Exploration of Foods and Foodomics

Open Exploration

Open Access Perspective

From data to nutrition: the impact of computing infrastructure and artificial intelligence

Pierpaolo Di Bitonto[1†], Michele Magarelli[1†], Pierfrancesco Novielli[1,2], Donato Romano[1,2], Domenico Diacono[2], Lorenzo de Trizio[1], Angelo Mariano[3], Claudia Zoani[4], Riccardo Ferrero[5], Alessandra Manzin[5], Maria De Angelis[1], Roberto Bellotti[2,6], Sabina Tangaro[1,2*]

[1]Department of Soil, Plant and Food Sciences, University of Bari Aldo Moro, 70126 Bari, Italy

[2]National Institute of Nuclear Physics, Bari Section, 70126 Bari, Italy

[3]Department of Energy Technologies and Renewable Sources (TERIN), ICT Division, National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), 70124 Bari, Italy

[4]National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), Department of Sustainability, Circularity and Adaptation to Climate Change of Production and Territorial Systems, Biotechnology and Agro-industry Division, Research Center Casaccia, 00123 Rome, Italy

[5]National Institute of Metrological Research (INRIM), 10135 Turin, Italy

[6]Interuniversity Department of Physics "M. Merlin", University of Bari Aldo Moro, 70126 Bari, Italy

[†]These authors contributed equally to this work.

***Correspondence:** Sabina Tangaro, Department of Soil, Plant and Food Sciences, University of Bari Aldo Moro, 70126 Bari, Italy. sabina.tangaro@uniba.it

Academic Editor: Jose Antonio Mendiola, Institute of Food Science Research (CIAL-CSIC), Spain

Received: May 20, 2024 **Accepted:** November 13, 2024 **Published:** December 3, 2024

Cite this article: Di Bitonto P, Magarelli M, Novielli P, Romano D, Diacono D, de Trizio L, et al. From data to nutrition: the impact of computing infrastructure and artificial intelligence. Explor Foods Foodomics. 2024;2:810–29. https://doi.org/10.37349/eff.2024.00063

Abstract

This article explores the significant impact that artificial intelligence (AI) could have on food safety and nutrition, with a specific focus on the use of machine learning and neural networks for disease risk prediction, diet personalization, and food product development. Specific AI techniques and explainable AI (XAI) are highlighted for their potential in personalizing diet recommendations, predicting models for disease prevention, and enhancing data-driven approaches to food production. The article also underlines the importance of high-performance computing infrastructures and data management strategies, including data operations (DataOps) for efficient data pipelines and findable, accessible, interoperable, and reusable (FAIR) principles for open and standardized data sharing. Additionally, it explores the concept of open data sharing and the integration of machine learning algorithms in the food industry to enhance food safety and product development. It highlights the METROFOOD-IT project as a best practice example of implementing advancements in the agri-food sector, demonstrating successful interdisciplinary collaboration. The project fosters both data security and transparency within a decentralized data space model, ensuring reliable and efficient data sharing. However, challenges such as data privacy, model interoperability, and ethical considerations remain key obstacles. The article also discusses the need for ongoing interdisciplinary collaboration between data scientists, nutritionists, and food technologists to effectively address these challenges. Future research should focus on refining AI models to improve their reliability and exploring how to integrate these technologies into everyday nutritional practices for better health outcomes.

Keywords

Food, food contaminants, computing infrastructure, data analysis, machine learning, artificial intelligence, microbiome, health risks

Introduction

The field of nutrition is undergoing a significant transformation driven by the growing importance of data and artificial intelligence (AI) [1]. These advances are revolutionizing the way we analyze, understand, and adapt dietary practices. AI techniques enable us to unlock the intricate network of relationships between food and health by enabling the analysis of vast amounts of data. These data can be diverse, including information from clinical studies, population surveys, biometric measurements, and individual diet behavior. By identifying meaningful relationships and patterns within this data, AI offers a powerful tool for personalizing dietary recommendations and optimizing nutritional strategies [1–3]. However, simply possessing significant computing power is not sufficient to fully unlock the potential of AI in nutrition. To efficiently manage and analyze the ever-growing volume of data, robust data analysis pipelines are essential. These pipelines require careful engineering to ensure continuous processing of new data streams. Data operations (DataOps) emerge as a key approach to address these challenges. It focuses on enhancing efficiency and collaboration within an organization's data management processes [4]. DataOps is applying development and operations (DevOps) principles to data management. It aims to automate data pipelines and improve collaboration between teams to deliver reliable and timely data for business needs. It focuses on collaboration, automation, and continuous processes to deliver software faster and more reliably. DataOps aims to integrate all aspects of the data lifecycle, including people, processes, and technologies. This holistic approach tackles the challenges associated with data collection, processing, management, and distribution. Similar to the DevOps approach used in software development, DataOps fosters collaboration and breaks down silos between teams. This streamlined communication facilitates a smoother and more responsive flow of data.

This article outlines the significant impact that AI has had on nutrition, highlighting the crucial role played by computing infrastructures in analyzing such data. It explores how AI can be utilized to enhance our understanding of the relationship between food and health, with particular emphasis on disease risk prediction, diet personalization, dietary behavior analysis, and the development of novel food products. Furthermore, it examines the pivotal role of high-performance computing (HPC) infrastructures and the DataOps approach in optimizing nutritional data analysis. Interdisciplinary collaboration is essential for driving advancements in AI within the field of nutrition. Data scientists, nutritionists, and food technologists each bring unique expertise that is crucial to the successful development and application of AI models. Data scientists contribute advanced machine learning (ML) and data analytics techniques, while nutritionists provide insights into human metabolism and dietary needs. Food technologists ensure that practical aspects of food production and quality control are taken into account. One successful example of such collaboration is the METROFOOD-IT project, which integrates expertise from these diverse fields to create AI-driven models to improve food safety, quality, and personalized nutrition. This interdisciplinary approach fosters innovation and ensures that AI models are scientifically robust and practically applicable across the agri-food sector.

Background: the interplay between food, health, and technology

Understanding the intricate relationship between food and health is essential. This relationship directly impacts individual well-being and quality of life. Diet and nutrition play a central role in preventing chronic diseases and maintaining optimal health. A balanced and varied diet provides the body with the essential nutrients it needs to function properly and protect itself against chronic conditions such as cardiovascular diseases, type 2 diabetes, obesity, and certain cancers. However, this relationship is complex and multifaceted, influenced by a variety of factors. These factors include accessibility and availability of healthy foods, cultural, socioeconomic, and environmental habits, as well as advertising and the prevalence of highly processed foods. In addition, scientific research and technological innovation play a vital role in the development of new strategies, treatments, and interventions to address challenges related to food and health. These advancements aim to promote a healthy and sustainable lifestyle for all, highlighting the importance of a holistic approach to nutrition and disease prevention [2].

The gut microbiota: a hidden player in health

The intricate relationship between food and health is closely intertwined with the human microbiota, the vast community of microorganisms residing within our bodies, primarily in the gut. This complex ecosystem, composed of bacteria, viruses, fungi, and other microorganisms, collectively known as the gut microbiome, plays a crucial role in our health, influencing numerous physiological, metabolic, and immunological aspects. This microbial community interacts with the human body in intricate ways, impacting physiological processes and immune responses. For instance, the gut microbiota aids in breaking down dietary fibers and other complex carbohydrates that human enzymes cannot digest on their own. This breakdown process produces short-chain fatty acids (SCFAs) and other metabolites that exert beneficial effects on intestinal health and systemic metabolism [3]. Diet has a fundamental influence on modulating the gut microbiota. The types and quantities of food we consume significantly impact the composition and diversity of the microbiota, with direct consequences for our health. For example, a diet rich in plant fibers promotes the growth of beneficial bacteria in the colon, while excessive intake of foods high in saturated fats and sugars can disrupt the balance of the microbiota and contribute to the development of pathological conditions. Conversely, the microbiota can influence our metabolic response to the foods we consume, contributing to digestion, nutrient absorption, and the production of bioactive metabolites that can have either positive or negative effects on our health [4]. Furthermore, a personalized approach to nutrition, based on understanding the individual's specific microbiota and metabolic needs, could represent the future of preventive and personalized medicine. This approach would allow individuals to optimize their health through targeted and individualized dietary choices. Various therapies targeting the microbiome are being explored to regulate the human microbiome, including dietary intervention, food supplements, antibiotics, probiotics, prebiotics, synbiotics, postbiotics, psychobiotics, bacteriophages, and fecal microbiota transplantation [5].

