BIG DATA FOR SUSTAINABLE CROP HEALTH: A CASE STUDY ON CGIAR

Barkavi A/P P Cheven (23093149)
Tan Kai Qi (23051355)
Ng Chee Kang (23051336)
Seh Chia Shin (17202838)

Faculty of Computer Science and Information Technology, Universiti Malaya
50603 Kuala Lumpur, Federal Territory of Kuala Lumpur, Malaysia

barkavicheven@gmail.com, 23051355@siswa.um.my, 23051336@siswa.um.my and 17202838@siswa.um.edu.my

Corresponding Author: Barkavi A/P P Cheven (23093149), Tan Kai Qi (23051355), Ng Chee Kang (23051336), Seh Chia Shin (17202838)
Received: 09 April 2025, 12 April 2025, Accepted: 13 April 2025
Published online: 14 April 2025

Abstract: This study explores the transformative impact of big data in promoting sustainable crop health through a focused case study on CGIAR, a global agricultural research organization. As agriculture faces challenges from climate change, pest outbreaks, and resource limitations, the integration of big data analytics emerges as a pivotal solution. The paper highlights how CGIAR harnesses vast datasets, including weather patterns, soil conditions, and crop genomics, to enhance decision-making, predict disease outbreaks, and optimize farming practices. By leveraging data-driven technologies and open-access platforms, CGIAR exemplifies how big data can be utilized to support global food security, improve crop resilience, and promote sustainable agricultural practices. This case study highlights the potential of data integration and machine learning in driving innovations that ensure long-term agricultural sustainability.

Keywords: Big data; CGIAR; Crop health; Agriculture

1. BACKGROUND STUDY

The Consultative Group on International Agricultural Research (CGIAR) stands as an unparalleled global consortium that has metamorphosed into the world's most extensive agricultural innovation network, orchestrating a symphony of scientific endeavors aimed at fortifying food security, alleviating poverty, and enhancing nutritional outcomes across the developing world [1]. Established in 1971 through the collaborative vision of the Rockefeller and Ford Foundations, this venerable institution has undergone a remarkable evolution, transcending its initial conceptualization to emerge as a multifaceted constellation of research centers, partnerships, and initiatives that collectively constitute a paradigmatic shift in agricultural research and development [2]. The organizational architecture of CGIAR encompasses a network of 15 independent research centers, strategically distributed across 70 countries, thereby creating an expansive geographical footprint that facilitates contextualized research responsive to diverse agroecological conditions and socioeconomic realities [1]. The genesis of CGIAR can be traced to the past period of the 1960s, characterized by predictions of widespread famine and agricultural insufficiency in developing regions [3]. This existential threat catalyzed the establishment of an institutional framework dedicated to the systematic application of scientific methodologies to agricultural challenges in resource-constrained environments. The evolutionary trajectory of CGIAR has been marked by successive waves of organizational reconfiguration, each representing a strategic response to emerging challenges and opportunities in the global agricultural landscape [4]. The most recent metamorphosis, initiated in 2020, has culminated in the "One CGIAR" framework, a transformative recalibration that consolidates governance structures, harmonizes research agendas, and optimizes resource allocation to maximize institutional impact and operational efficiency [5].

As for the geographical footprint and operational modalities, the geographical constellation of CGIAR's research centers constitutes a strategic distribution that encompasses diverse agroecological zones, from the arid landscapes of the Middle East to the tropical ecosystems of Southeast Asia, and from the highlands of Latin America to the savannahs of Sub-Saharan Africa [6]. This expansive footprint enables the organization to conduct contextualized research that addresses the heterogeneous challenges facing agricultural systems across the developing world. The operational modalities of CGIAR are characterized by a multidimensional approach that integrates field-based experimentation, laboratory analysis, computational modeling, and participatory methodologies, thereby creating a comprehensive research ecosystem that transcends conventional disciplinary boundaries [7].

Moreover, CGIAR's research portfolio encompasses a diverse array of thematic areas, including crop improvement, livestock systems, aquaculture, agroforestry, climate adaptation, nutrition, policy analysis, and gender equity, all unified by a common commitment to sustainable agricultural development [1]. This multifaceted research agenda is operationalized through a series of research programs and platforms that facilitate interdisciplinary collaboration and knowledge integration across the organizational network. The thematic orientation of CGIAR's research is guided by a strategic framework that prioritizes impact pathways aligned with the Sustainable Development Goals, particularly those related to zero hunger, poverty reduction, gender equality, climate action, and responsible consumption and production [8].

Furthermore, recognizing the transformative potential of digital technologies and data-driven approaches, CGIAR has embarked on an ambitious journey of digital transformation that permeates all aspects of its research and operational activities [9]. The CGIAR Platform for Big Data in Agriculture, launched in 2017, represents a cornerstone of this digital strategy, serving as a catalytic mechanism for harnessing the power of big data analytics, artificial intelligence, machine learning, and remote sensing to address complex agricultural challenges [10]. This platform facilitates the collection, integration, analysis, and visualization of diverse data streams, ranging from genomic sequences to satellite imagery, and from household surveys to meteorological measurements, thereby creating an unprecedented repository of agricultural intelligence that informs decision-making processes at multiple scales [1]. To be more profound, the digital ecosystem cultivated by CGIAR encompasses a constellation of innovative initiatives, including the development of mobile applications for agricultural advisory services, the deployment of sensor networks for environmental monitoring, the utilization of drone technology for crop surveillance, and the implementation of blockchain systems for supply chain transparency [11]. These technological interventions collectively constitute a paradigmatic shift in agricultural research methodologies, enabling the generation of insights and solutions that would be unattainable through conventional approaches. The organization's commitment to open data principles further amplifies the impact of these digital initiatives, with the GARDIAN platform serving as a centralized repository for CGIAR's research outputs, thereby democratizing access to agricultural knowledge and fostering collaborative innovation across geographical and institutional boundaries [12].

Besides, CGIAR’s operational philosophy is predicated on the recognition that complex agricultural challenges necessitate collaborative approaches that transcend institutional silos and disciplinary boundaries. Consequently, the organization has cultivated an extensive network of partnerships with diverse stakeholders, including national agricultural research systems, academic institutions, governmental agencies, non-governmental organizations, private sector entities, and farmer associations [1]. These collaborative relationships facilitate the co-creation of knowledge, the contextualization of research outputs, and the scaling of innovations, thereby enhancing the relevance, applicability, and impact of CGIAR's research portfolio. The partnership dynamics within the CGIAR ecosystem are characterized by a multidirectional flow of knowledge, resources, and expertise, creating synergistic relationships that leverage the complementary strengths of diverse actors within the agricultural innovation landscape [13]. This collaborative approach is exemplified by initiatives such as the Excellence in Breeding Platform, which brings together CGIAR centers, national partners, and private sector entities to accelerate genetic gains in crop improvement programs through the application of cutting-edge technologies and methodologies [14].

