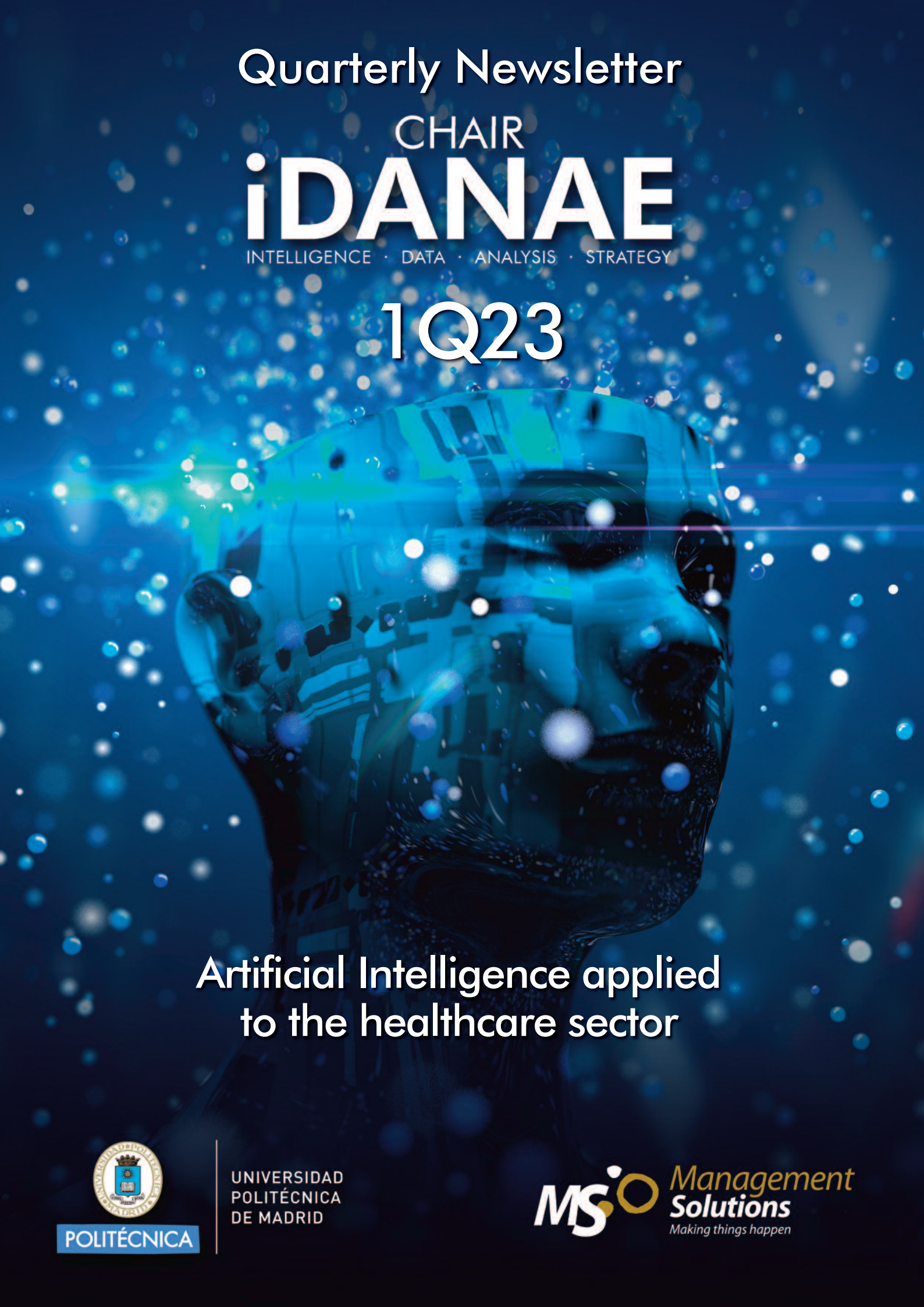
Quarterly Newsletter

CHAIR

iDANAE

INTELLIGENCE · DATA · ANALYSIS · STRATEGY

1Q23



Artificial Intelligence applied to the healthcare sector

POLITÉCNICA

UNIVERSIDAD POLITÉCNICA DE MADRID

MS Management Solutions

Making things happen

Introduction

The fast development of technology during last years has led to the generation of big amounts of data available from different sources of information, and the possibility of its treatment and analysis. In the healthcare sector, these data make it possible to carry out more effective studies for many applications, such as diagnosis, disease prevention, understanding of virus evolution, or pharmaceutical developments, among many others. Indeed, investments in healthcare technology can not only save lives but also make health systems more efficient.