The rise of ultra-processed foods and their impact on health

The increasing trend of urbanization, changing lifestyles, the demand for convenience, and aggressive marketing by food companies have significantly impacted food quality. Over the past decades, there has been a dramatic rise in the global consumption of processed foods [6]. Ultra-processed foods are food products that undergo extensive industrial processes and often contain artificial ingredients, chemical additives, and high amounts of sugars, saturated fats, and salt. These foods are frequently devoid of essential nutrients and rich in empty calories, making them unhealthy when consumed in excess. The NOVA classification system categorizes foods into groups based on the extent of processing for human consumption. Ultra-processed foods belong to one of the NOVA categories and typically have a long list of ingredients (five or more) [7]. These ingredients often include oils, salt, and preservatives commonly used in processed foods, but they also include additives that enhance and mask flavors and odors and alter the final consistency of the product. Excessive consumption of ultra-processed foods has been linked to various health risks, including an increased risk of obesity, type 2 diabetes, heart diseases (HDs), hypertension, and other chronic conditions. These foods can also contribute to digestive disorders, chronic inflammation, metabolic imbalances, and alterations in the gut microbiota. An unhealthy diet has a significant impact on health, surpassing the effects of alcohol, tobacco, drugs, and unsafe sexual practices. This dietary imbalance is especially problematic in African countries, where malnutrition and obesity co-exist, contributing to a dual burden of disease. Poor diet is also a major factor in the rise of noncommunicable diseases (NCDs) such as coronary HD (CHD) and type 2 diabetes. By contrast, adopting a healthy diet and lifestyle can reduce the genetic risk of CHD by nearly 50%, establishing diet quality as a crucial, modifiable risk factor for chronic diseases. Advancements in nutrition science have enabled comprehensive databases, such as USDA's FoodData Central and Denmark's Frida, which provide detailed nutritional profiles. These resources facilitate studies on single nutrients and their effects, leading to key findings, such as the negative impact of trans fats and the cardiovascular benefits of omega-3 fatty acids, legumes, and nuts. However, focusing solely on individual nutrients has limitations, as it fails to consider the interactions between compounds in whole foods. For instance, studies initially linked β-carotene to a higher risk of prostate cancer, but later research attributed this to specific foods like papaya, not β-carotene itself. This highlights the first paradigm: nutrients should be studied within the context of whole foods and their complex interactions. Traditional nutrition research, focused on vitamins and other essential micronutrients, has identified around 150 key nutrients tracked in most databases. However, foods contain over 139,000 other compounds, many of which have significant health implications, such as polyphenols. The second paradigm, therefore, is the importance of documenting the "dark matter" of nutrition—the diverse array of bioactive compounds that extend beyond well-known nutrients. The third paradigm involves understanding how food compounds interact with human proteins, influencing health through mechanisms that are best understood using a network approach. Unlike reductionist methods, network science enables a more holistic view of how these compounds modulate biological processes, including nutrient bioavailability, the food matrix, and interactions with commensal microbes. This network perspective aligns with evolutionary biology, as it considers how human diets have co-evolved with various life forms [8].

Fraud prevention and traceability

Fraud prevention and traceability are critical aspects of ensuring food quality, safety, and transparency, and maintaining consumer trust in the food supply chain. In recent years, there has been a significant focus on this field driven by technological advancements and increasing consumer demand for transparency and safety [9]. Consumers are increasingly interested in knowing the origin, supply chain, and production practices behind the food they eat. This access to information empowers them to make more informed choices, such as preferring products from suppliers who adopt sustainable practices or adhere to specific quality standards [10]. Traceability systems play a vital role in the prompt identification and removal of defective or contaminated products from the market, ensuring that consumers receive high-quality products that meet safety standards. The Internet of Things (IoT) and sensor technology have enabled the development of smart devices and systems that monitor various aspects of food production and distribution in real-time [11]. These technologies allow for continuous monitoring of environmental conditions, such as temperature, humidity, and location, throughout the food supply chain. This continuous monitoring helps to ensure the integrity and safety of food products from farm to fork [12]. Another important aspect to consider is the potential negative impact on human health of dietary exposure to food contaminants, like microplastics [5]. These can derive from food packaging and their release can be enhanced by high temperatures, material age, liquid contact, and mechanical stress. To prevent possible risks to human health, biocompatible and eco-friendly packaging methods could be used as alternatives to conventional plastic-based packaging. In parallel, highly-accurate and standardized detection methods should be implemented at industrial level in quality assessment tests, to improve production processes and reduce as much as possible food contamination from contact materials. Metrology, in combination with large-scale data analysis, can play a key role in this framework, providing reliable and traceable tools for the physico-chemical characterization of food samples and the extraction of relevant data for contaminant quantification [13]. An example of application is the determination of microplastic concentrations, size distributions, and polymer types in various food matrices, starting from reference materials for the accurate calibration.

Data and computing infrastructure

In today's data-driven world, organizations are increasingly reliant on robust data and computing infrastructure to effectively manage, analyze, and extract insights from vast amounts of information. This infrastructure plays a crucial role in enabling data-driven decision-making, optimizing operations, and driving innovation across various domains [14]. The smart data model is a conceptual framework that outlines the principles and strategies for building an intelligent and efficient data management system. It emphasizes the integration of data governance, data quality, data security, and data analytics to create a unified and actionable data ecosystem [15]. Core principles of the smart data model are listed below. Data governance (establishing clear ownership, policies, and procedures): it plays a pivotal role in ensuring the integrity, consistency, and compliance of data assets throughout their lifecycle. By establishing clear ownership, policies, and procedures, organizations can effectively manage data assets, ensuring adherence to regulatory requirements and fostering a culture of data stewardship. Data quality (maintaining accuracy, completeness, and relevance): it is important for reliable and trustworthy data analysis. The smart data model emphasizes the importance of maintaining data accuracy, completeness, and relevance throughout its lifecycle. This involves implementing data quality checks, data cleansing techniques, and data validation processes to ensure that data remains accurate, consistent, and fit for purpose. Data security (protecting sensitive data from unauthorized access): it is a critical aspect of the smart data model, safeguarding sensitive data from unauthorized access, modification, or destruction. Organizations must implement robust security measures, including access controls, encryption techniques, and data breach prevention strategies, to protect sensitive data and maintain the integrity of their data ecosystem. In AI applications for nutrition, data security is particularly crucial due to the sensitive nature of personal health information. Specific strategies include using end-to-end encryption for data transmission, anonymizing datasets to protect individual privacy, and implementing strong user authentication methods. Technologies such as homomorphic encryption, which allows data to be analyzed without being decrypted, and secure multi-party computation, which enables collaborative data analysis without sharing raw data, are also essential to ensure privacy in these applications. Furthermore, conducting regular audits and penetration testing can help identify vulnerabilities and bolster the security infrastructure. Privacy and ethical considerations for AI applications in nutrition require special attention to privacy and ethical concerns. The collection and processing of personal health data must comply with regulatory frameworks such as General Data Protection Regulation (GDPR) in the European Union. Ethical concerns include ensuring informed consent for data usage, minimizing bias in AI models to prevent discriminatory outcomes, and ensuring transparency in how data is used. Establishing clear data privacy policies and maintaining transparency with stakeholders about data collection and usage practices are critical to fostering trust. Additionally, adherence to privacy by design principles, where privacy considerations are integrated into the technology from the outset, further enhances the ethical use of data in AI-driven nutrition applications. Data analytics (leveraging advanced techniques for meaningful insights): it plays a central role in extracting meaningful insights from data. The smart data model encourages the utilization of advanced analytical techniques, such as ML, AI, and statistical analysis, to uncover patterns, trends, and anomalies within data sets. In the context of nutrition, AI algorithms must also be explainable and interpretable to ensure that healthcare professionals and consumers can trust the recommendations generated. Explainable AI (XAI) techniques help in understanding the rationale behind dietary suggestions and can address concerns about "black box" models, which are often opaque in their decision-making processes. The advancement of AI in the field of nutrition relies heavily on interdisciplinary collaboration. Bringing together data scientists, nutritionists, and food technologists is essential to ensure that AI models are both scientifically valid and practically applicable. Data scientists provide expertise in ML, data analytics, and computational methods, while nutritionists contribute critical insights into dietary needs, human metabolism, and health impacts. Food technologists play a key role in understanding the practical aspects of food production and quality. An example of successful collaboration is the METROFOOD-IT project, which integrates expertise from these various disciplines to develop AI-driven models that enhance food safety, quality, and personalized nutrition. This type of collaborative approach ensures that AI technologies are designed with a deep understanding of the complexities of human nutrition, leading to more effective and trustworthy outcomes. Additionally, regular workshops and collaborative sessions help in aligning the goals across disciplines, which is crucial for the successful implementation of AI innovations in the agri-food sector. Implementing a smart data model offers a multitude of advantages. It fosters improved data accessibility and usability, making high-quality information readily available to diverse stakeholders. This empowers collaboration and innovation based on a shared understanding of the data [16]. Furthermore, smart data models lead to enhanced operational efficiency. By streamlining data management processes, they eliminate data silos and duplication of efforts. This optimizes data utilization and reduces the burden on data management teams [17]. Smart data models also contribute to risk mitigation and compliance. By strengthening data security protocols and embedding privacy considerations at every stage, smart data models ensure that organizations comply with data privacy regulations, such as GDPR and Health Insurance Portability and Accountability Act (HIPAA). This minimizes risks associated with data breaches and unauthorized access [18]. Ultimately, smart data models empower organizations to make data-driven decisions. By providing reliable data insights, they enable informed decision-making, leading to improved overall business performance [19].

