Review
22

Egypt. J. Bot. **Vol. 62**, No. 2, pp. 291-303 (2022)

Egyptian Journal of Botany
http://ejbo.journals.ekb.eg/

Artificial Intelligence for Plant Genomics and Crop Improvement

Yasmin Hatem, Gehan Hammad#, Gehan Safwat
Faculty of Biotechnology; October University for Modern Sciences and Arts (MSA), Cairo, Egypt.

BECAUSE of rapidly increasing population growth rates, food scarcity has developed into a serious global problem. Furthermore, population growth is expected to reach nine billion by 2050, likely resulting in dramatic issues with the global food supply and accessibility. Numerous technologies are being developed to boost food production in agriculture to close the food gap and overcome obstacles such as climate change, water scarcity, disease, and pests. Understanding plant genomics may facilitate the identification, cloning, and sequencing of genes involved in resistance to adverse environmental influences. Numerous techniques for crop improvement have emerged over the last few decades, including tissue culture transformation and mutagenesis. Recently, artificial intelligence and machine learning have been integrated as a potential multidisciplinary approach to enhancing and improving the agriculture sector, including food, and this field is rapidly evolving. This review explores plant genomics as a solution to future food security concerns by examining the relationship of agriculture, food production, and artificial intelligence as a promising approach for determining the genome and its variations to genetically improve crops in future agriculture.

Keywords: Agriculture, Artificial intelligence, Crop improvement, Deep learning, Machine learning, Plant genomics.

Introduction

Today, the world faces threats to food security, primarily because of the world's rapid population growth, which is expected to reach nine billion by 2050. However, this large population will necessitate the production of numerous fuel, food, and fiber-derived products through agricultural systems (Tilman et al., 2011). To meet these rising demands, current crop productivity must be doubled by 2050, implying an annual increase in total factor productivity of 1.75%. Although total factor productivity growth is currently above 1.5% annually, it has slowed to 0.96% in developing countries, posing a significant obstacle to crop productivity improvement (Steensland & Zeigler, 2018).

Additionally, the agricultural sector generates a variety of critical demands, such as water, energy, wood products, and land for infrastructure development, urbanization, and waste disposal (Lal , 2008). Thus, the current objective is more complex than simply increasing food yields, as it entails a shift to a much more comprehensive system of crop growth that produces socially and environmentally sustainable outcomes (Godfray et al., 2010). By contrast, agricultural sustainability faces significant obstacles because of climate change, limited water resources, a shrinking workforce, and a scarcity of arable land. Consequently, it is vital to increase agricultural production systems' long-term viability and productivity (Schmidhuber & Tubiello, 2007). A well-performing, healthy agricultural sector is linked to the country's environmental, social, and economic well-being. Additionally, the challenges of increasing food production require increasing the productivity of low-yielding and complex farming systems while minimizing environmental and natural resource degradation (Saghir, 2018).

These issues can be addressed in two ways: improved crops or improved crop management. Crop improvement aims to create new crop cultivars with higher yields, higher quality,

#Corresponding author email: gmhammad@msa.edu.eg
Received 04/07/ 2021; Accepted 14 /04/ 2022
DOI: 10.21608/ejbo.2022.83200.1731
Edited by: Prof. Dr. Wafaa Amer, Faculty of Science, Cairo University, Giza, Egypt.

and more adaptability to various conditions, such as saline soils. However, improving crop management aims to advance farming concepts such as precision agriculture, which leverages technological advancements to reduce inputs (such as chemical and irrigation application) and increase output (such as productivity and quality) in agricultural production systems (Houle et al., 2010).

To address these issues, multivariate, unpredictable, and complex agricultural ecosystems must be better understood through continuous evaluation, monitoring, and measurement of a variety of physical phenomena and characteristics. This requires the analysis of a substantial amount of agricultural data (Kamilaris et al., 2017). Along with implementing new information and communication technologies for both short- and long-term farm/crop management, Kamilaris et al. (2016) suggested that existing management, larger-scale ecosystem observation, and decision-making tasks can be enhanced through situation, context, and location awareness. Bastiaanssen et al. (2000) advocated for the adoption of remote sensing for large-scale agricultural observation.

Likewise, Tyagi (2016) asserted that "smart" farming is critical for resolving agricultural production issues, particularly those related to environmental impact, food security, productivity, and sustainability (Gebbers & Adamchuk, 2010). In the mid-1980s, information-based agricultural production management systems emerged as a technique for administering the correct treatment at the proper time and location. The primary drivers are a better understanding of crop and soil variability, as well as the introduction of technologies such as microcomputers, geographic information systems, and global navigation satellite systems (Barrett, 2010). The first application of information-based agriculture was to adjust fertilizer distribution throughout an agricultural field in response to changing soil conditions. Since then, new techniques have emerged, including autonomous machinery, automatic steering of agricultural implements and processes, on-farm research, product traceability, and software for managing the total agricultural production system (Gebbers & Adamchuk, 2010).

Additionally, plant genomics has advanced tremendously in recent years, with the emergence of low-cost, high-throughput approaches for identifying multidimensional genome-wide molecular traits (Crick, 1970). Besides facilitating the collection of molecular phenotypes, genomics also predicts and explains them using reliable data mining methods. Recently, researchers discovered that deep learning is particularly effective at these tasks (Wang et al., 2020).

Plant genomics

The whole-genome sequencing of "small" genomes and the development of new technologies that speed up the cloning and sequencing processes facilitate the advancement of plant genomics research, complete DNA sequencing for certain species, and improve significant advancements in biotechnology and plant breeding (James, 2000). The following issues are expected to be prioritized: discovering new genes, which is critical for plant biotechnology development; identifying, cloning, and regulating chromosome pairing in polyploid plants; and sequencing genes responsible for resistance and variability to unfavorable environmental factors. As such, this provides breeders with additional options for optimizing the breeding process (Zelenin et al., 2001).

All other living organisms on the earth are considered to be a flow of data. Likewise, data flow analysis is essential for crop research and crop (Wang et al., 2020). By contrast, the data flow of the plants starts with the genomic DNA sequence and ends with observed agronomic traits or phenotypes. However, information is altered between these two points via translation and transcription, which Francis Crick referred to in 1957 as "the central dogma in molecular biology" (Crick, 1970). Additionally, to advance basic research and crop development methods, the data flow, which includes genomic DNA sequences and phenotypes, should be considered (Wang et al., 2020).