The multifaceted impact of CGIAR's research and development activities spans multiple dimensions, from the tangible outputs of improved crop varieties and livestock breeds to the intangible outcomes of enhanced institutional capacities and policy frameworks [15]. The organization's contributions to global food security are quantitatively substantiated by the widespread adoption of improved crop varieties developed by CGIAR centers, with an estimated 70% of wheat area and 45% of rice area in developing countries planted with CGIAR-derived varieties [16]. The economic impact of these agricultural innovations is equally impressive, with studies indicating that for every dollar invested in CGIAR research, approximately $17 in benefits are generated for developing countries, representing an exceptional return on investment in the realm of international development [17]. Beyond these quantifiable metrics, CGIAR's impact extends to broader developmental dimensions, including the enhancement of nutritional outcomes through biofortified crops, the mitigation of environmental degradation through sustainable land management practices, the empowerment of women through gender-responsive agricultural technologies, and the strengthening of resilience to climate variability through adaptive farming systems [1]. These multidimensional contributions collectively position CGIAR as an indispensable actor in the global effort to achieve sustainable agricultural development and food security in the face of escalating challenges such as population growth, climate change, and resource constraints.

Regarding the big data applications in agricultural research, the integration of big data methodologies within CGIAR's research framework represents a paradigmatic shift in agricultural science, enabling the organization to harness the power of data-driven insights to address complex challenges across the agricultural value chain [9]. The voluminous nature of agricultural data, characterized by its heterogeneity, velocity, and complexity, necessitates sophisticated analytical approaches that transcend conventional statistical methodologies. CGIAR's big data initiatives encompass a diverse array of applications, including predictive modeling for crop yield forecasting, genomic selection for accelerated breeding, remote sensing for environmental monitoring, and machine learning for pest and disease detection [10]. The organization's commitment to big data is operationalized through various platforms and initiatives, including the CGIAR Platform for Big Data in Agriculture, which serves as a centralized hub for data-related activities across the CGIAR network [1]. This platform facilitates the development of data standards, the enhancement of analytical capacities, the promotion of open data principles, and the cultivation of a data-driven culture within the agricultural research community. The platform's Communities of Practice, focusing on areas such as crop modeling, geospatial analysis, and socioeconomic data, provides collaborative spaces for researchers to exchange methodologies, share insights, and co-create solutions to common challenges in agricultural data science [12].

Apart from that, the CGIAR's application of big data in agricultural research is exemplified by initiatives such as the "Computer Vision for Crop Disease" project, which utilizes machine learning algorithms to analyze images of plant leaves for early detection of diseases, thereby enabling timely interventions that minimize crop losses and reduce reliance on chemical pesticides [5]. Similarly, the organization's work on digital soil mapping combines satellite imagery, ground-based measurements, and machine learning techniques to generate high-resolution maps of soil properties, providing valuable information for targeted interventions in soil fertility management and land use planning [18].

As CGIAR navigates the complex landscape of 21st-century agricultural challenges, its strategic vision is guided by a commitment to harnessing the transformative potential of digital technologies and data-driven approaches to accelerate agricultural innovation and enhance developmental impact [5]. The organization's future trajectory encompasses several key dimensions, including the expansion of digital advisory services to reach millions of smallholder farmers, the enhancement of predictive capabilities for climate-related risks, the acceleration of genetic gains through genomic selection, and the optimization of resource allocation through precision agriculture techniques [8]. The "One CGIAR" transformation represents a strategic recalibration aimed at enhancing the organization's capacity to address emerging challenges and capitalize on new opportunities in the agricultural research landscape [1]. This institutional metamorphosis is characterized by a unified governance structure, an integrated operational framework, and a harmonized research agenda, all designed to maximize the collective impact of CGIAR's diverse capabilities and resources. The digital dimension of this transformation is particularly significant, with investments in data infrastructure, analytical capabilities, and digital skills positioned as central elements of the organization's strategic roadmap.

In conclusion, CGIAR stands as a paragon of agricultural innovation and data-driven transformation, leveraging its extensive network, multidisciplinary expertise, and technological capabilities to address the complex challenges facing agricultural systems in the developing world. Through its integration of big data methodologies, digital technologies, and collaborative approaches, the organization continues to generate impactful solutions that enhance food security, reduce poverty, improve nutrition, and promote environmental sustainability across diverse agroecological and socioeconomic contexts. As the agricultural landscape evolves in response to emerging challenges such as climate change, population growth, and resource constraints, CGIAR's adaptive capacity and innovative spirit position as an indispensable actor in the global effort to achieve sustainable agricultural development and food security for future generations.

2. CHARACTERISTICS OF BIG DATA

The Consultative Group on International Agricultural Research (CGIAR) stands as a preeminent global partnership that unites international organizations engaged in research for a food-secure future. In the contemporary agricultural landscape, CGIAR has emerged as a vanguard institution leveraging big data analytics to revolutionize agricultural practices, enhance food security, and mitigate the deleterious impacts of climate change on global food systems [5]. The organization's strategic deployment of big data methodologies exemplifies the transformative potential of data-driven approaches in addressing multifaceted agricultural challenges across diverse geographical and socioeconomic contexts.

CGIAR's research initiatives generate and process colossal volumes of agricultural data that transcend traditional data management capabilities. The organization's data repositories encompass extensive genomic sequences from thousands of crop varieties and their wild relatives, with each genome containing billions of base pairs [19]. High-resolution satellite imagery spanning millions of hectares of agricultural land, with temporal resolutions capturing daily changes over decades [20], further contributes to this data deluge. Comprehensive climate datasets incorporating historical records and predictive models with petabyte-scale storage requirements [21] and phenotypic data from field trials conducted across 15 research centers in over 70 countries, generating terabytes of observational data annually [22], round out this massive data ecosystem. This voluminous data necessitates advanced computational infrastructure, with CGIAR's data centers processing approximately 2.5 petabytes of agricultural data annually, a figure that continues to expand exponentially as sensor networks and field monitoring systems proliferate across global agricultural landscapes [12].

The heterogeneity of CGIAR's data ecosystem represents a quintessential characteristic of big data in agricultural research. The organization integrates disparate data types including structured data: tabular datasets containing crop yield statistics, meteorological measurements, and soil composition analyses with precise schema definitions [23]. Unstructured data, including textual reports, farmer interviews, and indigenous knowledge repositories that require natural language processing for meaningful extraction [24], adds another layer of complexity. Semi-structured data, such as XML and JSON files containing sensor readings from IoT devices deployed across experimental farms with hierarchical data organization [25], must also be integrated. Spatial data representing land use patterns, crop distribution, and ecological zones with complex topological relationships [26] and temporal data tracking crop growth stages, disease progression, and climate patterns with varying granularity and periodicity [27] complete this diverse data landscape. CGIAR's Platform for Big Data in Agriculture has pioneered the integration of these heterogeneous data types through innovative data harmonization frameworks and ontological standards, enabling cross-domain analytics that were previously unattainable [22].