HPC in AI for nutrition

HPC infrastructures are essential for handling the vast and complex datasets involved in AI applications for nutrition. HPC systems provide the computational power needed to process large-scale nutritional data, run sophisticated AI algorithms, and conduct simulations that would otherwise be infeasible on standard computing systems. These infrastructures are particularly important in tasks such as disease risk prediction, diet personalization, and food product optimization, where the analysis of multi-dimensional data, such as genomic, microbiome, and epidemiological data, requires high-speed processing and substantial storage capacity. One example of an HPC system used in this context is ReCaS-Bari, which provides computational resources for analyzing large datasets in nutrition and food safety studies. Another prominent platform is the CRESCO cluster at ENEA, which supports advanced AI models for processing genomic and microbiome data to identify patterns that correlate diet with health outcomes. These HPC platforms enable the training of deep learning models, which can analyze data at unprecedented speeds, reducing the time needed for insights from weeks to hours. For example, in personalized nutrition research, HPC systems allow the integration of multiple datasets (e.g., dietary intake, gut microbiota, and genetic information) to provide tailored dietary recommendations. This computational power enhances the accuracy and timeliness of AI predictions, ultimately leading to more effective interventions in public health and personalized medicine.

A framework for data management and analytics

GAIA-X is a comprehensive data platform that provides a unified environment for managing, analyzing, and visualizing data at scale according to European data strategy [20, 21]. It offers a suite of tools and functionalities that cater to the needs of various data professionals, from data engineers to data scientists. Key features of GAIA-X: (1) data ingestion and integration: it seamlessly ingests data from various sources, including structured, semi-structured, and unstructured data formats; (2) data storage and management: it provides secure and scalable storage for large volumes of data, enabling efficient data organization and retrieval; (3) data processing and transformation: it offers robust data processing capabilities to clean, transform, and prepare data for analysis; (4) data analytics and visualization: it includes advanced data analytics tools and visualization capabilities to uncover insights and patterns from data; (5) ML and AI: it supports the integration of ML and AI algorithms for predictive modeling and intelligent data analysis. Benefits of using GAIA-X are: (1) accelerated data-driven insights: streamlines the process of extracting meaningful insights from data, enabling faster and more informed decision-making; (2) enhanced collaboration and innovation: fosters collaboration among data professionals, facilitating knowledge sharing and innovation; (3) scalable and cost-effective solution: provides a scalable and cost-effective data management solution that can accommodate growing data volumes and evolving business needs; (4)

empowered data professionals: equips data professionals with the tools and functionalities they need to effectively manage, analyze, and derive value from data.

The smart data model and data platform GAIA-X provide a solid foundation for organizations to build a robust and intelligent data management ecosystem. By adopting these frameworks and tools, organizations can effectively harness the power of data to drive innovation, optimize operations, and gain a competitive edge in the data-driven world.

Methods

Data analysis is crucial for extracting valuable insights from the agri-food sector. However, two key challenges hinder this process: data availability and traceability. Access to high-quality data is essential for any successful analysis. The open science approach, which promotes transparency and research sharing, plays a fundamental role in this area. Once data is available and traceable, advanced analysis techniques such as ML come into play. ML has proven extremely useful in the food industry, finding applications in various areas to enhance production, distribution, quality, and safety of food products. Next subsections will focus specifically on open science and how ML is used to analyze data in the agri-food sector.

Open science

Open science aims to increase the accessibility, reproducibility, and impact of scientific research by promoting collaboration, transparency, and inclusivity in the research process [22]. It involves sharing research findings, data, methodologies, and other research outputs openly and transparently, often using digital technologies and the internet. It is instead a much broader concept that also encompasses, for example, openness to raw and processed research data [open data (OD)], educational materials (open educational resources), the use of open methodologies throughout the research cycle (open methodology), the use of open-source software (open source), and the adoption of open practices even in peer review to verify the quality of scientific work (open peer review). Due to the vast quantity and complexity of data, as well as the speed at which this data is generated, the so-called findable, accessible, interoperable, and reusable (FAIR) principles have been implemented to ensure uniform data management methods and open access to the data [23, 24]. Data reusability is ensured, for example, through documentation with embedded instructions to maintain reusability while minimizing the number of required files. Additionally, the data is in a common format and can be read using widely available software (open-source or commercial). Data and metadata should be easily readable by both humans and machines. Specifically, the use of machine-readable metadata is crucial for the automatic discovery of datasets and services. Once the user has found the requested data, the accessibility of OD implies that data are easily retrievable and usable by anyone with an interest, without significant limitations. These data are made available without restrictions or with minimal restrictions on access and use, allowing users to utilize them for various purposes such as research, analysis, and application development. To ensure interoperability, individual datasets and results can be described using established field-specific vocabularies, standards, formats, and methodologies, such as GUM, OBO, DICOM, NetCDF, HDF5, CityGML, INSPEC, ISO 9001 [25–32]. The adoption of a FAIR approach to data management and the development of digital services supporting the agri-food sector will facilitate the development of data-driven methods and machine/deep learning models based on the availability of data to be analyzed with technologies based on big data analytics. Data acquisition is made possible through automation systems to maximize efficiency and result from reproducibility. The generated or collected data can be made available through defined applications and repositories that make the data freely accessible both through machine-to-human interactions and machine-to-machine interactions, using user interfaces and application programming interfaces (APIs) respectively. OD sharing is becoming increasingly relevant in the food industry, thanks to its benefits for supply chain efficiency, food safety, and sustainability. However, several barriers must be addressed, which can be overcome with targeted strategies. Some successful cases of OD sharing in the agri-food sector include: IBM Food Trust (IBM Food Trust. Retrieved from https://www.ibm.com/it-it/products/supply-chain-intelligence-suite/food-trust), Open Food Facts (Open Food Facts. About Open Food Facts, retrieved from https://world.openfoodfacts.

org), and Food Integrity (Food Integrity project. About Food Integrity, retrieved from https://www.foodintegrity.org/). The IBM Food Trust initiative, based on blockchain technology, is a significant example of OD sharing in the food industry. This platform allows supply chain stakeholders, from producers to retailers, to share real-time data regarding food traceability. Companies such as Walmart, Nestlé, and Unilever participate in this initiative. The benefits include increased transparency, reduced food waste, and improved food safety. For instance, by sharing data on Food Trust, Walmart reduced the time needed to trace the origin of a batch of mangoes from days to just seconds, improving crisis management during health scares. Open Food Facts is an open-source platform that collects and shares information about ingredients, nutritional values, and other characteristics of food products from around the world. This project is particularly important for consumers, as it allows full transparency on available products, promoting more informed and healthy food choices. Additionally, it enables companies to improve the quality of their products through shared data on consumer habits and preferences. The benefits introduced by this initiative include increased transparency, greater consumer engagement, and improved product quality. The Food Integrity project, funded by the European Union, has promoted the sharing of scientific data to tackle food fraud and improve food traceability. Involving over 60 organizations across Europe, the project created an accessible database that enhances quality control and the prevention of fraudulent practices in the food supply chain. Despite the success of some initiatives, OD sharing in the food industry faces several challenges. The most significant of these is the protection of intellectual property. Companies are often reluctant to share data for fear of losing competitive advantages. Data on production processes, recipes, and suppliers often represent strategic information that could be used by competitors to gain a market edge. Blockchain and other secure technologies can help ensure that shared data is only accessible to authorized parties, preserving confidentiality when necessary. For example, IBM Food Trust allows selective data sharing, where only relevant and non-sensitive information is visible to specific stakeholders. Another recurring challenge is data interoperability and standardization, which can facilitate sharing through open protocols. Organizations such as GS1 (GS1 Global. About GS1 standards, retrieved from https://www.gs1.org), which develops global standards for product information communication, are working to promote interoperable and universally accepted data formats. OD sharing may also pose security and privacy risks, especially when it comes to personal or commercially sensitive data. This is particularly relevant in the context of regulations like the GDPR in Europe, which protects consumers' personal data. In addition to adopting advanced security protocols, such as encryption, another approach is "data minimization". Companies can share only the data strictly necessary to achieve the sharing objective. Moreover, adopting frameworks like "data trust" can ensure responsible governance of shared data. OD sharing may require significant investments in IT infrastructure and qualified human resources to ensure the quality and security of shared data. Public-private partnerships can help reduce implementation costs. For example, projects like Food Integrity and Open Food Facts are partly funded by public grants and supported by volunteer communities, making data sharing more accessible.