The study of genotypes and genomes is critical in a wide variety of areas of contemporary plant research. van Dijk et al. (2021) reported that the DNA sequencing revolution has enabled the identification of the entire genetic material of various plants, including model organisms (e.g., *Arabidopsis thaliana*); academically significant species of trees, mosses, algae, and flowering plants; and economically important crops (e.g., maize, cotton, wheat, rice, and soy) (van Dijk

et al., 2021). Additionally, many species have population-wide genotype databases, allowing for the easy identification of genetic variation ranging from single-nucleotide polymorphisms and small deletions/insertions to gene copy number variation and genome structural variation (Torkamaneh et al., 2018). Furthermore, advances in assessing changes in the interactions and amounts of biochemical components in the cell, such as proteins, occur across the genome, metabolites, and RNA, resulting in "-omics" aggregated datasets. Moreover, several applications, such as automated and high-throughput quantification of the yield, growth, morphology, or development of plant organs, tissues, canopies, or entire plants, and the conditions under which the plants thrive, have become available in recent years (Zhao et al., 2019).

With the advent of the "big data" era in plant sciences, explaining or anticipating phenotypes from underlying genotypes under several environmental variables has become a challenge in fundamental and applied research and breeding applications (van Dijk et al., 2021). However, genotype variation results in changes in the biochemical composition of cells, which affect plant growth and organ formation. Ultimately, it combines agriculturally relevant characteristics, such as pest and stress tolerance and yield, with the environment. Analyzing the effects of environment and phenotypic variation due to genotypic variation provides key insights into controlling essential plant physiology and development procedures. Determining quality and yield traits from genotypes found in specific environments is critical for modern molecular plant breeding. Examining phenotypes collected at multiple levels, or connecting these phenotypes to genotypes, necessitates the assimilation and processing of ever-larger, heterogeneous, and noisy datasets (van Dijk et al., 2021).

These approaches, however, have limitations: the number of resources available in molecular phenotypes is unknown, making the mechanistic realization of the process from DNA sequences to terminal phenotypes difficult (Wang et al., 2020). Nevertheless, advancements in two fields of research, in particular, have begun to close this gap. One of them is association analysis, which connects molecular phenotypes to terminal phenotypes, such as the transcriptome-wide association study, which has a shorter information delivery channel and fewer data translation stages than genome-wide association research (Wainberg et al., 2019). The other advancement is the use of deep learning algorithms to infer molecular phenotypes from their molecular characteristics upstream or directly from genomic DNA sequences (Eraslan et al., 2019).

Artificial intelligence in the agricultural sector:

Various techniques, such as tissue culture mutagenesis and transformation, have been used for crop improvement. The study of functional genomics contributes to our understanding of the plant genome and enables its modification. Nanotechnology, RNA interference, and next-generation sequencing have emerged as promising new techniques for increasing crop yields in response to future demands (Rashid et al., 2017). Helping farmers and stakeholders make better decisions by implementing sustainable agriculture practices, particularly through the use of digital technologies such as cloud computing, the Internet of Things, and artificial intelligence (AI), is a critical choice for anticipating efficient solutions. Additionally, location intelligence technology frequently incorporates AI components (deep and machine learning algorithms) (Ben Ayed & Hanana, 2021).

However, AI is a creative technique that uses technology, including most computer systems, robotics, and digital equipment, to simulate human intelligence and capability processes (Patel et al., 2021). Furthermore, healthcare, finance, pharmaceutical research, retail, marketing, and intelligent process automation are among the industries that have grown the fastest in recent years; they all utilize AI (Ben Ayed & Hanana, 2021). Machine learning (ML) is a fundamental concept in AI that enables people to work more efficiently and creatively in various situations. ML uses mathematical and statistical tools to determine from datasets and make data-driven predictions to decisions. Supervised, unsupervised, and reinforcement learning are the three primary tasks of ML (Traore et al., 2017).

As illustrated in Fig. 1, the goal of supervised learning is to map variables such as DNA sequences to a selected output variable such as histone marks (Traore et al., 2017). There are two types of target variables: categorical (categorization) and continuous (categorization) (regression). Here are a few examples of supervised learning programs

that predict regulatory and nonregulatory regions in the maize genome (Mejía-Guerra & Buckler, 2019): mRNA expression levels (Washburn et al., 2019), plant stress phenotyping (Ghosal et al., 2018), macronutrient deficiency in tomatoes prediction (Tran et al., 2019), and *Arabidopsis thaliana* polyadenylation site prediction (Gao et al., 2018).

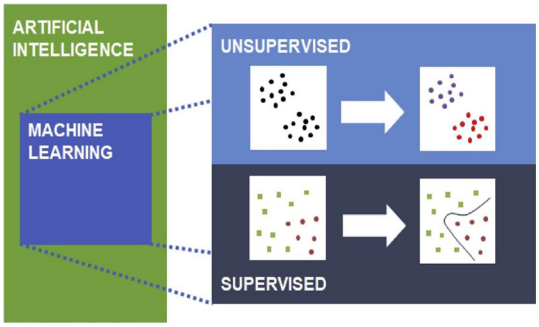

Fig. 1. Machine learning (ML) is an artificial intelligence (AI) subfield, a larger field [The distinction between supervised and unsupervised techniques is critical in ML. Supervised procedures make use of labeled input data. It maps variables such as DNA sequences to the output variable of choice, such as histone marks. Classification is one application in which distinct labels (red squares vs. green circles) are known. The model learns to determine the label for recent items. On the right, the curve depicts a decision boundary, which represents what the supervised model has learned. By contrast, unsupervised approaches look for groups or trends in data rather than assigning labels. It identifies hidden patterns within an unlabeled dataset. Clustering is a good example of an unsupervised algorithm because it can detect two groups. Adapted from: van Dijk et al. (2021)]

As illustrated in Fig. 1, unsupervised learning uses unlabeled datasets with no previous knowledge of the output and input variables and encompasses algorithms such as clustering, genetic algorithms, artificial neural networks, and deep learning. This unsupervised ML example discovers hidden patterns in an unlabeled dataset and is frequently used for data reduction and dimensionality exploration (Jordan & Mitchell, 2015). Additionally, numerous algorithms, including Q-learning and deep Q-learning, are used in the third type of ML task, referred to as reinforcement learning, for robot navigation, machine skill acquisition, and real-time decision-making (Mohri et al., 2018).