The velocity dimension of CGIAR's big data operations manifests in the rapid acquisition, processing, and dissemination of agricultural information across temporal scales. Real-time sensor networks deployed across experimental farms transmit environmental parameters at sub-minute intervals, generating continuous data streams that require immediate processing [25]. Automated weather stations integrated with CGIAR's data infrastructure provide hourly meteorological updates across diverse agroecological zones, enabling timely interventions during extreme weather events [21]. Satellite imagery with daily revisit capabilities delivers near-real-time monitoring of crop conditions, necessitating high-throughput processing pipelines to extract actionable insights within operational timeframes [20]. Mobile applications developed by CGIAR enable farmers to report pest outbreaks and disease symptoms instantaneously, creating dynamic data streams that inform rapid response mechanisms [24]. To accommodate these high-velocity data streams, CGIAR has implemented distributed computing architectures and stream processing frameworks that enable real-time analytics, reducing the latency between data acquisition and actionable insights from days to minutes [12].

The veracity dimension addresses the inherent uncertainties and quality concerns in CGIAR's agricultural big data ecosystem. Data quality assessment frameworks implemented across CGIAR centers establish standardized protocols for evaluating the accuracy, completeness, and consistency of agricultural datasets [28]. Automated validation algorithms detect anomalous values and inconsistencies in sensor readings, flagging potential errors for human review before incorporation into analytical pipelines [25]. Metadata standards adhering to FAIR principles (Findable, Accessible, Interoperable, Reusable) enhance the provenance tracking and reliability assessment of datasets shared across the CGIAR network [29]. Uncertainty quantification methodologies accompany predictive models, providing confidence intervals and reliability metrics for agricultural forecasts that inform risk management strategies [21]. Cross-validation techniques comparing satellite-derived indicators with ground-truth measurements establish accuracy assessments for remote sensing products used in crop monitoring applications [20]. CGIAR's commitment to data veracity is exemplified by its establishment of data quality certification processes that ensure research outputs meet rigorous standards before informing agricultural policies and interventions [22].

The variability dimension reflects CGIAR's capacity to accommodate fluctuating data patterns and evolving analytical requirements in agricultural research. Seasonal variations in data collection intensity, with peak periods during growing seasons generating up to five times the data volume compared to off-season periods, necessitate elastic computing resources [26]. Climate change-induced shifts in agricultural patterns create evolving data signatures that require adaptive analytical frameworks capable of detecting emerging trends and anomalies [27]. Market dynamics and policy changes influence farmer behavior and agricultural practices, introducing variability in socioeconomic datasets that inform CGIAR's intervention strategies [30]. Pest and disease outbreaks create sudden spikes in monitoring data and farmer reports, triggering automated scaling of computational resources to accommodate surge capacity requirements [25]. To address these variability challenges, CGIAR has implemented adaptive data processing pipelines with dynamic resource allocation, ensuring analytical continuity despite fluctuating data characteristics and processing demands [12].

The visualization dimension represents CGIAR's approach to rendering complex agricultural data comprehensible and actionable for diverse stakeholders. Interactive dashboards developed by CGIAR's data scientists transform multidimensional crop performance data into intuitive visual interfaces accessible to researchers, policymakers, and agricultural extension officers [22]. Geospatial visualization platforms integrate satellite imagery, climate data, and crop models to generate dynamic maps of agricultural productivity and vulnerability across temporal and spatial scales [20]. Network visualization tools elucidate complex relationships between genetic traits, environmental factors, and crop performance, facilitating the identification of resilient varieties for climate adaptation [19]. Temporal visualization techniques illustrate long-term trends in agricultural productivity and climate patterns, highlighting critical intervention points for sustainable intensification strategies [21]. CGIAR's visualization innovations extend beyond conventional approaches to include virtual reality applications that immerse stakeholders in simulated agricultural landscapes, enhancing comprehension of complex ecological interactions and intervention impacts [12].

The value dimension encapsulates CGIAR's ultimate objective: transforming big agricultural data into tangible benefits for global food systems and rural livelihoods. Precision agriculture recommendations derived from big data analytics have increased crop yields by 15-25% while reducing input costs by 10-20% across CGIAR's intervention sites in South Asia and Sub-Saharan Africa [23]. Early warning systems powered by real-time data integration have enabled preemptive responses to pest outbreaks, preventing estimated crop losses valued at $45-60 million annually across CGIAR's partner countries [25]. Climate-smart agriculture strategies informed by big data models have enhanced the resilience of over 8 million smallholder farmers to climate variability, reducing crop failure risks by 30-40% during extreme weather events [27]. Data-driven policy recommendations have influenced agricultural investment decisions and resource allocations in over 30 countries, optimizing the impact of limited resources on food security outcomes [30]. CGIAR's value proposition extends beyond immediate agricultural productivity to encompass broader societal benefits, including enhanced environmental sustainability, improved nutrition outcomes, and strengthened rural economies through data-empowered decision-making [5].

The seven dimensions of big data which are volume, variety, velocity, veracity, variability, visualization, and value collectively define CGIAR's innovative approach to agricultural research and development. By harnessing these multifaceted characteristics, CGIAR has established a transformative paradigm that transcends traditional agricultural research methodologies, enabling evidence-based interventions that address the complex challenges of global food security in the context of climate change and population growth. As agricultural systems continue to evolve in response to global challenges, CGIAR's big data ecosystem represents a vanguard approach to generating, integrating, and translating agricultural information into actionable knowledge. The organization's strategic investment in data infrastructure, analytical capabilities, and knowledge dissemination mechanisms exemplifies the transformative potential of big data in revolutionizing agricultural practices and enhancing food security across diverse geographical and socioeconomic contexts.

3. SIX PHASES OF BIG DATA

The growth of big data has significantly changed numerous industry sectors and institutions, including agriculture. CGIAR (Consultative Group on International Agricultural Research) is one of the institutions using big data for agricultural development. CGIAR able to enhance decision-making processes, agricultural productivity and disease detection across developing countries through the big data platform. This organization employs a comprehensive six-phase big data approach to address real-world challenges in agriculture, such as early disease detection, food security and sustainability [31].

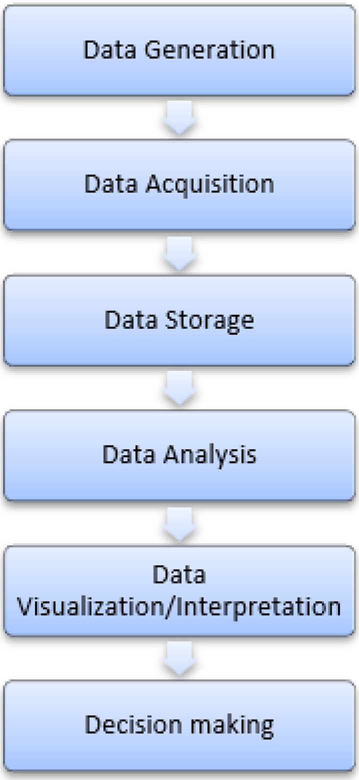


Figure 1. Six Phases of Big Data

Data Generation

CGIAR consists of raw agricultural data that is produced through various sources and methods. These include field experiments, sensor data, drone and satellite images, farmer reports and mobile applications [32]. CGIAR collaborates with local farmers and researchers to generate reliable and diverse datasets that can represent real-world agricultural conditions, Other output datasets from the result framework are aligned with 5 impact areas to achieve the Sustainable Development Goals (SDGs) [33]:

- Climate adaptation and mitigation
- Environmental health and biodiversity
- Gener quality, youth and social inclusion
- Nutrition, health and food security
- Poverty reduction, livelihoods and job

Based on the Kaggle dataset, it is generated by capturing high-resolution crop images from farms [34]. The dataset provides a comprehensive visual dataset that can allow researchers to be used for machine learning-based disease studies. The images with ID that show the disease symptoms allow researchers to study the crop disease by machine learning model based on the rust condition of the crop.