Practical application of DataOps and FAIR principles

The practical application of DataOps in the nutrition sector is realized through the automation and optimization of data pipelines that manage complex, multi-source datasets. For instance, in AI-driven nutrition research, DataOps techniques ensure that data from clinical trials, epidemiological studies, and public health databases are continuously integrated, cleaned, and pre-processed for analysis. This approach enhances the efficiency of data processing, reduces errors, and accelerates the delivery of insights. In the context of HPC infrastructures, DataOps allows for the seamless flow of data between storage, processing, and analysis, supporting real-time data analytics and AI model training. Similarly, the implementation of FAIR principles ensures that nutritional data is managed in a way that facilitates open science and interoperability. For example, nutritional datasets are stored in standardized formats that are easily accessible to both researchers and machines. This allows AI models to access and process data without the need for extensive manual intervention. In the METROFOOD-IT project, the adoption of FAIR principles has enabled the creation of an OD platform that supports the sharing of food quality and traceability data across different stakeholders in the agri-food sector, fostering collaboration and transparency. This open-access data model has proven crucial in integrating diverse datasets, from food production data to health outcomes, making it possible to train AI systems on reliable and standardized data. In real-world applications, the principles of DataOps and FAIR have been successfully implemented in projects like METROFOOD-IT, where they play a crucial role in ensuring efficient data flow and interoperability. In the METROFOOD-IT project, FAIR principles guide the management of large datasets related to food quality, safety, and traceability. By structuring data to be FAIR, the project ensures that the data can be shared seamlessly across different stakeholders, from food producers to researchers. For instance, food safety data generated from IoT sensors is stored in standardized formats that allow both human and machine interpretation, which enhances real-time decision-making. However, implementing these strategies comes with challenges. One major obstacle in adopting FAIR principles is the heterogeneity of data—different stakeholders often use different data formats, which complicates integration. To overcome this, METROFOOD-IT adopted common standards for metadata and data formats, ensuring that all participants adhered to a unified framework for data entry and retrieval. In the context of DataOps, the automation of data pipelines within METROFOOD-IT allowed for real-time processing of data from multiple sources. A key challenge here was ensuring data quality at each stage of the pipeline. This was addressed by implementing continuous monitoring and validation processes that automatically flagged any inconsistencies or errors in the data.

Machine learning

ML has proven to be extremely useful in the food industry, finding applications in various areas to enhance food production, distribution, quality, and safety. ML can also be employed to personalize individuals' diets, using data on their dietary preferences, nutritional needs, and physiological characteristics to suggest personalized meal plans that improve health and well-being. There are several types of data used in the field of ML in the food industry for example nutritional data, genomic and microbiome data, food contaminant, or environmental data [33]. Combining different types of data can provide a more comprehensive understanding of food processes and consumer behaviors, enabling the development of predictive models and innovative solutions to improve food production, distribution, and consumption. The data may include information on raw materials, production parameters, sensor readings, quality inspection results, and historical records. Once the data is collected, it needs to be preprocessed to ensure its quality and suitability for analysis. Data analysis is crucial for extracting valuable insights from the agri-food sector [34, 35]. The next step is to select appropriate ML models for the analysis task at hand. This may involve choosing between supervised, unsupervised, or semi-supervised learning approaches, depending on the nature of the data and the objectives of the analysis [36, 37]. In supervised learning, the model is trained on a labeled dataset, where each example is associated with an input and an output. During training, the model learns the mapping between inputs and outputs, allowing it to make predictions on new, unseen data. Common tasks in supervised learning include classification (assigning inputs to discrete categories) and regression (predicting continuous values) [38]. Using regression analysis, one could build predictive models to understand how different dietary factors contribute to health outcomes. This information can be used to inform public health interventions and dietary guidelines aimed at reducing the prevalence of obesity and related chronic diseases. An example of classification may include, for example, the origin of food products that are strictly connected with their quality like protected designation of origin (PDO) foods, which can be valuable for quality control and certification purposes in the food industry. In unsupervised learning, the model is trained on an unlabeled dataset, where only the input data is available. Common tasks in unsupervised learning include clustering and dimensionality reduction [39]. An example of clustering in the context of food and health could involve categorizing food products based on their nutritional profile and health effects, while dimensionality reduction techniques highlight relevant features in a complex dataset. For example, it is possible to identify, using ML models, a cluster of foods high in saturated fats and simple sugars, which are associated with a higher risk of obesity and HD, and another cluster of foods high in fiber and protein, which are considered healthier. Semi-supervised learning lies between supervised and unsupervised learning [40]. Once the models are selected, they need to be trained on the available data. During training, the models learn to identify patterns and relationships in the data that are indicative of production issues or quality deviations. After training, the models need to be evaluated to assess their performance and generalization ability. In ML, evaluation metrics are used to assess the performance of a model on a given dataset. The choice of evaluation metrics depends on the nature of the problem (e.g., classification, regression, and clustering) and the specific goals of the analysis. Once the models are trained and evaluated, they can be deployed into production systems for real-time monitoring and analysis. Employing ML in the realm of anti-fraud and food traceability offers multiple opportunities to enhance food safety and protect consumers from fraud and contamination. ML models can be trained to recognize patterns and anomalies in food-related data, enabling the identification of potential cases of food fraud, such as food counterfeiting, the addition of undeclared ingredients, fraudulent labeling, or the presence of contaminants, like microplastics released from packaging. Despite these advantages, the use of ML in food safety also presents several challenges. One significant issue is data quality and availability. ML relies on large, high-quality datasets, but in many cases, data in the food supply chain can be incomplete, inconsistent, or difficult to access. Poor data quality can lead to inaccurate predictions or missed contamination risks. Another challenge is the risk of overfitting, where the ML model becomes too tailored to the training data, leading to false positives or negatives. This can either cause unnecessary recalls or fail to detect real safety issues. Additionally, ML models, particularly deep learning, often lack interpretability. This "black box" nature makes it difficult for food safety managers and regulators to understand how decisions are made, leading to a lack of trust in the technology. There is also a dependency on digital infrastructure, such as IoT sensors and cloud computing, for ML to be effective. Companies without robust technological resources may struggle to implement ML solutions, creating a disparity in food safety practices across the industry. Privacy concerns also arise with the use of ML, especially when dealing with sensitive data from the supply chain or consumer health records in cases of contamination. Finally, the costs of implementing and maintaining ML systems can be high, which may be prohibitive for smaller companies. XAI plays a crucial role in food safety by providing transparency and interpretability in ML models used for various tasks, such as quality control, contamination detection, and risk assessment. XAI techniques can highlight anomalies or outliers in food production processes, indicating potential safety hazards [41]. XAI brings its own set of benefits to food safety, addressing some of the transparency issues associated with traditional ML. XAI increases trust by providing clear explanations for how predictions or decisions are made. In a field like food safety, where regulatory compliance is crucial, this transparency is essential. XAI helps companies and regulators understand the rationale behind safety actions, such as product recalls or flagged contamination risks. This also aids in regulatory compliance, as the decision-making process can be easily audited. XAI improves decision-making by allowing managers to understand the factors influencing predictions, leading to more informed actions. Additionally, XAI can help reduce bias by exposing the factors that drive model decisions. If a model is unfairly targeting certain suppliers or regions, XAI can highlight these biases and allow for corrections. However, XAI also has limitations. One drawback is the potential trade-off between explainability and accuracy. In some cases, simplifying a model to make it more interpretable may reduce its predictive power, which could limit its effectiveness in preventing contamination risks. Another issue is the complexity of implementing XAI. It can require additional computational resources and expertise, making it more costly and challenging to adopt. XAI can also slow down decision-making, as more time is needed to explain and interpret the model's outputs. In food safety, where quick action is often required, this delay can be problematic. Lastly, there is a risk of over-interpretation of XAI results. Stakeholders may focus too much on certain explanations, leading to decisions based on incomplete or misunderstood information.