However, AI technology has recently gained traction in the agri-food industry. It aids in decision-making processes, service development, and model identification in supply chain phases and agri-food applications. In agriculture, the primary goal of AI is to provide accuracy and predict decisions to maximize productivity while conserving resources (Patel et al., 2021). Thus, to address agricultural challenges, AI tools propose algorithms that can assess performance, forecast unexpected problems or occurrences, and identify patterns, such as water consumption and irrigation process management through the installation of intelligent irrigation systems (Suprem et al., 2013).

Additionally, AI is being used in applications such as fine-tuning automated machines for accurate pest or disease detection, weather forecasting, and a focus on analyzing unhealthy crops and enhancing the capacity for healthy crop production. Furthermore, AI simplifies the process of cultivating, harvesting, and marketing crops (Ben Ayed & Hanana, 2021). Thus, AI advancements have aided agro-based businesses in operating more efficiently by enhancing crop management practices, allowing many technology companies to invest in algorithms that benefit agriculture and address farmer challenges such as pest and weed infestations and climate change-related yield reductions (Sujatha et al., 2021).

Farmers can also use technology to stay up to date on weather forecasting data, which can help farmers improve yields and profitability while preventing crop damage. As a result of the analysis of the generated data, AI enables farmers to learn more and understand and then take action by implementing practices that allow them to make wise decisions (Crane-Droesch, 2018). Additionally, AI approaches are capable of monitoring soil management and health by identifying plant pests and diseases, as well as nutrient deficiencies and possible soil defects, by analyzing either flora patterns in farms or an image captured with a camera recognition tool (Ben Ayed & Hanana, 2021). AI technology has a significant functional benefit in environmental conservation by reducing pesticide usage. For instance, farmers could use AI techniques incorporating robots, ML, and computer vision to spray chemicals only where weeds are present, thereby managing weeds more efficiently and

accurately. This would reduce the amount of chemical substance sprayed across the entire field (Kamilaris & Prenafeta-Boldú, 2018).

ML algorithms are becoming increasingly important in the agriculture supply chain's core four clusters (preproduction, production, processing, and distribution) (Ahumada & Villalobos, 2009). ML technologies are used to forecast crop yield, irrigation requirements, and soil parameters during the preproduction stage. Additionally, ML could be used to identify diseases and forecast weather in the subsequent stage of the production phase. In the third cluster of the processing phase, ML algorithms are used to forecast production planning to achieve high and safe product quality. Finally, ML algorithms may benefit the distribution cluster, particularly in storage, customer analysis, and transportation (Ben Ayed & Hanana, 2021).

Deep learning for plant genomics

ML, the science of using programming to teach computers to learn from data, is frequently used in genomics to capture large datasets and generate new biological hypotheses. More expressive ML models are required to derive recent insights from the massive influx of genomics data. Deep learning has revolutionized industries such as natural language processing and computer vision by successfully utilizing massive datasets. Furthermore, deep learning is a relatively new technique for data analysis and image processing with many promising applications and enormous potential. Moreover, deep learning has been successfully applied in various industries, and it has recently entered the agricultural sector (Bioshop, 2016).

For example, deep learning enhances classical ML by increasing the model's complexity and modifying the data with various functions that allow for the hierarchical representation of data at multiple levels of abstraction (LeCun & Bengio, 1995; Schmidhuber, 2015). The automatic extraction of features from raw data is a critical benefit of deep learning, as higher-level features are formed by the components of lower-level features (LeCun et al., 2015). Because of the more complicated models used in deep learning, which enable massive parallelization, they can manage highly complex problems quickly and efficiently (Pan & Yang, 2010).

Numerous deep learning approaches have been evaluated in recent years for their performance in genomic prediction. However, the fundamental distinction between deep learning methods and traditional statistical learning methods is that those deep learning methods are considered nonparametric models, which provide a high degree of adaptability to complex data-output associations (Kononenko & Kukar, 2007).

There is compelling evidence that deep learning algorithms capture nonlinear patterns more effectively than traditional genome-based methods used in genomic selection (GS). These algorithms generate a meta-picture of GS performance and demonstrate how these tools may aid in the solution of difficult plant breeding problems. Additionally, as is common in gene selection-assisted breeding, deep learning algorithms can integrate data from multiple sources and have demonstrated the ability to improve prediction accuracy for large plant breeding datasets. Consequently, using deep learning techniques on large training and testing datasets is critical (Bernardo, 2008).

Furthermore, the progress being made today in modeling the molecular phenotype through various deep learning approaches and introducing their use to identify functional variations may be useful for crop genetic improvement. Deep learning models may be used in synthetic biology to discover novel beneficial genes (Ramstein et al., 2018). However, deep learning has been used to solve complex biological problems via large-scale data analysis in genomics, proteomics, transcriptomics, systems biology, and metabolomics (Xu & Jackson, 2019).

Four steps are typically included in a deep learning approach that uses molecular phenotypes and biological sequences as target and predictor variables. As illustrated in Fig. 2, (a) Biological sequences are the preprocessing of target variables and predictors, namely, encoding and retrieval of biological sequences, categorical or appropriate validation of predictor-target pairs, the numerical representation of molecular phenotypes, test sets, and training, typically taking into account evolutionary relationships between biological sequences (Washburn et al., 2019). (b) Training is the step that encompasses model construction and training, hyperparameter selection, and model architecture, as well as model training

using the training set. However, performance on the validation set must be monitored during training to determine when the model should be stopped to avoid both overfitting and underfitting. (c) Evaluation is the process of evaluating the performance of trained models on a new dataset (the test set). The area under the ROC curve is a statistic for evaluating model performance in classification problems, whereas R-squared is a metric for evaluating model performance in regression problems. (d) Interpretation is the process of utilizing saliency or feature attribution approaches to deduce functional elements from biological sequences (Wang et al., 2020).