Data Acquisition

The generated data is collected and centralized for further processing. CGIAR utilizes different digital tools and platforms for data gathering. The tools include mobile apps, Open Data Kit (ODK) and automated sensors. In the crop disease detection project, images and metadata, for example crop type, geographic location, and disease label are uploaded to centralized databases through cloud-based platforms. By collaborating with multiple agricultural institutions and research centers, CGIAR ensures that the data acquisition process is extensive and continuous, covering multiple countries and crop types.

Data Storage

In 2014, CGIAR collaborated with AWS to make its Global Circulation Models (GCM) data accessible and support research on climate change impacts on agriculture [35]. CGIAR also uses AWS architecture for projects such as Digital Earth Africa program based on the FAIR data principles [36]. Furthermore, CGIAR has leveraged cloud-based platforms such as Google Earth Engine, which operates on Google's cloud infrastructure, to process and analyze geospatial data on a global scale [37]. The effort reflects that the CGIAR's commitment to using cloud storage services to improve the accessibility and scalability of research data.

Data Analysis

At the next phase of data storage is data analysis, it is the phase where raw data is transformed into useful insights. CGIAR is using advanced analytical techniques such as statistical modeling, machine learning (ML) and deep learning (DL) algorithms. For example, convolutional neural networks (CNNs) are used to analyze crop images and identify disease patterns. By training models on thousands of labeled images, CGIAR can develop automated systems that detect diseases with high accuracy [34]. Furthermore, predictive analytics is used to forecast crop yields, disease outbreaks and the effects of climate change. The use of big data analytics helps CGIAR to derive meaningful conclusions that guide field practices and policy decisions. Another example from [38], the data such as soil moisture, average rainfall and soil nutrients are used to predicts whether the fertilizers will cause crop disease. Based on the research, the Hadoop system was used to store and analyze the data by using ML techniques.

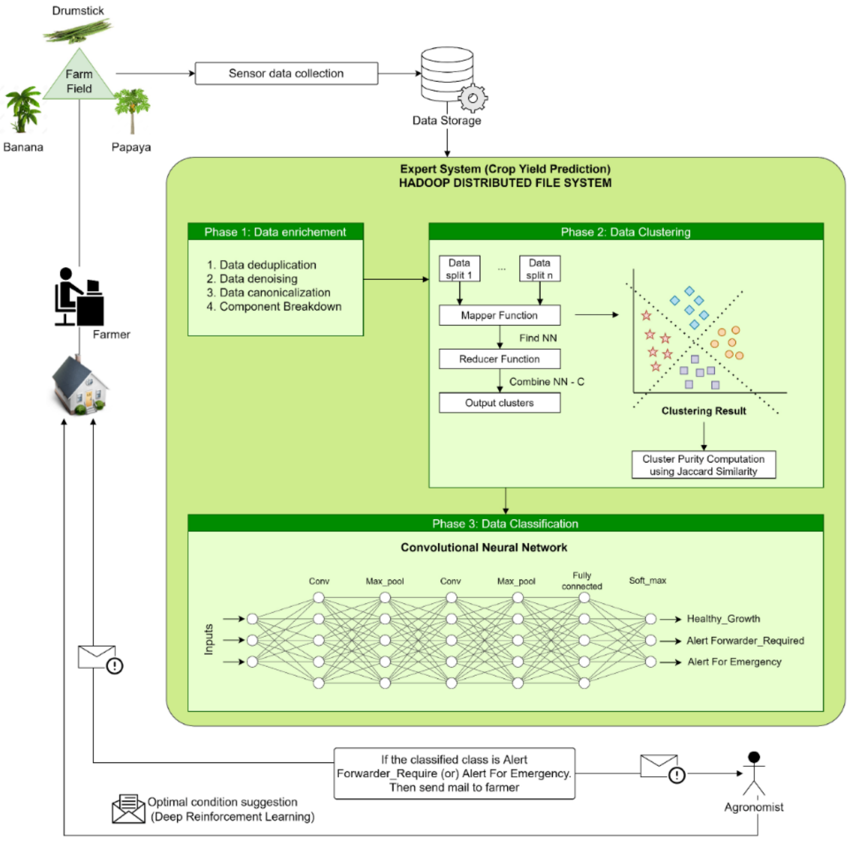
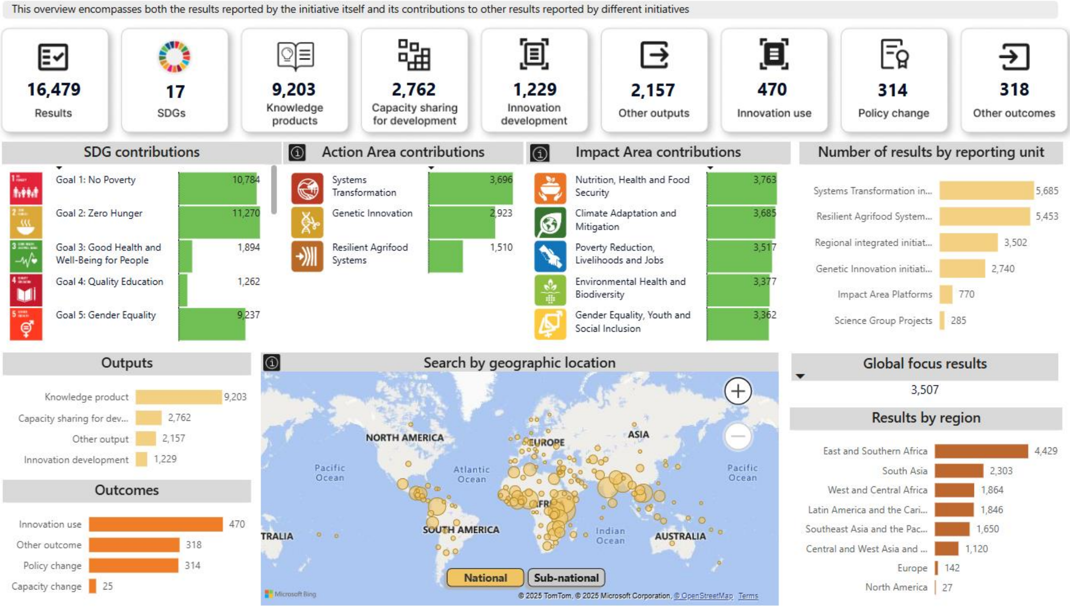


Figure 2. Big data architecture for the fertilizer analysis and yield prediction (Cravero et al., 2022)

Data Visualization and Interpretation

Once analyzed, data must be visualized effectively to make it understandable and actionable. CGIAR develops dashboards, interactive maps and infographics to present analysis results to farmers, researchers and policymakers. The most common tool used by CGIAR is its custom web applications. From the web application, the heat map from the result dashboard shows the projects' outcomes that have been worked on worldwide. These visualizations help stakeholders to track, visualize and communicate the progress and outcomes of CGIAR's research and innovation activities about global food security, climate resilience, and agricultural development. The dashboard showcases the global impact from CGIAR in crop productivity, climate change adaptation, gender equality and nutrition and food system for decision-making for a certain research or project.


