frontiers | Frontiers in Sustainable Food Systems

TYPE Original Research
PUBLISHED 08 March 2023
DOI 10.3389/fsufs.2022.1013701

OPEN ACCESS

EDITED BY
P. V. Vara Prasad,
Kansas State University, United States

REVIEWED BY
Ignacio Antonio Ciampitti,
Kansas State University, United States
Prasanna H. Gowda,
Agricultural Research Service (USDA),
United States

*CORRESPONDENCE
Jaron Porciello
✉ jaron.porciello@nd.edu

SPECIALTY SECTION
This article was submitted to Land, Livelihoods and Food Security, a section of the journal Frontiers in Sustainable Food Systems

RECEIVED 07 August 2022
ACCEPTED 28 December 2022
PUBLISHED 08 March 2023

CITATION
Porciello J, Lipper L and Ivanina M (2023) Using machine learning to evaluate 1.2 million studies on small-scale farming and post-production food systems in low- and middle-income countries. *Front. Sustain. Food Syst.* 6:1013701. doi: 10.3389/fsufs.2022.1013701

Using machine learning to evaluate 1.2 million studies on small-scale farming and post-production food systems in low- and middle-income countries

Jaron Porciello[1]*, Leslie Lipper[2] and Maryia Ivanina[3]

[1]Lucy Family Institute for Data and Society, University of Notre Dame, Notre Dame, IN, United States, [2]Department of Agriculture and Resource Economics, University of California, Berkeley, Berkeley, CA, United States, [3]EPAM Systems, Minsk, Belarus

Recent developments have emphasized the need for agrifood systems to move beyond a production-oriented approach to recognize agriculture as part of a broader agrifood system that prioritizes livelihoods, social equity, diets, and climate and environmental outcomes. At the same time, the knowledge base for agriculture is growing exponentially. Using artificial intelligence and machine learning approaches, we reviewed more than 1.2 million publications from the past 20 years to assess the current landscape of agricultural research taking place in low- and middle-income countries. The result is a clearer picture of what research has been conducted on small-scale farming and post-production systems from 2000 to the present, and where persistent evidence gaps exist. We found that the greatest focus of the literature is on economic outcomes, such as productivity, yield, and incomes. There is also some emphasis on identifying and measuring environmental outcomes. However, noticeable data gaps exist for agricultural research focused on nutrition and diet, and gender and inclusivity.

KEYWORDS

agrifood systems, small scale farmers, evidence gaps, low-and middle-income countries, machine-learning, data science

1. Introduction

Decision-making is best informed by an up-to-date and comprehensive review of evidence on a particular topic. The ability to optimize insights from existing agricultural knowledge, and especially research that has explored agriculture's links to impacts on critical issues such as nutrition, climate change, and biodiversity, is key to informing ongoing policy decisions. While data-driven decision-making is widely promoted, especially in the context of complex development issues, the agrifood systems community still lacks critical data and tools that make summarizing data accessible and easily understandable.

Recent work and investment are helping to change this situation, especially for data collection at the country level. Programs like the 50 × 30 initiative[1] are closing the country-level data gap in agriculture and helping measure progress toward the Sustainable Development Goals (SDGs) by building strong nationally representative survey programs. The Food and Agriculture

1 https://www.50x2030.org/

Organization of the United Nations (FAO), likewise, supports the International System for Agricultural Science and Technology[2] to collect the data needed to measure SDGs (Lowder et al., 2021).

Solutions to domain-specific knowledge areas such as agriculture and livelihoods, environment and natural resource management, nutrition and health, and human capital and education are often found within the scientific literature. Expert knowledge, often in the form of scientific papers and other written analysis, is key to developing these solutions, as decisions need to be taken by integrating multiple information sources, incorporating accumulated experience, and weighing uncertainty. At the same time, the amount of available information is increasing exponentially—estimates suggest that human knowledge is doubling every 10–15 years—which makes it increasingly difficult to provide evidence-based interventions while avoiding the risk of confirmation bias or cherry-picking (Bornmann and Mutz, 2015; Bornmann et al., 2021).

Natural language processing (NLP) and machine learning can be highly effective at uncovering insights from large and representative datasets, helping us to make better use of the data in existing scientific publications. NLP is a branch of artificial intelligence that deals with the interpretation and manipulation of human language by computers. Machine learning is the use computers to learn and adapt without following explicit instructions by using algorithms and statistical models to analyze and draw inferences from patterns in data. Both machine learning and NLP approaches are designed to handle classification tasks with speed and accuracy, especially in datasets that lack metadata (Gil et al., 2014).

Recent work has allowed NLP to generate performing information extraction and summarization using relevant data from various sources. Such approaches have transformed how we can approach text-based classification. Pre-trained transform models such as Bidirectional Encoder Representations from Transformers (BERT), SciBERT and named-entity recognition with BERT are highly adept at capturing the context-dependent meaning of words even before additional training for other tasks that require expert input in the form of training data (Devlin et al., 2018; Beltagy et al., 2019; Luoma and Pyysalo, 2020). This can save significant time and money while delivering new insights.

Allowing for better understanding of the degree to which data and analyses are capturing systematic interactions is one of the most important features of ML and NLP approaches. This study reports on the use of machine learning to process and analyze 1.2 million summaries of past publications from a representative dataset of agricultural research focused on low- and middle-income countries. Its primary aim is the summarization of data to inform a series of open-ended questions that are difficult to answer because the data are scattered across millions of individual studies. These questions include:

- Who are the user groups included within studies?
- What are the most-studied interventions and outcomes by researchers?
- What is the research output across low- and middle-income countries?
- How much of the research is targeted at solutions for small-scale farmers and other agricultural actors vs. laboratory studies or other controlled environments?

2. Methods

2.1. Approach: Mapping 1.2 million studies in agriculture

Recent work in measuring the output of overall scientific growth across certain fields has primarily focused on the comprehensiveness of large databases, such as Dimensions, Scopus, Web of Science and Microsoft Research (Bornmann et al., 2021). We targeted CABI's CAB Abstracts in part because of CABI's mission to identify and aggregate research from low- and middle-income countries, making it among the best databases in the world for our purposes. Similar analyses to ours, focused on agriculture and regional specific agricultural components, such as rice research in low- and middle-income countries, indicates the suitability of CAB Abstracts for such analyses (Rafols et al., 2020; Amarante et al., 2021).

We obtained 1.3 million citation records from data partner CAB Abstracts using the search strategy: (de: "climate") OR (de:biodiversity) OR (de: farm*) OR (de: agricultur*) OR (de:crop) OR [de:("food policy" or "agricultural sector" or "food security" or sustainabilit*" or "environment" or "nutrition" or "product*" or "yield" OR "hunger" or "agricultural policy" or "development aid")] yr:[2000 TO 2021].

We reduced 1.3 to 1.2 million by removing duplicate citations to produce our final dataset for analysis. No further reduction, using more specific inclusion criteria, was initiated was this effort. Artificial intelligence-assisted techniques were used to summarize abstracts by the categories are shown in Figure 1. NLP for text extraction and large-scale machine-learning language models were used to model the data for tasks associated with the identification of study user population, interventions, outcomes, geography, and crop type, among other elements. *A priori* determination of the categories was done in consultation with the expert-assembled Commission on Sustainable Agriculture Intensification (CoSAI). The prioritization on some specific tasks by the CoSAI groups enabled a more focused approach for the machine-learning.

