) to find the most critical features taken by the model to make predictions. Finally, we will evaluate and compare the algorithms according to their results' reproducibility and execution time.

Table 6.1: Melanoma dataset description after oversampling class imbalance.

Data	#Observations	Distribution of observations
Train	2,000	374/Melanoma, 1372/Nevus, 254/Seborrheic-keratosis
Validation	150	30/Melanoma, 78/Nevus, 42/Seborrheic-keratosis
Test	600	117/Melanoma, 393/Nevus, 90/Seborrheic-keratosis

Thus, in the preprocessing step, the distribution of classes were equalized using random oversampling initially. Oversampling solely can lead to overoptimism in prediction. Assuming the training data is split into train and validation sets. It is expected that some images in the training data appear in the validation set, since there exist multiple replicated images as a result of random oversampling the minority classes. As such overfitting could happen where the model prediction will be high in training data, but very low in unseen data. Here data augmentation could alleviate overfitting. The data augmentation in this study is done through rescaling, rotating, width-shift, height-shift, and horizontal-flip augmentation. The pipeline in Figure 6.2 follows by applying pre-trained ResNet [343] convolutional Deep Learning model and saving the best weights. Then, model agnostic post-hoc explainers (SHAP with Deep and Gradient explainers, LIME with three well-known segmentation algorithms) are used to evaluate the results based on reproducibility of the results and execution time.

Reproducibility means the ability of the method to successfully reproduce same explanations in multiple runs. Likewise, execution time refers to the elapsed time starting from creating the explainer until calculating the explanation and generating the pictorial results. Table 6.1 also shows the main characteristics of the Melanoma dataset prior to oversampling and augmentation. After oversampling the distribution of each class in training set is equal to 1,372 so that the the entire training set contains 4,116 observations.

6.4.1 Evaluation

This section provides related information for calculated metrics. All the experiments have been conducted in a virtualization environment running on a private high-performance cluster computing platform. Our virtualization platform is hosted in the computational environment mentioned in Chapter 2 Section 2.5. Concretely, this platform is made up of a CPU with Intel(R) Xeon(R) Gold 6130 @ 2.10GHz, maximum 2 TB of HDD, maximum 64 GB of RAM, and Ubuntu 20.04.3

Table 6.2: Description of four selected samples for experimentation.

Test observation	Real label	Melanoma	Nevus	Seborrheic-keratosis
Sample 1	Nevus	0.31	0.57	0.12
Sample 2	Melanoma	1.00	0.00	0.00
Sample 3	Nevus	0.00	1.00	0.00
Sample 4	Seborrheic-keratosis	0.00	0.00	1.00

LTS(GNU/Linux 5.4.0-1049-kvm x86_64).

Since it is impossible to illustrate the entire test samples four test samples were selected to investigate the reproducibility and execution time analysis as explained in Table 6.2, so that for each sample the real labeling and the prediction of deep learning are shown.

6.4.2 Evaluation of LIME

Figure 6.3 illustrates the reproducibility of LIME using three well-known segmentation algorithm namely, quick shift, Simple Linear Iterative Clustering (SLIC) and felzenswalb. Quick shift uses approximation of kernelized mean-shift and it belongs to the family of local mode-seeking algorithms. SLIC uses k-means which is a simpler clustering method in comparison with the clustering method in quick shift. In contrast, felzenswalb uses a graph-based approach for image segmentation.

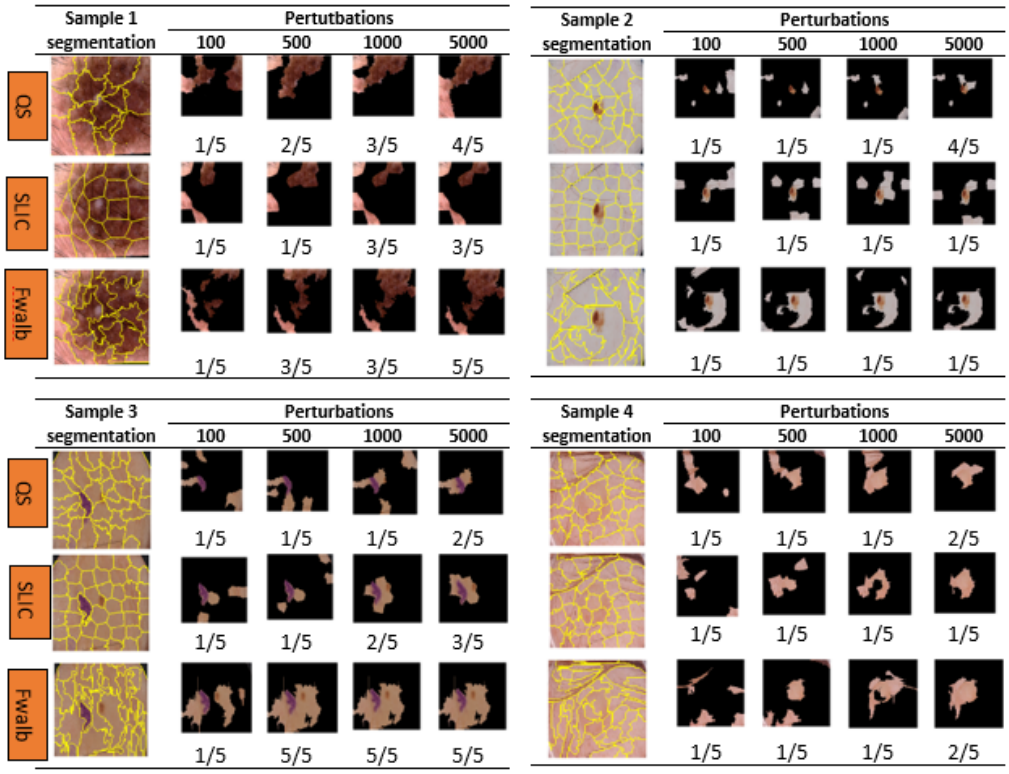


Figure 6.3: Strict analysis of LIME reproducibility by increasing number of perturbations.

Figure 6.3 is the result of 5 multiple runs of LIME algorithm for 5 top features with different

number of perturbations regarding each of the four images in Table 6.2. Figure 6.3 illustrate original segmentation for each image, so that the segmentation algorithms are tuned to contain roughly same number of segments for each algorithm, to have a fair comparison between them. Under each image is a fraction that shows how many times LIME is able to generate exactly same 5 top features in 5 multiple runs using each segmentation algorithm. For example, 4/5 for sample 1 with quick shift algorithm and 5,000 perturbations means the result of LIME in 4 runs from 5 runs are exactly same. As such, sample 1 achieves 1/5 for 100 perturbations using quick shift algorithm, which means that there are five unique results so that one of them is selected randomly.

It is noteworthy that quick shift and SLIC have relatively the same segmentation trend compared with felzenswalb, so this last sometimes results in segments with sizes that vary greatly, as in sample 2 and sample 3. This may affect the reproducibility of LIME either positively in sample 3 or negatively in sample 2.

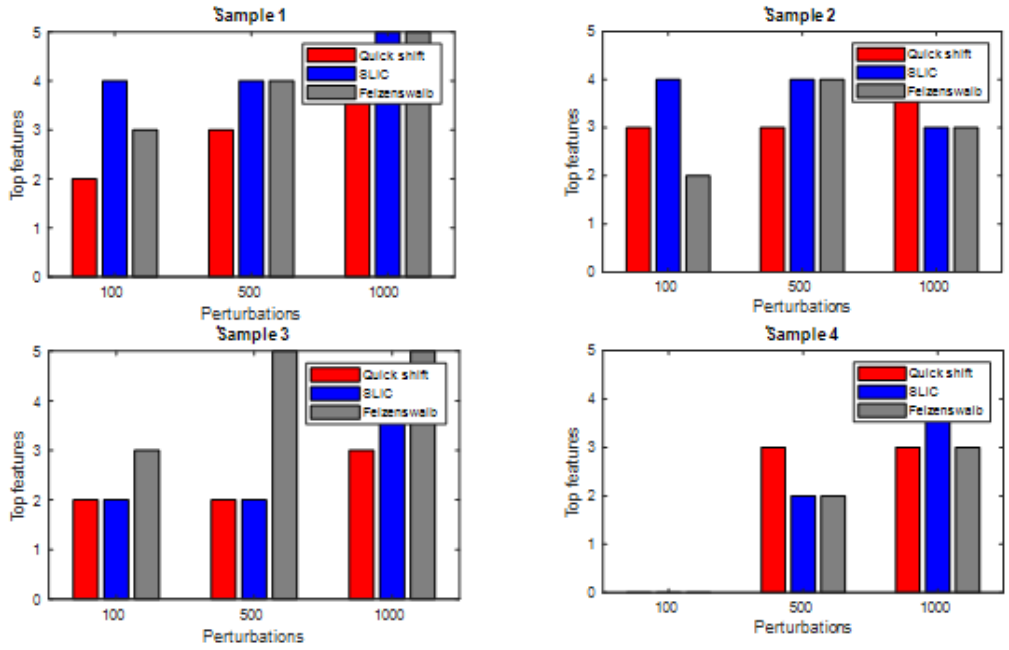


Figure 6.4: Gentle analysis of LIME reproducibility by increasing number of perturbations.

While Figure 6.3 analyzes reproducibility strictly, Figure 6.4 checks the reproducibility of LIME more gently by calculating the number of features in each perturbation (100, 500, and 1,000) that have also been observed when the perturbation is 5,000. Figure 6.4 shows that as the number of perturbation increases from 100 to 1,000, more features from that perturbation are observed within 5,000 perturbation. If two superpixels are equally good at explaining, LIME may pick an arbitrary one which sometimes result in not reproducible explanations. Figure 6.4 shows that by increasing the number of perturbation, LIME converges to reproducibility.

Recalling that good segmentation often depends on the application, illustrations in Figure 6.3 and 6.4 show that the reproducibility in LIME mostly increases while the number of perturbation increases from 500 to 5,000 using any segmentation algorithm (the default number of perturbation in LIME is 1,000). While increasing number of perturbations has a positive effect in reproducibility of LIME, another approach is to fix the random seed to initialize the random number generator. This way, using any number of perturbations the explainability results are same. Nonetheless,

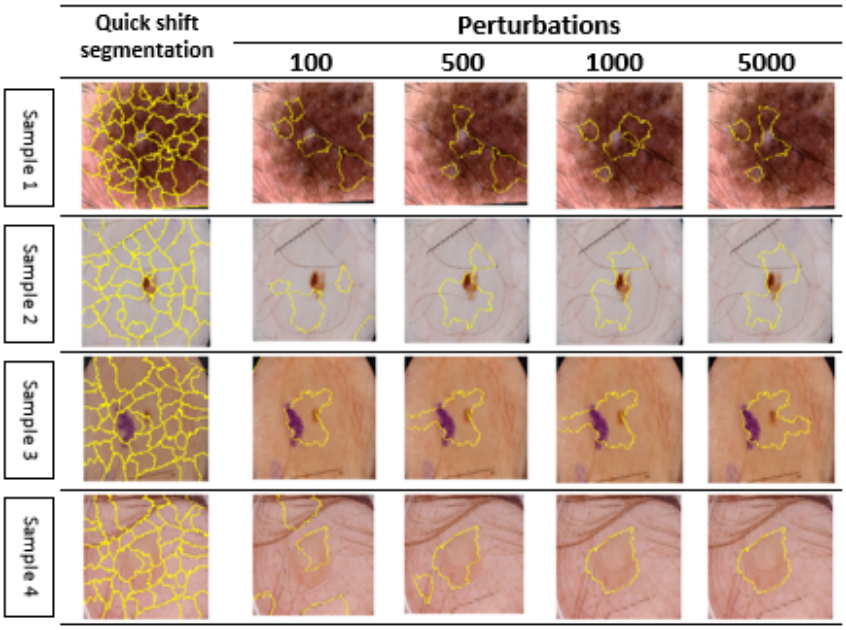


Figure 6.5: Reproducibility analysis of LIME using fixed random seed and variable number of perturbations.

greater number of perturbations together with fixed random seed result in better accuracy as well. Figure 6.5 shows how successfully LIME recognizes regions contributed to target label by increasing the number of perturbations and using fixed random seed. This last figure also reveals that LIME intelligently did not recognize *mm* scale and hair as effective features, but considers the stain in sample 3 within 5 most important superpixels.

6.4.3 Evaluation of SHAP

As commented before, there exists two variations of SHAP optimized for deep learning, namely gradient explainer and deep explainer. The SHAP kernel explainer could also be used because it works for all models, but is slower than the other model type-specific algorithms, as it makes no assumptions about the model type. Thus, to avoid redundancy of figures with same results and for the sake of hardware limitation (passing more than 100 background data was unreasonably expensive), the reproducibility of SHAP has been tested using solely with gradient explainer, shown in Figure 6.6. Generally, pink pixels contribute to the model output and blue pixels contribute not being of that class. The intensity of color shows the intensity of contribution. Since gradient and deep explainer explains the prediction using pixels and not superpixels, it is difficult to trace the reproducibility numerically as it was done for LIME.