This overview encompasses both the results reported by the initiative itself and its contributions to other results reported by different initiatives

16,479 Results
17 SDGs
9,203 Knowledge products
2,762 Capacity sharing for development
1,229 Innovation development
2,157 Other outputs
470 Innovation use
314 Policy change
318 Other outcomes
SDG contributions
Goal 1: No Poverty 10,784
Goal 2: Zero Hunger 11,270
Goal 3: Good Health and Well-Being for People 1,894
Goal 4: Quality Education 1,262
Goal 5: Gender Equality 9,237
Action Area contributions
Systems Transformation 3,696
Genetic Innovation 2,923
Resilient Agrifood Systems 1,510
Impact Area contributions
Nutrition, Health and Food Security 3,763
Climate Adaptation and Mitigation 3,685
Poverty Reduction, Livelihoods and Jobs 3,517
Environmental Health and Biodiversity 3,377
Gender Equality, Youth and Social Inclusion 3,362
Number of results by reporting unit
Systems Transformation in... 5,685
Resilient Agrifood System... 5,453
Regional integrated initiat... 3,502
Genetic Innovation initiati... 2,740
Impact Area Platforms 770
Science Group Projects 285
Outputs
Knowledge product 9,203
Capacity sharing for dev... 2,762
Other output 2,157
Innovation development 1,229
Outcomes
Innovation use 470
Other outcome 318
Policy change 314
Capacity change 25
Search by geographic location
NORTH AMERICA
EUROPE
ASIA
Pacific Ocean
Atlantic Ocean
AFRICA
SOUTH AMERICA
Indian Ocean
AUSTRALIA
TRALIA
National
Sub-national
Microsoft Bing
© 2025 TomTom, © 2025 Microsoft Corporation, © OpenStreetMap Terms
Global focus results
3,507
Results by region
East and Southern Africa 4,429
South Asia 2,303
West and Central Africa 1,864
Latin America and the Cari... 1,846
Southeast Asia and the Pac... 1,650
Central and West Asia and ... 1,120
Europe 142
North America 27

Figure 3. CGIAR's Result Dashboard [31]

Decision Making

The last phase of the big data which involves decision-making based on data visualization. At CGIAR, big data allows stakeholders to make decisions such as when to trigger farmers on the potential disease risks, where to implement pest control measures and how to breed disease-resistant crop varieties. Besides, the policy decisions from the government may also be influenced by CGIAR's findings or data, especially in regions having food insecurity. By integrating data-driven recommendations into everyday farming and policy frameworks, CGIAR able to enhance resilience, productivity and sustainability in agriculture.

In summary, the six phases of big data, like data generation, acquisition, storage, analysis, visualization and decision-making, are effectively utilized by CGIAR to address agricultural challenges. With real-world datasets like the crop disease detection project, CGIAR is showing the advantage of big data usage that can transform farming practices, improve food security and enhance global agricultural sustainability to meet the SDG goals. These efforts highlight how importance of data-driven innovation in addressing the complex and dynamic needs of modern agriculture.

4. THE NECESSITY OF BIG DATA SOFTWARE

4.1 Why Big Data Software is Essential in Crop Health

Agriculture today faces a complex and dynamic landscape, where ensuring crop health is not just a matter of monitoring leaves for discoloration but involves understanding a vast web of interconnected data. Modern crop health monitoring encompasses an enormous scale of information, from multispectral satellite imagery and sensor-based soil moisture readings to drone footage and weather forecasts. This data is heterogeneous, high in volume, and fast-changing, making timely interventions crucial for preventing crop loss. Yet, the agricultural sector continues to struggle with fragmented and unstandardized data, limiting the full potential of digital transformation. As noted by the World Economic Forum, agricultural data today lacks interoperability and often remains siloed, making analysis and collaboration difficult. To truly capitalize on the promise of emerging technologies like AI and remote sensing, this data must be efficiently collected, processed, and integrated.

Traditional crop monitoring techniques, such as manual field inspections or sample-based lab analyses, are proving inadequate for today's agricultural demands. These approaches are labor-intensive, time-consuming, and inherently limited in scope, often providing only a partial snapshot of field conditions. Moreover, their subjectivity and inconsistency introduce room for error, while their inability to deliver real-time insights delays responses to pest outbreaks, nutrient deficiencies, or water stress. In large-scale agricultural operations like those supported by CGIAR, such limitations can result in significant productivity losses and misallocation of resources. The need for non-invasive, scalable, and continuous monitoring solutions has never been more urgent.

Big data software offers a transformative solution to these challenges by enabling scalable, timely, accurate, and cost-effective crop health monitoring. For organizations like CGIAR, whose mission focuses on achieving food security and sustainable agriculture, this capability is crucial. Big data platforms integrate and analyze inputs from satellites, IoT devices, field cameras, and historical yield data to detect patterns and predict threats. For instance, early disease detection models powered by machine learning can process thousands of leaf images from different regions in real-time, alerting farmers before outbreaks spread. These systems significantly reduce costs associated with manual labor, improve the accuracy of yield forecasts, and allow real-time decision-making, thereby enhancing agricultural resilience in the face of climate change, pest invasions, and market volatility. As highlighted in the NASSCOM report, unlocking the potential of AI and data could generate $65 billion in economic value in India alone [39], underscoring its relevance on a global scale.

The complexity, urgency, and scale of crop health management today make big data software not just a beneficial tool, but a necessary infrastructure, particularly for mission-driven institutions like CGIAR.

4.2 Real-World Example: The CGIAR Project

To illustrate the transformative potential of big data software in agriculture, the CGIAR-sponsored Computer Vision for Crop Disease competition offers a compelling real-world example. Hosted on Zindi in collaboration with the Computer Vision for Agriculture (CV4A) Workshop at ICLR, the challenge focused on a critical agricultural issue: wheat rust detection in sub-Saharan Africa, particularly in Ethiopia and Tanzania. Wheat rust is a fast-spreading fungal disease that drastically reduces crop yields and endangers food security in vulnerable regions. Early and accurate detection is crucial, but traditional monitoring methods struggle to scale to regional or national levels.

The dataset for the challenge contained thousands of labeled images of wheat leaves, sourced both from public platforms (like Google Images) and directly from the field through partnerships with organizations such as CIMMYT. Each image was annotated to indicate whether the plant was healthy, infected with stem rust, or infected with leaf rust. While some images featured both types of rust, participants were tasked with building machine learning models that could classify the image based on the dominant condition, outputting probabilities for each category.

