agriculture

Review

Harnessing AI-Powered Genomic Research for Sustainable Crop Improvement

Elżbieta Wójcik-Gront [1], Bartłomiej Zieniuk [2] and Magdalena Pawełkowicz [3,*]

[1] Department of Biometry, Institute of Agriculture, Warsaw University of Life Sciences-SGGW, 159 Nowoursynowska Str., 02-776 Warsaw, Poland; elzbieta_wojcik-gront@sggw.edu.pl

[2] Department of Chemistry, Institute of Food Sciences, Warsaw University of Life Sciences-SGGW, 159C Nowoursynowska Str., 02-776 Warsaw, Poland; bartlomiej_zieniuk@sggw.edu.pl

[3] Department of Plant Genetics, Breeding and Biotechnology, Institute of Biology, Warsaw University of Life Sciences-SGGW, 159 Nowoursynowska Str., 02-776 Warsaw, Poland

* Correspondence: magdalena_pawelkowicz@sggw.edu.pl

Abstract: Artificial intelligence (AI) can revolutionize agriculture by enhancing genomic research and promoting sustainable crop improvement. AI systems integrate machine learning (ML) and deep learning (DL) with big data to identify complex patterns and relationships by analyzing vast genomic, phenotypic, and environmental datasets. This capability accelerates breeding cycles, improves predictive accuracy, and supports the development of climate-resilient, high-yielding crop varieties. Applications such as precision agriculture, automated phenotyping, predictive analytics, and early pest and disease detection demonstrate AI's ability to optimize agricultural practices while promoting sustainability. Despite these advancements, challenges remain, including fragmented data sources, variability in phenotyping protocols, and data ownership concerns. Addressing these issues through standardized data integration frameworks, advanced analytical tools, and ethical AI practices will be critical for realizing AI's full agricultural potential. This review provides a comprehensive overview of AI-powered genomic research, highlights the role of big data in training robust AI models, and explores ethical and technological considerations for sustainable agricultural practices.

Keywords: artificial intelligence; machine learning; deep learning; crop improvement; genomic study; big data

check for updates

Citation: Wójcik-Gront, E.; Zieniuk, B.; Pawełkowicz, M. Harnessing AI-Powered Genomic Research for Sustainable Crop Improvement. *Agriculture* **2024**, *14*, 2299. https://doi.org/10.3390/agriculture14122299

Academic Editor: Rodomiro Ortiz

Received: 29 November 2024
Revised: 12 December 2024
Accepted: 13 December 2024
Published: 14 December 2024

1. Introduction

Artificial intelligence (AI) is revolutionizing genomic selection and agriculture by enhancing the efficiency and precision of breeding programs. This integration of AI encompasses machine learning (ML) and deep learning (DL) techniques, which analyze vast datasets to identify patterns and relationships between genotypes and phenotypes. The advent of high-throughput genotyping technologies allows for collecting extensive genetic information across numerous markers, facilitating the development of predictive models that can significantly improve the selection of superior genotypes [1,2]. Genomic selection (GS) leverages these advancements by utilizing data from many genetic markers to estimate breeding values without pinpointing specific gene locations. This method, first articulated by Meuwissen et al. [3] in 2001, has gained traction as a powerful tool in crop improvement, particularly in response to the challenges posed by population growth and climate change [4]. By accurately predicting genetic potential, AI reduces the number of breeding cycles needed to develop new crop varieties, saving time, labor, and resources [3]. AI-driven genomic selection not only accelerates the breeding process but also enhances prediction accuracy, thereby addressing complex agricultural traits more effectively than traditional methods. As AI continues to evolve, its applications in agriculture promise to bridge the gap between food production demands and sustainable practices, ultimately leading to more resilient agricultural systems capable of meeting future challenges [5].

AI significantly enhances the efficiency of crop improvement techniques through various innovative applications that leverage data analysis, machine learning, and predictive modeling. Here are some key ways in which AI contributes to this field, like precision agriculture, accelerated breeding processes, predictive analytics, early disease and pest detection, and integration of genomics and phenomics. Researchers have used AI models to analyze genomic data and identify heat-tolerance genes in wheat, leading to the development of varieties that maintain yield under high temperatures [6]. AI techniques, such as Random Forest and Convolutional Neural Networks, have been employed to predict drought resilience in maize by integrating genomic and environmental datasets. These models have enabled the selection of drought-tolerant lines with a 30% yield improvement under water-scarce conditions [7]. In India, AI-driven genomic studies have identified salt-tolerance genes in rice, leading to the development of varieties capable of thriving in saline soils. This innovation has benefited farmers in coastal regions, improving food security and livelihoods [8]. AI facilitates precision agriculture by analyzing large datasets related to weather patterns, soil conditions, and crop health. This allows farmers to make informed decisions about irrigation, fertilization, and pest management, ultimately optimizing resource use and maximizing yields [9]. AI-driven tools used in rice and wheat breeding have reduced water usage by 20% while maintaining or increasing yields [10]. AI technologies also accelerate the breeding process by automating phenotyping, which involves observing and selecting the most promising crop varieties based on their growth characteristics. For instance, AI systems can analyze thousands of images of crops to identify traits that contribute to resilience and productivity [1,9]. AI-powered predictive analytics enable farmers to forecast crop yields and assess the impact of various factors on production. This capability helps in planning sowing and harvesting schedules more effectively, thereby enhancing overall productivity and profitability [1,2]. AI improves early disease and pest detection through image recognition and real-time monitoring, enabling proactive crop protection to minimize yield losses [11]. Moreover, AI links genomic data with phenotypic traits, allowing for rapid identification of genes associated with desirable traits. This integration accelerates the development of improved crop varieties adapted to specific environmental conditions [9]. AI-driven studies on wild relatives of staple crops, such as wild wheat and barley, have identified genes for disease resistance and abiotic stress tolerance. For instance, genes from wild barley have been used to improve resistance to leaf rust in cultivated varieties, safeguarding yields against fungal diseases [12]. Overall, the incorporation of AI into crop improvement techniques not only enhances efficiency but also contributes to sustainable agricultural practices by optimizing resource use and minimizing waste.