Experience-based medicine is being replaced by an evidence-based and patient-centered approach. The data-driven identification of disease states and treatment options is a crucial challenge for precision medicine. Artificial intelligence (AI) is emerging as a new critical tool to improve predictive capabilities for disease diagnosis and treatment outcomes in both the laboratory and clinic [1] [2].

AI, and more generally machine learning (ML), has a significant potential and should be further expanded to achieve innovative progress that can transform this field of knowledge. However, this potential of development requires the application of modelling approaches that can help researchers gain a better understanding of biological processes and systems. In addition, these approaches need to address current problems, such as working with multiple heterogeneous data types, insufficient or low-quality data, the need for interpretability and explanability of the results, or the use of alternative learning approaches.

Being able to handle large amounts of data is a challenge from the technical point of view. To have enough data to obtain useful insights from it, different sources of information are usually involved (e.g., information from hospitals, wearables, pharma industry, etc.). Furthermore, it is required to have a technical knowledge of Big Data and Machine Learning techniques to be able to address the complex challenges that the healthcare sector has dealt with for a long time. The modelling of biological and chemical data could bring important advances in complex problems of the industry, such as drug discovery or pharmacovigilance.

A good example that shows the advantage of the use of big amounts of data as well as the increase in computing speed is the fast development of the COVID-19 vaccines [3]. Moreover, there are a lot of advantages in drug discovery. For instance, these tools can be very important on the identification of antibiotics. As an example, they have been proven to be very useful to discover certain antibacterial treatment, a task defined as a priority by the World Health Organization [4].

These examples show that, although the technical effort in terms of Big Data and Data Science is substantial, the benefits obtained could be priceless. For example, the pharmaceutical industry revenues are constantly growing, reaching 1,482 billion dollars worldwide in 2022 [5]. Furthermore, public initiatives like the Next Generation EU funds are investing economic resources and funds to improve health public services.

Nevertheless, it is essential to consider the importance of accuracy, reliability, and interpretation of results for its application in this sector. Potential errors in models can lead to severe consequences such as incorrect diagnoses, which can negatively impact health, result in wasted funds on ineffective treatments, or lead to fruitless research. Consequently, it is vital for medical professionals and researchers to utilize these results as a tool. Furthermore, the ethical aspect of the use of these algorithms must be considered (for example, who is responsible for a mistake within the decision-making process for a diagnosis, the doctor, or the algorithm?). These aspects include not only the responsibility, but the complete life cycle of the model [6].

In this document, some of the different types of data repositories available in the healthcare sector are reviewed. Then, three different successful use cases are explained to illustrate the possibilities of the application of AI techniques and specific data types. Later, the general challenges and opportunities in the field are discussed. Finally, main conclusions are presented.

Some data repositories

Within the healthcare sector, the process of data generation occurs across different units and geographies, not necessarily connected (for example, different hospitals, research groups, studies, etc.). To take advantage of the data produced and make data exploitation, a proper process of data collection is necessary. The lack of sharing of information between different companies has led to the creation of public initiatives to collect and share data.

For example, the European Union is working on the European Health Data Space, which will allow individuals to control and use their health data in their home country or in other Member States by fostering a true single market for digital health services and products and by providing a consistent, reliable and efficient framework for using health data for research, innovation, policy development and regulatory activities, while ensuring full compliance with the EU's strict data protection rules.

There may be many different sources of data repositories. For example, digital medical records, data from intensive unit care, treatments and monitoring, drug catalogs, labor data, social networks, wearables, omics, or images. Each of these many types of data have their own specificities. As an example, some of these types of repositories can be highlighted: diseases type catalogs, image repositories, omics data and wereables.

- ***Diseases type catalogs.*** There are currently thousands of diseases catalogued according to different standards, databases, and vocabularies. The total number of diseases is very difficult to determine. Part of this catalogues is sometimes associated with the defining characteristics of the diseases. Using the information on these characteristics is vital to understand how the diseases look and behave.