2.2. Machine-learning to identify agricultural interventions, outcomes, and study design types

Identifying interventions, outcomes, study design types and more is normally undertaken during an evaluation of the evidence on a specific topic, such as part of an impact assessment or a systematic review, by domain experts looking through thousands of underlying original research papers. A well-trained machine-model can accelerate the labeling of many of these tasks. This study further contributes to exploring the role of computation to accelerate evidence and impact synthesis work in agriculture and climate change scientific publication datasets (Porciello et al., 2020; Callaghan et al., 2021).

Training data assembled from collaborative coding from previous exercises, including more than 2,500 high-quality papers from across

2 https://agris.fao.org/agris-search/index.do

10.3389/fsufs.2022.1013701

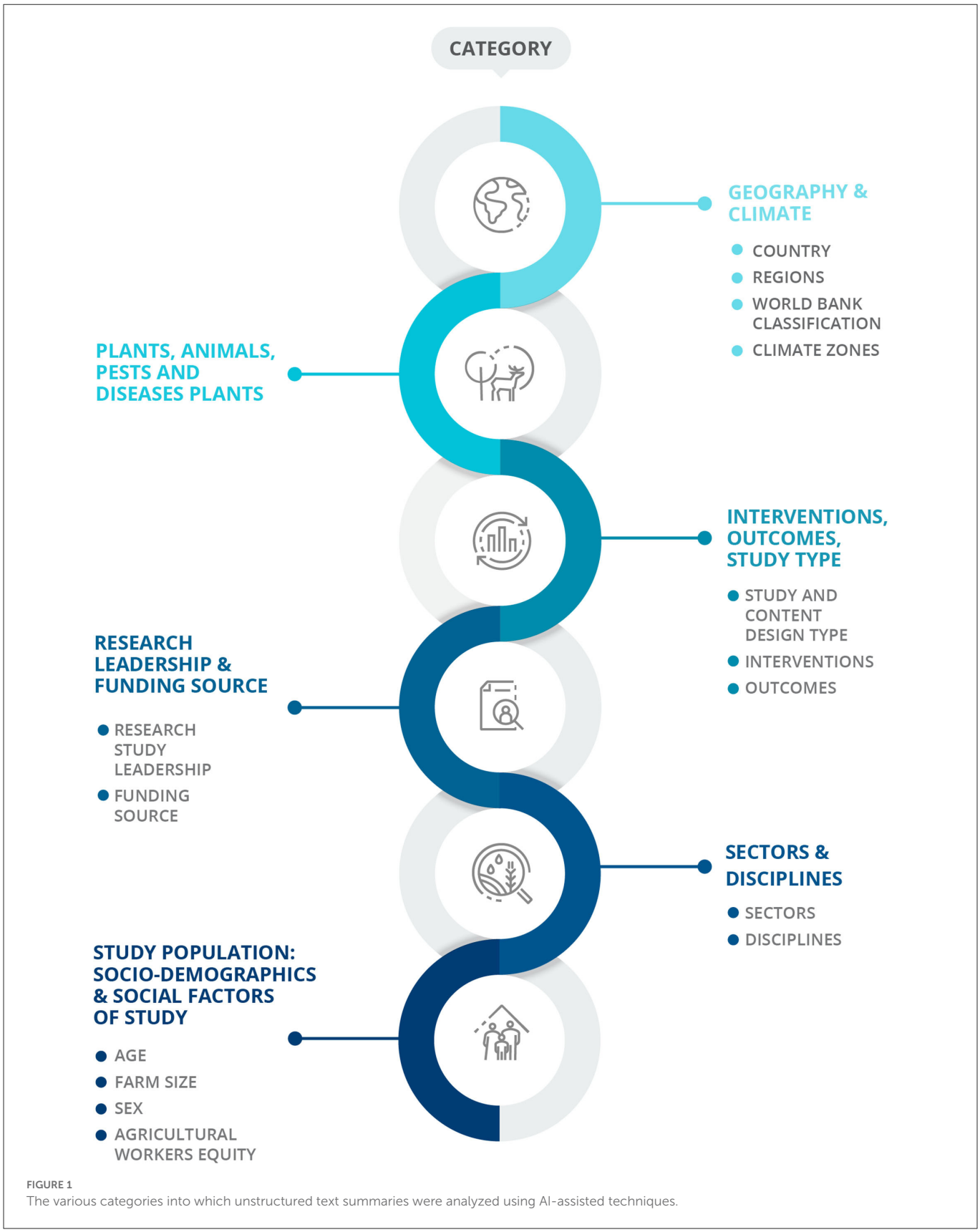

FIGURE 1
The various categories into which unstructured text summaries were analyzed using AI-assisted techniques.

the Ceres2030: Sustainable Solutions to End Hunger project, was used to enhance an artificial intelligence pipeline that supports classification and information extraction tasks (identified in Figure 1) for agriculture and related areas in international development (Acevedo et al., 2020; Baltenweck et al., 2020; Bizikova et al., 2020; Liverpool-Tasie et al., 2020; Nature, 2020a; Piñeiro et al.,

2020; Porciello et al., 2020; Ricciardi et al., 2020; Stathers et al., 2020). In addition, the underlying models been continuously trained on tasks supporting diverse development literature as a result of other partnerships, including use in new domains such as water, hygiene, and sanitation, digital agriculture, and development and humanitarian assistance, and all of which required the identification of outcomes and interventions (Garbaro et al., 2020; Jardine, 2021; Porciello and Ivanina, 2021; Porciello et al., 2022).

Unlike health and medical sector, which maintains an International Classification of Health Interventions[3] through the World Health Organization (WHO), agrifood systems lack a similar standardized taxonomy of interventions. One most powerful structured collections of agricultural concepts, terms, definitions, and relationships—FAO's AGROVOC—defines an intervention simply as a "controlled price" (AGROVOC: AGROVOC Multilingual Thesaurus, n.d.). This definition is a sparse interpretation of the range of potential activities that can be used to support policies and programs to improve agrifood systems. Other organizations, including the OECD recommend expanding the interpretation beyond price interventions to include more agricultural, humanitarian and development sector activities (OECD, 2019).

We developed a proxy to inform how to approach an unstructured text corpus to identify literature that describes interventions but importantly, without necessarily using the term intervention. Training of the model for interventions included searching articles and summary data for synonyms of intervention and enhanced using Word2vec. Word2vec was chosen because of its more than decade-long history of performing NLP tasks to find syntactic and semantic similarities of words. Word2vec's shallow language model is appropriate for small and relatively heterogeneous datasets such as ours, and it has low computational costs, taking <1 day to learn high-quality word vectors from a 1.6-billion-word dataset. Similar models, such as Global Vectors (GloVe), could be used in conjunction with or instead of Word2vec with similar results, although training time might slightly increase (Sharma et al., 2017, p. 2). Using pre-trained Google News and Wikipedia Word2vec models, similar concepts to interventions for the agricultural domain were identified, including "program or programme," "strategy," and "government initiative" (Porciello et al., 2020). Next, to surface all potential and specific interventions, we incorporated a semi-unsupervised model-based approach *via* coreference resolution models to support NLP tasks by linking noun phrases with entities in the text. A training dataset that broadly represented how interventions were described in the literature as technological, socioeconomic, and ecosystem service interventions was applied. More description about these categories is provided in the results section. Next, we sought to surface and label how more specific interventions, such as drip-irrigation or solar-irrigation, could be represented and labeled as part of a narrow cluster of interventions, such as "irrigation" interventions.