Deep learning and the central dogma

Gene properties and DNA

The shape of the DNA molecule plays a critical role in determining the transcription factor's DNA-binding specificity (Lai et al., 2019). Accessible data types include chromatin accessibility assays such as MNase-seq, FAIRE, and DNase-seq, as well as other genomic assays such as RNA-seq expression and microarray. Similarly, gene expression profiles, ChIP-seq data, ampDAP-seq (which uses amplified and thus demethylated DNA as histone and substrate modifications), and DAP-seq (DNA affinity purification sequencing) are available to investigate the molecular mechanisms underlying gene expression (Zampieri et al., 2019). To evaluate these massive datasets, several deep learning approaches for modeling transcription factor DNA-binding specificity were developed. Additionally, certain algorithms based on deep learning have been developed to predict transcription factor binding *in vivo*. DeepBind, for example, can learn numerous motifs to deduce the binding sites of RNA and DNA-binding proteins (Alipanahi et al., 2015). However, because the datasets are readily available, such methods were performed on human cell lines or tissues (Wang et al., 2020).

Finding critical genomic regulatory regions in maize species is considered a difficult task because of the presence of numerous repeated elements and large intergenic areas. To address these concerns, techniques such as k-mer grammars use natural language processing to precisely and cost-effectively designate regulatory areas in maize lines (Mejía-Guerra & Buckler, 2019). ML models are effective in various areas of plant biology. They can be developed using a variety of sequencing data, either individually or in combination, and additional data, such as DNase I hypersensitivity data, to improve *in vivo* transcription binding site prediction (Qin & Feng, 2017).

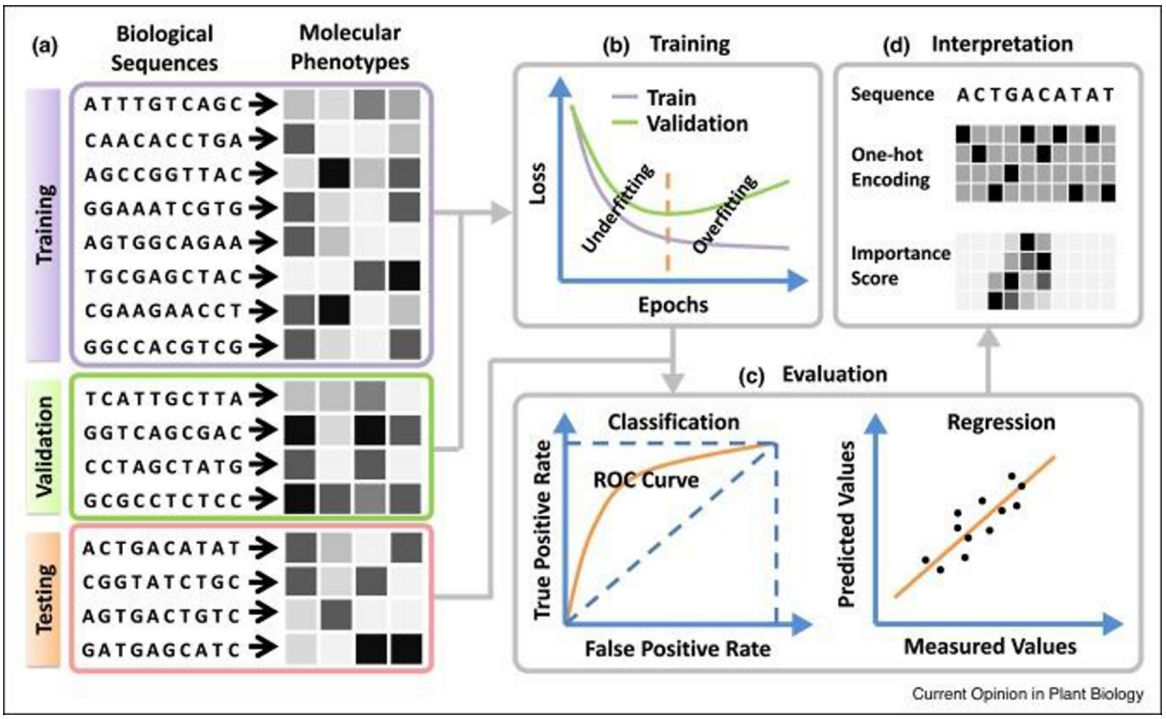

Fig. 2. The workflow of deep learning in genomics [Adapted from: Wang et al., 2020]

Protein properties

The tertiary structure of any protein has a direct effect on its function. Secondary structure, model properties, interresidual contact, solvent accessibility, signal peptides, transmembrane topology, contact maps, disorder-to-order transition, protein disorders, backbone dihedrals, enzyme dynamics, and protein interaction sites can all be used to deduce a protein's tertiary structure (Wang et al., 2020). DeepNovo was developed to extract key amino acid characteristics from a de novo peptide sequence using a convolutional neural network (CNN) algorithm (Tran et al., 2017). Google's AlphaFold has generated considerable interest by utilizing AI to forecast the tertiary structure of proteins (Evans et al., 2018). By using interresidue contact mapping and relative solvent accessibility, raw multiple sequence alignments were used in deep learning algorithms to predict secondary structure (Mirabello & Wallner, 2018). Although deep learning algorithms have demonstrated promise in various fields, their predictive power for protein–protein interaction (PPI) prediction has been limited by noisy data and insufficient coverage. However, a new model that uses sequence information to predict PPIs and homodimeric interactions is being developed as a defect prevention-based process (Hashemifar et al., 2018).

Data and model sharing

There are several ways to share animals and plant data. The Kipoi repository was recently launched to make sharing and reusing predictive genomics models easier (Avsec et al., 2019). This is accomplished by using Findable, Accessible, Interoperable, and Reusable concepts (Wilkinson et al., 2016). According to Washburn et al. (2019), most Kipoi models developed for human genomes or animal and plant genomics datasets can be easily retrained or even applied directly to plants (including models that can predict biochemical properties of proteins).

However, some plant-specific concerns must be addressed when dealing with them. For example, when comparing Sorghum and Maize relative gene expression levels, maize's tetraploidy may present significant challenges (Washburn et al., 2019). Additionally, plant species with polyploidy and significant tandem gene duplication may have an imbalanced measurement of gene expression, resulting in lower-quality training datasets. Furthermore, because the size of genome elements (exons, introns, and the distance between promoters and enhancers) varies significantly between plant and animal species, model structures must be reoptimized (Wang et al., 2020).