Results

The success of AI applications in nutrition is highly dependent on interdisciplinary collaboration. In the METROFOOD-IT case study, for example, the cooperation between data scientists, nutritionists, and food technologists has been instrumental in developing accurate predictive models for food authenticity and safety. This collaboration ensured that AI technologies were effectively integrated with practical knowledge of food production and nutrition, leading to impactful innovations in the field. The success of METROFOOD-IT in advancing AI-driven innovations in nutrition can be attributed to a comprehensive implementation strategy. Infrastructure investments in HPC systems and cloud platforms provided the computational backbone needed to manage and analyze vast amounts of data. Simultaneously, workforce training initiatives ensured that stakeholders at all levels had the skills required to engage with the new technologies effectively. The focus on regulatory compliance, particularly in the management of personal health data, ensured that the project adhered to legal frameworks, maintaining trust and transparency among all participants. Together, these strategies have created a robust, scalable, and compliant ecosystem for AI-driven nutrition research.

Application of ML models

ML is transforming the food industry in numerous areas, from production chains to consumer experiences. ML plays a crucial role in determining the origin of food, significantly enhancing traceability across the entire supply chain. ML algorithms process these complex datasets to map the entire journey of a food product from production to the consumer. This allows for the precise identification of a product's origin and helps detect anomalies that might indicate fraud or contamination issues. One of the most promising combinations is between blockchain and ML. Blockchain provides an immutable record of transactions related to a food item, while ML analyzes this data to predict potential food fraud, such as ingredient substitution or tampering with origin information. Furthermore, it can verify product authenticity by comparing collected data with official claims, ensuring that a product labeled as organic or from a specific geographical region is indeed what it claims to be. ML is also used to identify the chemical and physical characteristics that distinguish food products based on their origin. For instance, analyzing isotopic compositions or mineral levels can confirm a product's authenticity, ensuring that it comes from the region indicated on the label, as required for products with PDO. Additionally, with the integration of IoT technology, connected devices can monitor food conditions, such as temperature and humidity, during transport and storage. ML analyzes this real-time data to ensure that products are kept in optimal conditions, enhancing transparency and reliability in the supply chain. Finally, ML can detect trends and patterns in data related to the food supply chain, identifying potential risks or problematic areas, such as suppliers not adhering to quality standards. Predictive analytics also allow companies to anticipate future issues related to food supply or quality, enabling more proactive management of production and distribution.

Mozzarella di Bufala PDO analyzing microbiota case of study

The case study of Mozzarella di Bufala Campana PDO has been considered by examining the composition of the microbiota in each sample [42, 43]. Mozzarella di Bufala Campana is a soft, fresh, stretched-curd cheese traditionally produced in the provinces of Caserta and Salerno (Italy). Production also takes place in selected localities of the metropolitan city of Naples, as well as in southern Lazio, northern Apulia, and the municipality of Venafro in Molise. Mozzarella di Bufala Campana is often known as "white gold" in homage to the cheese's prized nutritional and taste qualities. It was granted PDO status in 1996. PDO is a certification that guarantees the authenticity and quality of food products linked to specific geographical regions. ML can play a significant role in reinforcing the integrity and efficiency of PDO certification by analyzing complex data related to geographical origin, production methods, and product quality. Three different supervised ML algorithms have been compared and the best classifier model is represented by random forest with an area under the curve (AUC) value of 0.93 and the top accuracy of 0.87. ML models effectively classify origin, offering innovative ways to authenticate regional products and support local economies. Further research can explore microbiota analysis and extend applicability to diverse food products and contexts for enhanced accuracy and broader impact. The use of microbiota to determine the origin of food is an innovative approach that leverages the unique microbial communities associated with specific geographic regions, production environments, or even individual farms. This method is based on the idea that different environments (such as soil, water, air, and the surfaces of plants or animals) host distinct microbial populations. These microbial "signatures" can serve as biological markers to trace the geographic and production origins of food products. Microbiota-based techniques, combined with advanced technologies like ML and DNA sequencing, are increasingly being used to improve food authenticity, traceability, and safety. Each environment has a unique microbiota composition. These microbial communities are shaped by local factors like climate, soil type, altitude, and farming practices. For example, grapes grown in different wine-producing regions carry distinct microbial communities from the soil and air, which influence the fermentation process and, ultimately, the wine's flavor profile. By sequencing the DNA of microbes found in the food, scientists can create a microbial "fingerprint" that correlates with a specific region or environment. This fingerprint can then be used to verify the geographic origin of the product. Microbiota analysis can help verify the provenance of high-value products, supporting certifications like PDO, or organic labels. This can protect consumers from fraudulent products and ensure that producers are fairly compensated for their geographically unique products. Food safety is another area where microbiota can be beneficial. By identifying microbial contamination patterns, authorities can trace contamination back to its source more quickly, improving outbreak management and reducing public health risks. Despite its promise, using microbiota to determine food origin also presents challenges. One limitation is variability in microbial communities. Microbial populations can fluctuate based on seasonal changes, weather conditions, or variations in agricultural practices, which can make it harder to establish a consistent microbial fingerprint for certain regions or products. Additionally, the presence of similar microbial species in different geographic areas can complicate efforts to pinpoint a product's exact origin, particularly if the regions share environmental conditions. Another challenge is the need for advanced technologies and expertise. The process of extracting, sequencing, and analyzing microbial DNA requires sophisticated equipment and specialized knowledge. While the cost of DNA sequencing has decreased, it remains a barrier to widespread adoption in smaller or less technologically advanced regions. Finally, the presence of contaminants or external microbes picked up during transport, handling, or processing can obscure the original microbial fingerprint. For instance, if food products are handled in multiple locations or exposed to different environments, the microbial signature may become mixed, making it harder to trace the true origin.

METROFOOD-IT: an agri-food innovation engine

The METROFOOD-IT research infrastructure, acting as Italian national node of the European METROFOOD-RI [European Strategy Forum on Research Infrastructures (ESFRI) domain: health and food], whose strengthening and implementation is funded under the National Recovery and Resilience Plan (NRRP, NextGenerationEU), plays as a catalyst for progress in the agri-food sector. It champions the digital transformation of production and distribution chains, prioritizing traceability, environmental responsibility, and open communication. The infrastructure's core mission is to establish a lasting model of support services for the entire agri-food supply chain, with a particular focus on small and medium-sized businesses. In particular, METROFOOD-IT aims to create a digital ecosystem that facilitates the collection, analysis, and sharing of data related to food quality, safety, and traceability. By leveraging advanced metrological standards, IoT sensors, spectrometry, and hyperspectral imaging, it seeks to guarantee the integrity of products throughout the supply chain, from raw materials to finished goods. Moreover, the project integrates blockchain technology and cloud infrastructures to provide transparent and immutable food traceability, ensuring consumers have access to reliable information regarding food origins and authenticity. METROFOOD-IT is built on the belief that transparent production and distribution processes are key to a more sustainable and productive agri-food system. The project tackles vulnerabilities like fraud and adulteration by verifying and communicating the origins and authenticity of ingredients and finished products. It also works to improve food quality and safety through enhanced controls and defense strategies, integrating suitable tools across all stages of the supply chain. For instance, by utilizing IoT sensors to monitor real-time parameters like temperature and humidity, METROFOOD-IT has optimized agricultural practices, reducing waste, and improving product quality. In addition, AI and ML algorithms are employed to analyze large data sets, allowing for predictive modeling of food quality and shelf-life. Furthermore, the project is committed to empowering policymakers and regulatory bodies to promote food transparency and educate consumers, ultimately encouraging healthier and more environmentally responsible dietary choices [44, 45]. This is achieved through a combination of digital and physical infrastructure, which provides policymakers with accurate, real-time data, supporting informed decisions regarding food safety and sustainability. Educational campaigns and consumer tools, enabled by the project's data collection and analysis capabilities, are also part of this effort to raise awareness and influence healthier dietary habits. The infrastructure, distributed across Italy, provides services that bridge the gap between research, innovation, industry players, and consumers. It aims to cultivate a sustainable and innovative agri-food sector, ensuring food safety, promoting healthy habits, and offering solutions for a circular bioeconomy. Among the project's preliminary results, the improvement of food traceability through blockchain technology and the personalization of consumer products stand out as examples of how digital tools can directly benefit the agri-food system. These efforts are complemented by the use of advanced data analytics, which has led to the development of customized food products and services, tailored to individual nutritional needs and preferences. METROFOOD-IT is actively integrating the digital and physical aspects of its infrastructure to offer services that digitize agri-food systems, guarantee quality, safety, traceability, transparency, sustainability, and resilience. This work will culminate in the creation of an OD platform, a cloud-based infrastructure for data collection and dissemination, and a federated ICT (information and communication technology) solution for long-term access to FAIR data, fostering open science practices. As a best practice in the use of AI, blockchain, and advanced sensor technologies, METROFOOD-IT is revolutionizing the agri-food sector, contributing to a safer, more transparent, and environmentally friendly future.