- ***Image repositories:*** Medical imaging data, such as X-rays, CT scans, MRI scans, and ultrasound images, can be used for data science projects in the healthcare sector. A well-known initiative for public download of images medical images regarding cancer is The Cancer Imaging Archive (TCIA) [7].

- ***Omics data:*** Omics data refers to large-scale biological data generated by high-throughput technologies that study different types of biomolecules, including genomics, transcriptomics, proteomics, metabolomics and epigenomics. Analyzing and integrating omics data can provide a more comprehensive view of biological systems and aid in drug development and disease prevention. Many initiatives have been taken with omics data. Examples are: Paired Omics Data Platform [8], European Genome-Phenome Archive [9], Chinese CNSA [10], or The Cancer Genome Atlas Program (TCGA) [11].

- ***Wearables***: these data refer to data generated by wearable devices, such as smartwatches, fitness trackers and other portable sensors that are designed to track and monitor various physiological and activity-related parameters of an individual, such as heart rate, steps taken, sleep patterns, and calories burned. Nowadays, it is usual to find cases where the sensors incorporated in smartwatches have saved lives by identifying anomalous health metrics. A good example of wearable data collection project is the Open Wearables Initiative (OWEAR), which is a collaboration designed to promote the effective use of high-quality, sensor-generated measures of health in clinical research through the open sharing of algorithms and data sets [12].

Successful Applications of Data Science to Healthcare

There are many different use cases where the AI has been successfully applied within the healthcare sector. Some of the cases show the use of different data repositories and modelling techniques, and the diversity of results that can be obtained. In this section, three successful use cases will be discussed to illustrate the possibilities of the application of AI in healthcare sector:

- Drug discovery and repositioning
- Biomedical imaging and Pharmaceutical Research
- Omics data modelling

Drug discovery and repositioning

Drug discovery for the treatment of diseases caused from pathogenic microorganisms is a long and expensive process [13]. It means analyzing data from disease catalogues and omics data showing molecules that could produce the inhibition of microbial. This involves the screening of a significant amount of synthetic chemical libraries. The analyzed information is hosted in databases such as ChEMBL or ZINC. Then, the molecules which show biological activity, more commonly known as "hits", are identified. The process continues with the optimization of the active compound to increase the biological activity [14].

Recently, the training of Deep Learning models has led to accelerate the task of finding the active molecules and it is useful to discover new antibacterial treatments [4]. They help to reduce considerably the time spent and to improve the accuracy of the predictions. For example, in Stokes et al. 2022 [4], a deep neural network can identify the so-called active molecule Halicin. This molecule is effective in the treatment of bacterial infections such the ones produced by Clostridioides difficile, a microorganism that could cause an inflammation of the colon. Furthermore, it has shown that it can be useful against Acinetobacter Baumannii bacteria, qualified by the World Health Organization as a priority pathogen against which new antibiotics are needed.

The method implemented is based on two stages:

1. First, data from different databases which have information about active molecules were binarized to build the input

data set. The binarization was performed to indicate whether the molecules were hit or not.

2. Then, these data were used to train a binary classification model based on a deep neural network that predicts the probability of whether a new compound will inhibit the bacteria growthy, based on its structure. To feed the model, the graph representation of a molecule is transformed into a continuous vector.

In addition, these new computational tools have been applied to identify drug candidates against other infectious microorganisms such as parasites. For instance, in Neves et al. 2020 [15], artificial intelligent approaches are used to find therapeutics solutions to Malaria, a severe tropical disease caused by parasites of Plasmodium genus. As it occurs with other pathogenic microorganisms, parasites can mutate and show an increased resistance to drugs. Therefore, it is needed to identify new more effective drugs which replace the old ones.

Neves et al. 2020 [15] propose the deployment of quantitative structure-activity relationships (QSAR) modeling. QSAR models are capable to establish relationships between the chemical structure attributes and biological activity of compounds. The methodology goes as follows:

- First, the chemical characteristics are converted to molecular descriptors. They will be the independent variables of the problem. The biological activity will be transformed into the dependent variable of the problem.

- Second, a specific predictive model is used. Linear regression models had been the preferred ones until new powerful techniques appeared. Bayesian neural networks, Support Vector Machine, Random Forests and Deep Neural Networks have now replaced this statistical method.