The *nsamples* parameter in gradient explainer (by default = 200) indicates the number of samples are taken to compute the expectation and shows accuracy of explanation. This gives better estimates of SHAP values as the *nsamples* increases, which leads to low variate estimation of the SHAP values, however the execution time increases. Figure 6.6 shows that as the *nsamples* increases from 100 to 5,000 the explainability becomes a bit more reproducible, which is less obvious in sample 1 because the deep learning model is not completely sure about its prediction. Figure 6.6 also shows that gradient explainer considers the stain in sample 3 same as LIME in Figure 6.5.

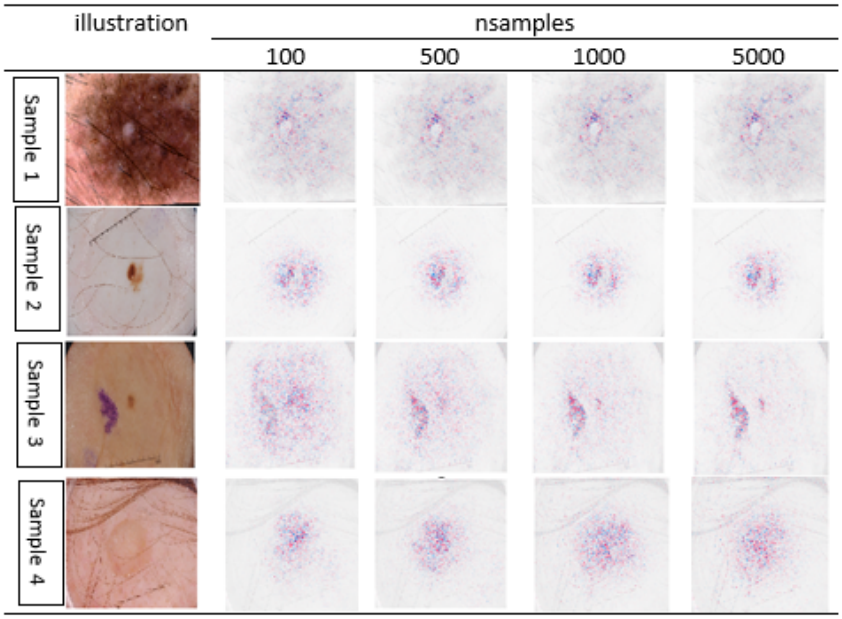


Figure 6.6: Visual reproducibility analysis of SHAP gradient explainer.

The gradient explainer in Figure 6.6 uses the entire 4,116 images in train set as a background data (the random seed in calculation of SHAP values is set to 42).

6.4.4 Computational effort

From the point of view of the computational effort, Figure 6.7 compares LIME (using quick shift) and SHAP gradient explainer in terms of execution time, so that N is the number of perturbation and *nsamples* for LIME and SHAP, respectively.

It is clear that LIME spends less amount of time for explainability as N increases, while SHAP gradient explainer is almost three times slower than LIME. It is noteworthy mentioning that changing the segmentation algorithm does not have a considerable difference in execution time of LIME. SLIC is very competitively faster than quick shift and also quick shift is very closely faster than felzenswalb. Thus, Figure 6.7 the better performance in terms of execution time of LIME using quick shift as a moderate segmentation algorithm. Technically speaking, LIME has more reproducibility power and is almost much faster than SHAP gradient on Melanoma dataset. Thus, there are sufficient engineering justifications to use LIME for explainability of deep learning on melanoma dataset for a single prediction rather than SHAP gradient explainer.

6.5 Conclusion

This chapter is devoted to investigate the explainability of Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) in order to help in the differential diagnosis of pigmented skin lesions. The evaluation criteria focuses on the reproducibility of the results, as well as the execution time. Three variations of LIME (using three well-known segmentation algorithms) are used and gradient explainer is selected for SHAP. From the engineering point of view, and in the context of the experiments conducted in this study, LIME performs faster than

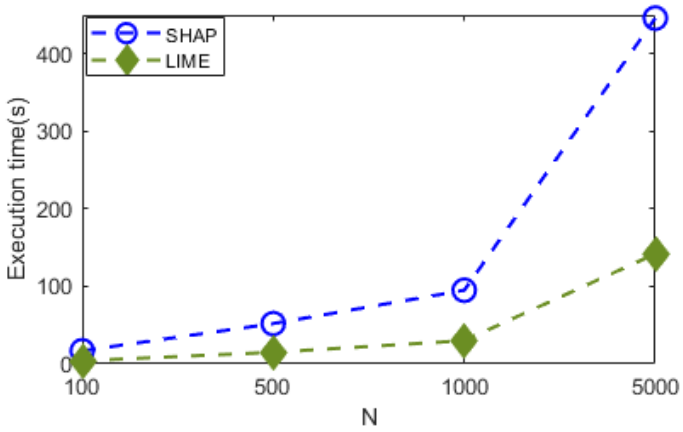


Figure 6.7: Efficiency of LIME Vs SHAP. Execution time versus N, where N represents the number perturbations in the case of LIME, and the number of samples for SHAP. Should be noted that LIME performs better than SHAP in terms of computational effort.

SHAP. The idea is that while acceptable results are achieved by LIME in the case of differential diagnosis of pigmented skin lesions, there is no need to use SHAP because of its expensive efficiency. LIME works with super pixels and the reproducibility of results were more controllable than SHAP gradient explainer. Thus, it can be concluded that XAI methods show potentials in providing interpretable results for the specific case of pigmented skin lesions classification, in the context of Melanoma cancer diagnosis. Specifically, LIME shows better performance than SHAP gradient explainer in terms of reproducibility and execution time.

Part III
Final observations

This chapter collects the final ideas of this PhD Thesis dissertation. Section 7.1 includes the conclusions from all the experiments and discussions conducted in supporting publications. Then, in Section 7.2, we present the future lines of investigation that we plan to research from the latter works.

Chapter 7

Conclusions and future works

7.1 Conclusions

Biomedical research is constantly growing, as many technological advances have been made to expand databases, from next-generation sequencing to other biomedical data available, such as biomedical imaging, annotation metadata, IoT resources, etc. Due to the large amount of clinical data generated at different scales and in multiple dimensions, more comprehensive analyses can be performed to improve patients' quality of life and prevent or detect diseases. These data can be combined with other problem-related data sets. Such data integration can be beneficial since studying such a diverse heterogeneous data set from different sources of information can reveal interesting patterns or additional information than if the data were analyzed independently.

In a nutshell, and as far as the subject matter of this work is concerned, the integration and transmission of data among clinical organizations are of the utmost importance. Moreover, in the last decades, the development of Artificial Intelligence algorithms has played an essential role in biomedical research. It has demonstrated outstanding abilities for interpreting and analyzing large scales of clinical data and developing predictive models.

The present Thesis addresses a key challenge nowadays: the importance of integrating, combining and analyzing several sources of heterogeneous clinical data to support physicians in their decision-making. This approach refers to the combination of several of these clinical analyses in an integrated way. This clinical data integration allows more complex analyses with great potential to result in relevant breakthroughs in many life science fields. Some examples include identifying novel prognostic or diagnostic disease biomarkers.

In concrete, this Thesis probes into the main research questions and challenges formulated in the motivation of this work related to the implementation of precision medicine or AI-supported healthcare. Thus, this Thesis answers the question of how far we are (technologically speaking) from resolving these outstanding issues and also attempts to improve the existing data ecosystem to the point where these issues are addressed. In summary, the main contributions of this Thesis are shown as follows[1]:

- **Chapter 3. Contribution to flexible management and analysis of heterogeneous biomedical data**. In the context of clinical data management systems, there are continuing limitations in the data acquisition process due to the heterogeneous nature of clinical data. Most systems found in the state-of-the-art can not cover these limitations. Also, few of these

[1]The relationship between the conclusions and the research questions formulated in the motivation of this Thesis (Chapter 1) is quite self-explanatory. However, these relationships are highlighted with the corresponding references in parenthesis, e.g. (Q1), (Ch1), etc.

tools allow the analysis of clinical data for disease diagnosis. Answering the question Q1 exposed in the motivation of this Thesis, the necessity to align clinical data to innovative AI analysis strategies has stimulated developments in data integration, analytics tools implementation and knowledge representation in heterogeneous and predictive decision-making analysis. In this regard, we have designed and implemented a new software tool for flexible management and analysis of biomedical data from multiple sources (FIMED) to cover the first challenge (Ch1) and the abovementioned limitations. It allows the integration of clinical trials in different formats for their subsequent analysis. FIMED internally implements a workflow with several components for data collection and management with adaptability to multiple trials, data analysis and data visualization. In this sense, FIMED allows the clinical researcher to perform a complete clinical assay workflow since it allows the design of customized and easily modifiable electronic forms thanks to the flexibility provided by its NoSQL database engine. This way, FIMED allows the data collection process to be carried out incrementally without redefining the schema. It also provides fast data availability and allows the storage of different samples associated with the patient to provide additional information and even to perform more exhaustive analyses.

FIMED includes several analysis tools to assist the clinical expert in detecting diseases or searching for new biomarkers. These include gene expression analysis algorithms, gene regulatory network analysis, and data visualization to annotate gene functionality and identify core genes. Implementing these analysis tools was our first attempt to answer the research question (Q2) and meet the second challenge (Ch2) proposed in this Thesis. The motivation was to improve the inference and interpretation of complex biological networks to extract relevant information for disease diagnosis. Besides, FIMED has been validated in a real-world use case with authentic expression data from metastatic Melanoma patients. It is worth noting that FIMED has been updated several times and is under continuous maintenance. In its latest version (FIMED 2.0), new optimization analysis and visualization tools were added to reconstruct gene regulatory networks to facilitate disease diagnosis for the clinical expert, as exposed in (Ch2). This new version allows the practitioner to use four network construction methods: data assimilation, linear interpolation, tree-based ensemble or Gradient Boosting Machine regression. In addition, research has been done to perform an ensemble of these algorithms for reconstructing gene regulatory networks to obtain more stable and reliable results. It is worth noting that FIMED has been designed to offer mechanisms to extend the software with new components to broaden its functionalities. It has been evident how this tool can be easily integrated into different use cases, making FIMED a robust clinical research tool for data management, analysis and visualization in clinical assays in other studied diseases. Apart from the public instance provided[2], the project can be deployed by IT administrators in any health information system, ensuring higher protection of the clinical data.

- **Chapter 4. Contribution to the reconstruction of gene regulatory networks with multi-objective optimization**. In the context of GRNs reconstruction, clinical experts can investigate the functionalities of biological molecules such as genes, proteins, and RNA and their organization in living organisms. In the same way, practitioners can observe the behavior and organization of the components of living cells and see the relationship between them in molecular processes to decipher the mechanisms of multiple genetic diseases such as cancer or diabetes. In this regard, numerous recent studies have intensified the use of particle swarm optimisers to deal with the inference of GRNs. However, there is still a lack of proposals based on multi-objective formulations. Hence, the primary motivation of this contribution

[2] https://khaos.uma.es/fimed/

is to cover the challenge (Ch3) proposed in Chapter 1, in which new computational and experimental methodologies are required to explain biological networks due to the complexity of the high-dimensional nature of biomedical data. In this sense, this work proposes to apply and evaluate a representative set of multi-objective particle swarm optimisers (MOPSOs), which use different archiving strategies (hyper-volume and aggregation) and, consequently, different approaches for the selection of leaders in the context of the inference of GRNs. Therefore, this work attempt to obtain unbiased conclusions concerning which of them (and other related MOPSOs) could be used by experts in studies in silico/in vivo to find new possible gene interactions taking part in genetic regulations.

The optimal tuning of parameters in S-Systems is currently dealt with metaheuristics for continuous optimization. In this regard, a series of MOPSO variants, namely OMOPSO [36], MOPSO [37], VEPSO [38], SMPSO [39], DMOPSO [40] and MOPSOHv [41], have been adapted to deal (for the first time) with the inference of GRNs. These techniques have been selected as they constitute a heterogeneous set of multi-objective optimizers, performing different learning procedures and inducing different behaviors. In this sense, a thorough experimental comparison has been carried out on gene expression data from benchmarking networks of the DREAM3 and DREAM4 challenges [42] based on real organisms (E.Coli and Yeast). Finally, meaningful experiments are conducted to infer networks from IRMA in vivo samples and from Melanoma cancer samples of actual patients. The ability to reproduce biological behavior is assessed in terms of algorithmic convergence and diversity and in terms of precision of inferred networks regarding gold standards.