Deep learning for plant breeding

The elimination of deleterious alleles is a critical component of crop breeding in modern management approaches and environmental adaptation. However, because crop species interact with their environments and genotypes more frequently than livestock species, it is possible that predicting allele impacts (whether detrimental, adaptive, or beneficial) in crop species will be more difficult. This difficulty could be overcome by using environment-specific models or models that incorporate environmental parameters as additional inputs. Consequently, the functional variations predicted by deep learning algorithms are likely to be critical in the next breeding era, when crop improvement will rely heavily on genome editing (Wang et al., 2020).

More importantly, there is no restriction on known advantageous variants found in nature when using this breeding-by-editing method. Instead, the use of deep learning models provides an unlimited amount of freedom to develop unique advantageous alleles based on an "understanding" of the biological processes of interest. For example, Rodríguez-Leal et al. (2017) altered the (SlCLV3) promotor of the tomato CLAVATA3 gene to optimize inflorescence branching and fruit size. Because of the lack of functional annotations for the SlCLV3 promoter, the CRISPR/Cas9 system was used to perform saturation promoter mutagenesis, followed by the selection of mutants with favorable inflorescence and fruit features. However, developing a deep learning model that uses the SlCLV3 promoter sequence to predict gene expression levels will facilitate the identification of critical cis-elements on the promoter at single-nucleotide resolution, perform model-guided promoter editing, and quantify their loss-of-function effects on SlCLV3 gene expression (Wang et al., 2020).

In synthetic biology, another technique for creating unique genetic components with specific functionalities is to use generative models. For example, after determining the mutation space of existing promoters, models can be trained to produce new promoters with spatiotemporal

specificity (Wang et al., 2020). Despite increased interest in generative models as variational autoencoders and generative adversarial networks, their applications in synthetic biology remain limited. For instance, they utilize generative adversarial networks (GANs) to generate antimicrobial peptides encoded by synthetic DNA sequences (Gupta & Zou, 2018). However, using generative models to produce novel DNA elements, regulatory circuits, or genes with desired functions and then applying them to crop improvement show significant promise (Wang et al., 2020).

Advantages and applications of deep learning in agriculture

Fruit counting (Rahnemoonfar & Sheppard, 2017), obstacle detection (Christiansen et al., 2016), and image categorization are considered to be emerging areas of research. Furthermore, several articles discuss forecasting future characteristics, such as weather conditions (Sehgal et al., 2017), corn yield (Kuwata & Shibasaki, 2015), and on field soil moisture content (SMC) (Song et al., 2016).

For example, SMC is critical in meteorology, climate change, and agriculture. Consequently, soil moisture affects water migration through infiltration and runoff. However, accurate modeling of the spatiotemporal distributions of surface soil moisture in farmlands is critical (Sadler et al., 2005). Moreover, it can be used with cutting-edge irrigation technologies such as lateral moving irrigators and center pivots (Kim et al., 2008).

To adopt sustainable water management and reduce the increased irrigation water demand in arid areas, a stable simulated technique is required to predict the complicated dynamics of SMC. However, the SMC in agricultural regions is heterogeneous in both place and time because of evapotranspiration, soil superposition, precipitation, terrain, soil management, irrigation, and fertilization. The validation results of this study demonstrated that by incorporating field data and environmental variables, a deep learning-based macroscopic cellular automata (MCA) model was successfully able to indicate the spatial-temporal patterns of SMC. Additionally, because environmental variables are increasingly being used to model soil moisture, deep belief network techniques may aid in calibrating MCA that are dependent on SMC calculations, thereby providing a new tool for canal irrigation system SMC monitoring (Song et al., 2016).

Additionally, deep learning techniques, such as deep CNNs, are used to estimate the yield. However, farmers can make more informed decisions about cultivation practices, disease control, and the size of the harvest workforce when they know the exact number of fruits, flowers, and trees. The current yield estimation method, which relies on manually counting flowers or fruits, is inefficient and inconvenient for large fields. Automated yield estimation based on robotic agriculture is a viable option in this case. The network is trained on synthetic data before being evaluated on real-world data (Rahnemoonfar & Sheppard, 2017).

Another advantage of deep learning is the ability to train models on synthetic data and then use them to solve real-world problems. Identifying maize and weeds in fields is a good example (Slaughter et al., 2008). Herbicide application is the preferred method of weed control because of its lower cost and greater effectiveness than mechanical weeding. However, because of environmental concerns, the government is increasing pressure on farmers to reduce their reliance on herbicides (Dyrmann et al., 2016). Knowing which weed species are prevalent in agricultural fields is critical for site-specific weed management; thus, plant species can be recognized in colored images by using a deep CNN (Slaughter et al., 2008). The construction and training of deep CNNs to discriminate between numerous plant species aimed to establish a hierarchy of self-learned features dependent on less abstract information from the network's earlier layers (Dyrmann et al., 2016).

There are additional agricultural problems that deep learning can help solve. However, these issues include leaf and soil nitrogen content, seed identification, water stress detection, irrigation, disease or defect detection on food, water erosion assessment, crop-hail damage, herbicide use, greenhouse monitoring, contaminant detection, and pest detection. Thus, deep learning entails prediction or categorization, image analysis, and computer vision, or, more broadly, data analysis (Kamilaris & Prenafeta-Boldú, 2018).

Additionally, there are two strategies for

agricultural production system enhancement: crop improvement and crop management. Conversely, crop improvement aims to create new crop cultivars that are higher in quality, yield, and adaptability to various environmental conditions. Crop management aims to improve farming principles through precision agriculture, which leverages technological advancements to decrease input (e.g., chemical and irrigation application) and increase output (e.g., quality and productivity), for agricultural production systems (e.g., data science techniques, automation, and sensing). In the last 5 years, gantry-based, tower-based, ground mobile, satellite-based, and low- and high-altitude aerial systems have significantly improved phenotyping capabilities and throughput (Jiang & Li, 2020).