- Finally, the results predicted by the model are applied to carry out the virtual screening (VS). VS is a usual procedure in drug development and is based on testing all necessary hypotheses before conducting clinical trials.

Furthermore, the characteristics of diseases can be utilized for drug repositioning, which involves trying to use drugs that have already been approved and are in use to try to treat a different disease from their original indication. This can be used both to find treatments for diseases that do not have them, and to find therapeutic alternatives to diseases that already have drugs associated with them, which can also be very useful.

This may be done using the so-called "disease networks": graphs where diseases can be related through their characteristics under the assumption that two diseases share some of these elements. This is discussed for example in the DISNET project. The aim of DISNET project is to integrate in a single database information from public and heterogeneous sources to allow the analysis and understanding of the relationships among diseases. As part of the knowledge that will be derived from this understanding, the project aims to create disease networks that can be analyzed to create new strategies in terms of drug repurposing, mainly applied to rare diseases[1].

Biomedical imaging and Pharmaceutical Research

Computer vision is used in healthcare sector to perform many tasks. Applications like tumor detection, automatic cell counting, or rapid disease early detection are used to assist healthcare professionals.

In pharmaceutical industry, computer vision is used to solve different problems and to improve the efficiency of certain processes. Different techniques from image analysis and biological imaging can be used in processes like clinical trials or preclinical studies. Recently, Deep Learning techniques like Convolutional Neural Networks (CNN) have become the preferred method for scientists to carry out image processing tasks, because of the efficiency and good results they offer.

One interesting activity in healthcare sector is to investigate composition of cells, since it enables to analyze, for instance, tumor characteristics. However, it represents a challenging task, as the cells usually show an heterogenous nature. In this context, the mentioned deep learning-based methods are used to carry out image-analysis tasks through cell images. These activities are, for example, hit identification [16], phenotypic screening [17] or phenotype classification [18].

CNNs have replaced conventional automatic analysis techniques [19]. Before the emergence of CNNs, image processing methods were applied to identifying the region of interest (ROI). Then, different features were extracted and converted to features, i.e., a series of vectors. Finally, these features are used to train a statistical classification model such as linear discriminant analysis (LDA) or a support vector machine (SVM). Thus, information about the effects of drugs and diseases can be quantified.

[1]See https://medal.ctb.upm.es/disnet/ for further information

However, CNNs does not need to define these features; they can learn the images characteristics by themselves. CNNs not only give better results in classification and segmentation tasks, but also it enables to perform other duties such as the improvement on the resolution of images or the elimination of image noise. In addition, these new techniques reduce the high computational cost that involves the extraction of features and the selection of analysis methods (figure 1).

To sum up, it is known that CNNs provide more useful data for activities realized on biological field than traditional approaches. That is, the high accuracy that presents these methods and the efficiency mean an important advance in drug discovery.

Omics data modelling

Omics data covers information from multiple biodata sets [20]: genomics, phenomics, epigenomics, pharmacogenomics, transcriptomics, proteomics, metabolomics, lipidomics, etc.

From a genomic modelling perspective, one of the potential sources to be exploited by data science field is the analysis of nucleotides strands viruses. Viruses are infectious microorganisms whose main components are nucleic acids biomolecules. There are viruses that are ADN-based and others which are ARN-based. Apart from having distinct structures, these viruses differ also on mutation rate, as well as how and where they carry out processes like replication.