Facilitating data sharing and utilization in the agri-food sector

The METROFOOD-IT e-infrastructure represents an advanced technological ecosystem designed to facilitate data sharing and utilization in the agri-food sector. This infrastructure is based on a robust and scalable architecture that integrates services, data, and linking infrastructures to offer a unified environment for data management, analysis, and validation. The METROFOOD-IT e-infrastructure offers a wide range of services aimed at end-users with diverse needs. These services include: smart sensors, which measure physical and chemical properties using various methodologies, such as spectroscopy, electrochemical sensors, and biosensors; HPC facilities, which provide advanced computational capabilities for large-scale data analysis, AI, and numerical modeling; food tracking, which enhances food production quality and traceability through data management solutions, biomolecular sensors, and self-calibration tools; co-design and co-creation, which facilitate rapid prototyping, experimentation, and demonstration of innovative solutions for the agri-food sector; education, which offers training courses and educational materials through an e-learning platform; web portal, which provides access to data and services through the OD and ICT integration component. The data infrastructure supports the services described above and acts as an interface for their integration within the METROFOOD-IT ecosystem. This infrastructure includes a service-specific backend system managed by individual project partners and a data integration that aggregates data from the backends of different services, ensuring syntactic and semantic data uniformity (see Figure 1).

Agri-food OD and ICT integration play a crucial role in the integration of services and data within the METROFOOD-IT ecosystem. OD and ICT are based on a high-performance hardware/software infrastructure optimized for HPC whose main components are hosted at the ReCaS-Bari data center and are federated with ENEA CRESCO HPC clusters. Its main functions include: (1) large dataset management: easily store, process, and archive large amounts of data; (2) data integration: aggregates and processes data from various services using customized pre-processing pipelines; (3) smart data model: implements a standardized data model for the agri-food sector, facilitating data sharing and use; (4) information extraction: employs statistical algorithms and multivariate techniques to extract useful information from images and numerical data; (5) complex models: applies ML models and neural networks for high-

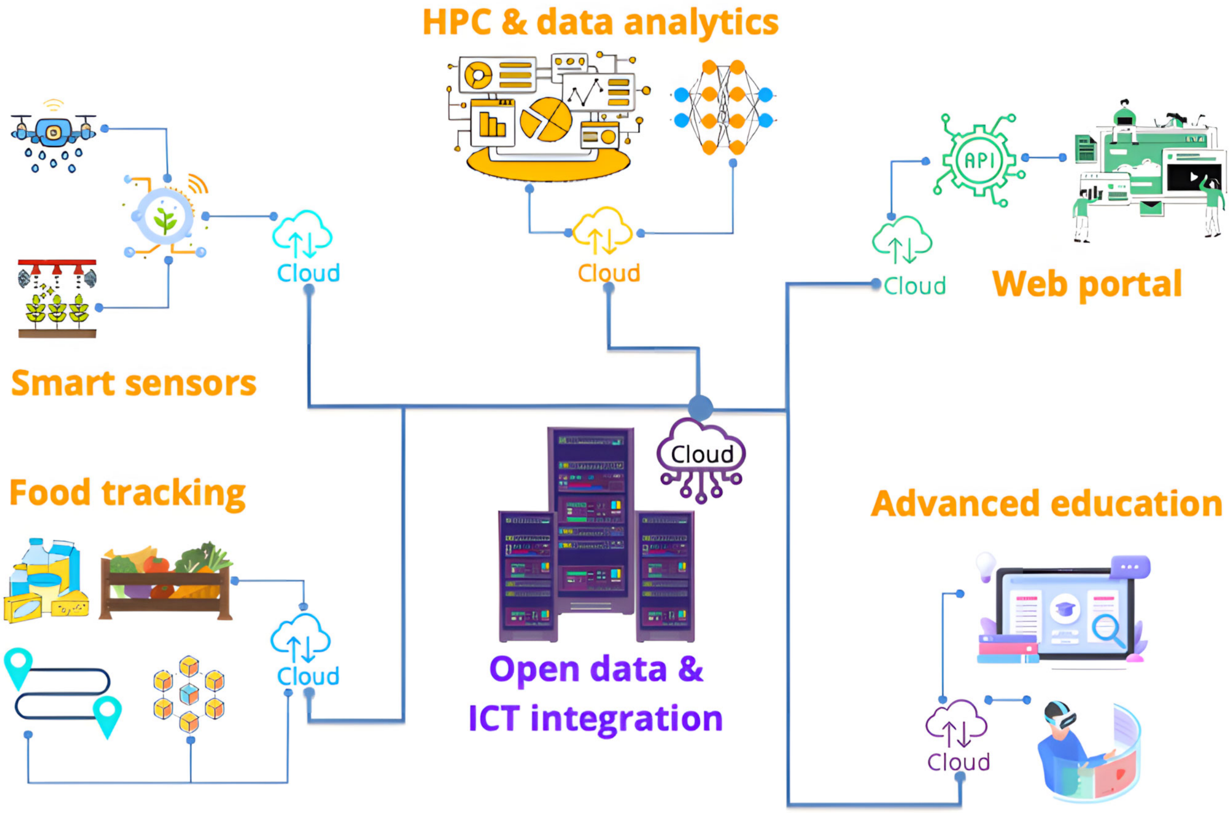

Figure 1. Data infrastructure. API: application programming interface; HPC: high-performance computing; ICT: information and communication technology

dimensional data analysis; (6) smart data model: data sharing in a complex scenario like agri-food requires a well-structured data model that ensures the syntactic and semantic uniformity of data from diverse sources. The smart data model, developed within the smart data models initiative (SDMI), fulfills these requirements by defining the following: (1) schema: the technical structure of the model, specifying data types and organization; (2) specification: detailed documentation for human users explaining the model and data semantics; (3) uniform resource identifier (URI): URIs for each attribute or entity associated with the model, ensuring complete data retrievability; (4) payload examples: practical examples of how data should be structured within the model. The METROFOOD-IT project should simplify the access to the data from both a generator perspective and a user perspective. The adoption of the "agri-food" smart data model facilitates true integration of the infrastructure on an international scale, laying the foundation for the effective utilization of collected data in the agri-food sector [41].

Infrastructure investments

In the METROFOOD-IT project, significant investments were made in both hardware and software infrastructures to support data sharing and analysis across the agri-food sector. HPC systems, such as ReCaS-Bari and the ENEA CRESCO HPC clusters, were deployed to handle the large volumes of data generated by IoT sensors, metrological devices, and AI models. These systems offer the computational capacity necessary for real-time data processing and large-scale AI model training. Furthermore, cloud-based infrastructures were integrated to ensure scalability and flexibility, allowing the project to accommodate increasing data volumes without compromising performance. To maximize the efficiency of data management, METROFOOD-IT adopted a distributed infrastructure model, which links local and national nodes, creating a decentralized but unified system for data processing and storage. This distributed architecture reduces bottlenecks and ensures that data is always available to stakeholders, no matter their geographic location. A critical component of the METROFOOD-IT implementation strategy was workforce training. Given the complexity of AI models, HPC infrastructures, and FAIR data management principles, targeted training programs were developed for various stakeholders, including researchers, data scientists, and food industry professionals. These programs focused on equipping personnel with the necessary skills to use data integration tools, cloud platforms, and AI-driven data analysis systems. To ensure accessibility, METROFOOD-IT deployed an e-learning platform, offering courses that covered topics such as data governance, AI model deployment, and secure data sharing. This workforce training initiative was essential in ensuring that all project participants, regardless of their technical background, could contribute effectively to the data-driven innovations of the project. Continuous professional development sessions and workshops were held to keep teams updated on the latest technologies and methodologies. Ensuring compliance with regulatory frameworks, such as the GDPR in the European Union, was a central focus of the METROFOOD-IT project. The collection and processing of personal health data related to nutrition required strict adherence to data privacy and security protocols. To address these regulatory requirements, METROFOOD-IT implemented end-to-end encryption for data transmission, ensuring that sensitive information was protected from unauthorized access. In addition, the project developed a comprehensive data governance framework that outlined specific protocols for data storage, access, and sharing. These protocols ensured that all personal data was anonymized where possible, and data subjects were given full transparency and control over how their data was used. Regular audits and security assessments were conducted to ensure that all data management practices remained in compliance with regulatory standards. To support ethical considerations, METROFOOD-IT established clear informed consent procedures, ensuring that participants in research and data collection activities were fully aware of how their data would be used and had the option to opt out at any stage.