Plant phenotyping is critical for improving plant breeding programs, managing agricultural systems, and comprehending the interactions of plants with their environment. Imaging techniques have demonstrated a significant potential for increasing plant phenotyping over the last 5 years, resulting in a greater emphasis on imaging-based plant phenotyping (Jiang & Li, 2020). Furthermore, as the volume of image data grows, it is critical to developing robust analytical techniques capable of reliably and rapidly eliciting phenotypic features. Image sensors have grown in popularity because of their high capacity for collecting complex features. 2D imaging techniques (e.g., spectral, thermal, and red, green and blue color imaging) can provide spatial information and an additional data dimension about a scene, such as spectral data extracted from spectral images. Additionally, 3D imaging (e.g., LiDAR) can provide a three-dimensional structure of a scene, which can deduce the object's morphological characteristics (length, volume, and area). Likewise, 2.5D imaging (e.g., depth camera), like 2D imaging, preserves the imaging plane's structure information and records a scene's depth information, which can help reconstruct the scene's 3D structure. The phenotyping of plant morphology, development, postharvest quality, and physiology have been accomplished using imaging-based technologies (Jiang & Li, 2020).

Precision crop management (PCM) is a technique for agricultural management that maximizes profitability while protecting the environment by targeting crop and soil inputs based on field needs. By contrast, PCM has been hampered by a lack of widely disseminated information on soil and crop conditions (Moran et al., 1997). Another application of AI in crop improvement is PROLOG, which evaluates a farm system's operational behavior using machinery capacity, meteorological data, labor information, and available operators, tools, and tractors. Additionally, it determines gross revenue, crop yields, and net profit for individual fields and the entire farm (Lal et al., 1992). To use AI in cucumber harvesting, individual software and hardware components of the robot, such as the manipulator, autonomous vehicle, two detection computer vision systems, end-effector, and 3D imaging of the environment and the fruit, are used. Finally, a manipulator control strategy ensures collision-free motions during harvesting (Henten et al., 2002).

Conclusion

The predicted overpopulation in the coming years poses a serious challenge because it will be difficult to provide them with sufficient food. More importantly, climate change, technological advancement, water scarcity, and other environmental challenges will all work together to reduce food production. Fortunately, crop improvement technologies are the most critical component of agricultural productivity growth. The use of digital technology at many stages of the supply chain, such as sensors, AI, remote satellite data and automation of farm machinery, ML for improved crop monitoring, and water, is critical for agriculture food product traceability.

Association mapping has been used in wild plant populations to identify genetic loci connected to terminal traits or molecular phenotypes. Consequently, it is reasonable to believe that combining models that "understand" the information transfer from DNA to molecular phenotypes with association mapping studies that link molecular phenotypes to terminal characteristics will aid in prioritizing causative variants.

Furthermore, ML may be used in the future in plant research to predict which regions of the genome can be edited to achieve a desired phenotype in the case of genetic modification or to improve crops and agricultural production systems by assessing crop performance *in vivo*, on the field, or in the greenhouse to provide optimal

local growth conditions. However, there has been a considerable rise in the use of deep learning in agriculture, as it is deemed a promising approach for determining the genome and its variations to genetically improve crops in future agriculture.

Competing interests: The authors report no conflicts of interest regarding this work.

Authors' contributions: The authors contributed equally to the completion of the review, the idea was incepted by GS, the approach, research for literature was prepared by GH, YH was the major contributor of writing which was reviewed and edited by both GH and GS.

Ethics approval: Not applicable.

References

Ahumada, O., Villalobos, J.R. (2009) Application of planning models in the agri-food supply chain: A review. *European Journal of Operational Research*, **196**(1), 1–20.

Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, **33**(8), 831–838.

Avsec, Ž., Kreuzhuber, R., Israeli, J., Xu, N., Cheng, J., Shrikumar, A., et al. (2019) The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nature Biotechnology*, **37**(6), 592–600.

Barrett, C. (2010) Measuring Food Insecurity. Science, 327(5967), new series, 825-828. Retrieved June 22, 2021, from http://www.jstor.org/stable/40509899

Bastiaanssen, W.G.M., Molden, D.J., Makin, I.W. (2000) Remote Sensing for Irrigated Agriculture: Examples from Research and Possible Applications. In: "*Agricultural Water Management*", Elsevier, Vol. 46(2), pp. 137-155.

Ben Ayed, R., Hanana, M. (2021) Artificial intelligence to improve the food and agriculture sector. *Journal of Food Quality*, **2021**, 1–7. https://doi.org/10.1155/2021/5584754

Bernardo, R. (2008) Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Science*, **48**(5), 1649–1664.

Bishop, C.M. (2016) Pattern Recognition and Machine Learning (Springer, New York).

Christiansen, P., Sørensen, R., Skovsen, S., Jaeger, C.D., Jørgensen, R.N., Karstoft, H., et al. (2016) Towards autonomous plant production using fully convolutional neural networks, In: *International Conference on Agricultural Engineering*, Aarhus, pp. 1–8.

Crane-Droesch, A. (2018) Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, **13**(11), 114003.

Crick, F. (1970) Central dogma of molecular biology. *Nature*, **227**(5258), 561–563.

Dyrmann, M., Karstoft, H., Midtiby, H.S. (2016) Plant species classification using deep convolutional neural network. *Biosystems Engineering*, **151**, 72–80.

Eraslan, G., Avsec, Ž., Gagneur, J., Theis, F.J. (2019) Deep learning: New computational modelling techniques for genomics. *Nature Reviews Genetics*, **20**(7), 389–403.

Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, G., Qin, C., et al. (2018) De novo structure prediction with deep-learning based scoring. In: *Thirteenth Critical Assessment of Techniques for Protein Structure Prediction* (Abstracts).

Gao, X., Zhang, J., Wei, Z., Hakonarson, H. (2018) DeepPolyA: A Convolutional Neural Network Approach for Polyadenylation Site Prediction. *IEEE Access*, **6**, 24340–24349.

Gebbers, R., Adamchuk, V.I. (2010) Precision agriculture and food security. *Science*, **327**(5967), 828–831.

Ghosal, S., Blystone, D., Singh, A.K., Ganapathysubramanian, B., Singh, A., Sarkar, S. (2018) An explainable deep machine vision framework for plant stress phenotyping. *Proceedings of the National Academy of Sciences*, **115**(18), 4613–4618.

Godfray, H.C., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., Toulmin, C. (2010) Food Security: The Challenge of Feeding 9 Billion People. *Science*, **327**(5967), 812–818.