This task, though framed as an image classification problem, revealed the real-world complexities of agricultural data. Preprocessing was necessary to clean, normalize, and augment the imagery data, especially given variations in lighting, camera angles, and image quality. Additionally, storage and access of large-scale image datasets, along with the need for computationally intensive model training and evaluation, highlighted the importance of robust big data infrastructure.

Building a disease classification model at this scale is only feasible through the use of scalable machine learning pipelines, cloud storage, and distributed computing. Big data software frameworks such as TensorFlow, PyTorch, or Apache Spark (for data handling) are essential to process large volumes of heterogeneous image data efficiently and in a reproducible manner.

Ultimately, projects like CGIAR's wheat rust challenge underscore how big data technologies enable novel, scalable approaches to disease monitoring and agricultural risk management. Such initiatives not only push the boundaries of AI research but also serve CGIAR's broader mission: to improve food security and resilience in the face of climate change, pests, and declining crop yields.

4.3 Key Big Data Software

4.3.1 Machine Learning Platforms

In CGIAR-like agricultural applications, machine learning platforms such as TensorFlow and PyTorch play a central role in automating image-based crop and disease diagnostics. These platforms are essential for training and deploying deep convolutional neural networks (CNNs) to identify plant diseases from leaf images, enabling high-throughput screening without the need for expert manual inspection.

One study utilized AlexNet and GoogLeNet architectures to classify 14 crop species and 26 diseases using over 54,000 images [40]. The best model achieved a remarkable 99.35% accuracy, demonstrating the power of deep learning in agricultural disease detection. Importantly, transfer learning significantly improved model performance, highlighting the value of pre-trained models in data-scarce environments, a common scenario in developing regions served by CGIAR.

Furthermore, the study emphasizes the importance of model interpretability and continual learning, ensuring that solutions remain adaptive to new diseases and crop types. These platforms thus serve not only as analytical engines but also as the technological backbone for scalable, mobile diagnostic tools targeting smallholder farmers.

4.3.2 Cloud Storage & Distributed Computing

Big data in agriculture involves vast volumes of image, sensor, and geospatial data, necessitating robust cloud storage and distributed computing infrastructures. Platforms like Amazon Web Services (AWS), Google Cloud Platform (GCP), Apache Hadoop, and Apache Spark enable efficient data ingestion, parallel processing, and global access for CGIAR's research network.

A comprehensive survey [41] on deep learning in agriculture showed that distributed computing environments are critical for:

- Scaling model training across terabytes of heterogeneous agricultural datasets,
- Accelerating analysis through parallelized computation (e.g., Spark for real-time plant health monitoring),
- Enabling global collaboration, where scientists across different continents can access shared datasets and model outputs.

Cloud infrastructures also support automated data pipelines, essential for continuous monitoring of crops, integrating weather data, remote sensing inputs, and farmer-reported outcomes.

This architecture forms the digital core of modern agricultural innovation, allowing CGIAR to efficiently process and analyze complex, multi-modal data for informed decision-making and rapid intervention.

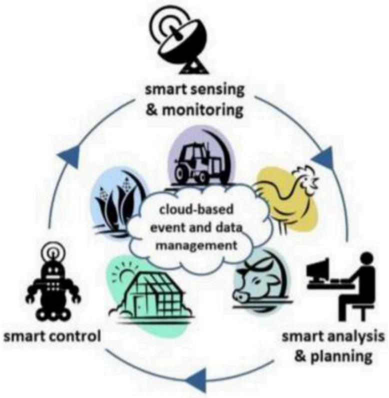


Figure 4. The smart farm conceptual framework

4.3.3 Geospatial and Visualization Tools

Geospatial tools are fundamental to spatially-aware agricultural analytics, enabling CGIAR and its partners to visualize and understand phenomena like disease outbreaks, soil degradation, and climate impacts.

In the proposed Geospatial Cloud Framework, platforms like WebGIS, QGIS, and Google Earth Engine integrate satellite imagery, drone-collected data, and IoT signals to support precision agriculture and conservation. These tools allow:

- Visualization of spatiotemporal disease spread across regions,
- Overlaying of vegetation health indices with intervention hotspots,
- Simulation of policy outcomes via digital twins for sustainable practices.

The study on this framework envisions a global-to-local feedback loop, where data from smart farms inform regional models and vice versa [42]. Such a system aligns with CGIAR’s mission to support both on-the-ground farming and high-level policy decisions, fostering a data ecosystem that is as scalable as it is context-sensitive.

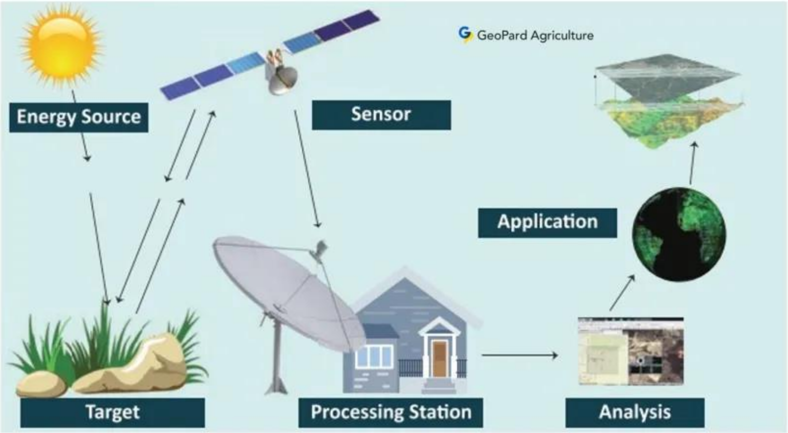


Figure 5. Applications of (GIS) Geoinformatics in Agriculture

4.4 Translates Big Data into Action

In a world where agricultural threats are becoming increasingly unpredictable and severe, big data software has moved from a luxury to a necessity. Plant disease outbreaks are not only a scientific challenge but also a humanitarian and economic one, threatening food security and farmer livelihoods globally. Traditional tools simply cannot respond fast enough, nor operate at the scale needed to manage these complex, transboundary risks. Big data software fills this gap by offering scalable, real-time insights, transforming how we detect, understand, and respond to crop health crises.

CGIAR's work illustrates this transformation in action. With its global reach and science-driven mission, CGIAR is leveraging big data to develop early warning systems, precision agriculture methods, and collaborative research platforms. These efforts are powered by a growing ecosystem of tools and technologies, such as machine learning models for disease detection, geospatial mapping for environmental monitoring, and cloud-based infrastructure for managing massive datasets.

Importantly, CGIAR is not starting from scratch. It has already built key platforms like GARDIAN (Global Agricultural Research Data Innovation & Acceleration Network), which harvests and connects research outputs across CGIAR institutions to enable rapid data discovery and reuse. Meanwhile, CG Labs provides an open, cloud-based analytical environment where researchers can run models, visualize data, and collaborate in real time. These initiatives demonstrate that CGIAR isn't just adopting big data tools, it is actively developing and deploying them to serve both the research community and real-world stakeholders.