- **Chapter 5. Contribution to time series streaming data analysis with biomedical data from sensors devices**. In biomedical time series analysis, Deep Learning techniques are considered powerful tools and enable extracting the most predictive features from complex datasets. A key feature that differentiates Deep Learning from other Machine Learning techniques is its ability to learn representations directly from structures without using predefined structure descriptors. This ability eliminates the need for conventional feature selection and reduction processes. However, Deep Learning approaches require large amounts of data and labeling these data is costly and time-consuming. Data labeling is challenging when dealing with real-world problems in uncontrolled environments and even more so when working on Big Data use cases where a minimal amount of labeled data and a massive amount of unlabeled data are considered. In this sense, the main contribution of this work is to address the third challenge of this Thesis (Ch3), in which we propose new AI strategies to deal with small labeled and no labeled clinical datasets. An excellent way to address these problems is to adopt a semi-supervised approach, which can employ unlabeled data with a small number of labeled examples. Therefore, a streaming semi-supervised HAR strategy has been proposed for monitoring overweight patients in a real-world healthcare system involving a data fusion task of accelerometer-sensorised data from labeled/unlabeled samples. Specifically, this work aimed to classify the daily activities of 300 patients, equivalent to 30 TBs of private raw movement data. However, no labeled data were available in our dataset. For this reason, a set of state-of-the-art datasets in the HAR problem environment has been collected and integrated to use as publicly labeled data. For the data integration, a thorough methodology based on interpolation, normalization, data resampling, and class imbalance techniques have been used since the data have been collected from different sources or devices, in other formats and at distinct sampling frequencies (related to the first challenge, Ch1).

  In addition, a preliminary study has been performed to observe which datasets had a similar distribution to our unlabeled dataset. Then a semi-supervised CNN-Encoder-Decoder model was trained with public labeled and private unlabeled data, which can learn the most relevant features of the unlabeled data and then use it to classify the activities. In this regard,

the knowledge extraction from the unlabeled data through the unsupervised part of the model (Encoder-Decoder) is stored and used as a starting point for the model training in the supervised part. A thorough experimentation has been conducted for model selection and validation, where this strategy has been evaluated with varying amounts of unlabeled data. The resulting analysis workflow is deployed on a cluster of Spark nodes, so the continuous classification of 30 TBs sensor data is predicted for a group of patients. The proposed HAR streaming deep-learning approach properly classifies movement patterns in real-time conditions, which is crucial for long-term daily patient monitoring. It represents a step forward to meet the challenges identified in a recent survey [6], which mainly consists of the generation of real-time activity recognition platforms and the development of more accurate unsupervised modelling for this problem. Therefore, we can conclude that our data collection and integration strategy, together with our semi-supervised deep learning on Spark stream processing, is a solution in this direction.

- **Chapter 6. Contribution to explainable artificial intelligence for biomedical image classification**. In the context of AI in medicine, numerous AI techniques have recently achieved great predictive success for many biomedical applications. However, in many cases, explaining the clinical outcome of highly complex models is challenging. For this reason, this work proposes to study and develop additional techniques that allow us to clarify the results of these black-box models, which is essential for the clinical domain where decisions will affect patients' lives, as discussed in the fourth challenge of this Thesis (Ch4). The main contribution is to provide the clinical expert with the ability to interpret the results obtained by the algorithms. In this regard, we have developed a methodology to assess the quality of explainability algorithms against a set of metrics in the case of a Melanoma image classification dataset. First, a pre-trained algorithm (Resnet) classified melanoma skin cancer images for early disease detection. Second, two of the most widely used post-hoc explainability algorithms (LIME and SHAP) were applied to explain and validate the results obtained by the classifier. These algorithms returned, as a result, the Melanoma images with the most critical features (super/pixels) of the image to perform the prediction. Finally, at the core of this work, we set out to experimentally and technically evaluate these algorithms in terms of reproducibility and runtime results. In both cases, LIME performs better than SHAP for this use case.

In summary, the publications related to this research have introduced solutions based on new algorithmic strategies and software tools that contribute to the development and improvement of precision medicine. In this respect, the present Thesis provides new software solutions and AI strategies that allow the clinical researcher to collect and integrate patients' clinical information and perform comprehensive analyses. Each of the scientific contributions outlined in this Thesis address some of the outstanding issues in the biomedical data ecosystem raised in this work's motivation. In this sense, Table 7.1 shows the relationships between published articles and the main challenges presented in this Thesis in a biomedical data ecosystem. From the clinical data management point of view, we have designed FIMED, which allows flexible collection, integration and management of biomedical data from multiple sources.

Moreover, we have proposed AI and optimization techniques and further analyzed them to provide the researcher with comprehensive analysis tools for disease detection or new drug development. From the point of view of applications, we have addressed several real-world problems in healthcare areas (such as gene expression data analysis for cancer detection, reconstruction of GRNs with multi-objective particle swarm optimizers, real-time patient monitoring through a deep learning approach with sensor-collected information, biomedical images classification using CNN and interpretability of the result employing XAI algorithms). Finally, these proposals provide

quality results and are intended to be more flexible and robust than solutions addressing the same kind of problem in the literature, showing the usefulness of our proposals to address issues that could arise in academia and industry.

Table 7.1: Summary of the relationships between the published papers and the main challenges in a biomedical data ecosystem.

Reference	Challenge (s)
FIMED [30]	Ch1, Ch2
MOPSOs GRNs [31]	Ch2
DL HAR [32]	Ch1, Ch3
FIMEDV2 [33]	Ch1, Ch2
XAI Melanoma [34]	Ch4

7.2 Future work

As future research lines in general, we plan to continue this proposal of integration and analysis of clinical data from different trials with the primary objective of improving the access and integration of heterogeneous health data. We also aim to enhance the existing analysis techniques by considering integrating data from different sources. In this sense, we want to continue working on the design and development of strategies that allow us to address the problems in a biomedical data ecosystem and thus continue to improve on the challenges that we defined in the motivation of this Thesis towards achieving precision and personalized medicine.

Furthermore, considering different areas of knowledge related to the contributions proposed in these studies, different lines of research have been identified for future work. This section presents some of the most outstanding ones:

- As future work for the first challenge (Ch1), we plan to continue upgrading FIMED to ensure future compatibility with more use cases. Thus, we will consider the adaptability to more gene expression file formats, other diseases and the integration with other analytical tools or algorithms (advanced GRNs algorithms, new ML approaches for real-time sensor data analysis and clinical image analysis, and new XAI algorithms). Moreover, we plan to continue researching new ways of integrating data into AI. Multi-modal integration allows more complex analyses with the potential for suitable breakthroughs in multiple biomedical fields. In this sense, we plan to develop algorithms that explicitly address the diversity of data and combine them by inferring a single model, as mentioned in the Figure in the motivation of this Thesis. This strategy attempts to integrate the data sources within the construction of the predictive model to combine multi-omics data, biomedical images and clinical patient information into a single robust model. This research line addresses the boundaries posed by the conventional approach of guiding ML analysis independently by combining a diverse data set and extracting significant conclusions from the integrated data.

- As future activities for the second challenge (Ch2), we are interested in adapting different optimizers, such as Differential Evolution, with specific parameters and operators for the efficient reconstruction of GRNs. To this end, modern auto-configuration techniques would aid in finding accurate tuning for GRNs. Besides, the design of new encoding and co-evolutionary strategies seems to be an optimistic line to enhance the predictive power of algorithms. In this regard, developing distributed parallel approaches could improve large-scale networks' performance. From the perspective of network modelling, we also plan to

work on new approaches requiring fewer parameters than S-System to be tuned. In addition, we want to study the integration of heterogeneous multi-omics data for the inference of GRNs. Although DNA micro-array data are commonly employed for network inference, the reconstruction of GRNs using only micro-array data is fundamentally limited since the information value of such data is constrained by technological and biological aspects. We consequently suggest researching more advanced techniques to more accurately reconstruct the structure and dynamics of GRNs by combining additional types of biological data, such as data from alternative experiments in a different format and databases.

- As a future research line related to the third challenge (Ch3), a semi-supervised approach has been proposed in this Thesis to help leverage unlabeled data from our dataset along with labeled data collected and integrated from the literature. Admittedly, the results are promising, as we have leveraged the knowledge from the unlabeled data to aid model learning. However, we plan to continue working on new strategies to improve the quality of the results. We plan to develop a more robust methodology based on transfer learning techniques for integrating accelerometry data from different sources and formats. We intend to make an automatic flow that chooses those datasets from state-of-the-art with the most similar data distribution to our private dataset and integrates them into a single dataset ready to be used (Ch1). We also plan to develop AI strategies using machine learning models to assist in automatic data labeling.

- As a future research line to cover the fourth challenge (Ch4), further research is required to support AI decision-making techniques by applying XAI methods. In this sense, we intend to investigate new metrics to evaluate the results obtained from XAI algorithms. As a first step, we propose that XAI results should be reproducible and replicable. Hence, the training model should produce consistent results, and also, the model should perform consistently even when trained with different samples of data. In addition, we plan to design novel techniques for the visualization of massive data to obtain accurate and comprehensible interpretations for the human expert. Moreover, from an algorithmic point of view, XAI's future work is to approach the explainability of deep learning on Melanoma data set through improving LIME, as well as to tacking with other different kind of medical image datasets.

List of Tables

List of Figures

Bibliography

[1] J. Xuan, Y. Yu, T. Qing, L. Guo, and L. Shi. "Next-generation sequencing in the clinic: promises and challenges". *Cancer letters* 340.2 (2013), pp. 284–295.

[2] E. D. Green and M. S. Guyer. "Charting a course for genomic medicine from base pairs to bedside". *Nature* 470.7333 (2011), pp. 204–213.

[3] W. Warwick, S. Johnson, J. Bond, G. Fletcher, P. Kanellakis, et al. "A framework to assess healthcare data quality". *The European Journal of Social & Behavioural Sciences* (2015).

[4] O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya. "Deep learning for healthcare applications based on physiological signals: A review". *Computer methods and programs in biomedicine* 161 (2018), pp. 1–13.

[5] M. A. Al-Garadi, M. S. Khan, K. D. Varathan, G. Mujtaba, and A. M. Al-Kabsi. "Using online social networks to track a pandemic: A systematic review". *Journal of biomedical informatics* 62 (2016), pp. 1–11.

[6] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu. "Deep learning for sensor-based activity recognition: A survey". *Pattern Recognition Letters* 119 (2019), pp. 3–11.

[7] E. De-La-Hoz-Franco, P. Ariza-Colpas, J. M. Quero, and M. Espinilla. "Sensor-based datasets for human activity recognition–a systematic review of literature". *IEEE Access* 6 (2018), pp. 59192–59210.

[8] P. Bet, P. C. Castro, and M. A. Ponti. "Fall detection and fall risk assessment in older person using wearable sensors: a systematic review". *International journal of medical informatics* (2019).

[9] D. A. Raptis, T. Mettler, M. A. Fischer, M. Patak, M. Lesurtel, D. Eshmuminov, O. De Rougemont, R. Graf, P.-A. Clavien, and S. Breitenstein. "Managing multicentre clinical trials with open source". *Informatics for Health and Social Care* 39.2 (2014), pp. 67–80.

[10] B. Krishnankutty, S. Bellary, N. B. Kumar, and L. S. Moodahadu. "Data management in clinical research: an overview". *Indian journal of pharmacology* 44.2 (2012), p. 168.

[11] Z. Lu and J. Su. "Clinical data management: Current status, challenges, and future directions from industry perspectives". *Open Access Journal of Clinical Trials* 2 (2010), pp. 93–105.

[12] D. M. Dilts and A. B. Sandler. "Invisible barriers to clinical trials: the impact of structural, infrastructural, and procedural barriers to opening oncology clinical trials". *Journal of Clinical Oncology* 24.28 (2006), pp. 4545–4552.

[13] H. Leroux, S. McBride, and S. Gibson. "On selecting a clinical trial management system for large scale, multi-centre, multi-modal clinical research study." *HIC*. 2011, pp. 89–95.

[14] B. Choi, S. Drozdetski, M. Hackett, C. Lu, C. Rottenberg, L. Yu, D. Hunscher, and D. Clauw. "Usability comparison of three clinical trial management systems". *AMIA Annual Symposium Proceedings*. Vol. 2005. American Medical Informatics Association. 2005, p. 921.

[15] J Holmes, L Sacchi, R Bellazzi, et al. "Artificial intelligence in medicine". *Ann R Coll Surg Engl* 86 (2004), pp. 334–8.

[16] Y. Mintz and R. Brodie. "Introduction to artificial intelligence in medicine". *Minimally Invasive Therapy & Allied Technologies* 28.2 (2019), pp. 73–81.

[17] S. Huang, J. Yang, S. Fong, and Q. Zhao. "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges". *Cancer letters* 471 (2020), pp. 61–71.

[18] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman. "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities". *Information Fusion* 50 (2019), pp. 71–91.

[19] L. Rundo, C. Militello, S. Vitabile, G. Russo, E. Sala, and M. C. Gilardi. "A survey on nature-inspired medical image analysis: a step further in biomedical data integration". *Fundamenta Informaticae* 171.1-4 (2020), pp. 345–365.