Gupta, A., Zhou, J. (2018) Feedback GAN (FBGAN) for DNA: A novel feedback-loop architecture for optimizing protein functions. arXiv:1804.01694 [q-bio.GN]. In: *Thirteenth Critical Assessment of Techniques for Protein Structure Prediction* (Abstracts).

Hashemifar, S., Neyshabur, B., Khan, A.A., Xu, J. (2018) Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics*, **34**(17), i802–i810.

Henten, E.V., Hemming, J., Tuijl, B.V., Kornet, J., Meuleman, J., Bontsema, J., Os, E. (2002) An Autonomous Robot for Harvesting Cucumbers in Greenhouses. *Autonomous Robots,* **13**, 241-258.

Houle, D., Govindaraju, D.R.,Omholt, S. (2010) Phenomics: The next challenge. Nature reviews. *Genetics*, **11**(12), 855–866.

James, R. (2000) Differentiating genomics companies. *Nature Biotechnology*, **18**(2), 153–155.

Jiang, Y., Li, C. (2020) Convolutional neural networks for image-based high-throughput plant phenotyping: A review. *Plant Phenomics*, **2020**, 1–22. https://doi.org/10.34133/2020/4152816

Jordan, M.I., Mitchell, T.M. (2015) Machine learning: Trends, perspectives, and prospects. *Science*, **349**(6245), 255–260.

Kamilaris, A., Prenafeta-Boldú, F.X. (2018) Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, **147**, 70–90.

Kamilaris, A., Gao, F., Prenafeta-Boldú, F., Ali, M.I. (2016) Agri-IoT: A semantic framework for Internet of Things-enabled smart farming applications. 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), pp. 442-447.

Kamilaris, A., Kartakoullis, A., Prenafeta-Boldú, F.X. (2017) A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, **143**(1), 23–37.

Kim, Y., Evans, R.G., Iversen, W.M. (2008) Remote sensing and control of an irrigation system using a distributed wireless sensor network. In: "*IEEE Transactions on Instrumentation and Measurement*", Vol. 57, no. 7, pp. 1379-1387, July 2008, doi: 10.1109/TIM.2008.917198.

Kononenko, I., Kukar, M. (2007) "*Machine Learning and Data Mining: Introduction to Principles and Algorithms*". London: Horwood Publishing.

Kuwata, K., Shibasaki, R. (2015) Estimating crop yields with deep learning and remotely sensed data. *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. https://doi.org/10.1109/igarss.2015.7325900

Lai, X., Stigliani, A., Vachon, G., Carles, C., Smaczniak, C., Zubieta, C., Kaufmann, K., Parcy, F. (2019) Building transcription factor binding site models to understand gene regulation in plants. *Molecular Plant*, **12**(6), 743–763.

Lal, H., Jones, J.W., Peart, R.M., Shoup, W.D. (1992) FARMSYS—A whole-farm machinery management decision support system. *Agricultural Systems*, **38**(3), 257–273.

Lal, R. (2008) Soils and sustainable agriculture. A review. *Agronomy for Sustainable Development*, **28**(1), 57–64.

LeCun, Y., Bengio, Y. (1995) Convolutional networks for images, speech, and time series. Handbook Brain Theory Neural Networks 3361 (10).

LeCun, Y., Bengio, Y., Hinton, G. (2015) Deep learning. *Nature*, **521**(7553), 436–444.

Mohri, M., Rostamizadeh, A., Talwalka, A. (2018) "*Foundations of Machine Learning*", MIT Press, Cambridge, MA, USA, https://cs.nyu.edu/~mohri/mlbook/.

Mejia-Guerra, M.K., Buckler, E.S. (2019) A k-mer grammar analysis to uncover maize regulatory architecture. *BMC Plant Biology*, **19**(1). https://doi.org/10.1186/s12870-019-1693-2

Mirabello, C., Wallner, B. (2018) RawMSA: End-to-end Deep Learning Makes Protein Sequence Profiles and Feature Extraction obsolete. https://doi.org/10.1101/394437

Moran, M.S., Inoue, Y., Barnes, E.M. (1997) Opportunities and limitations for image-based remote sensing in precision crop management. *Remote Sensing of Environment*, **61**(3), 319–346.

Pan, S.J., Yang, Q. (2010) A Survey on transfer learning. *IEEE Transactions on Knowledge and*

Data Engineering, **22**(10), 1345–1359.

Patel, G., Rai, A., Das, N., Singh, R. (2021) Eds. in Smart Agriculture: Emerging Pedagogies of Deep Learning, Machine Learning and Internet of Things, CRC Press, Boca Raton, FL, USA, 1st ed.

Qin, Q., Feng, J. (2017) Imputation for transcription factor binding predictions based on deep learning. *PLOS Computational Biology*, **13**(2). https://doi.org/10.1371/journal.pcbi.1005403

Rahnemoonfar, Maryam, Sheppard, C. (2017) Deep count: Fruit counting based on deep simulated learning. *Sensors (Basel, Switzerland)*, **17**(4), 905. 10.3390/s17040905.

Ramstein, G.P., Jensen, S.E., Buckler, E.S. (2018) Breaking the curse of dimensionality to identify causal variants in Breeding 4. *Theoretical and Applied Genetics*, **132**(3), 559–567.

Rashid, B., Tariq, M., Khalid, A., Shams, F., Ali, Q., Ashraf, F., Ghaffar, I., et al. (2017) Crop improvement: New approaches and modern techniques. *Plant Gene and Trait*. https://doi.org/10.5376/pgt.2017.08.0003

Rodríguez-Leal, D., Lemmon, Z.H., Man, J., Bartlett, M.E., Lippman, Z.B. (2017) Engineering quantitative trait variation for crop improvement by genome editing. *Cell*, **171**(2). https://doi.org/10.1016/j.cell.2017.08.030

Sadler, E., Evans, R.G., Stone, K.,Camp, C.R. (2005) Opportunities for conservation with precision irrigation. *Journal of Soil and Water Conservation,* **60**, 371–379.

Saghir, J. (2014) Global challenges in agriculture and the World Bank's response in Africa. *Food and Energy Security*, **3**(2), 61–68.