Ultimately, the integration of big data software into CGIAR's strategy is about more than technical innovation. It's about making these capabilities accessible to the people who need them most: farmers on the front lines of climate and disease challenges, and policymakers tasked with making informed, timely decisions. By bridging the gap between data science and practical application, CGIAR is ensuring that agricultural innovation benefits not just labs and journals, but entire communities, accelerating the path toward sustainable, secure food systems for the future.

5. KEY OBSTACLES FOR BIG DATA SOLUTION

1. Data Quality and Standardization
   Agricultural data is frequently inconsistent across sources, unstructured, and heterogeneous. The data quality varies due to CGIAR gathering information from multiple sources, which include sensors such as satellites and drones, crop types, and manual input, which lead to inconsistent data quality. For example:
   - The Kaggle dataset is useful, but it contains problems with image quality, inconsistent labelling, and might not be representative of all crop conditions globally.
   - Variations in disease severity, lighting, and image resolution increase the difficulty of training robust machine learning (ML).

2. Infrastructure and connectivity challenge
   A large number of CGIAR field sites are located in low-income nations with poor internet connectivity and limited access to high-performance computing resources. For instance:
   - Transmitting large datasets such as high-resolution crop materials can be a challenge for central processing.
   - Device limitations prevent real-time processing or mobile based machine learning applications.

3. Lack of skilled personnel challenge
   Agronomy, data science, and remote sensing are interdisciplinary skills that are frequently lacking in CGIAR centres. This makes it difficult to implement big data analytics and AI in agriculture. For example:
   - The process of teaching agricultural research in deep learning and data wrangling takes a longer time.
   - A meaningful interpretation of ML model outputs required domain expertise, which a generic data scientist might not have.

   Agronomic knowledge is necessary to avoid false positives or negatives in real-world settings in order to interpret the findings of disease detection models in the Kaggle dataset.

4. Privacy and Data Ownership Issues
   Information gathered from smallholder farmers raises concerns regarding ethical use, consent, data ownership, and ethical usage. For instance:
   - If the benefits are unclear, farmers might be reluctant to share data.
   - Regulatory frameworks for data protection are weak in many of the nations where CGIAR operates.

   There might be worries about bias if disease data is used for commercial seed/pesticide recommendations.

5. Integration with existing agricultural practices
   Big data solutions need to be easily incorporated into conventional farming systems and must be actionable. For instance:
   - AI-based recommendations might not be understood or trusted by farmers.
   - Practical decision making tools and model outputs may not be compatible.

   For example, a model that uses lead photos to predict maize disease needs to combine with advisory systems that inform farmers about the best time and method to react with considering local context.

6. Limited Localized Datasets
   Although labelled images of crop diseases are available in the Kaggle dataset, it might not include regional variants specific to certain agro-ecological zones. For example:
   - ML models that have been trained on global datasets might not perform well in local contexts.
   - Annotation and data collection must be done locally, which requires a lot of effort and coordination.

6. CONCLUSION

The CGIAR case study illustrates how big data can revolutionize agricultural research and development, particularly in addressing global challenges such as food security, climate change, and resource scarcity. Through the implementation of a six-phase big data framework, CGIAR has effectively created a digital ecosystem that provides timely and actionable insights to farmers, researchers, and policymakers. It utilizes advanced technologies like artificial intelligence, remote sensing, cloud infrastructure, and machine learning models to enable precision agriculture that improves yield forecasting, optimizes resource allocation, and monitors crop health.

CGIAR's commitment to open data, interdisciplinary collaboration, and strategic partnerships further strengthens its ability to deliver impactful solutions. Projects like the Computer Vision for Crop Disease show how AI and big data software can minimize agricultural losses, lower input costs, and detect diseases early, especially in vulnerable areas. Cloud-based tools, geospatial analytics, and FAIR data principles have been integrated to create a scalable infrastructure that facilitates global collaboration and real-time interventions.

CGIAR remains a leader in implementing data-driven approaches to sustainable agriculture, despite multiple obstacles such as inconsistent data quality, limited connectivity in low-income regions, and the shortage of skilled personnel, CGIAR remains at the forefront of implementing data-driven strategies for sustainable agriculture. Its ambitious "One CGIAR" transformation is a reflection of an institutional realignment to accommodate new demands through harmonized research agendas and unified governance.

Ultimately, CGIAR is a prime example of how big data can transform unprocessed data into insightful knowledge that drives the growth of sustainable agriculture. CGIAR ensures that innovations are not only scientifically sound but also socially inclusive, economically feasible, and environmentally resilient by bridging the gap between research and practice. This encourages agriculture to thrive in the future despite the rapidly changing global environment.

7. REFERENCES

[1] *CGIAR Annual Report*, 2023. [Online]. Available: https://annualreport.cgiar.org/2023#section--id-4

[2] D. G. Dalrymple, “International agricultural research as a global public good: Concepts, the CGIAR experience and policy issues,” *J. Int. Dev.*, vol. 20, no. 3, pp. 347–379, 2008.

[3] P. L. Pingali, “Green revolution: Impacts, limits, and the path ahead,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109, no. 31, pp. 12302–12308, 2012.

[4] D. Byerlee and J. K. Lynam, “The development of the international agricultural research system: A historical perspective,” *World Dev.*, vol. 135, p. 105080, 2020.

[5] CGIAR, *CGIAR 2030 Research and Innovation Strategy: Transforming food, land, and water systems in a climate crisis*. CGIAR System Organization, 2022.

[6] S. Özgediz, *The CGIAR at 40: Institutional evolution of the world's premier agricultural research network*. CGIAR Fund Office, 2012.

[7] A. Hall, J. Dijkman, B. Taylor, L. Williams, and J. Kelly, *One CGIAR: Leveraging systems thinking for more effective agricultural research for development*. CGIAR Independent Science and Partnership Council, 2018.

[8] A. Steiner *et al.*, “Actions to transform food systems under climate change,” CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS), 2020.

[9] S. J. Janssen *et al.*, “Towards a new generation of agricultural system data, models and knowledge products: Information and communication technology,” *Agric. Syst.*, vol. 155, pp. 200–212, 2021.

[10] D. Jiménez *et al.*, “From observation to information: Data-driven understanding of on-farm yield variation,” *PLoS One*, vol. 14, no. 4, p. e0215641, 2019.

[11] U. Deichmann, A. Goyal, and D. Mishra, “Will digital technologies transform agriculture in developing countries?” *Agric. Econ.*, vol. 47, no. S1, pp. 21–33, 2016.

[12] E. Arnaud *et al.*, “The ontologies community of practice: A CGIAR initiative for big data in agrifood systems,” *Patterns*, vol. 1, no. 7, p. 100105, 2020.

[13] D. J. Spielman, J. Ekboir, and K. Davis, “The art and science of innovation systems inquiry: Applications to Sub-Saharan African agriculture,” *Technol. Soc.*, vol. 31, no. 4, pp. 399–405, 2021.

[14] J. M. Ribaut and M. Ragot, “Modernising breeding for orphan crops: Tools, methodologies, and beyond,” *Planta*, vol. 250, no. 3, pp. 971–977, 2019.