Next, the model was trained to identify outcomes. Unlike interventions, there are standardized definitions for outcomes (Table 1 in Results). The model was trained to detect when an outcome was mentioned and had a relationship to narrow classes from the intervention. A single example consists of a sentence, an intervention from the ontology and/or plant, animal product from the AGROVOC dictionary, and an outcome from the sentence. When the model detects an outcome is connected with a particular intervention in the context of a sentence, it labels the citation with the appropriate outcome based on the general definition.

Both rule-based and transformer-based models were used for this task with similar results. A rule-based support-vector machines (SVM) was used in a semi-unsupervised approach to organize studies according to NLP-derived intervention, outcome, and study design type taxonomies. An SVM–*k* nearest neighbors–stochastic gradient boosting approach was used for classifying specific interventions, where all the supporting content (in this case, summary data) is examined in a vector space. The SVM is a supervised classification algorithm that learns by example to discriminate among two or more given classes of data, and they work well with high-dimensional data especially for smaller datasets. In addition, BERT-based models are designed for sentence level and token-level tasks and are useful for identifying relationships in small pieces of text. BERT models including base BERT, Roberta, Albert, SciBERT, and DistillBERT were tested. DistilBERT Named Entity Recognition (NER) uses the BERT architecture but performs knowledge distillation during the pre-training, allowing for lighter, faster and cheaper transformer model, and reduces the size of a BERT model by 40%. Due to the size of the labeled dataset, models were trained by freezing all layers (which is responsible for encoding the text) except the last two layers (where classification occurs).

Finally, study design types also lack common definitions. These were labeled using expert data and the transformer model SciBERT, which has been pre-trained on scientific articles (Beltagy et al., 2019). For other tasks, text extraction models, including pre-trained spaCy, specialized dictionaries, and ontologies of AGROVOC and the National Agricultural Library Thesaurus, were used to identify and label geography, plants, animals, diseases, research leadership and funding, and study populations.

3. Results

One of the most useful ways to report the findings of this analysis is through an evidence gap map (Figure 2), a visual and interactive tool that provides an overview of all evidence collected on a particular issue (Vincent et al., 2022). Evidence gap maps enable policy makers and practitioners to review findings, explore the quality of the existing evidence, and make evidence-based decisions in international development policy and practice. They also identify key "gaps" where little or no research has been published (Snilstveit et al., 2016).

The key components of an evidence gap map are interventions and outcomes. The evidence gap map identifies the most frequently studied interventions as determined by a threshold of at least 10,000 articles and categorizes them into one of three broad categories of agricultural research (socioeconomic, technological, and ecosystem services). Importantly, an evidence gap map does not prioritize or claim there is a single intervention that is "a silver-bullet" to support agricultural development outcomes. Rather, the intention is to surface volumes of research and where more, and less, emphasis has been placed.

3 https://www.who.int/standards/classifications/international-classification-of-health-interventions

10.3389/fsufs.2022.1013701

TABLE 1 Outcomes, descriptions and definitions.

Outcome class	Description	Specific outcomes	Definition
Economic growth	Growth across all agriculture or food systems sectors and subsectors that improve the lives of farmers and food systems actors and their families through increases in income, productivity, employment, and practice change.	Income amount	Change in income
		Income diversity	Change in sources of income
		Productivity	Change in on-farm crop, labor or livestock productivity or value-chain productivity
		Yield	Change in yield from crop, livestock or foraging
		Adoption	Change in a user's adoption of management or technology related to other agricultural outcomes
		Market efficiency	Change in decision-making based on available, relevant market information
Healthy people	Ensuring reliable access to a sufficient quantity of affordable, nutritious food and food consumed is represented by different food groups	Dietary diversity	Change in dietary adequacy, including nutrient intake, nutrient adequacy index, and food-based diet quality index
		Food access	Change in an individuals' or households' ability to access food
		Food availability	Change in availability of food
		Malnutrition	Change in malnutrition status
		Wasting and stunting	
		Nutritious food availability	Change in availability or access to nutritious food
Healthy planet	The process of incentiving practices that emphasize environmental and planetary health	Environmental Sustainability	Change in sustainability of natural resource management such as water, forest or soil management e.g., reduced soil erosion, reduced tree cover loss or increased tree cover,
		Climate mitigation	Change in greenhouse gas emissions
		Change in capacity to adapt to the impacts of climate change	Adaptation and behavior change that respond specifically to impacts of climate change
		Biodiversity	Change in biological resources at genetic, species or ecosystem level (on farm or off-farm)
Gender & Inclusivity	The process of improving the terms of participation in society, particularly for people who are disadvantaged, through enhancing opportunities, access to resources, voice and respect for rights. This is measured through resulting from the support and inclusive design of all people, but in particular traditionally marginalized groups such as women and people with disabilities, as well as through increased decision-making.	Increased Knowledge	Change in knowledge about agriculture or food systems related content
		Women's empowerment	Change in women's ability to influence and make decisions independently
		Women's access to resources	Change in women's access to resources (e.g., credit, or inputs)
		Social inclusion	Change in obstacles that limit agency and decision-making capacity
		Social learning	Change in knowledge and practices through group and community engagement

Technological interventions constitute the use of practices and technologies (both direct and indirect) to support agricultural production and food systems (Acevedo et al., 2020; FAO, 2022a,b). Indirect uses include underlying technology such as biotechnology to improve seeds, whereas direct would be use of irrigation, mechanization, and inputs such as fertilizer. Socioeconomic interventions include market and finance interventions that contribute to accessing markets, credit or other financial products