METROFOOD-IT a possible scenario for data space application

The organization of the data management infrastructure within METROFOOD-IT could be based on data spaces, a form of multilateral organization implemented to achieve shared objectives. Within the data space, three distinct roles emerge: (1) federator: operates as a neutral entity, ensuring data sharing integrity and data ecosystem sustainability. Its key functions include managing the service portfolio, the decentralization level, and business services; (2) data provider: publishes data sources, identifies and registers participants, and manages data exchange; (3) data consumer: searches for data sources, identifies and authorizes participants, and uses the data for stated purposes. Data spaces do not require physical data integration, as data remain within respective proprietary repositories; integration occurs at a semantic level using shared vocabularies. Data payloads are exchanged exclusively between data providers and data consumers. Within the data ecosystem, three operational levels emerge: (1) data-driven services: enable the provision of data services such as fraud prevention, food tracking, and quality analysis, while simultaneously acquiring new data to enhance the ecosystem; (2) data objects: abstract the data level by creating logical datasets, such as a “digital twin” of a PDO production, consumer type, or geographical area, characterized by the data ecosystem; (3) GAIA-X: the software infrastructure that organizes, distributes, and builds value-added data services.

Federation services are provided by the GAIA-X architecture, which offers a federated catalog of distributed services such as sovereign data exchange, identity and trust management, and compliance services. GAIA-X is designed to ensure integration between architectures and processes supporting the data space. The evolution of roles within the data space progresses in parallel with the data ecosystem’s growth and complexity. As the ecosystem becomes more complex, it demands increasingly sophisticated federation services.

Conclusions

The field of nutrition is undergoing a significant transformation driven by the growing importance of data and AI. This article has explored the profound impact of AI on our understanding of food and health, highlighting its potential for personalized dietary recommendations, disease risk prediction, the impact of food contaminants, and the development of novel food products. HPC infrastructures, coupled with AI-driven models, are instrumental for efficiently managing and analyzing the increasing volume of nutritional data. The smart data model provides a framework for building an intelligent and efficient data management system, while GAIA-X offers a comprehensive platform for data management, analysis, and visualization. OD sharing and FAIR principles are essential for ensuring data accessibility, interoperability, and reusability. ML algorithms play a vital role in various aspects of the food industry, from production and quality control to personalized nutrition. The METROFOOD-IT research infrastructure represents an excellent application scenario for experimenting with the use of data spaces in the agri-food sector. Its e-infrastructure offers a range of services and data resources, facilitating data sharing and utilization to promote a more sustainable, transparent, and efficient food system. The infrastructure's data space model fosters collaboration between stakeholders and ensures data security through a decentralized approach. Future research directions and technological advancements: moving forward, several areas warrant further research to fully unlock the transformative potential of AI in nutrition. Future research should explore advanced AI techniques, such as reinforcement learning and generative models, to develop more sophisticated dietary planning tools and predictive models for public health. Emphasis should also be placed on developing XAI and interpretable AI models to increase trust and transparency, particularly in healthcare and nutrition applications. Research on integrating AI with other emerging technologies (such as blockchain for food traceability and IoT for real-time monitoring of dietary habits) can further enhance the value of AI in the agri-food sector. Additionally, research into overcoming ethical and privacy challenges, particularly in data-sharing environments, is crucial. Policy implications and strategic recommendations: policymakers have a critical role to play in shaping the future of AI in nutrition. Policies must be developed to ensure the ethical collection, storage, and use of personal health data, with clear guidelines for AI model validation and the prevention of algorithmic bias. Regulations that encourage interdisciplinary collaboration among data scientists, healthcare professionals, and food technologists will facilitate the adoption of AI solutions that are both scientifically robust and socially acceptable. Incentivizing OD initiatives and ensuring compliance with privacy frameworks such as GDPR and HIPAA will be crucial for creating a transparent data-sharing culture that benefits consumers, researchers, and the food industry alike. Roadmap for stakeholders: to accelerate AI-driven advancements in nutrition, a strategic roadmap is recommended for stakeholders. Key actions include fostering public-private partnerships to fund research and development projects, encouraging interdisciplinary research teams to address complex nutritional issues, and developing educational programs to train the next generation of professionals in AI and nutrition. Industry stakeholders should focus on adopting FAIR principles and implementing best practices in DataOps to ensure data quality and availability, which are pivotal for training effective AI models. Collaboration between technology providers, the food industry, and government bodies is also essential to build resilient and responsive data infrastructures. In conclusion, the integration of data science and AI holds immense potential for revolutionizing the field of nutrition. By leveraging these advancements, we can create a future where personalized dietary strategies and a data-driven approach to food production contribute to improved health outcomes for all. However, realizing this future will require continued investment in technology, well-crafted policy frameworks to support ethical data use, and sustained interdisciplinary collaboration. The journey towards a truly AI-powered, transparent, and efficient nutrition ecosystem is ongoing, and strategic, well-coordinated efforts across sectors are key to achieving meaningful progress.

Abbreviations

AI: artificial intelligence

CHD: coronary heart disease

DataOps: data operations

DevOps: development and operations

FAIR: findable, accessible, interoperable, and reusable

GDPR: General Data Protection Regulation

HDs: heart diseases

HIPAA: Health Insurance Portability and Accountability Act

HPC: high-performance computing

ICT: information and communication technology

IoT: Internet of Things

ML: machine learning

OD: open data

PDO: protected designation of origin

URI: uniform resource identifier

XAI: explainable artificial intelligence

Declarations

Acknowledgments

Authors would like to thank the resources made available by ReCaS, a project funded by the MIUR (Italian Ministry for Education, University and Research) in the "PON Ricerca e Competitività 2007–2013-Azione I-Interventi di rafforzamento strutturale" *PONa*3_00052, Avviso 254/Ric, University of Bari.

Author contributions

PDB and MM equally contributed to: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing—original draft, Writing—review & editing, Visualization. PN, DR, DD, LdT, A Mariano, CZ, RF, A Manzin, MDA, and RB: Writing—review & editing. ST: Conceptualization, Methodology, Validation, Writing—original draft, Writing—review & editing, Visualization, Supervision, Project administration, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The author declares that there are no conflicts of interest.

Ethical approval

Not applicable.

Consent to participate

Not applicable.

Consent to publication

Not applicable.

Availability of data and materials

Not applicable.

Funding

Part of this paper is funded by METROFOOD-IT funding from the European Union-NextGenerationEU, PNRR-mission 4 "education and research" component 2: from research to business, investment 3.1: fund for the realization of an integrated system of research and innovation infrastructures [IR0000033] (D.M. Prot. n.120 del 21/06/2022). The study sponsor contributed to all aspects of the research, including the study design, data collection, analysis, and interpretation, as well as the writing of the report and the decision to submit the paper for publication.