[20] *Gartner*. https://www.gartner.com. Accessed: 2022-07-30.

[21] D. J. Duffy. "Problems, challenges and promises: perspectives on precision medicine". *Briefings in bioinformatics* 17.3 (2016), pp. 494–504.

[22] L. J. Frey. "Data integration strategies for predictive analytics in precision medicine". *Personalized Medicine* 15.6 (2018), pp. 543–551.

[23] L. A. Denson, M. Curran, D. P. McGovern, W. A. Koltun, R. H. Duerr, S. C. Kim, R. B. Sartor, F. A. Sylvester, C. Abraham, E. F. de Zoeten, et al. "Challenges in IBD research: precision medicine". *Inflammatory bowel diseases* 25.Supplement_2 (2019), S31–S39.

[24] T. Hulsen, S. S. Jamuar, A. R. Moody, J. H. Karnes, O. Varga, S. Hedensted, R. Spreafico, D. A. Hafler, and E. F. McKinney. "From big data to precision medicine". *Frontiers in medicine* (2019), p. 34.

[25] A. Alyass, M. Turcotte, and D. Meyre. "From big data analysis to personalized medicine for all: challenges and opportunities". *BMC medical genomics* 8.1 (2015), pp. 1–12.

[26] D. G. Barrett, A. S. Morcos, and J. H. Macke. "Analyzing biological and artificial neural networks: challenges with opportunities for synergy?" *Current opinion in neurobiology* 55 (2019), pp. 55–64.

[27] G. Adam, L. Rampášek, Z. Safikhani, P. Smirnov, B. Haibe-Kains, and A. Goldenberg. "Machine learning approaches to drug response prediction: challenges and recent progress". *NPJ precision oncology* 4.1 (2020), pp. 1–10.

[28] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". *Information fusion* 58 (2020), pp. 82–115.

[29] N. Scarpato, A. Pieroni, L. Di Nunzio, and F. Fallucchi. "E-health-IoT universe: A review". *management* 21.44 (2017), p. 46.

[30] S. Hurtado, J. García-Nieto, I. Navas-Delgado, and J. F. Aldana-Montes. "FIMED: Flexible management of biomedical data". *Computer Methods and Programs in Biomedicine* 212 (2021), p. 106496.

[31] S. Hurtado, J. García-Nieto, I. Navas-Delgado, A. J. Nebro, and J. F. Aldana-Montes. "Reconstruction of gene regulatory networks with multi-objective particle swarm optimisers". *Applied Intelligence* 51.4 (2021), pp. 1972–1991.

[32] A. P. I. N.-D. Sandro Hurtado José García-Nieto. "Human Activity Recognition From Sensorised Patient's Data in Healthcare: A Streaming Deep Learning-Based Approach". *International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI)* (2022).

[33] S. Hurtado, J. García-Nieto, and I. Navas-Delgado. "A Service for Flexible Management and Analysis of Heterogeneous Clinical Data". *International Work-Conference on Bioinformatics and Biomedical Engineering*. Springer. 2022, pp. 227–238.

[34] S. Hurtado, H. Nematzadeh, J. García-Nieto, M.-Á. Berciano-Guerrero, and I. Navas-Delgado. "On the Use of Explainable Artificial Intelligence for the Differential Diagnosis of Pigmented Skin Lesions". *International Work-Conference on Bioinformatics and Biomedical Engineering*. Springer. 2022, pp. 319–329.

[35] S. Huratdo-Requena, C. Barba-González, M. Rybiński, F. J. Barón-López, J. Wärnberg, I. Navas-Delgado, and J. F. Aldana-Montes. "Análisis de datos de acelerometría para la detección de tipos de actividades". *Jornadas de Ingeniería del Software y Bases de Datos (In press)*. 2018.

[36] M. R. Sierra and C. A. C. Coello. "Improving PSO-Based Multi-objective Optimization Using Crowding, Mutation and epsilon-Dominance". *Evolutionary Multi-Criterion Optimization, Third International Conference, EMO 2005, Guanajuato, Mexico, March 9-11, 2005, Proceedings*. 2005, pp. 505–519.

[37] C. A. Coello Coello, G. Toscano Pulido, and M. Salazar Lechuga. "Handling Multiple Objectives With Particle Swarm Optimization". *IEEE Transactions on Evolutionary Computation* 8.3 (2004), pp. 256–279.

[38] K. Parsopoulos, D. Tasoulis, and M. Vrahatis. "Multiobjective Optimization Using Parallel Vector Evaluated Particle Swarm Optimization". *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2004)*. Vol. 2. Innsbruck, Austria: ACTA Press, 2004, pp. 823–828.

[39] A. J. Nebro, J. J. Durillo, J. Garcia-Nieto, C. A. Coello Coello, F. Luna, and E. Alba. "SMPSO: A new PSO-based metaheuristic for multi-objective optimization". *IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making*. 2009, pp. 66–73.

[40] S. Zapotecas Martínez and C. A. Coello Coello. "A Multi-objective Particle Swarm Optimizer Based on Decomposition". *2011 Genetic and Evolutionary Computation Conference (GECCO'2011)*. Dublin, Ireland: ACM Press, 2011, pp. 69–76.

[41] A. Nebro, J. Durillo, and C. C. Coello. "Analysis of leader selection strategies in a MOPSO". *IEEE Cong. on Evol. Comp. (CEC),* 2013, pp. 3153–3160.

[42] R. J. Prill, D. Marbach, J. Saez-Rodriguez, P. K. Sorger, L. G. Alexopoulos, X. Xue, N. D. Clarke, G. Altan-Bonnet, and G. Stolovitzky. "Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges". *PLoS ONE* 5.2 (Feb. 2010), pp. 1–18.

[43] M. R. Cowie, J. I. Blomster, L. H. Curtis, S. Duclaux, I. Ford, F. Fritz, S. Goldman, S. Janmohamed, J. Kreuzer, M. Leenay, et al. "Electronic health records to facilitate clinical research". *Clinical Research in Cardiology* 106.1 (2017), pp. 1–9.

[44] R. Kohli and S. S.-L. Tan. "Electronic Health Records". *Mis Quarterly* 40.3 (2016), pp. 553–574.

[45] A. Shahnaz, U. Qamar, and A. Khalid. "Using blockchain for electronic health records". *IEEE Access* 7 (2019), pp. 147782–147795.

[46] J.-S. Park, S. J. Moon, J.-H. Lee, J.-Y. Jeon, K. Jang, and M.-G. Kim. "The first step to the powers for clinical trials: a survey on the current and future Clinical Trial Management System". *Translational and Clinical Pharmacology* 26.2 (2018), p. 86.

[47] W. Kuchinke, C. Ohmann, Q. Yang, N. Salas, J. Lauritsen, F. Gueyffier, A. Leizorovicz, C. Schade-Brittinger, M. Wittenberg, Z. Voko, et al. "Heterogeneity prevails: the state of clinical trial data management in Europe-results of a survey of ECRIN centres". *Trials* 11.1 (2010), pp. 1–10.

[48] M. Gerritsen, O. Sartorius, F. vd Veen, and G. Meester. "Data management in multi-center clinical trials and the role of a nation-wide computer network. A 5 year evaluation." *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association. 1993, p. 659.

[49] L. E. Chai, S. K. Loh, S. T. Low, M. S. Mohamad, S. Deris, and Z. Zakaria. "A review on the computational approaches for gene regulatory network construction". *Computers in biology and medicine* 48 (2014), pp. 55–65.

[50] R. D. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. J. Pugh, H. McDonald, R. Varhol, S. J. Jones, and M. A. Marra. "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing". *Biotechniques* 45.1 (2008), pp. 81–94.

[51] S. Min, B. Lee, and S. Yoon. "Deep learning in bioinformatics". *Briefings in bioinformatics* 18.5 (2017), pp. 851–869.

[52] D. Xie, L. Zhang, and L. Bai. "Deep learning in visual computing and signal processing". *Applied Computational Intelligence and Soft Computing* 2017 (2017).

[53] A. Pal, A. Chaturvedi, U. Garain, A. Chandra, and R. Chatterjee. "Severity grading of psoriatic plaques using deep CNN based multi-task learning". *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE. 2016, pp. 1478–1483.

[54] N. C. Codella, Q.-B. Nguyen, S. Pankanti, D. A. Gutman, B. Helba, A. C. Halpern, and J. R. Smith. "Deep learning ensembles for melanoma recognition in dermoscopy images". *IBM Journal of Research and Development* 61.4/5 (2017), pp. 5–1.

[55] M. I. Razzak, S. Naz, and A. Zaib. "Deep learning for medical image processing: Overview, challenges and the future". *Classification in BioApps* (2018), pp. 323–350.

[56] M. Bageshwari, P. Adurkar, and A. Chandrakar. "Clinical database: Rdbms v/s newer technologies (NoSQL and XML database); Why look beyond Rdbms and consider the newer" (2014).

[57] N. Jatana, S. Puri, M. Ahuja, I. Kathuria, and D. Gosain. "A survey and comparison of relational and non-relational database". *International Journal of Engineering Research & Technology* 1.6 (2012), pp. 1–5.

[58] M. Sharma, V. D. Sharma, and M. M. Bundele. "Performance analysis of RDBMS and no SQL databases: PostgreSQL, MongoDB and Neo4j". *2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*. IEEE. 2018, pp. 1–5.

[59] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton. "Big data: the management revolution". *Harvard business review* 90.10 (2012), pp. 60–68.

[60] A. Davoudian, L. Chen, and M. Liu. "A survey on NoSQL stores". *ACM Computing Surveys (CSUR)* 51.2 (2018), pp. 1–43.

[61] P. Shah, R. Adurkar, S. Desai, S. Kadakia, and K. Bhowmick. "Covid-19 database management: A non-relational approach (nosql and xml)". *Intelligent Data Communication Technologies and Internet of Things*. Springer, 2021, pp. 483–497.

[62] M. Ercan and M. Lane. "An Evaluation of the suitability of NoSQL databases for distributed EHR systems". ACIS. 2014.

[63] J. L. Carlson. *Redis in action*. Manning Publications Co., 2013.

[64] R. Cattell. "Scalable SQL and NoSQL data stores". *ACM Sigmod Record* 39.4 (2011), pp. 12–27.

[65] N. Leavitt. "Will NoSQL databases live up to their promise?" *Computer* 43.2 (2010).

[66] L. George. *HBase: the definitive guide: random access to your planet-size data*. " O'Reilly Media, Inc.", 2011.

[67] A. Abouzeid, K. Bajda-Pawlikowski, D. Abadi, A. Silberschatz, and A. Rasin. "HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads". *Proceedings of the VLDB Endowment* 2.1 (2009), pp. 922–933.

[68] E. Dede, B. Sendir, P. Kuzlu, J. Hartog, and M. Govindaraju. "An evaluation of Cassandra for Hadoop". *Cloud Computing (CLOUD), 2013 IEEE Sixth International Conference on*. IEEE. 2013, pp. 494–501.

[69] K. Banker. *MongoDB in action*. Manning Publications Co., 2011.

[70] F. Provost and T. Fawcett. "Data science and its relationship to big data and data-driven decision making". *Big data* 1.1 (2013), pp. 51–59.

[71] S. Sivasubramanian. "Amazon dynamoDB: a seamlessly scalable non-relational database service". *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. 2012, pp. 729–730.

[72] D. Ostrovsky, Y. Rodenski, and M. Haji. *Pro couchbase server*. Springer, 2015.

[73] M. Lal. *Neo4j graph data modeling*. Packt Publishing Ltd, 2015.

[74] O. Erling and I. Mikhailov. "Virtuoso: RDF support in a native RDBMS". *Semantic web information management*. Springer, 2010, pp. 501–519.

[75] *Stardog*. https://www.stardog.com/. Accessed: 2022-09-01.

[76] K. K.-Y. Lee, W.-C. Tang, and K.-S. Choi. "Alternatives to relational database: Comparison of NoSQL and XML approaches for clinical data storage". *Computer Methods and Programs in Biomedicine* 110.1 (2013), pp. 99–109. ISSN: 0169-2607.

[77] M. Chen, S. Mao, and Y. Liu. "Big data: A survey". *Mobile networks and applications* 19.2 (2014), pp. 171–209.

[78] R. Bhardwaj, A. Sethi, and R. Nambiar. "Big data in genomics: An overview". *2014 IEEE International Conference on Big Data (Big Data)*. IEEE. 2014, pp. 45–49.

[79] B. Kayyali, D. Knott, and S. Van Kuiken. "The big-data revolution in US health care: Accelerating value and innovation". *Mc Kinsey & Company* 2.8 (2013), pp. 1–13.

[80] D. Kiritsis. "Closed-loop PLM for intelligent products in the era of the Internet of things". *Computer-Aided Design* 43.5 (2011), pp. 479–501.

[81] S. Li, L. Xu, X. Wang, and J. Wang. "Integration of hybrid wireless networks in cloud services oriented enterprise information systems". *Enterprise Information Systems* 6.2 (2012), pp. 165–187.