10.3389/fsufs.2022.1013701

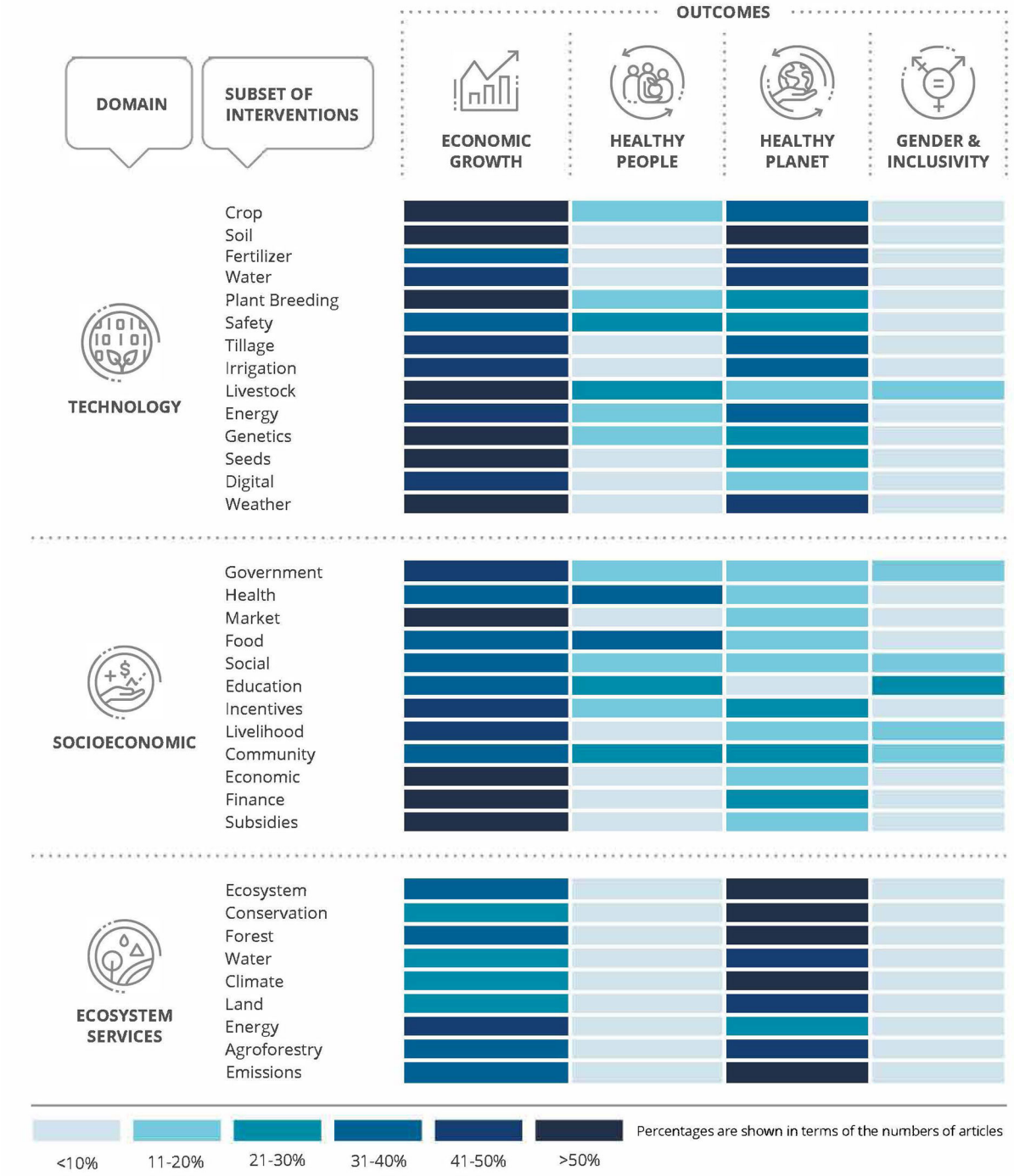

FIGURE 2
An intervention and outcome evidence gap map identifying the most frequently studied interventions and associated outcomes.

or investments in value chain development, as well as interventions that increase knowledge or awareness, transfer skills, and build capacities such as education (Liverpool-Tasie et al., 2020). This category also includes policy and government interventions, such as government, funder, or other organizational programs and policies to support farmers and agri-food system actors through incentives, or direct support, and includes interventions to improve inclusion of women and other marginalized groups (Barrett et al., 2020). Ecosystem services interventions focus on improving ecosystem services with regulating and supporting functions such as clean air, nutrient cycling, pollination, erosion control, carbon storage and more (Piñeiro et al., 2020). Additional analysis can be conducted to further sub-divide the categories for additional, discrete analysis.

The evidence gap map in Figure 2 shows the frequency of interventions per outcome, expressed as a percentage across the literature. For instance, over 50% of plant breeding interventions

in the literature are associated with outcomes related to economic growth, whereas 11–20% are associated with nutrition outcomes, 21–30% with environmental outcomes, and <10% with women's empowerment and inclusion. Table 1 provides outcome descriptions and definitions.

The highest reported outcome is economic, such as productivity, yield, and incomes, in the literature. This reflects the fact that agricultural research and innovation literature has been largely focused on improving productivity of a small number of crops rather than focusing on other important aspects of crop research, such as dietary diversity (Serraj and Pingali, 2018). Some emphasis has been placed on on identifying and measuring environmental outcomes, including water use and health, across many of the intervention categories, especially those focused on ecosystem services.

Where the data gaps are more noticeable are regarding agricultural research focused on nutrition and diet, and women's empowerment and other inclusivity outcomes mentioned in the literature, such as increased knowledge obtained through training and education programs. For the latter, the gaps are widespread across all intervention categories.

Figure 3 provides a regional level overview of the publication trends focused on specific crops mentioned in title and abstract data. Table 2 provides a breakdown of the specific crops included in each category and their inclusion was determined *a priori* through consultation (as referenced in the introduction). Generalized terms such as cover crops, livestock feed crops, container plants, bee plants, beverage crops, and oils were excluded from the mapping because it was unclear from the summary what crops they referred to, and because they totaled fewer than 25,000 mentions. Each study was labeled with multiple labels, meaning that more than one relevant label could be applied. For instance, if a study focused on wheat, maize, and rice in Vietnam and Thailand, then the study would be counted as "1" in all subsequent categories.

China, Brazil, and India lead the way in publishing research outputs, but different countries and regions come into focus depending on the target crops, as highlighted by the maps in Figure 3. Perhaps as expected, countries that are home to a major international research center, such as the International Maize and Wheat Improvement Center in Mexico or the International Rice Research Institute in the Philippines, have a higher prevalence of research related to the specific crops being studied. Other grains that are important for food security, such as millet and sorghum, have a smaller cumulative total of around 10,000 articles.

The findings on study design types by research categories (Figure 4) show research activities that report on non-human experiments, such as field trials, laboratory, and simulation studies. A total of six labels were created to identify study population types: field study, experimental study, simulation/modeling study, narrative/review study, laboratory study, and observational studies. Each citation received only one study type. The categories along the Y axis are CABI Codes. CABI Codes is an index of 23 major subject areas related to the area of the citation, each with their own set of sub-codes (https://www.cabdirect.org/help/about-cabicodes.html). CABI codes are added by the vendor when an article is included in CAB Abstract database. This provides an existing, manually curated index of research topics that does not rely on machine-learning. The subject area of agricultural economics has the largest number of observational studies, followed by field crops, meteorology and climate, and water resources.

Finally, a multi-label approach to capture information about the study population communities, including when studies mention descriptions about age, sex, affiliation with indigenous communities or other, and agricultural workers, including farmers. Despite a generalized, multi-labeling approach, the data collection and reporting on user populations is very weak. Only about 25% of studies reported any information about a population of study. Though there may be widespread acknowledgment that women, farming communities and others in the agricultural workforce face significant challenges, there is a risk they will be undermined in these types of global assessments by weak data collection practices regarding demographics and other specific descriptions and/or underreporting in the literature (Teeken et al., 2018).