[15] M. Renkow and D. Byerlee, “The impacts of CGIAR research: A review of recent evidence,” *Food Policy*, vol. 35, no. 5, pp. 391–402, 2010.

[16] M. A. Lantican *et al.*, *Impacts of international wheat improvement research, 1994–2014*. CIMMYT, 2016.

[17] J. M. Alston, P. G. Pardey, and X. Rao, *The payoff to investing in CGIAR research: An estimate of the return on investment*. SoAR Foundation, 2020.

[18] T. Hengl *et al.*, “Soil nutrient maps of Sub-Saharan Africa: Assessment of soil nutrient content at 250 m spatial resolution using machine learning,” *Nutr. Cycl. Agroecosyst.*, vol. 109, no. 1, pp. 77–102, 2017.

[19] M. Halewood *et al.*, “Using genomic sequence information to achieve the crop diversity system management goals,” *Glob. Food Secur.*, vol. 18, pp. 9–14, 2018.

[20] D. Jiménez *et al.*, “A scalable scheme to implement data-driven agriculture for small-scale farmers,” *Glob. Food Secur.*, vol. 30, p. 100559, 2021.

[21] B. M. Campbell *et al.*, “Urgent action to combat climate change and its impacts (SDG 13): Transforming agriculture and food systems,” *Curr. Opin. Environ. Sustain.*, vol. 34, pp. 13–20, 2019.

[22] CGIAR Platform for Big Data in Agriculture, *Annual Report 2020: Digital dynamism for adaptive food systems*. CGIAR, 2021.

[23] D. Jiménez *et al.*, “From observation to information: Data-driven understanding of on-farm yield variation,” *PLoS One*, vol. 15, no. 4, p. e0229881, 2020.

[24] J. Porciello, S. Coggins, G. Otunba-Payne, and E. Mabaya, “A systematic scoping review: How are farmers using digital services in low-and middle-income countries?” *NPJ Digit. Med.*, vol. 4, no. 1, pp. 1–14, 2021.

[25] S. Delerce *et al.*, “Assessing weather-yield relationships in rice at local scale using data mining approaches,” *PLoS One*, vol. 15, no. 5, p. e0233246, 2020.

[26] T. S. Rosenstock, A. Nowak, and E. Girvetz, *The climate-smart agriculture papers: Investigating the business of a productive, resilient and low emission future*. Springer Nature, 2019.

[27] A. M. Loboguerrero *et al.*, “Food and earth systems: Priorities for climate change adaptation and mitigation for agriculture and food systems,” *Sustainability*, vol. 11, no. 5, p. 1372, 2018.

[28] S. Leonelli, R. P. Davey, E. Arnaud, G. Parry, and R. Bastow, “Data management and best practice for plant science,” *Nat. Plants*, vol. 3, no. 6, pp. 1–4, 2017.

[29] M. D. Wilkinson *et al.*, “The FAIR guiding principles for scientific data management and stewardship,” *Sci. Data*, vol. 3, no. 1, pp. 1–9, 2016.

[30] J. Porciello, M. Ivanina, M. Islam, S. Einarson, and H. Hirsch, “Accelerating evidence-informed decision-making for the Sustainable Development Goals using machine learning,” *Nat. Mach. Intell.*, vol. 2, no. 10, pp. 559–565, 2020.

[31] *CGIAR Results Dashboard*, [Online]. Available: https://www.cgiar.org/food-security-impact/results-dashboard/. Accessed: Apr. 6, 2025.

[32] *CGIAR BIG DATA Platform*, [Online]. Available: https://bigdata.cgiar.org/. Accessed: Apr. 7, 2025.

[33] CGIAR System Organization, *CGIAR Performance and Results Management*, 2022.

[34] CGIAR, *Computer Vision for Crop Disease*. [Online]. Available: https://www.kaggle.com/datasets/shadabhussain/cgiar-computer-vision-for-crop-disease/data. Accessed: Apr. 7, 2025.

[35] AWS News Blog, “Earth Science on AWS with new CGIAR and Landsat Public Data Sets,” *Amazon Web Services*, [Online]. Available: https://aws.amazon.com/blogs/aws/earth-science-data-sets-on-aws/. [Accessed: Apr. 8, 2025].

[36] CGIAR, “Digital twin for water management handed over for trial in the Limpopo River Basin,” *CGIAR Newsroom*, [Online]. Available: https://www.cgiar.org/news-events/news/digital-twin-handover-limpopo/. [Accessed: Apr. 8, 2025].

[37] A. M. Basel, K. T. Nguyen, E. Arnaud, and A. C. W. Craparo, “The foundations of big data sharing: A CGIAR international research organization perspective,” *Front. Environ. Sci.*, vol. 11, 2023. [Online]. Available: https://doi.org/10.3389/fenvs.2023.1107393

[38] S. Anna Alex and A. Kanavalli, “Intelligent computational techniques for crops yield prediction and fertilizer management over big data environment,” *Int. J. Innov. Technol. Explor. Eng. (IJITEE)*, vol. 12, pp. 2278–3075, 2019. [Online]. Available: https://doi.org/10.35940/ijitee.L2622.1081219

[39] Nasscom, “Unlocking value from data and AI - The India Opportunity,” *Nasscom Publications*, [Online]. Available: https://nasscom.in/knowledge-center/publications/unlocking-value-data-and-ai-india-opportunity. [Accessed: Apr. 8, 2025].

[40] S. P. Mohanty, D. P. Hughes, and M. Salathé, “Using deep learning for image-based plant disease detection,” *Front. Plant Sci.*, vol. 7, 2016. [Online]. Available: https://doi.org/10.3389/fpls.2016.01419

[41] A. Kamilaris and F. X. Prenafeta-Boldú, “Deep learning in agriculture: A survey,” *Comput. Electron. Agric.*, vol. 147, pp. 70–90, 2018. [Online]. Available: https://doi.org/10.1016/j.compag.2018.02.016

[42] J. A. Delgado, N. M. Short, D. P. Roberts, and B. Vandenberg, “Big data analysis for sustainable agriculture on a geospatial cloud framework,” *Front. Sustain. Food Syst.*, vol. 3, 2019. [Online]. Available: https://doi.org/10.3389/fsufs.2019.00054

[43] E. Beza, P. Reidsma, P. M. Poortvliet, et al., “Review of big data applications in agriculture: Status and challenges,” *Agric. Syst.*, vol. 155, pp. 78–90, 2018.

[44] I. M. Carbonell, “The ethics of big data in agriculture,” *Internet Policy Rev.*, vol. 5, no. 1, 2016.

[45] A. Cravero, S. Pardo, P. Galeas, J. López Fenner, and M. Caniupán, “Data type and data sources for agricultural big data and machine learning,” *Sustainability*, vol. 14, no. 23, p. 16131, 2022. [Online]. Available: https://doi.org/10.3390/SU142316131

[46] S. Wolfert, L. Ge, C. Verdouw, and M. J. Bogaardt, “Big data in smart farming – A review,” *Agric. Syst.*, vol. 153, pp. 69–80, 2017.