[82] S. Li, L. D. Xu, and S. Zhao. "The internet of things: a survey". *Information systems frontiers* 17.2 (2015), pp. 243–259.

[83] M. Haghi Kashani, M. Madanipour, M. Nikravan, P. Asghari, and E. Mahdipour. "A systematic review of IoT in healthcare: Applications, techniques, and trends". *Journal of Network and Computer Applications* 192 (2021), p. 103164. ISSN: 1084-8045.

[84] D. Wilson. "An overview of the application of wearable technology to nursing practice". *Nursing forum*. Vol. 52. 2. Wiley Online Library. 2017, pp. 124–132.

[85] A. Ahmed, R. Latif, S. Latif, H. Abbas, and F. A. Khan. "Malicious insiders attack in IoT based multi-cloud e-healthcare environment: a systematic literature review". *Multimedia Tools and Applications* 77.17 (2018), pp. 21947–21965.

[86] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. "Spark: Cluster Computing with Working Sets". *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*. HotCloud'10. USENIX Association, 2010, pp. 10–10.

[87] B. Laurie and P. Laurie. *Apache: The definitive guide*. " O'Reilly Media, Inc.", 2003.

[88] S. Salloum, R. Dautov, X. Chen, P. X. Peng, and J. Z. Huang. "Big data analytics on Apache Spark". *International Journal of Data Science and Analytics* 1.3 (2016), pp. 145–164.

[89] M. Zaharia. *An architecture for fast and general data processing on large clusters*. Morgan & Claypool, 2016.

[90] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi, et al. "Spark sql: Relational data processing in spark". *Proceedings of the 2015 ACM SIGMOD international conference on management of data*. 2015, pp. 1383–1394.

[91] M. Zaharia, T. Das, H. Li, T. Hunter, S. Shenker, and I. Stoica. "Discretized streams: Fault-tolerant streaming computation at scale". *Proceedings of the twenty-fourth ACM symposium on operating systems principles*. 2013, pp. 423–438.

[92] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, and S. Owen. "Mllib: Machine learning in apache spark". *The Journal of Machine Learning Research* 17.1 (2016), pp. 1235–1241.

[93] J. E. Gonzalez. "From graphs to tables the design of scalable systems for graph analytics." *WWW (Companion Volume)*. 2014, pp. 1149–1150.

[94] A. M. Turing. "Computing machinery and intelligence". *Parsing the turing test*. Springer, 2009, pp. 23–65.

[95] M. Minsky. "Steps toward artificial intelligence". *Proceedings of the IRE* 49.1 (1961), pp. 8–30.

[96] G. Huang, G.-B. Huang, S. Song, and K. You. "Trends in extreme learning machines: A review". *Neural Networks* 61 (2015), pp. 32–48. ISSN: 0893-6080.

[97] M. Chiang and T. Zhang. "Fog and IoT: An overview of research opportunities". *IEEE Internet of things journal* 3.6 (2016), pp. 854–864.

[98] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew. "Deep learning for visual understanding: A review". *Neurocomputing* 187 (2016), pp. 27–48.

[99] H. Nguyen, L.-M. Kieu, T. Wen, and C. Cai. "Deep learning methods in transportation domain: a review". *IET Intelligent Transport Systems* 12.9 (2018), pp. 998–1004.

[100] S Alshahrani and E. Kapetanios. "Are deep learning approaches suitable for natural language processing?" *International Conference on Applications of Natural Language to Information Systems*. Springer. 2016, pp. 343–349.

[101] S. Schaal. "Is imitation learning the route to humanoid robots?" *Trends in cognitive sciences* 3.6 (1999), pp. 233–242.

[102] K.-H. Yu, A. L. Beam, and I. S. Kohane. "Artificial intelligence in healthcare". *Nature biomedical engineering* 2.10 (2018), pp. 719–731.

[103] P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov. "Applications of deep learning in biomedicine". *Molecular pharmaceutics* 13.5 (2016), pp. 1445–1454.

[104] G. Rong, A. Mendez, E. Bou Assi, B. Zhao, and M. Sawan. "Artificial Intelligence in Healthcare: Review and Prediction Case Studies". *Engineering* 6.3 (2020), pp. 291–301. ISSN: 2095-8099.

[105] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang. "Artificial intelligence in healthcare: past, present and future". *Stroke and vascular neurology* 2.4 (2017).

[106] N Murali[1] and N Sivakumaran. "Artificial intelligence in healthcare–a review" (2018).

[107] E. Diaz-Flores, T. Meyer, and A. Giorkallos. "Evolution of Artificial Intelligence-Powered Technologies in Biomedical Research and Healthcare". Springer, 2022.

[108] L. Rubinger, A. Gazendam, S. Ekhtiari, and M. Bhandari. "Machine learning and artificial intelligence in research and healthcare," *Injury* (2022).

[109] E. Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[110] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[111] Y. Anzai. *Pattern recognition and machine learning*. Elsevier, 2012.

[112] O. Chapelle, B. Scholkopf, and A. Zien. "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]". *IEEE Transactions on Neural Networks* 20.3 (2009), pp. 542–542.

[113] J. E. Van Engelen and H. H. Hoos. "A survey on semi-supervised learning". *Machine Learning* 109.2 (2020), pp. 373–440.

[114] Y. Reddy, P Viswanath, and B. E. Reddy. "Semi-supervised learning: A brief review". *Int. J. Eng. Technol* 7.1.8 (2018), p. 81.

[115] N. N. Pise and P. Kulkarni. "A survey of semi-supervised learning methods". *2008 International conference on computational intelligence and security*. Vol. 2. IEEE. 2008, pp. 30–34.

[116] V. J. Prakash and D. L. Nithya. "A survey on semi-supervised learning techniques". *arXiv preprint arXiv:1402.4645* (2014).

[117] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, A. M. Umar, O. U. Linus, H. Arshad, A. A. Kazaure, U. Gana, and M. U. Kiru. "Comprehensive review of artificial neural network applications to pattern recognition". *IEEE Access* 7 (2019), pp. 158820–158846.

[118] G. Hinton, Y. LeCun, and Y. Bengio. "Deep learning". *Nature* 521.7553 (2015), pp. 436–444.

[119] G. E. Hinton, S. Osindero, and Y.-W. Teh. "A fast learning algorithm for deep belief nets". *Neural computation* 18.7 (2006), pp. 1527–1554.

[120] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". *nature* 521.7553 (2015), pp. 436–444.

[121] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean. "A guide to deep learning in healthcare". *Nature medicine* 25.1 (2019), pp. 24–29.

[122] A Krizhevsky, I Sutskever, and G Hinton. *ImageNet classification with deep convolutional neural networks. Advances in Neural Informa-tion Processing System 25 (NIPS 2012). 2012.* 2017.

[123] I. R. I. Haque and J. Neubert. "Deep learning approaches to biomedical image segmentation". *Informatics in Medicine Unlocked* 18 (2020), p. 100297.

[124] L. Pauly, D Hogg, R Fuentes, and H Peel. "Deeper networks for pavement crack detection". *Proceedings of the 34th ISARC*. IAARC. 2017, pp. 479–485.

[125] H. Gholamalinezhad and H. Khosravi. "Pooling methods in deep neural networks, a review". *arXiv preprint arXiv:2009.07485* (2020).

[126] M. F. A. Hady and F. Schwenker. "Semi-supervised learning". *Handbook on Neural Information Processing* (2013), pp. 215–239.

[127] M. Chen, Z. Xu, K. Weinberger, and F. Sha. "Marginalized denoising autoencoders for domain adaptation". *arXiv preprint arXiv:1206.4683* (2012).

[128] A. Gogna and A. Majumdar. "Semi supervised autoencoder". *International Conference on Neural Information Processing*. Springer. 2016, pp. 82–89.

[129] G. Zhang, Y. Liu, and X. Jin. "A survey of autoencoder-based recommender systems". *Frontiers of Computer Science* 14.2 (2020), pp. 430–450.

[130] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. "Extracting and composing robust features with denoising autoencoders". *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1096–1103.

[131] M. Chen, K. Weinberger, F. Sha, and Y. Bengio. "Marginalized denoising auto-encoders for nonlinear representations". *International conference on machine learning*. PMLR. 2014, pp. 1476–1484.

[132] D. P. Kingma and M. Welling. "Auto-encoding variational bayes". *arXiv preprint arXiv:1312.6114* (2013).

[133] A. J. Kulkarni and K. Tai. "Probability collectives: a multi-agent approach for solving combinatorial optimization problems". *Applied Soft Computing* 10.3 (2010), pp. 759–771.

[134] G. N. Vanderplaats. *Numerical optimization techniques for engineering design: with applications*. Vol. 1. McGraw-Hill New York, 1984.

[135] T. Bäck. *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. Oxford University Press, 1996.

[136] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, 1989.

[137] A.-L. Barabasi and Z. N. Oltvai. "Network biology: understanding the cell's functional organization". *Nature reviews genetics* 5.2 (2004), pp. 101–113.

[138] W. Yan, W. Xue, J. Chen, and G. Hu. "Biological networks for cancer candidate biomarkers discovery". *Cancer Informatics* 15 (2016), CIN–S39458.

[139] P. K. Kreeger and D. A. Lauffenburger. "Cancer systems biology: a network modeling perspective". *Carcinogenesis* 31.1 (2010), pp. 2–8.

[140] P. R. Somvanshi and K. Venkatesh. "A conceptual review on systems biology in health and diseases: from biological networks to modern therapeutics". *Systems and synthetic biology* 8.1 (2014), pp. 99–116.

[141] S. Jin, X. Zeng, F. Xia, W. Huang, and X. Liu. "Application of deep learning methods in biological networks". *Briefings in bioinformatics* 22.2 (2021), pp. 1902–1917.

[142] G.-W. Li and X. S. Xie. "Central dogma at the single-molecule level in living cells". *Nature* 475.7356 (2011), pp. 308–315.

[143] J. Ruan, A. K. Dean, and W. Zhang. "A general co-expression network-based approach to gene expression analysis: comparison and applications". *BMC systems biology* 4.1 (2010), pp. 1–21.

[144] W. W. Wasserman and A. Sandelin. "Applied bioinformatics for the identification of regulatory elements". *Nature Reviews Genetics* 5.4 (2004), pp. 276–287.

[145] S. G. Potkin, F. Macciardi, G. Guffanti, J. H. Fallon, Q. Wang, J. A. Turner, A. Lakatos, M. F. Miles, A. Lander, M. P. Vawter, et al. "Identifying gene regulatory networks in schizophrenia". *Neuroimage* 53.3 (2010), pp. 839–847.

[146] E. Liu, L. Li, and L. Cheng. "Gene Regulatory Network Review". *Encyclopedia of Bioinformatics and Computational Biology*. Ed. by S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach. Oxford: Academic Press, 2019, pp. 155–164. ISBN: 978-0-12-811432-2.

[147] W.-P. Lee and W.-S. Tzou. "Computational methods for discovering gene networks from expression data". *Briefings in bioinformatics* 10.4 (2009), pp. 408–423.

[148] T. I. Lee and R. A. Young. "Transcriptional regulation and its misregulation in disease". *Cell* 152.6 (2013), pp. 1237–1251.

[149] C. A. Jackson, D. M. Castro, G.-A. Saldi, R. Bonneau, and D. Gresham. "Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments". *elife* 9 (2020), e51254.

[150] Z.-P. Liu. "Towards precise reconstruction of gene regulatory networks by data integration". *Quantitative Biology* 6.2 (2018), pp. 113–128.

[151] L. F. Iglesias-Martinez, W. Kolch, and T. Santra. "BGRMI: A method for inferring gene regulatory networks from time-course gene expression data and its application in breast cancer research". *Nature, Scientific Reports* 6.37140 (2016).

[152] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke. "Gene regulatory network inference: data integration in dynamic models—a review". *Biosystems* 96.1 (2009), pp. 86–103.

[153] J. García-Nieto, A. J. Nebro, and J. F. Aldana-Montes. "Inference of gene regulatory networks with multi-objective cellular genetic algorithm". *Computational Biology and Chemistry* 80 (2019), pp. 409 –418. ISSN: 1476-9271.

[154] N. N. Hitoshi Iba. *Evolutionary Computation in Gene Regulatory Network Research*. Wiley, Series in Bioinformatics, 2016. ISBN: 978-1-118-91151-8.

[155] L. Liu and J. Liu. "Inferring gene regulatory networks with hybrid of multi-agent genetic algorithm and random forests based on fuzzy cognitive maps". *Applied Soft Computing* 69 (2018), pp. 585 –598. ISSN: 1568-4946.

[156] Abhishek and S. Singh. "Article: A Gene Regulatory Network Prediction Method using Particle Swarm Optimization and Genetic Algorithm". *International Journal of Computer Applications* 83.12 (2013). Full text available, pp. 32–37.