Schmidhuber, F., Tubiello, N. (2007) Global food security under climate change. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 104, no. 50, pp. 19703–19708.

Schmidhuber, J. (2015) Deep learning in neural networks: An overview. *Neural Networks*, **61**, 85–117.

Sehgal, V.K., Singh, M., Jain, N., Pathak. H. (2017) Climate change and variability: Mapping vulnerability of agriculture using Geospatial Technologies. In: "*Agriculture under Climate Change: Threats, Strategies and Policies*", ed. V.V. 74. Belavadi, N.

Slaughter, D.C., Giles, D.K., Fennimore, S.A., Smith, R.F. (2008) Multispectral machine vision identification of lettuce and weed seedlings for automated weed control. *Weed Technology*, **22**(2), 378–384.

Song, X., Zhang, G., Liu, F., Li, D., Zhao, Y., Yang, J. (2016) Modeling spatio-temporal distribution of soil moisture by deep learning-based cellular automata model. *Journal of Arid Land*, **8**(5), 734–748.

Steensland, A., Zeigler, M. (2018) Global Agricultural Productivity Report, https://globalagriculturalproductivity.org/.

Sujatha, R., Chatterjee, J.M., Jhanjhi, N.Z., Brohi, S.N. (2021) Performance of deep learning vs machine learning in plant leaf disease detection. *Microprocessors and Microsystems*, **80**, 103615. https://doi.org/10.1016/j.micpro.2020.103615

Suprem, A., Mahalik, N., Kim, K. (2013) A review on application of technology systems, standards and interfaces for agriculture and food sector. *Computer Standards & Interfaces*, **35**(4), 355–364.

Tilman, D., Balzer, C., Hill, J., Befort, B.L. (2011) Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences*, **108**(50), 20260–20264.

Torkamaneh, D., Boyle, B., Belzile, F. (2018) Efficient genome-wide genotyping strategies and data integration in crop plants. *Theoretical and Applied Genetics*, **131**(3), 499–511.

Tran, N.H., Zhang, X., Xin, L., Shan, B., Li, M. (2017) *De novo* peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences*, **114**(31), 8247–8252.

Tran, T.-T., Choi, J.-W., Le, T.-T., Kim, J.-W. (2019) A comparative study of deep CNN in forecasting and classifying the macronutrient deficiencies on development of tomato plant. *Applied Sciences*, **9**(8), 1601. https://doi.org/10.3390/app9081601

Traore, B.B., Kamsu-Foguem, B., Tangara, F. (2017)

Data mining techniques on satellite images for discovery of risk areas. *Expert Systems with Applications*, **72**, 443–456.

Tyagi, A.C. (2016) Towards a second Green Revolution. *Irrigation and Drainage*, **65**(4), 388–389.

van Dijk, A.D., Kootstra, G., Kruijer, W., de Ridder, D. (2021) Machine learning in plant science and plant breeding. *IScience*, **24**(1), 101890. https://doi.org/10.1016/j.isci.2020.101890

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., Björkegren, et al. (2019) Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*, **51**(4), 592–599.

Wang, H., Cimen, E., Singh, N., Buckler, E. (2020) Deep learning for plant genomics and crop improvement. *Current Opinion in Plant Biology*, **54**, 34–41.

Washburn, J.D., Mejia-Guerra, M.K., Ramstein, G., Kremling, K.A., Valluru, R., Buckler, E.S., Wang, H. (2019) Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proceedings of the National Academy of Sciences*, **116**(12), 5542–5549.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**(1). https://doi.org/10.1038/sdata.2016.18

Xu, C., Jackson, S.A. (2019) Machine learning and complex biological data. *Genome Biology*, **20**(1). https://doi.org/10.1186/s13059-019-1689-0

Zampieri, G., Vijayakumar, S., Yaneske, E., Angione, C. (2019) Machine and deep learning meet genome-scale metabolic modeling. *PLOS Computational Biology*, **15**(7) https://doi.org/10.1371/journal.pcbi.1007084

Zelenin, A.V., Badaeva, E.D., Muravenko, O.V. (2001) Introduction into plant genomics. molecular biology, **35**(3), 285–293.

Zhao, C., Zhang, Y., Du, J., Guo, X., Wen, W., Gu, S., Wang, J., Fan, J. (2019) Crop phenomics: Current status and perspectives. *Frontiers in Plant Science*, **10**. https://doi.org/10.3389/fpls.2019.00714 .

الذكاء الاصطناعي في مجال تحليل جينوم النباتات وتحسين المحاصيل.

ياسمين حاتم، چيهان حماد، چيهان صفوت
كلية التكنولوجيا الحيوية - جامعة اكتوبر للعلوم الحديثة والآداب- جيزة- مصر.

ادت الزيادة السريعة في معدلات النمو السكاني إلى ندرة الغذاء مما جعلها مشكلة عالمية خطيرة. علاوة على ذلك، من المتوقع أن يصل النمو السكاني إلى تسعة مليارات بحلول عام 2050، مما سيؤدي على الأرجح إلى مشاكل في الإمدادات الغذائية العالمية وإمكانية الوصول إليها. يتم تطوير العديد من التقنيات لتعزيز إنتاج الغذاء في الزراعة لسد الفجوة الغذائية والتغلب على العقبات مثل تغير المناخ وندرة المياه والأمراض والآفات. قد يسهل فهم الجينوم النباتي تحديد واستنساخ وتسلسل الجينات المشاركة في مقاومة التأثيرات البيئية الضارة. لقد ظهرت العديد من التقنيات لتحسين المحاصيل على مدى العقود القليلة الماضية، بما في ذلك تحويل زراعة الأنسجة والطفرات. في الآونة الأخيرة، تم دمج الذكاء الاصطناعي والتعلم الآلي كنهج متعدد التخصصات محتمل لتعزيز وتحسين قطاع الزراعة، بما في ذلك الغذاء، وهذا المجال يتطور بسرعة. وهذه المراجعة توضح علم الجينوم النباتي كحل لمخاوف الأمن الغذائي في المستقبل من خلال دراسة العلاقة بين الزراعة وإنتاج الغذاء والذكاء الاصطناعي كنهج واعد لتحديد الجينوم وتنوعاته لتحسين المحاصيل وراثيًا في الزراعة المستقبلية.