4. Discussion

4.1. Prioritizing research gaps

The way we think about agriculture is currently undergoing a major shift away from a focus on production and toward a broader understanding that puts agriculture in the larger context of an agrifood system with complex interactions between food production, processing, consumption, nutrition, social change, and climate change (Barrett et al., 2020; Lipper et al., 2020). This shift implies a need to rethink the role of agricultural research and development efforts, and push for innovations that go beyond productivity. There is a corresponding urgency to identify priority investments (Reardon et al., 2019; Laborde et al., 2020). To do so, however, we must have an adequate and accessible evidence base for understanding agricultural innovations and their potential in the context of a transformation.

Integrated approaches across interventions are more effective in achieving gains across the entire food system. Therefore, the relative scarcity of research emphasizing diet, nutrition, and women's empowerment relative to the long-standing priorities of productivity and yield in agricultural research should not necessarily lead us to conclude that some areas of research only need to "catch up" to others. Simply focusing on expanding the literature in one of the relatively under-researched areas will not address the yawning gap of evidence on the interactions that occur across various outcomes.

However, not all areas where there is a dearth of research can be treated equally or with the same urgency. There are many areas of research where we have gaps in the evidence on the impact of interventions on specific outcomes (Figure 3) but identifying where significant trade-offs between outcomes can arise from interventions is key in the context of analyzing the food system and its interactions (Fuso Nerini et al., 2018; Kroll et al., 2019). For example, the lack of research on fruits, vegetables, and more nutritious grains such as millet and sorghum (Figure 3), as well as accompanying post-harvest storage to ensure safety and reduce loss, is a gap in our understanding relevant not only to improving diets and addressing micro-nutrient deficiencies, but to gender and inclusivity, given the high rates of female participation in horticultural and post-harvest activities (Kennedy et al., 2017; Nordhagen, 2021).

There is too little data being reported in agrifood systems literature about study populations, and the impacts and uptake of innovations across small-scale farmers and their communities. Better identification of relevant characteristics of the people and

10.3389/fsufs.2022.1013701

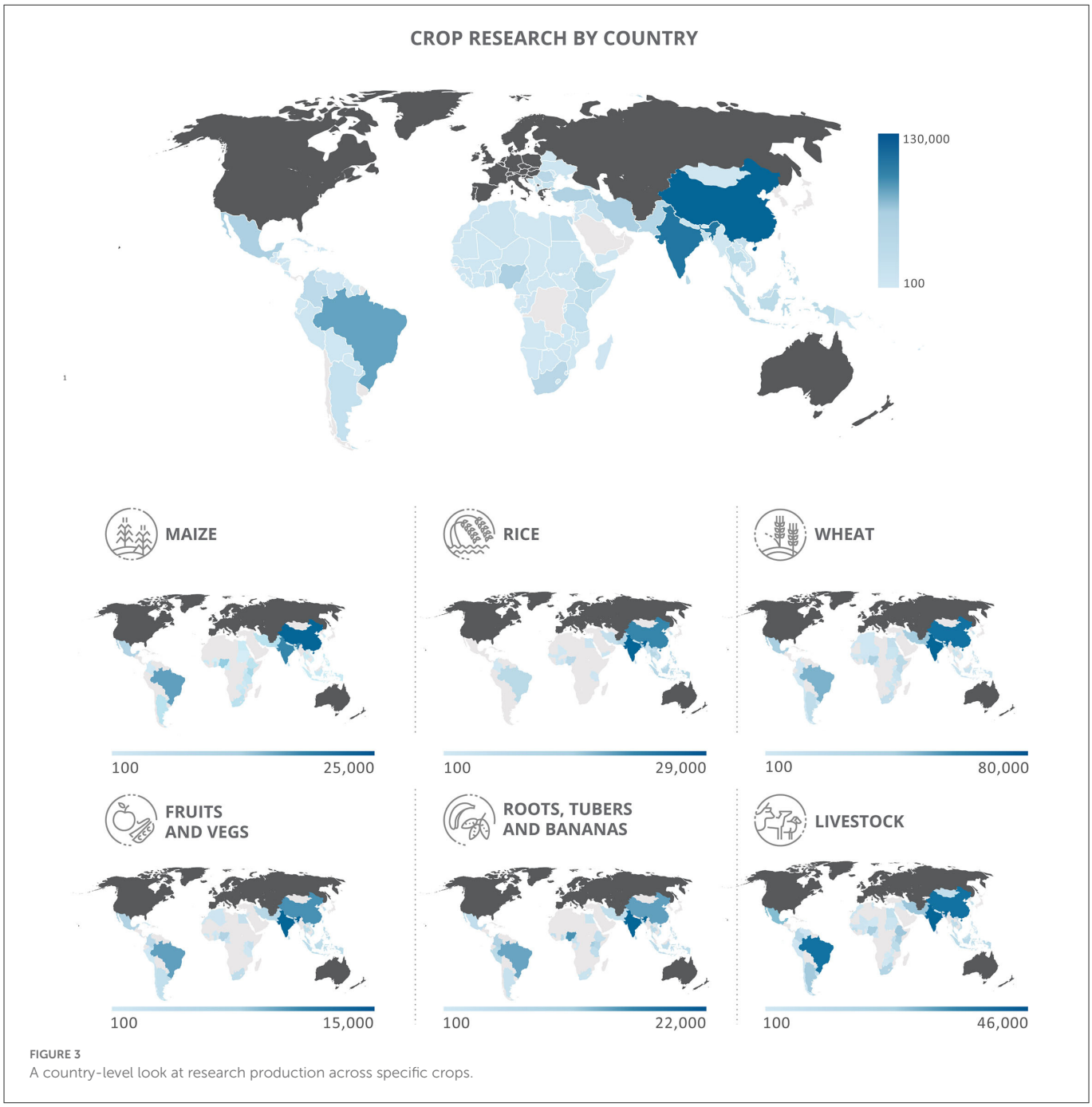

FIGURE 3
A country-level look at research production across specific crops.

TABLE 2 Crops per category.

Maize	Rice	Wheat	Fruits and vegetables	Roots, tubers, and bananas	Livestock
Maize (all variations, e.g., flint maize)FAO Cereal/Maize Sweet Corn *Zea mays*	Rice (all variations, e.g., wetland rice) Cereal/Rice *Oyrza* (all variations)	Wheat (all variations, e.g., winter wheat) Cereal/Wheat *Triticum* (all variations)	More than 100 individual fruit and vegetable crops were searched. The full list is included in the appendix.	Banana (all variations, including cooking banana) Cassava Yuca Yam Sweet potato Potato (all variations) Turnip Taro Rutabaga	Cattle Swine/Pigs Poultry (e.g., chickens) Domesticated Buffalo Sheep Goat Yak Zebu Horse Camel