[157] B. Jana, S. Mitra, and S. Acharyya. "Repository and Mutation based Particle Swarm Optimization (RMPSO): A new PSO variant applied to reconstruction of Gene Regulatory Network". *Applied Soft Computing* 74 (2019), pp. 330 –355. ISSN: 1568-4946.

[158] A. Khan, S. Mandal, R. K. Pal, and G. Saha. "Construction of Gene Regulatory Networks Using Recurrent Neural Networks and Swarm Intelligence". *Scientifica* 2016 (2016), Article ID 1060843 14 pages.

[159] R. Sultana, D. Showkat, M. Samiullah, and A. R. Chowdhury. "Reconstructing Gene Regulatory Network with Enhanced Particle Swarm Optimization". *Neural Information Processing*. Ed. by C. K. Loo, K. S. Yap, K. W. Wong, A. Teoh, and K. Huang. Cham: Springer International Publishing, 2014, pp. 229–236. ISBN: 978-3-319-12640-1.

[160] R. Xu, D. W. II, and R. Frank. "Inference of Genetic Regulatory Networks with Recurrent Neural Network Models Using Particle Swarm Optimization". *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4.4 (2007), pp. 681–692. ISSN: 1545-5963.

[161] M. Savageau. *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*. Addison-Wesley Educational Publishers Inc, 2010. ISBN: 0201067382, 978-0201067385.

[162] S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi, and M. Tomita. "Dynamic modeling of genetic networks using genetic algorithm and S-system". *Bioinformatics* 19.5 (2003), pp. 643–650.

[163] J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood. "Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation". *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE. 2018, pp. 1–8.

[164] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. "Key challenges for delivering clinical impact with artificial intelligence". *BMC medicine* 17.1 (2019), pp. 1–9.

[165] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol. "AI in health and medicine". *Nature Medicine* 28.1 (2022), pp. 31–38.

[166] P. Hall. "On the art and science of machine learning explanations". *arXiv preprint arXiv:1810.02909* (2018).

[167] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang. "XAI—Explainable artificial intelligence". *Science robotics* 4.37 (2019), eaay7120.

[168] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. "Explaining explanations: An overview of interpretability of machine learning". *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE. 2018, pp. 80–89.

[169] F. Doshi-Velez and B. Kim. "Towards a rigorous science of interpretable machine learning". *arXiv preprint arXiv:1702.08608* (2017).

[170] G. Vilone and L. Longo. "Notions of explainability and evaluation approaches for explainable artificial intelligence". *Information Fusion* 76 (2021), pp. 89–106. ISSN: 1566-2535.

[171] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation". *journal of Computational and Graphical Statistics* 24.1 (2015), pp. 44–65.

[172] S. Atakishiyev, H. Babiker, N. Farruque, R. Goebel, M. Kima, M. H. Motallebi, J. Rabelo, T. Syed, and O. R. Zaïane. "A multi-component framework for the analysis and design of explainable artificial intelligence". *arXiv preprint arXiv:2005.01908* (2020).

[173] B. Babic, S. Gerke, T. Evgeniou, and I. G. Cohen. "Beware explanations from AI in health care". *Science* 373.6552 (2021), pp. 284–286.

[174] A. Adadi and M. Berrada. "Explainable AI for healthcare: from black box to interpretable models". *Embedded Systems and Artificial Intelligence*. Springer, 2020, pp. 327–337.

[175] J. Petch, S. Di, and W. Nelson. "Opening the black box: the promise and limitations of explainable machine learning in cardiology". *Canadian Journal of Cardiology* (2021).

[176] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. "Definitions, methods, and applications in interpretable machine learning". *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22071–22080.

[177] G. Vilone and L. Longo. "Notions of explainability and evaluation approaches for explainable artificial intelligence". *Information Fusion* 76 (2021), pp. 89–106.

[178] Y. Zhang, Y. Weng, and J. Lund. "Applications of Explainable Artificial Intelligence in Diagnosis and Surgery". *Diagnostics* 12.2 (2022), p. 237.

[179] N. Burkart and M. F. Huber. "A survey on the explainability of supervised machine learning". *Journal of Artificial Intelligence Research* 70 (2021), pp. 245–317.

[180] M. T. Ribeiro, S. Singh, and C. Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[181] S. M. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017.

[182] S. Hart. "Shapley value". *Game theory*. Springer, 1989, pp. 210–216.

[183] M. Ou, R. Ma, J. Cheung, K. Lo, P. Yee, T. Luo, T. Chan, C. H. Au, A. Kwong, R. Luo, et al. "Database. bio: a web application for interpreting human variations". *Bioinformatics* 31.24 (2015), pp. 4035–4037.

[184] J. S. Beckmann and D. Lew. "Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities". *Genome Medicine* 8.134 (2016), pp. 2 –11.

[185] S. K. Gill, A. F. Christopher, V. Gupta, and P. Bansal. "Emerging role of bioinformatics tools and software in evolution of clinical research". *Perspectives in clinical research* 7.3 (2016), p. 115.

[186] W. L. Schulz, B. G. Nelson, D. K. Felker, T. J. Durant, and R. Torres. "Evaluation of relational and NoSQL database architectures to manage genomic annotations". *Journal of biomedical informatics* 64 (2016), pp. 288–295.

[187] V. Bianchi, A. Ceol, A. G. E. Ogier, S. de Pretis, E. Galeota, K. Kishore, P. Bora, O. Croci, S. Campaner, B. Amati, M. J. Morelli, and M. Pelizzola. "Integrated Systems for NGS Data Management and Analysis: Open Issues and Available Solutions". *Frontiers in Genetics* 7 (2016), p. 75. ISSN: 1664-8021.

[188] G. M. Weber, K. D. Mandl, and I. S. Kohane. "Finding the missing link for big biomedical data". *Jama* 311.24 (2014), pp. 2479–2480.

[189] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. "The human genome browser at UCSC". *Genome research* 12.6 (2002), pp. 996–1006.

[190] M. M. Astrahan, M. W. Blasgen, D. D. Chamberlin, K. P. Eswaran, J. N. Gray, P. P. Griffiths, W. F. King, R. A. Lorie, P. R. McJones, J. W. Mehl, et al. "System R: relational approach to database management". *ACM Transactions on Database Systems (TODS)* 1.2 (1976), pp. 97–137.

[191] R. Elmasri. *Fundamentals of database systems*. 2017.

[192] E. Rahm and P. A. Bernstein. "An Online Bibliography on Schema Evolution". *SIGMOD Rec.* 35.4 (Dec. 2006), 30–31. ISSN: 0163-5808.

[193] P. Payne, A. W. Greaves, and T. J. Kipps. "CRC Clinical Trials Management System (CTMS): an integrated information management solution for collaborative clinical research." *AMIA... Annual Symposium proceedings. AMIA Symposium*. Vol. 2003. American Medical Informatics Association. 2003, pp. 967–967.

[194] R. Li, H. He, R. Wang, S. Ruan, T. He, J. Bao, J. Zhang, L. Hong, and Y. Zheng. "TrajMesa: A Distributed NoSQL-Based Trajectory Data Management System". *IEEE Transactions on Knowledge and Data Engineering* (2021), pp. 1–1.

[195] A. Rafique, D. Van Landuyt, and W. Joosen. "PERSIST: Policy-Based Data Management Middleware for Multi-Tenant SaaS Leveraging Federated Cloud Storage". *Journal of Grid Computing* 16 (2018), 165–194.

[196] A. Rafique, D. Van Landuyt, E. Heydari Beni, B. Lagaisse, and W. Joosen. "CryptDICE: Distributed data protection system for secure cloud data storage and computation". *Information Systems* 96 (2021), p. 101671. ISSN: 0306-4379.

[197] M. Ahmadian, F. Plochan, Z. Roessler, and D. C. Marinescu. "SecureNoSQL: An approach for secure search of encrypted NoSQL databases in the public cloud". *International Journal of Information Management* 37.2 (2017), pp. 63–74. ISSN: 0268-4012.

[198] J. Ainsworth and R. Harper. "The PsyGrid Experience: using web services in the study of schizophrenia". *International Journal of Healthcare Information Systems and Informatics (IJHISI)* 2.2 (2007), pp. 1–20.

[199] C. Brandt, A. M. Deshpande, C. Lu, G. Ananth, K. Sun, R. Gadagkar, R. Morse, C. Rodriguez, P. L. Miller, and P. M. Nadkarni. "TrialDB: A web-based Clinical Study Data Management System." *AMIA Annual Symposium Proceedings*. Vol. 2003. American Medical Informatics Association. 2003, pp. 794–794.

[200] M. Cavelaars, J. Rousseau, C. Parlayan, S. de Ridder, A. Verburg, R. Ross, G. R. Visser, A. Rotte, R. Azevedo, J.-W. Boiten, et al. "OpenClinica". *Journal of clinical bioinformatics*. Vol. 5. S1. Springer. 2015, S2.

[201] P. Cramon, Å. K. Rasmussen, S. J. Bonnema, J. B. Bjorner, U. Feldt-Rasmussen, M. Groenvold, L. Hegedüs, and T. Watt. "Development and implementation of PROgmatic: A clinical trial management system for pragmatic multi-centre trials, optimised for electronic data capture and patient-reported outcomes". *Clinical Trials* 11.3 (2014), pp. 344–354.

[202] P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J. G. Conde. "Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support". *Journal of biomedical informatics* 42.2 (2009), pp. 377–381.

[203] R. Krenn. "Design and Development of a Web-Based Clinical Trial Management System". PhD thesis. May 2014.

[204] L. Nguyen, A. Shah, M. Harker, H. Martins, M. McCready, A. Menezes, D. O. Jacobs, and R. Pietrobon. "DADOS-Prospective: an open source application for Web-based prospective data collection". *Source code for biology and medicine* 1.1 (2006), p. 7.

[205] D. Venizeleas, M. Linzbach, and C. Ohmann. "PhOSCo (Pharma Open Source Community): Open Source für klinische Studien". *Dtsch Arztebl International* 101.19 (2004), [19]–.

[206] C. A. Brandt, S. Argraves, R. Money, G. Ananth, N. M. Trocky, and P. M. Nadkarni. "Informatics tools to improve clinical research study implementation". *Contemporary clinical trials* 27.2 (2006), pp. 112–122.

[207] S. Kaur, I. Singh, et al. "Artificial intelligence based clinical data management systems: A review". *Informatics in Medicine Unlocked* 9 (2017), pp. 219–229.

[208] J. Shah, D. Rajgor, S. Pradhan, M. McCready, A. Zaveri, and R. Pietrobon. "Electronic data capture for registries and clinical trials in orthopaedic surgery: open source versus commercial systems". *Clinical Orthopaedics and Related Research®* 468.10 (2010), pp. 2664–2671.

[209] A. Nourani, H. Ayatollahi, and M. S. Dodaran. "Clinical Trial Data Management Software: A Review of the Technical Features". *Reviews on Recent Clinical Trials* 14.3 (2019), pp. 1–10. ISSN: 1574-8871.

[210] J. Muller, K. Heiss, and R. Oberhoffer. "Implementation of an open adoption research data management system for clinical studies". *BMC Res Notes* 10.252 (2017), pp. 1–10.

[211] K. Chodorow. "MongoDB: the definitive guide: powerful and scalable data storage". O'Reilly Media, Inc., 2013, p. 193. ISBN: 978-1-449-38156-1.

[212] J. Daemen and V. Rijmen. *The design of Rijndael: AES-the advanced encryption standard*. Belgium: Springer Science & Business Media, 2013.

[213] P. Mestdagh, P. Van Vlierberghe, A. De Weer, D. Muth, F. Westermann, F. Speleman, and J. Vandesompele. "A novel and universal method for microRNA RT-qPCR data normalization". *Genome biology* 10.6 (2009), R64.

[214] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. "Inferring regulatory networks from expression data using tree-based methods". *PloS one* 5.9 (2010), e12776.

[215] A. Irrthum, L. Wehenkel, P. Geurts, et al. "Inferring regulatory networks from expression data using tree-based methods". *PloS one* 5.9 (2010), e12776.

[216] T. Moerman, S. Aibar Santos, C. Bravo González-Blas, J. Simm, Y. Moreau, J. Aerts, and S. Aerts. "GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks". *Bioinformatics* 35.12 (2018), pp. 2159–2161.

[217] I. Navas-Delgado, J. García-Nieto, E. López-Camacho, M. Rybinski, R. Lavado, M. Á. B. Guerrero, and J. F. Aldana-Montes. "VIGLA-M: visual gene expression data analytics". *BMC bioinformatics* 20.4 (2019), p. 150.

[218] C. A. Torres-Cabala and J. L. Curry. "Genetics of Melanoma". Springer-Verlag New York, 2016.

[219] K. Glass, C. Huttenhower, J. Quackenbush, and G.-C. Yuan. "Passing messages between biological networks to refine predicted interactions". *PloS one* 8.5 (2013), e64832.