Copyright

References

1. Stahl BC, Antoniou J, Bhalla N, Brooks L, Jansen P, Lindqvist B, et al. A systematic review of artificial intelligence impact assessments. Artif Intell Rev. 2023;12799–831. [DOI] [PubMed] [PMC]
2. Jacobs DR, Tapsell LC. Food synergy: the key to a healthy diet. Proc Nutr Soc. 2013;72:200–6. [DOI] [PubMed]
3. Martin-Gallausiaux C, Marinelli L, Blottière HM, Larraufie P, Lapaque N. SCFA: mechanisms and functional importance in the gut. Proc Nutr Soc. 2021;80:37–49. [DOI] [PubMed]
4. Novielli P, Romano D, Magarelli M, Bitonto PD, Diacono D, Chiatante A, et al. Explainable artificial intelligence for microbiome data analysis in colorectal cancer biomarker identification. Front Microbiol. 2024;15:1348974. [DOI] [PubMed] [PMC]
5. Jadhav EB, Sankhla MS, Bhat RA, Bhagat DS. Microplastics from food packaging: An overview of human consumption, health threats, and alternative solutions. Environ Nanotechnology Monit Manag. 2021;16:100608. [DOI]
6. Elizabeth L, Machado P, Zinöcker M, Baker P, Lawrence M. Ultra-Processed Foods and Health Outcomes: A Narrative Review. Nutrients. 2020;12:1955. [DOI] [PubMed] [PMC]
7. Gibney MJ, Forde CG, Mullally D, Gibney ER. Ultra-processed foods in human health: a critical appraisal. Am J Clin Nutr. 2017;106:717–24. [DOI] [PubMed]
8. Menichetti G, Barabási A, Loscalzo J. Decoding the Foodome: Molecular Networks Connecting Diet and Health. Annu Rev Nutr. 2024;44:257–88. [DOI] [PubMed]
9. Farm to Fork strategy [Internet]. [cited 2024 Nov 13]. Available from: https://food.ec.europa.eu/horizontal-topics/farm-fork-strategy_en
10. Consumer trust in the food chain: exploring barriers and motivations [Internet]. [cited 2024 Nov 13]. Available from: https://eit.europa.eu/sites/default/files/18199_citizen_participation_forum_report.pdf
11. Luthra S, Mangla SK, Garg D, Kumar A. Internet of Things (IoT) in Agriculture Supply Chain Management: A Developing Country Perspective. In: Dwivedi YK, Rana NP, Slade EL, Shareef MA, Clement M, Simintiras AC, editors. Emerging Markets from a Multidisciplinary Perspective: Challenges, Opportunities and Research Agenda. Cham: Springer; 2018. pp. 209–20.
12. Rejeb A, Keogh JG, Zailani S, Treiblmaier H, Rejeb K. Blockchain Technology in the Food Industry: A Review of Potentials, Challenges and Future Research Directions. Logistics. 2020;4:27. [DOI]
13. Sorbo A, Pucci E, Nobili C, Taglieri I, Passeri D, Zoani C. Food Safety Assessment: Overview of Metrological Issues and Regulatory Aspects in the European Union. Separations. 2022;9:53. [DOI]
14. Wood B, Robinson E, Baker P, Paraje G, Mialon M, van Tulleken C, et al. What is the purpose of ultra-processed food? An exploratory analysis of the financialisation of ultra-processed food corporations and implications for public health. Global Health. 2023;19:85. [DOI] [PubMed] [PMC]
15. Ataei P, Litchfield AT. Big data reference architectures, a systematic literature review [Internet]. c2020 [cited 2024 Nov 13]. Available from: https://aisel.aisnet.org/acis2020/30
16. How to achieve smart data sharing [Internet]. Gartner, Inc.; c2024 [cited 2024 Nov 13]. Available from: https://www.gartner.com/smarterwithgartner/how-to-achieve-smart-data-sharing
17. Hashemi SK, Mirtaheri SL, Greco S. Fraud detection in banking data by machine learning techniques. IEEE Access. 2022;11:3034–43. [DOI]
18. Sliusar V, Akulyonok M, Andrianov A, Sliusar M, Tikhonov M. Algorithmic support for risk assessment in electronic production management. Proceedings of 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus); 2021 Jan 26–29; St. Petersburg, Moscow, Russia. IEEE; 2021. pp. 2262–5.
19. Olaoye F, Potter K. Business intelligence (bi) and analytics software: Empowering data-driven decision-making [Internet]. EasyChair; c2012–2024 [cited 2024 Nov 13]. Available from: https://easychair.org/publications/preprint/cR8v

20. Otto B. A federated infrastructure for european data spaces. Commun. 2022;65:44–5. [DOI]
21. National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Board on Research Data and Information; Committee on Toward an Open Science Enterprise. Open Science by Design: Realizing a Vision for 21st Century Research. Washington (DC): National Academies Press (US); 2018.
22. Murray-Rust P. Open data in science. Ser Rev. 2008;34:52–64. [DOI]
23. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. [DOI] [PubMed] [PMC]
24. Barker M, Chue Hong NP, Katz DS, Lamprecht AL, Martinez-Ortiz C, Psomopoulos F, et al. Introducing the FAIR Principles for research software. Sci Data. 2022;9:622. [DOI] [PubMed] [PMC]
25. Simmonds EG, Adjei KP, Andersen CW, Hetle Aspheim JC, Battistin C, Bulso N, et al. Insights into the quantification and reporting of model-related uncertainty across different disciplines. iScience. 2022; 25:105512. [DOI] [PubMed] [PMC]
26. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25:1251–5. [DOI] [PubMed] [PMC]
27. Abouakil D, Heurix J, Neubauer T. Data models for the pseudonymization of dicom data. Proceedings of 2011 44th Hawaii International Conference on System Sciences; 2011 Jan 4–7; Kauai, HI, USA. IEEE; 2011. pp. 1–11.
28. Russell RK, Hartnett E, Caron JL. Netcdf-4: Software implementing an enhanced data model for the geosciences. Proceedings of 22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology; 2006 Jan 31.
29. Folk M, Heber G, Koziol Q, Pourmal E, Robinson D. An overview of the hdf5 technology suite and its applications. In: Baumann P, Howe B, Orsborn K, Stefanova S. Proceedings of the EDBT/ICDT 2011 workshop on array databases; 2011 Mar 25; New York, United States. ACM; 2011. pp. 36–47.
30. Ledoux H, Arroyo Ohori K, Kumar K, Dukai B, Labetski A, Vitalis S. CityJSON: a compact and easy-to-use encoding of the CityGML data model. Open Geospatial Data Softw Stand. 2019;4:4. [DOI]
31. Moody DL, Shanks GG. Improving the quality of data models: empirical validation of a quality management framework. Inf Systems. 2023;28:619–50. [DOI]
32. Hinrichs H, Aden T. An iso 9001: 2000 compliant quality management system for data integration in data warehouse systems. In: Theodoratos D, Hammer J, Jeusfeld M, Staudt M, editors. Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2001); 2001 Jun 4; Interlaken, Switzerland. 2001. pp. 1–12.
33. Papoutsoglou G, Tarazona S, Lopes MB, Klammsteiner T, Ibrahimi E, Eckenberger J, et al. Machine learning approaches in microbiome research: challenges and best practices. Front Microbiol. 2023;14: 1261889. [DOI] [PubMed] [PMC]
34. Singh AV, Varma M, Rai M, Singh SP, Bansod G, Laux P, et al. Advancing predictive risk assessment of chemicals via integrating machine learning, computational modeling, and chemical/nano-quantitative structure-activity relationship approaches. Adv Intell Syst. 2024;6:2300366. [DOI]
35. Singh AV, Shelar A, Rai M, Laux P, Thakur M, Dosnkyi I, et al. Harmonization Risks and Rewards: Nano-QSAR for Agricultural Nanomaterials. J Agric Food Chem. 2024;72:2835–52. [DOI] [PubMed]
36. Badillo S, Banfai B, Birzele F, Davydov II, Hutchinson L, Kam-Thong T, et al. An Introduction to Machine Learning. Clin Pharmacol Ther. 2020;107:871–85. [DOI] [PubMed] [PMC]
37. Monaco A, Pantaleo E, Amoroso N, Lacalamita A, Lo Giudice C, Fonzino A, et al. A primer on machine learning techniques for genomic applications. Comput Struct Biotechnol J. 2021;19:4345–59. [DOI] [PubMed] [PMC]
38. Nasteski V. An overview of the supervised machine learning methods. Horizons. b. 2017;4:56. [DOI]

39. Shutaywi M, Kachouie NN. Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. Entropy (Basel). 2021;23:759. [DOI] [PubMed] [PMC]

40. van Engelen JE, Hoos HH. A survey on semi-supervised learning. Mach Learn. 2020;109:373–440. [DOI]

41. Bellotti R, Cerello P, Tangaro S, Bevilacqua V, Castellano M, Mastronardi G, et al. Distributed medical images analysis on a grid infrastructure. Futur Gener Comput Syst. 2007;23:475–84. [DOI]

42. Magarelli M, Novielli P, De Filippis F, Magliulo R, Di Bitonto P, Diacono D, et al. Explainable artificial intelligence and microbiome data for food geographical origin: the Mozzarella di Bufala Campana PDO Case of Study. Front Microbiol. 2024;15:1393243. [DOI] [PubMed] [PMC]

43. Magarelli M, Di Bitonto P, De Filippis F, Novielli P, Magliulo R, Diacono D. Securing origin integrity through machine learning analysis of mozzarella di bufala pdo microbiome. Proceedings of 2024 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4. 0 & IoT); 2024 May 29–31; Firenze, Italy. IEEE; 2024. pp. 38–42.

44. Singh AV, Bhardwaj P, Upadhyay AK, Pagani A, Upadhyay J, Bhadra J, et al. Navigating regulatory challenges in molecularly tailored nanomedicine. Explor BioMat-X. 2024;1:124–34. [DOI]

45. Singh AV, Bansod G, Mahajan M, Dietrich P, Singh SP, Rav K, et al. Digital Transformation in Toxicology: Improving Communication and Efficiency in Risk Assessment. ACS Omega. 2023;8: 21377–90. [DOI] [PubMed] [PMC]