10.3389/fsufs.2022.1013701

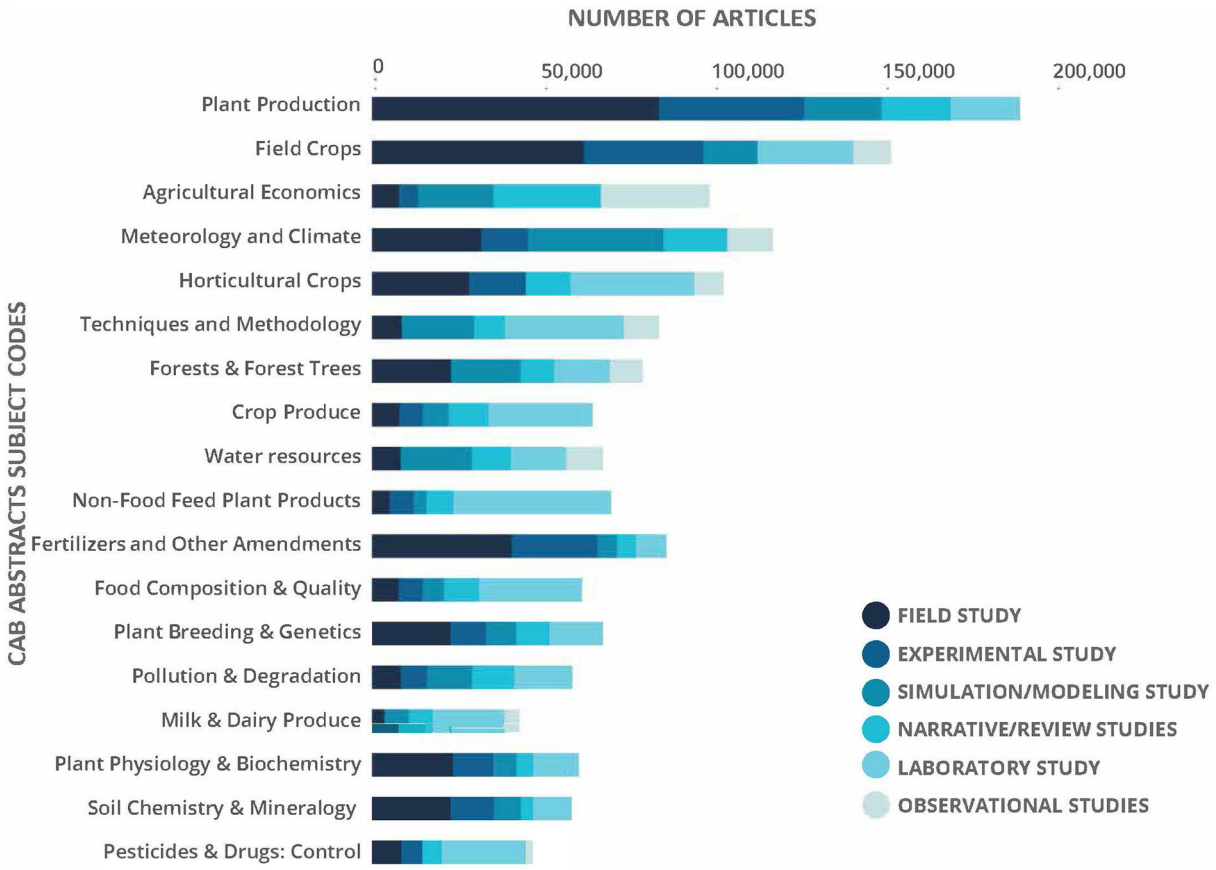

FIGURE 4
Study design and research topic areas.

communities involved in agricultural activities is essential to understanding the outcomes of interventions and the interactions that arise across different outcomes. Part of the issue is the extremely ambiguous descriptions of farmers and agricultural workers. These descriptions rarely include contextual clues about the type or size of farm they work on. Similar gaps were reported in another evidence analysis, which found that only 2–3% of studies across a portfolio of scoping reviews reported on the conditions and interventions of farmers in low- and middle-income countries (Nature, 2020b). Given that the emphasis of SDG 2 focuses on the conditions of poor farmers in low- and middle-income countries, high-impact, applied research to identify and report on successful programs across all outcomes in low- and middle-income countries is urgent.

Equally important for future of research is the capture of social equity and sociodemographic details that could underscore how barriers are systematic for some communities and not for others. Socioeconomic status, race, class, and gender can create interdependent systems of discrimination that reinforce the exclusion of some groups—particularly, but not only, women—from the benefits of certain programs and innovations. The ability to look at social factors as a system is essential to avoid tendencies to overgeneralize and assign certain characteristics to entire groups, such as elderly, youth or women (Sumberg and Hunt, 2019). A recent scoping review focused on digital agriculture identified that fewer than 30% of all studies reported socioeconomic and demographic data (Porciello et al., 2022). This shortcoming is of particular concern in the context of assessing multiple and potentially interacting outcomes from agricultural research. In a 2020 review of literature on factors influencing the adoption of sustainable agriculture, farmer characteristics—including asset levels, experience and risk preferences—were a key factor in explaining farmers' behavior, particularly where there were potential trade-offs between environmental and economic outcomes (Piñeiro et al., 2020). In discussing the reasons for the lack of progress in transforming small-scale agriculture, Woodhill et al. (2020) cite a lack of understanding of the diversity of characteristics and contexts of small-scale farmers is reported as a major factor. Here, again, the issue of multiple and potentially competing outcomes from agricultural change was important. As we look toward the future of research prioritization, equity outcomes need to become more pronounced (Davis et al., 2022; Laderchi et al., 2022).

In this respect agricultural and food systems studies fall well behind other disciplines, such as medicine and health. Coordinating bodies in health and medicine, such as Cochrane draft guidance and minimum standards for synthesis conduct, develop methodologies and training capacity, and commission and publish high-quality reviews. The absence of such coordination and synthesis in agricultural sciences has contributed to the evidence gaps mapped in this study. These gaps should no longer be ignored. Simply focusing on expanding the literature in one of the relatively under-researched areas will not address the yawning gap of evidence on the interactions that occur across various outcomes with interventions into any one piece of the system. Assessing progress on the myriad of impacts of what, where, when and why are often commissioned individually by

donors with little opportunities for coordination. Moreover, despite the existence of gaps in data collection, such as the absence of sociodemographic data about farmers that we have highlighted above, the lack of an organizing body means that there currently exists no group to champion for long-term change in research practices, methodologies for synthesis conduct, and data collection.

The aim of this study is to uncover relevant insights across primary studies and used only summary title, abstract and other available metadata. However, what authors choose to emphasize in the title, abstract and other summary data is influenced by various editorial decisions between themselves and the journals publishing the materials. For instance, some journals may ask authors to refrain from mentioning too many details in the abstract, such as the user population of study, countries of focus, or specific plants. Access to the full text is needed to evaluate the claims made in the summary data, such as whether the interventions and outcomes recognized in the abstract are substantially supported with high-quality data in the study (Garbaro et al., 2020; Porciello and Ivanina, 2021; Porciello et al., 2022).

Evidence from the Covid-19 Open Research Dataset (CORD-19) demonstrates the value obtaining copyright and permissions clearance from commercial publishers to support text mining and NLP research on scientific papers. CORD-19 is an open access collection of more than one million scientific papers published between March 13, 2020–June 2, 2022 related to coronavirus with the full-text available for text-mining of nearly 370 K papers (Wang et al., 2020). The opportunity to read and rapidly discover insights from primary scientific research during Covid-19 is useful to all scientists and policy-makers, and CORD-19 computational tools for text-mining delivered additional, rapid insight on internationally collaborative work, and the contributions of funders, countries, institutions, and fields throughout the pandemic (Wagner et al., 2022).