[220] M. L. Kuijjer, M. G. Tung, G. Yuan, J. Quackenbush, and K. Glass. "Estimating Sample-Specific Regulatory Networks". *iScience* 14 (2019), pp. 226–240. ISSN: 2589-0042.

[221] W.-P. Lee and Y.-T. Hsiao. "Inferring gene regulatory networks using a hybrid GA-PSO approach with numerical constraints and network decomposition". *Information Sciences* 188 (2012), pp. 80 –99. ISSN: 0020-0255.

[222] L. Palafox, N. Noman, and H. Iba. "Reverse Engineering of Gene Regulatory Networks Using Dissipative Particle Swarm Optimization". *Evolutionary Computation, IEEE Transactions on* 17.4 (2013), pp. 577–587. ISSN: 1089-778X.

[223] T. Akutsu. "Identification of genetic networks by strategic gene disruptions and gene over-expressions under a boolean model". *Theoretical Computer Science* 298.1 (2003), pp. 235 –251. ISSN: 0304-3975.

[224] Y. N. Kaznessis. "Models for synthetic biology". *BMC Systems Biology* 1.1 (2007), p. 47.

[225] A. Huynh-Thu and G. Sanguinetti. "Combining tree-based and dynamical systems for the inference of gene regulatory networks". *Bioinformatics* 31.10 (2015), pp. 1614–1622.

[226] K. Raza and M. Alam. "Recurrent neural network based hybrid model for reconstructing gene regulatory network". *Computational Biology and Chemistry* 64 (2016), pp. 322 –334. ISSN: 1476-9271.

[227] A. Sirbu, H. J. Ruskin, and M. Crane. "Comparison of evolutionary algorithms in gene regulatory network model inference". English. *BMC Bioinformatics* 11.1, 59 (2010).

[228] N. Noman and H. Iba. "Inferring Gene Regulatory Networks using Differential Evolution with Local Search Heuristics". *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4.4 (2007), pp. 634–647. ISSN: 1545-5963.

[229] M. S. Nobile and H. Iba. "A double swarm methodology for parameter estimation in oscillating Gene Regulatory Networks". *2015 IEEE Congress on Evolutionary Computation (CEC)*. 2015, pp. 2376–2383.

[230] Y. Chen and X. Zou. "Inferring Gene Regulatory Network Using An Evolutionary Multi-Objective Method". *arXiv:1512.05055, Cornell University Library* (2016).

[231] P.-K. Liu and F.-S. Wang. "Inference of biochemical network models in S-system using multiobjective optimization approach". *Bioinformatics* 24.8 (2008), p. 1085.

[232] C. Spieth, F. Streichert, N. Speer, and A. Zell. "Multi-objective Model Optimization for Inferring Gene Regulatory Networks". *Evolutionary Multi-Criterion Optimization*. Ed. by C. Coello Coello, A. Hernández Aguirre, and E. Zitzler. Vol. 3410. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2005, pp. 607–620. ISBN: 978-3-540-24983-2.

[233] Y. Zhang, D.-w. Gong, and J. Cheng. "Multi-Objective Particle Swarm Optimization Approach for Cost-Based Feature Selection in Classification". *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 14.1 (Jan. 2017), 64–75. ISSN: 1545-5963.

[234] X. Song, Y. Zhang, Y. Guo, X. Sun, and Y. Wang. "Variable-size Cooperative Coevolutionary Particle Swarm Optimization for Feature Selection on High-dimensional Data". *IEEE Transactions on Evolutionary Computation* Early Access (2020), pp. 1–1.

[235] Y. Zhang, D.-W. Gong, and Z. Ding. "A bare-bones multi-objective particle swarm optimization algorithm for environmental/economic dispatch" ().

[236] X. Cai. "A Multi-objective Gp-pso Hybrid Algorithm for Gene Regulatory Network Modeling". AAI3358776. PhD thesis. Manhattan, KS, USA, 2009. ISBN: 978-1-109-17856-2.

[237] E. O. Voit. *Computational Analysis of Biochemical Systems. A Practical Guide for Biochemists and Molecular Biologists*. Cambridge University Press, 2000. ISBN: 9780521780872.

[238] K.-Y. Tsai and F.-S. Wang. "Evolutionary optimization with data collocation for reverse engineering of biological networks". *Bioinformatics* 21.7 (2005), pp. 1180–1188.

[239] D. Tominaga, N. Koga, and M. Okamoto. "Efficient Numerical Optimization Algorithm Based on Genetic Algorithm for Inverse Problem". *Proceedings of the 2Nd Annual Conference on Genetic and Evolutionary Computation*. GECCO'00. Las Vegas, Nevada: Morgan Kaufmann Publishers Inc., 2000, pp. 251–258. ISBN: 1-55860-708-0.

[240] S. Kimura, K. Ide, A. Kashihara, M. Kano, M. Hatakeyama, R. Masui, N. Nakagawa, S. Yokoyama, S. Kuramitsu, and A. Konagaya. "Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm". *Bioinformatics* 21.7 (2005), pp. 1154–1163.

[241] P.-K. Liu and F.-S. Wang. "Inference of biochemical network models in S-system using multiobjective optimization approach". *Bioinformatics* 24.8 (2008), pp. 1085–1092.

[242] K. Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*. New York, NY, USA: John Wiley & Sons, Inc., 2001.

[243] J. Kennedy and R. Eberhart. "Particle swarm optimization". *IEEE IJCNN*. Vol. 4. 1995, 1942–1948 vol.4.

[244] M. Reyes-Sierra and C. A. Coello Coello. "Multi-Objective Particle Swarm Optimizers: A Survey of the State-of-the-Art". *International Journal of Computational Intelligence Research* 2.3 (2006), pp. 287–308.

[245] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II". *IEEE Transactions on Evolutionary Computation* 6.2 (2002), pp. 182–197.

[246] E. Zitzler and L. Thiele. "Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach". *IEEE Trans. on Evol. Comp.* 3.4 (1999), pp. 257–271.

[247] Q. Zhang and H. Li. "MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition". *IEEE Trans. on Evol. Comp.* 11.6 (2007), pp. 712–731.

[248] J. D. Schaffer. "Multiple Objective Optimization with Vector Evaluated Genetic Algorithms". *Proceedings of the 1st International Conference on Genetic Algorithms, Pittsburgh, PA, USA, July 1985*. 1985, pp. 93–100.

[249] P. K. Tripathi, S. Bandyopadhyay, and S. K. Pal. "Multi-Objective Particle Swarm Optimization with time variant inertia and acceleration coefficients". *Information Sciences* 177.22 (2007), pp. 5033 –5049. ISSN: 0020-0255.

[250] A. J. Nebro, J. J. Durillo, and M. Vergne. "Redesigning the jMetal Multi-Objective Optimization Framework". *Genetic and Evolutionary Computation Conference (GECCO 2015) Companion*. 2015, pp. 1093–1100.

[251] H. Ishibuchi, H. Masuda, and Y. Nojima. "Sensitivity of performance evaluation results by inverted generational distance to reference points". *2016 IEEE Congress on Evolutionary Computation (CEC)*. 2016, pp. 1107–1114.

[252] M. Clerc and J. Kennedy. "The particle swarm - explosion, stability, and convergence in a multidimensional complex space". *IEEE Transactions on Evolutionary Computation* 6.1 (2002), pp. 58 –73. ISSN: 1089-778X.

[253] T. Bartz-Beielstein. *Experimental Research in Evolutionary Computation: The New Experimentalism (Natural Computing Series)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 3540320261.

[254] M. Birattari, Z. Yuan, P. Balaprakash, and T. Stützle. "F-Race and Iterated F-Race: An Overview". *Experimental Methods for the Analysis of Optimization Algorithms*. Ed. by T. Bartz-Beielstein, M. Chiarandini, L. Paquete, and M. Preuss. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 311–336.

[255] A. E. Eiben and S. K. Smit. "Evolutionary Algorithm Parameters and Methods to Tune Them". *Autonomous Search*. Ed. by Y. Hamadi, E. Monfroy, and F. Saubion. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 15–36.

[256] V. A. Tatsis and K. E. Parsopoulos. "Dynamic parameter adaptation in metaheuristics using gradient approximation and line search". *Applied Soft Computing* 74 (2019), pp. 368 –384. ISSN: 1568-4946.

[257] J. J. Durillo, J. García-Nieto, A. J. Nebro, C. A. C. Coello, F. Luna, and E. Alba. "5th Int. Conf. Evol. Multi-Criterion Optimization". Springer, 2009. Chap. Multi-Objective Particle Swarm Optimizers: An Experimental Comparison, pp. 495–509.

[258] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, 2007. ISBN: 1584888148, 9781584888147.

[259] J. J. Durillo, A. J. Nebro, C. A. C. Coello, J. Garcia-Nieto, F. Luna, and E. Alba. "A Study of Multiobjective Metaheuristics When Solving Parameter Scalable Problems". *IEEE Transactions on Evolutionary Computation* 14.4 (2010), pp. 618–635. ISSN: 1089-778X.

[260] I. Cantone, L. Marucci, F. Iorio, M. A. Ricci, V. Belcastro, M. Bansal, S. Santini, M. di Bernardo, D. di Bernardo, and M. P. Cosma. "A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches". *Cell* 137.1 (2009), pp. 172 –181. ISSN: 0092-8674.

[261] J. Pirgazi and A. R. Khanteymoori. "A robust gene regulatory network inference method base on Kalman filter and linear regression". *PLOS ONE* 13.7 (July 2018), pp. 1–17.

[262] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550.

[263] D. Klopfenstein, L. Zhang, B. S. Pedersen, F. Ramírez, A. W. Vesztrocy, A. Naldi, C. J. Mungall, J. M. Yunes, O. Botvinnik, M. Weigel, et al. "GOATOOLS: A Python library for Gene Ontology analyses". *Scientific reports* 8.1 (2018), p. 10872.

[264] C. A. Tilford and N. O. Siemers. "Gene set enrichment analysis". *Protein Networks and Pathway Analysis*. Springer, 2009, pp. 99–121.

[265] M. S. Brady, D. D. Eckels, S. Y. Ree, K. E. Schultheiss, and J. S. Lee. "MHC class II-mediated antigen presentation by melanoma cells." *Journal of immunotherapy with emphasis on tumor immunology: official journal of the Society for Biological Therapy* 19.6 (1996), pp. 387–397.

[266] J. E. Lee, J. D. Reveille, M. I. Ross, and C. D. Platsoucas. "HLA-DQB1* 0301 association with increased cutaneous melanoma risk". *International journal of cancer* 59.4 (1994), pp. 510–513.

[267] I. Navas-Delgado, J. García-Nieto, E. López-Camacho, M. Rybinski, R. Lavado, M. Á. Berciano Guerrero, and J. F. Aldana-Montes. "VIGLA-M: visual gene expression data analytics". *BMC Bioinformatics* 20.4 (2019), p. 150.

[268] K. González, J. Fuentes, and J. L. Márquez. "Physical inactivity, sedentary behavior and chronic diseases". *Korean journal of family medicine* 38.3 (2017), p. 111.

[269] W. L. Haskell, S. N. Blair, and J. O. Hill. "Physical activity: health outcomes and importance for public health policy". *Preventive medicine* 49.4 (2009), pp. 280–282.

[270] A. Bourke, J. O'brien, and G. Lyons. "Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm". *Gait & posture* 26.2 (2007), pp. 194–199.

[271] F. Bagala, C. Becker, A. Cappello, L. Chiari, K. Aminian, J. M. Hausdorff, W. Zijlstra, and J. Klenk. "Evaluation of accelerometer-based fall detection algorithms on real-world falls". *PloS one* 7.5 (2012).

[272] F. M. Palechor, A. De la Hoz Manotas, P. A. Colpas, J. S. Ojeda, R. M. Ortega, and M. P. Melo. "Cardiovascular Disease Analysis Using Supervised and Unsupervised Data Mining Techniques." *JSW* 12.2 (2017), pp. 81–90.

[273] D. Arifoglu and A. Bouchachia. "Activity recognition and abnormal behaviour detection with recurrent neural networks". *Procedia Computer Science* 110 (2017), pp. 86–93.

[274] G. Kalouris, E. I. Zacharaki, and V. Megalooikonomou. "Improving CNN-based activity recognition by data augmentation and transfer learning". *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*. Vol. 1. IEEE. 2019, pp. 1387–1394.

[275] A. Papagiannaki, E. I. Zacharaki, K. Deltouzos, R. Orselli, A. Freminet, S. Cela, E. Aristodemou, M. Polycarpou, M. Kotsani, A. Benetos, et al. "Meeting challenges of activity recognition for ageing population in real life settings". *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE. 2018, pp. 1–6.

[276] C. A. Ronao and S.-B. Cho. "Human activity recognition with smartphone sensors using deep learning neural networks". *Expert systems with applications* 59 (2016), pp. 235–244.