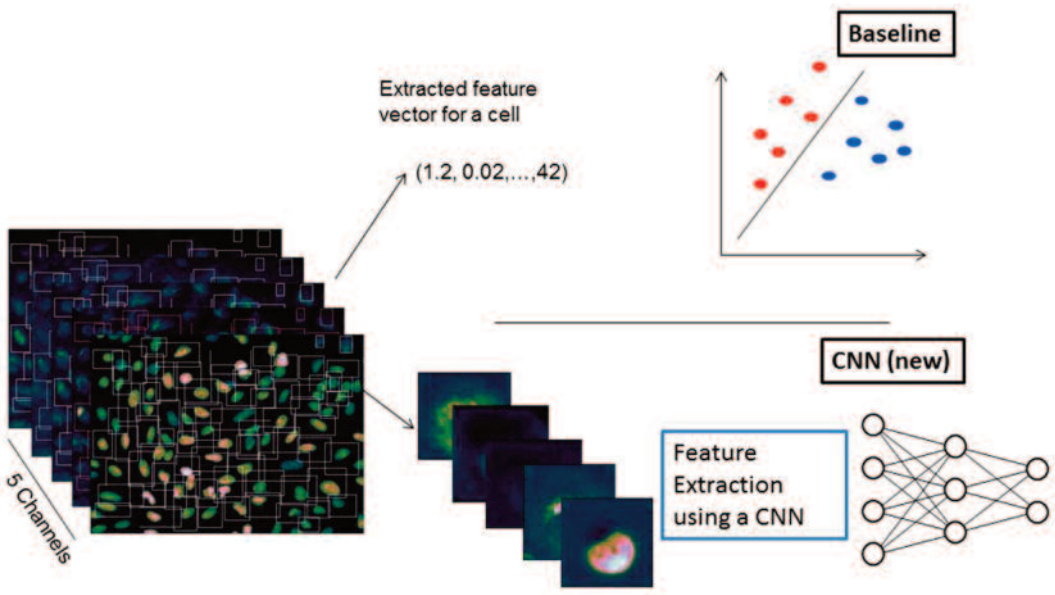
Two characteristics of viruses are the high mutation and replication rates. As virus replicates inside the host, the nucleotides sequence changes. Consequently, some of its attributes can be modified, such as adaptation capacity or virulence. Thus, the understanding of this viral evolution has been a complex task for scientists over the years.

Therefore, modelling the complex behavior of viruses could bring interesting benefits for pharmaceutical industry: identifying patterns and providing information about dynamics of virus evolution could lead to design more effective antiviral therapies. Machine learning tools can be used to carry out this task. Several works use them and propose different approaches to deal with this problem.

Data sets used as input of Machine Learning and Deep Learning models are based on nucleotide sequences. In these sequences of nucleotides, four types of nucleotides can be found, each associated to a unique letter. Thus, these sequences can be treated in two different ways:

- Nucleotides as characters: it does not require any preprocessing, as the sequence is introduced as input directly.
- Nucleotides as numerical codification result: in this scenario, each nucleotide is transformed into numerical vectors.

Figure 1. Feature extraction: handcrafted feature extraction against automated CNN extraction. Source: Dürr et al. 2016 [8].



In addition, nucleotide strands can be converted into amino acids sequences. It is carried out following the Standard Genetic Code (SGC). This code assigns specific amino acids to groups of three nucleotides, meaning that each sequence of three nucleotides will correspond to a specific amino acid. As a result, the data set will consist of sequences of amino acids instead of nucleotide sequences. These data sets could be interesting if the objective is, for example, to predict what three-dimensional structure will adopt a protein [21].

Other techniques under study are the use of neural networks to predict mutations in nucleotides sequences. Mostafa et al. 2016 [22] propose to feed the neural network with a sequence as input, and to achieve as output the immediately following sequence of the next generation. Therefore, the sequences in the training set are sorted in time according with the viral behavior dynamics, i.e., time series data. For the training of this model, each nucleotide letter is encoded using one-hot encoding.

This training finishes when the accuracy reaches 70%. However, this technique shows some drawbacks:

- Neural networks are black-box models. It means that these models attempt to approximate a complex function of which it is not possible to get insights about. Thus, it is not possible to obtain rules or patterns from the model.
- Computational complexity. Encoding the nucleotide sequences following the previous scheme leads to increase four times the length of the sequences.
- To train the model, a large data set with many sequences is needed. Additionally, simulating a virus' behavior in the laboratory for studying its fitness landscape is a time-consuming and costly process.

Finally, another interesting research line is developing by Delgado et al. [23]. In this paper, a machine learning technique is used to find out similarities between different populations of hepatitis C virus. Here, Self-Organized Maps method is implemented [24]. This unsupervised machine learning technique provides a 2D-dimensional representation preserving the topological structure of the data. Unlike other unsupervised methods, SOM allows to visualize the results in a 2D format. This makes it a powerful technique since it can cluster and reduce high-dimensional data. Results given by the SOM suggests that as the virus mutates, the number of clusters increases. That is, it can be found more differences between nucleotide sequences as the virus develops in the cells.