A demand-driven approach to obtaining access to critical research is relevant for the agrifood community considering the current, global food crisis (Laborde and Glover, 2022). For instance, recent research of over 1.2 million children in 44 low-and middle-income countries suggests that experiencing the current crisis of food inflation increases both the risks of stunting and wasting in children under 5, including infants, as well as decreased diet quality for older children (Headey and Ruel, 2022). Greater visibility of critical agrifood research, complemented with computation tools to extract and classify "what works" and major gaps in the evidence base is urgently needed to help policymakers implement relevant policies that may mitigate disastrous consequences, especially for vulnerable populations.

5. Conclusion

Using machine-learning to analyze and quantify data gaps in agricultural research allows for greater understanding of the degree to which data and analyses are capturing systematic interactions. These approaches are current unavailable through other means, including expensive subscription databases. This approach to define important concepts like interventions can be especially useful in disciplines like agriculture and food systems, where well-coordinated, standardized evidence synthesis is lacking. Machine learning approaches enable us to perform close readings of a large, representative dataset and provide descriptive details that can be used to inform research agendas and prioritization. Studies like this are necessarily limited in the observations and analysis based on what we can glean from summary data, given that full-text analysis of more than one million papers requires extensive processing time. In this study, the capture mentions of interventions and their outcomes presents a useful "birds-eye view" for future interrogations of the data, but both access and additional evaluation of the underlying studies is needed to support whether the identified interventions and outcomes are consistent with the findings of each study. Still, such approaches allow opportunities to track research over time to create a global monitoring and evaluation framework.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: CAB Direct, GitHub.

Author contributions

JP oversaw study design, data analysis and methodology, and manuscript development. MI provided coding and computation support. LL contributed to data review and manuscript development. All authors contributed to the article and approved the submitted version.

Funding

This research was commissioned by the Commission for Sustainable Agriculture Intensification (CoSAI) and supported by funders contributing to the CGIAR Trust Fund.

Acknowledgments

We thank Mary O'Connor, Cristina Ashby, Martin Parr, and Andy Robinson of CABI for early review of previous versions of this manuscript and for providing access to the dataset. Thank you to Julia Compton for engaging and thoughtful questions and relevant feedback throughout the study design and analysis. We acknowledge and appreciate the role of COSAI advisory board and their feedback and support on early analysis.

Conflict of interest

MI was employed by EPAM Systems, Minsk, Belarus.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Acevedo, M., Pixley, K., Zinyengere, N., Meng, S., Tufan, H., Cichy, K., et al. (2020). A scoping review of adoption of climate-resilient crops by small-scale producers in low- and middle-income countries. *Nat. Plants* 6, 1231–1241. doi: 10.1038/s41477-020-00783-z

AGROVOC: AGROVOC Multilingual Thesaurus (n.d.). Available online at: https://www.fao.org/agrovoc/ (accessed July 31, 2022).

Amarante, V., Burger, R., Chelwa, G., Cockburn, J., Kassouf, A., McKay, A., et al. (2021).Underrepresentation of developing country researchers in development research. *Appl. Econ. Lett.* 1–6. doi: 10.1080/13504851.2021.1965528

Baltenweck, I., Cherney, D., Duncan, A., Eldermire, E., Lwoga, E. T., Labarta, R., et al. (2020). A scoping review of feed interventions and livelihoods of small-scale livestock keepers. *Nat. Plants* 6, 1242–1249. doi: 10.1038/s41477-020-00786-w

Barrett, C. B., Benton, T. G., Cooper, K. A., Fanzo, J., Gandhi, R., Herrero, M., et al. (2020). Bundling innovations to transform agri-food systems. *Nat. Sustain.* 12, 974–976. doi: 10.1038/s41893-020-00661-8

Beltagy, I., Lo, K., and Cohan, A. (2019). *SciBERT: A Pretrained Language Model for Scientific Text*. ArXiv:1903, 10676. [Cs]. http://arxiv.org/abs/1903.10676

Bizikova, L., Nkonya, E., Minah, M., Hanisch, M., Turaga, R. M. R., Speranza, C. I., et al. (2020). A scoping review of the contributions of farmers' organizations to smallholder agriculture. *Nat. Food* 1, 620–630. doi: 10.1038/s43016-020-00164-x

Bornmann, L., Haunschild, R., and Mutz, R. (2021). Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Hum. Soc. Sci. Commun.* 8, 1. doi: 10.1057/s41599-021-00903-w

Bornmann, L., and Mutz, R. (2015). Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inf. Sci. Technol.* 66, 2215–2222. doi: 10.1002/asi.23329

Callaghan, M., Schleussner, C.-F., Nath, S., Lejeune, Q., Knutson, T. R., Reichstein, M., et al. (2021). Machine-learning-based evidence and attribution mapping of 100, 000 climate impact studies. *Nat. Climate Change* 11, 1–7. doi: 10.1038/s41558-021-01168-6

Davis, B., Lipper, L., and Winters, P. (2022). Do not transform food systems on the backs of the rural poor. *Food Sec.* 14, 729–740. doi: 10.1007/s12571-021-01214-3

Devlin, J., Chang, M.-W., and Lee, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. ArXiv:1810, 04805. [Cs]. http://arxiv.org/abs/1810.04805

FAO. (2022a). *AGROVOC Multilingual Thesaurus*. Available online at: https://agrovoc.fao.org/browse/agrovoc/en/ (accessed July 31, 2022).

FAO. (2022b). *The State of Food and Agriculture 2022: Leveraging Agricultural Automation for Transforming Agrifood Systems*. Rome: FAO.

Fuso Nerini, F., Tomei, J., To, L. S., Bisaga, I., Parikh, P., Black, M., et al. (2018). Mapping synergies and trade-offs between energy and the sustainable development goals. *Nat. Energy* 3, 10–15. doi: 10.1038/s41560-017-0036-5

Garbaro, A., Carneiro, B., Charpenteir, A., Cerulli, G., Porciello, J., Ivanina, M., et al. (2020). *Leveraging Artificial Intelligence and Big Data for IFAD* 2, 0. Available online at: https://www.ifad.org/documents/38714170/42863493/athena.pdf/68906a2a-afde-7f9c-02b7-468227706111b (accessed July 31, 2022).

Gil, Y., Greaves, M., and Hendler, J. (2014). Amplify scientific discovery with artificial intelligence. *Science* 346, 171–172. doi: 10.1126/science.1259439

Headey, D. D., and Ruel, M. T. (2022). *Food Inflation and Child Undernutrition in Low and Middle Income Countries*. Washington, DC: International Food Policy Research Institute (IFPRI). doi: 10.2499/p15738coll2.136457

Jardine, B. (2021). *Case Study from USAID's Center for Water Security, Sanitation, and Hygiene.* Research Technical Assistance Center (December 29 2021). Available online at: https://www.rtachesn.org/resources/machine-learning-for-programmatic-synthesis/ (accessed July 31 2022).

Kennedy, G., Stoian, D., Hunter, D., Kikulwe, E., Termote, C., Burlingame, B., et al. (2017). "Food biodiversity for healthy, diverse diets," in *Bioversity International. Mainstreaming Agrobiodiversity in Sustainable Food Systems: Scientific Foundations for an Agrobiodiversity Index* (Rome, Italy: Bioversity International). p. 23–52.