[277] Y. Saez, A. Baldominos, and P. Isasi. "A comparison study of classifier algorithms for cross-person physical activity recognition". *Sensors* 17.1 (2017), p. 66.

[278] T. Lv, X. Wang, L. Jin, Y. Xiao, and M. Song. "Margin-Based Deep Learning Networks for Human Activity Recognition". *Sensors* 20.7 (2020), p. 1871.

[279] F. Cruciani, A. Vafeiadis, C. Nugent, I. Cleland, P. McCullagh, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui. "Feature learning for Human Activity Recognition using Convolutional Neural Networks". *CCF Transactions on Pervasive Computing and Interaction* 2.1 (2020), pp. 18–32.

[280] J. R. Kwapisz, G. M. Weiss, and S. A. Moore. "Activity recognition using cell phone accelerometers". *ACM SIGKDD Explorations Newsletter* 12.2 (2011), pp. 74–82.

[281] A. Reiss and D. Stricker. "Introducing a new benchmarked dataset for activity monitoring". *2012 16th International Symposium on Wearable Computers*. IEEE. 2012, pp. 108–109.

[282] R. Chereshnev and A. Kertész-Farkas. "Hugadb: Human gait database for activity recognition from wearable inertial sensor networks". *International Conference on Analysis of Images, Social Networks and Texts*. Springer. 2017, pp. 131–141.

[283] M. Zhang and A. A. Sawchuk. "USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors". *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 2012, pp. 1036–1043.

[284] D. Balabka. "Semi-Supervised Learning for Human Activity Recognition Using Adversarial Autoencoders". *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. UbiComp/ISWC '19 Adjunct. London, United Kingdom: Association for Computing Machinery, 2019, 685–688. ISBN: 9781450368698.

[285] O. D. Lara and M. A. Labrador. "A survey on human activity recognition using wearable sensors". *IEEE communications surveys & tutorials* 15.3 (2012), pp. 1192–1209.

[286] A. Subasi, K. Khateeb, T. Brahimi, and A. Sarirete. "Human activity recognition using machine learning methods in a smart healthcare environment". *Innovation in Health Informatics*. Ed. by M. D. Lytras and A. Sarirete. Next Gen Tech Driven Personalized MedSmart Healthcare. Academic Press, 2020, pp. 123 –144. ISBN: 978-0-12-819043-2.

[287] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz. "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine". *International workshop on ambient assisted living*. Springer. 2012, pp. 216–223.

[288] L. Bao and S. S. Intille. "Activity recognition from user-annotated acceleration data". *International conference on pervasive computing*. Springer. 2004, pp. 1–17.

[289] W. Wu, S. Dasgupta, E. E. Ramirez, C. Peterson, and G. J. Norman. "Classification accuracies of physical activities using smartphone motion sensors". *Journal of medical Internet research* 14.5 (2012), e130.

[290] Y. Chen and Y. Xue. "A deep learning approach to human activity recognition based on single accelerometer". *2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE. 2015, pp. 1488–1492.

[291] N. Y. Hammerla, S. Halloran, and T. Plötz. "Deep, convolutional, and recurrent models for human activity recognition using wearables". *arXiv preprint arXiv:1604.08880* (2016).

[292] M. Zeng, T. Yu, X. Wang, L. T. Nguyen, O. J. Mengshoel, and I. Lane. "Semi-supervised convolutional neural networks for human activity recognition". *2017 IEEE International Conference on Big Data (Big Data)*. 2017, pp. 522–529.

[293] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu. "Sensor-based activity recognition". *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.6 (2012), pp. 790–808.

[294] Z. Hussain, M. Sheng, and W. E. Zhang. "Different approaches for human activity recognition: A survey". *arXiv preprint arXiv:1906.05074* (2019).

[295] A. Bulling, U. Blanke, and B. Schiele. "A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors". 46.3 (2014). ISSN: 0360-0300.

[296] *Activinsights*. https://activinsights.com/. Accessed: 2022-09-07.

[297] J. R. Kwapisz, G. M. Weiss, and S. A. Moore. "Activity Recognition Using Cell Phone Accelerometers". *ACM SIGKDD Explorations Newsletter* 12.2 (Mar. 2011), 74–82. ISSN: 1931-0145.

[298] A. Reiss, M. Weber, and D. Stricker. "Exploring and extending the boundaries of physical activity recognition". *2011 IEEE International Conference on Systems, Man, and Cybernetics*. 2011, pp. 46–50.

[299] M. Zhang and A. A. Sawchuk. "USC-HAD: A Daily Activity Dataset for Ubiquitous Activity Recognition Using Wearable Sensors". *ACM International Conference on Ubiquitous Computing (Ubicomp) Workshop on Situation, Activity and Goal Awareness (SAGAware)*. Pittsburgh, Pennsylvania, USA, 2012.

[300] R. Chereshnev and A. Kertész-Farkas. "HuGaDB: Human Gait Database for Activity Recognition from Wearable Inertial Sensor Networks". *Analysis of Images, Social Networks and Texts*. Ed. by W. M. van der Aalst, D. I. Ignatov, M. Khachay, S. O. Kuznetsov, V. Lempitsky, I. A. Lomazova, N. Loukachevitch, A. Napoli, A. Panchenko, P. M. Pardalos, A. V. Savchenko, and S. Wasserman. Cham: Springer International Publishing, 2018, pp. 131–141. ISBN: 978-3-319-73013-4.

[301] H. He and E. A. Garcia. "Learning from imbalanced data". *IEEE Transactions on knowledge and data engineering* 21.9 (2009), pp. 1263–1284.

[302] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[303] K. T. Nguyen, F. Portet, and C. Garbay. "Dealing with Imbalanced data sets for Human Activity Recognition using Mobile Phone sensors". 2018.

[304] S. Ertekin, J. Huang, L. Bottou, and L. Giles. "Learning on the border: active learning in imbalanced data classification". *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 2007, pp. 127–136.

[305] D. A. Cieslak, T. R. Hoens, N. V. Chawla, and W. P. Kegelmeyer. "Hellinger distance decision trees are robust and skew-insensitive". *Data Mining and Knowledge Discovery* 24.1 (2012), pp. 136–158.

[306] L. G. Fahad, S. F. Tahir, and M. Rajarajan. "Activity recognition in smart homes using clustering based classification". *2014 22nd International Conference on Pattern Recognition*. IEEE. 2014, pp. 1348–1353.

[307] K. Kunze, P. Lukowicz, H. Junker, and G. Tröster. "Where am i: Recognizing on-body positions of wearable sensors". *International Symposium on Location-and Context-Awareness*. Springer. 2005, pp. 264–275.

[308] S. Ioffe and C. Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". *arXiv preprint arXiv:1502.03167* (2015).

[309] Z. Zhang. "Improved adam optimizer for deep neural networks". *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*. IEEE. 2018, pp. 1–2.

[310] J. Snoek, H. Larochelle, and R. P. Adams. "Practical bayesian optimization of machine learning algorithms". *Advances in neural information processing systems*. 2012, pp. 2951–2959.

[311] A. D. Antar, M. Ahmed, and M. A. R. Ahad. "Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: a review". *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE. 2019, pp. 134–139.

[312] S. Slim, A. Atia, M Elfattah, and M Mostafa. "Survey on human activity recognition based on acceleration data". *International Journal of Advanced Computer Science and Applications* 10 (2019), pp. 84–98.

[313] A. Gupta, K. Gupta, K. Gupta, and K. Gupta. "A Survey on Human Activity Recognition and Classification". *2020 International Conference on Communication and Signal Processing (ICCSP)*. IEEE. 2020, pp. 0915–0919.

[314] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. "Using mobile phones to determine transportation modes". *ACM Transactions on Sensor Networks (TOSN)* 6.2 (2010), pp. 1–27.

[315] F. Foerster, M. Smeja, and J. Fahrenberg. "Detection of posture and motion by accelerometry: a validation study in ambulatory monitoring". *Computers in human behavior* 15.5 (1999), pp. 571–583.

[316] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". *Information Fusion* 58 (2020), pp. 82–115. ISSN: 1566-2535.

[317] M. A. Gulum, C. M. Trombley, and M. Kantardzic. "A Review of Explainable Deep Learning Cancer Detection Models in Medical Imaging". *Applied Sciences* 11.10 (2021).

[318] L. Edwards and M. Veale. "Enslaving the algorithm: from a "right to an explanation" to a "right to better decisions"?" *IEEE Security & Privacy* 16.3 (2018), pp. 46–54.

[319] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K. Müller. "Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond". *CoRR* abs/2003.07631 (2020).

[320] S. Knapič, A. Malhi, R. Saluja, and K. Främling. "Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain". *Machine Learning and Knowledge Extraction* 3.3 (2021), pp. 740–770.

[321] S. Banerjee, S. K. Singh, A. Chakraborty, A. Das, and R. Bag. "Melanoma Diagnosis Using Deep Learning and Fuzzy Logic". *Diagnostics* 10.8 (2020). ISSN: 2075-4418.

[322] A. Naeem, M. S. Farooq, A. Khelifi, and A. Abid. "Malignant Melanoma Classification Using Deep Learning: Datasets, Performance Measurements, Challenges and Opportunities". *IEEE Access* 8 (2020), pp. 110575–110597.

[323] M. T. Ribeiro, S. Singh, and C. Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". Association for Computing Machinery, 2016.

[324] S. M. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, 4768–4777. ISBN: 9781510860964.

[325] A. Malhi, S. Knapic, and K. Främling. "Explainable agents for less bias in human-agent decision making". *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Springer. 2020, pp. 129–146.

[326] S. Knapič, A. Malhi, R. Saluja, and K. Främling. "Explainable artificial intelligence for human decision support system in the medical domain". *Machine Learning and Knowledge Extraction* 3.3 (2021), pp. 740–770.

[327] K. Främling12. "Contextual importance and utility in R: the 'ciu'package" (2021).

[328] J. Petch, S. Di, and W. Nelson. "Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology". *Canadian Journal of Cardiology* 38.2 (2022), pp. 204–213. ISSN: 0828-282X.

[329] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. Díaz-Rodríguez. "Explainable artificial intelligence (xai) on timeseries data: A survey". *arXiv preprint arXiv:2104.00950* (2021).

[330] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis". *Medical Image Analysis* (2022), p. 102470.

[331] S. M. Humphries, A. M. Notary, J. P. Centeno, M. J. Strand, J. D. Crapo, E. K. Silverman, D. A. Lynch, and G. E. of COPD (COPDGene) Investigators. "Deep learning enables automatic classification of emphysema pattern at CT". *Radiology* 294.2 (2020), pp. 434–444.

[332] P. Zhu and M. Ogino. "Guideline-based additive explanation for computer-aided diagnosis of lung nodules". *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer, 2019, pp. 39–47.

[333] S. M. Lundberg and S.-I. Lee. "A unified approach to interpreting model predictions". *Advances in neural information processing systems* 30 (2017).

[334] B. H. van der Velden, M. H. Janse, M. A. Ragusi, C. E. Loo, and K. G. Gilhuijs. "Volumetric breast density estimation on MRI using explainable deep learning regression". *Scientific Reports* 10.1 (2020), pp. 1–9.

[335] A. Malhi, T. Kampik, H. Pannu, M. Madhikermi, and K. Främling. "Explaining machine learning-based classifications of in-vivo gastral images". *2019 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE. 2019, pp. 1–7.

[336] K. Hauser, A. Kurz, S. Haggenmüller, R. C. Maron, C. von Kalle, J. S. Utikal, F. Meier, S. Hobelsberger, F. F. Gellrich, M. Sergon, et al. "Explainable artificial intelligence in skin cancer recognition: A systematic review". *European Journal of Cancer* 167 (2022), pp. 54–69.

[337] M. Sadeghi, P. K. Chilana, and M. S. Atkins. "How users perceive content-based image retrieval for identifying skin images". *Understanding and interpreting machine learning in medical image computing applications*. Springer, 2018, pp. 141–148.

[338] M. Sadeghi, P. Chilana, J. Yap, P. Tschandl, and M. S. Atkins. "Using content-based image retrieval of dermoscopic images for interpretation and education: A pilot study". *Skin Research and Technology* 26.4 (2020), pp. 503–512.

[339] S. Jiang, H. Li, and Z. Jin. "A visually interpretable deep learning framework for histopathological image-based skin cancer diagnosis". *IEEE Journal of Biomedical and Health Informatics* 25.5 (2021), pp. 1483–1494.

[340] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images". *Nature medicine* 25.8 (2019), pp. 1301–1309.

[341] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, et al. "Human–computer collaboration for skin cancer recognition". *Nature Medicine* 26.8 (2020), pp. 1229–1234.

[342] P. V. Molle, M. D. Strooper, T. Verbelen, B. Vankeirsbilck, P. Simoens, and B. Dhoedt. "Visualizing convolutional neural networks to improve decision support for skin lesion classification". *Understanding and interpreting machine learning in medical image computing applications*. Springer, 2018, pp. 115–123.

[343] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.