Challenges and opportunities

Although there are many examples of successful use cases, health data is highly fragmented, and is affected by many barriers that include, among others, heterogeneous international ethical and legal compliance frameworks, as well as issues of data ownership, trust, and traceability. This undefined data landscape hides untapped potential that, when uncovered, can lead to significant improvements in health and healthcare innovation through innovative data science. In addition, the healthcare sector faces many challenges due to the complexity and diversity of biomedical data. Some of the most relevant issues in the field are the following:

- ***Data Quality:*** the quality of health data can vary significantly depending on the source and the method of collection. Cleaning, aggregating, and structuring this data into a usable format is a major challenge.

- ***Privacy and Security:*** health data is sensitive, and privacy concerns are a major challenge in the sector. Data must be securely stored, and access restricted to authorized individuals.

- ***Interoperability:*** health data is often stored in different formats and systems that do not easily communicate with each other. Integrate and harmonize data from various sources to make it more useful for analysis is crucial for later use.

- ***Implementation:*** even with simple data science techniques, implementing changes in the healthcare sector can be challenging. Data scientists must work with healthcare professionals to ensure that their findings are transformed into meaningful actions that improve patient outcomes.

However, the routine nature of data collection in the health system provides an invaluable and continuously updated source of information that can be used for self-improvement of health systems and to rapidly inform the decision-making processes to improve health care and public health strategies in general. In addition, the use of data science techniques offers several major business opportunities for companies in the industry. Indeed, according to OECD, data and digital technologies in health can result in a 30 per cent decrease in waste and inefficiencies (almost €400 billion per year) [25]. As an example, it is worth mentioning the following:

- Data management and analytics: as the quality of health data can vary widely and be complex, companies specializing in health data management and analytics can help healthcare organizations to collect, clean, and analyze their data to improve patient outcomes. These companies can provide data solutions that address data quality, privacy, and interoperability issues to help healthcare providers make more informed decisions.

- Development and application of machine learning and AI technologies: AI techniques can help achieve new discoveries, and healthcare providers can offer more personalized and effective treatments to their patients.

- Medical devices and wearables: these devices collect data on patient health and can be used to improve the quality of health data. This data helps healthcare professionals to provide more personalized care. In addition, medical devices and wearables can provide healthcare providers with real-time data on patient health, enabling early detection of potential health issues and the development of more personalized treatment plans.

In conclusion, the challenges facing the healthcare sector in utilizing data science techniques offer several major business opportunities. Companies that can provide data management and analytics solutions, develop machine learning and AI technologies, or provide healthcare consulting services, among many others, may have the potential for significant growth and success in the healthcare sector.

Conclusions

The advances in technology and the availability of new sources of data has opened big advances in the healthcare sector. The increase of speed in computation has brought the possibility of applying complex data science algorithm to health data. This provides insights that can be useful for finding new drugs or in disease prevention.

Examples of data repositories useful to obtain valuable outputs for healthcare research include diseases type catalogs, image repositories, omics data and wearables, among many others. The use of advanced modelling techniques capitalizes from all these different data repositories. Public initiatives are being developed to make sources of information and algorithms openly accessible to healthcare researchers.

In summary, the exploitation of health data and the application of AI in the healthcare sector are complex tasks that require expert knowledge and have some big challenges, but also create new business opportunities that can be exploited.

Bibliography

[1] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[2] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," 2015, pp. 234-241.

[3] R. Patel, M. Kaki and V. S. Potluri, "A comprehensive review of SARS-CoV-2 vaccines: Pfizer, Moderna & Johnson & Johnson," HUMAN VACCINES & IMMUNOTHERAPEUTICS , vol. 18, no. 1, 2022.

[4] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French and L. A. Carfrae, "A Deep Learning Approach to Antibiotic Discovery," Cell Press, vol. 180, no. 4, pp. 68-702, 2020.

[5] "Pharmaceutical market worldwide revenue 2001-2022 | Statista," 3 Feb 2023. [Online]. Available: https://www.statista.com/statistics/263102/pharmaceutical-market-worldwide-revenue-since-2001/.

[6] iDanae, "Ethics and Artificial Intelligence. Quarterly newsletter 4Q19," iDanae Chair, 2019.

[7] "The Cancer Imaging Archive," [Online]. Available: https://www.cancerimagingarchive.net/.