Kroll, C., Warchold, A., and Pradhan, P. (2019). Sustainable development goals (SDGs): are we successful in turning trade-offs into synergies? *Palgrave Commun.* 5, 1–11. doi: 10.1057/s41599-019-0335-5

Laborde, D., and Glover, J. (2022). *No End in Sight Yet for the Global Food Price Crisis*. IFPRI: International Food Policy Research Institute. Available online at: https://www.ifpri.org/blog/no-end-sight-yet-global-food-price-crisis (accessed December 21, 2022).

Laborde, D., Porciello, J., Smaller, C., and Murphy, S. (2020). *Ceres2030: Sustainable Solutions to End Hunger Summary Report* [Report]. Ceres2030. Available online at: https://ecommons.cornell.edu/handle/1813/72799 (accessed July 31, 2022).

Laderchi, C., Kanbur, R., and Winters, P. (2022). *Inclusion as a Key Goal in the Transformation of Food Systems*. Food Systems Economic Commission Blog. Available online at: https://www.foodsystemeconomics.org/post/inclusion-as-a-key-goal-in-the-transformation-of-food-systems (accessed march 3, 2022).

Lipper, L., DeFries, R., and Bizikova, L. (2020). Shedding light on the evidence blind spots confounding the multiple objectives of SDG 2. *Nat. Plants* 6, 1203–1210. doi: 10.1038/s41477-020-00792-y

Liverpool-Tasie, L. S. O., Wineman, A., Young, S., Tambo, J., Vargas, C., Reardon, T., et al. (2020). A scoping review of market links between value chain actors and small-scale producers in developing regions. *Nat. Sustain.* 3, 10. doi: 10.1038/s41893-020-00621-2

Lowder, S. K., Sánchez, M. V., and Bertini, R. (2021). Which farms feed the world and has farmland become more concentrated? *World Develop.* 142, 105455. doi: 10.1016/j.worlddev.2021.105455

Luoma, J., and Pyysalo, S. (2020). *Exploring Cross-sentence Contexts for Named Entity Recognition with BERT*. (arXiv:2006, 01563). doi: 10.18653/v1/2020.coling-main.78

Nature (2020a). *Nature Portfolio: Sustainable Solutions to End Hunger*. Available online at: https://www.nature.com/collections/dhiggjeagd (accessed July 31, 2022).

Nature (2020b). Ending hunger: science must stop neglecting smallholder farmers. *Nature* 586, 336. doi: 10.1038/d41586-020-02849-6

Nordhagen, S. (2021). *Gender Equity and Reduction of Post-Harvest Losses in Agricultural Value Chains*. Global Alliance for Improved Nutrition Working Paper #20. Geneva, Switzerland. doi: 10.36072/wp.20

OECD. (2019). *Digital Opportunities for Better Agricultural Policies*. Paris: OECD. doi: 10.1787/571a0812-en

Piñeiro, V., Arias, J., Dürr, J., Elverdin, P., Ibáñez, A. M., Kinengyere, A., et al. (2020). A scoping review on incentives for adoption of sustainable agricultural practices and their outcomes. *Nat. Sustain.* 3, 809–820. doi: 10.1038/s41893-020-00617-y

Porciello, J., Coggins, S., and Mabaya, E. (2022). Digital agriculture services in low- and middle-income countries: a systematic scoping review. *Global Food Security* 34, 100640. doi: 10.1016/j.gfs.2022.100640

Porciello, J., and Ivanina, M. (2021). *The Role of Machine Learning in Programmatic Assessment: A Case Study from USAID's Center for Water Security, Sanitation, and Hygiene*. Research Technical Assistance Center. Available online at: https://www.rtachesn.org/resources/machine-learning-for-programmatic-synthesis/ (accessed July 31, 2022).

Porciello, J., Ivanina, M., Islam, M., and Einarson, S. (2020). Accelerating evidence-informed decision-making for the Sustainable Development Goals using machine learning. *Nat. Machine Intell.* 2, 10. doi: 10.1038/s42256-020-00235-5

Rafols, I., Ciarli, T., and Chavarro, D. (2020). *Under-Reporting Research Relevant to Local Needs in the Global South*. Database biases in the representation of knowledge on rice. SocArXiv. doi: 10.31235/osf.io/3kf9d

Reardon, T., Lu, L., and Zilberman, D. (2019). Links among innovation, food system transformation, and technology adoption, with implications for food policy: overview of a special issue. *Food Policy* 83, 285–288. doi: 10.1016/j.foodpol.2017.10.003

Ricciardi, V., Wane, A., Sidhu, B. S., Godde, C., Solomon, D., McCullough, E., et al. (2020). A scoping review of research funding for small-scale farmers in water scarce regions. *Nat. Sustain.* 3, 836–844. doi: 10.1038/s41893-020-00623-0

Serraj, R., and Pingali, P. (2018). *Agriculture and Food Systems to 2050: Global Trends, Challenges and Opportunities (Vol. 2)*. Singapore: World Scientific. doi: 10.1142/11212

Sharma, Y., Agrawal, G., and Jain, P. (2017). "Vector representation of words for sentiment analysis using GloVe," in *2017 International Conference on Intelligent Communication and Computational Techniques (ICCT)*, p. 279–284. doi: 10.1109/INTELCCT.2017.8324059

Snilstveit, B., Vojtkova, M., Bhavsar, A., and Stevenson, J. (2016). Evidence and Gap Maps: a tool for promoting evidence informed policy and strategic research agendas. *J. Clin. Epidemiol.* 79, 120–129. doi: 10.1016/j.jclinepi.2016.05.015

Stathers, T., Holcroft, D., Kitinoja, L., Mvumi, B. M., English, A., Omotilewa, O., et al. (2020). A scoping review of interventions for crop postharvest loss reduction in sub-Saharan Africa and South Asia. *Nat. Sustain.* 3, 821–835. doi: 10.1038/s41893-020-00622-1

Sumberg, J., and Hunt, S. (2019). Are African rural youth innovative? Claims, evidence and implications. *J. Rural Stud.* 69, 130–136. doi: 10.1016/j.jrurstud.2019.05.004

Teeken, B., Olaosebikan, O., Haleegoah, J., Oladejo, E., Madu, T., Bello, A., et al. (2018). Cassava trait preferences of men and women farmers in nigeria: implications for breeding. *Econ. Botany* 72, 263–277. doi: 10.1007/s12231-018-9421-7

Vincent, A., Huang, C., and Lal, T. (2022). *Mapping evidence to strengthen impact on resilience and food security. International Initiative for Impact Evaluation (July 21 2022)*. Available online at: https://www.3ieimpact.org/blogs/mapping-evidence-strengthen-impact-resilience-and-food-security (accessed July 31, 2022).

Wagner, C. S., Cai, X., and Zhang, Y. (2022). One-year in: COVID-19 research at the international level in CORD-19 data. *PLoS ONE* 17, e0261624. doi: 10.1371/journal.pone.0261624

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., et al. (2020). *CORD-19: The COVID-19 Open Research Dataset* (arXiv:2004, 10706). arXiv. http://arxiv.org/abs/2004.10706

Woodhill, J., Hasnain, S., and Griffith, A. (2020). *What Future for Small-Scale Farmers?* Oxford: Environmental Change Institute.