[8] "Paired Omics Data Platform," [Online]. Available: https://pairedomicsdata.bioinformatics.nl/.

[9] M. A. e. a. Freeberg, "The European Genome-phenome Archive in 2021," Nucleic Acids Research, 2022.

[10] X. e. a. Guo, "CNSA: a data repository for archiving omics data," The Journal of Biological Databases and Curation, 2020.

[11] "The Cancer Genome Atlas Program," [Online]. Available: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

[12] "Open Wearables Initiative," [Online]. Available: https://www.owear.org/.

[13] Y. Zhang, T. Ye, X. Hui, M. Juhas and J. Li, "Deep Learning Driven Drug Discovery: Tackling Severe Acute Respiratory Syndrome Coronavirus 2," Frontiers in Microbiology, vol. 12, 2021.

[14] F. Saldívar-González, F. D. Prieto-Martínez and J. L. Medina-Franco, "Descubrimiento y desarrollo de fármacos: un enfoque computacional," Educación Química, vol. 28, pp. 51-58, 2018.

[15] B. J. Neves, R. C. Braga, V. M. Alves, M. N. N. Lima, G. C. Cassiano, E. N. Muratov and e. a. , "Deep Learning-driven research for drug," PLoS Computational Biology, vol. 16, no. 2, pp. e1007025-e1007046, 2020.

[16] L. David, J. Arús-Pous, J. Karlsson, O. Engkvist, E. . J. Bjerrum, T. Kogej and e. a. , "Applications of Deep-Learning in Exploiting Large-Scale and Heterogeneous Compound Data in Industrial Pharmaceutical Research," Frontiers in Pharmacology, vol. 10, 2019.

[17] D. Siegismund, V. Tolkachev, S. Heyse, B. Sick, O. Duerr and S. Steigele, "Developing Deep Learning Applications for Life Science and Pharma Industry," Drug Res, vol. 68, no. 06, pp. 305-310, 2018.

[18] O. Dürr and B. Sick, "Single-Cell Phenotype Classification Using Deep Convolutional Neural Networks," Journal of Biomolecular Screening, vol. 21, no. 9, pp. 998-1003, 2016.

[19] M. Boutros, F. Heigwer and C. Laufer, "Microscopy-Based High-Content Screening," Cell, vol. 163, no. 6, pp. 1314-1325, 2015.

[20] A. B. C. N. D. e. a. Monte, "Improved drug therapy: triangulating phenomics with genomics and metabolomics," Human Genomics, 2014.

[21] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov and O. Ronneberger, "Highly accurate protein structure prediction with AlphaFold," Nature, vol. 596, p. 583–589, 2021.

[22] S. A. Mostafa, M. E. Hassanien and A. Mostafa, "The prediction of virus mutation using neural networks and rough set techniques," EURASIP J Bioinform Syst Biol, vol. 10, 2016.

[23] S. Delgado, C. Perales, C. García-Crespo, M. E. Soria, I. Gallego, A. I. de Ávila and e. a. , "A Two-Level, Intramutant Spectrum Haplotype Profile ofHepatitis C Virus Revealed by Self-Organized Maps," Microbiology Spectrum, vol. 9, no. 3, pp. e01459-21, 2021.

[24] T. Kohonen, "The self-organizing map," Proceedings of the IEEE, vol. 78, no. 9, pp. 1464 - 1480, 1990.

[25] OECD, "Health in the 21st Century. Putting data to work for stronger health systems," OECD Health Policy Studies, Paris, 2019.

Authors

Ernestina Menasalvas (UPM)
Manuel Ángel Guzmán (Management Solutions)
Sergio Ruiz (Management Solutions)
Javier Muñoz (Management Solutions)
Mario Hernando (Management Solutions)

UNIVERSIDAD POLITÉCNICA DE MADRID

The Universidad Politécnica de Madrid is a multi-sector and multi-disciplinary Public Law Entity, which carries out teaching, research and scientific and technological development activities.

www.upm.es

Management Solutions
Making things happen

Management Solutions is an international consulting firm, focused on business, finance, risk, organization, technology and process consulting, operating in more than 50 countries and with a team of over 3,300 professionals working for more than 1,500 clients worldwide.

www.managementsolutions.com

For more information visit

blogs.upm.es/catedra-idanae/