The aim of this review article is to provide a comprehensive overview of the integration of AI in genomic research for sustainable agriculture. This paper examines the role of big data in training robust AI models that enhance crop improvement and discusses the ethical implications of using AI.

2. Principles of AI—Predictive Modeling

Predictive modeling has evolved from simple statistical tools into a sophisticated field powered by machine learning and deep learning. Beginning with foundational methods like linear and logistic regression, the discipline matured through the advent of decision trees, ensemble methods, and deep neural networks. These advancements have expanded the ability to analyze complex, high-dimensional datasets, leading to genomic selection and agricultural application breakthroughs. By understanding the historical context and examining the principles, strengths, and limitations of various predictive algorithms, readers can appreciate how each method contributes to the broader tapestry of data-driven decision-making. The progression of predictive modeling—from early statistical methods to advanced machine learning and deep learning frameworks—highlights the field's transformative potential. Today's algorithms can handle diverse data types, identify intricate patterns, and offer more accurate predictions. Techniques like cross-validation ensure that models are reliable and generalizable, while methods such as autoencoders and VAEs facilitate data exploration and hypothesis generation. Ultimately, this chapter underscores how predictive modeling is a key driver in modern research, enabling more informed decisions in agriculture, genomics, and beyond and setting the stage for further innovation.

2.1. Advanced Predictive Algorithms

The concept of predictive modeling traces its origins to early statistical techniques developed in the 20th century, such as linear regression and logistic regression [13]. These methods laid the foundation for predictive analytics by establishing relationships between variables to forecast outcomes. The concept of decision trees can be traced back to earlier statistical methods, such as discriminant analysis proposed by Ronald Fisher in 1936 [14]. However, it was not until the late 1970s that researchers began formalizing decision tree methodologies. In the 1960s, advancements in computing enabled the development of decision tree algorithms, such as the Classification and Regression Tree (CART) introduced by Breiman et al. in 1984, which became an important moment in predictive modeling [15]. CART utilizes a binary tree structure where each internal node represents a decision based on a predictor variable, leading to branches that represent possible outcomes. The terminal nodes (or leaves) contain the predicted outcomes for classification or regression tasks. This recursive partitioning approach helps create models that can effectively capture complex relationships within data. The emergence of ML in the late 1990s marked a significant leap forward. Algorithms such as Support Vector Machines (SVMs) [16] and ensemble learning techniques like Random Forest [17] introduced a level of predictive accuracy and robustness previously unattainable with traditional methods. SVMs are designed to find the optimal hyperplane that separates data points of different classes in a high-dimensional space. The algorithm focuses on maximizing the margin between the closest data points of each class, known as support vectors. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions (for classification) or the mean prediction (for regression). Machine learning models, such as Random Forest and deep learning architectures, have demonstrated their potential to revolutionize crop yield prediction. These models offer significantly more accurate forecasts by analyzing diverse factors like soil properties, climate, and historical yield data than traditional methods like regression trees [18,19]. This approach helps improve accuracy and control overfitting, making it particularly effective for complex datasets. The 2010s saw the rise in deep learning powered by advances in computational capabilities and the availability of large datasets. Neural networks evolved into multi-layered architectures capable of capturing intricate patterns in complex datasets, revolutionizing predictive modeling. Techniques like Convolutional Neural Networks (CNNs) for image data [20] and Recurrent Neural Networks (RNNs) for sequential data [21] emerged, paving the way for applications in genomics, healthcare, and beyond. CNNs were designed for handwritten digit recognition. This model utilized convolutional layers to preserve the spatial structure of images, allowing for more effective feature extraction compared to traditional neural networks that flattened input data. RNNs are a solution for processing sequential data. Unlike traditional feedforward networks, RNNs maintain a hidden state that captures information from previous inputs, allowing them to consider temporal dependencies. AI has fundamentally transformed predictive modeling by enabling the analysis of vast, high-dimensional datasets. Unlike traditional methods, AI models can learn non-linear relationships, interactions between features, and hierarchical structures, which are particularly useful in fields like genomics and agriculture [22]. Machine learning algorithms, including Random Forest, Gradient Boosting Machines (GBMs), and XGBoost (Extreme Gradient Boosting), have become staples in predictive modeling due to their ability to handle missing data, non-linear relationships, and feature importance ranking [23]. GBMs build models sequentially, where each new tree corrects errors made by previously trained trees. This technique focuses on minimizing the loss function by optimizing the model iteratively. XGBoost is an advanced implementation of gradient boosting that enhances speed and performance. It incorporates regularization techniques to prevent overfitting and optimize model performance. These models are widely applied in genomic selection to predict plant traits, such as drought resistance or yield, based on genetic markers [3]. Deep learning has further advanced predictive modeling by handling complex relationships that were previously difficult to model. Architectures like CNNs excel in processing spatial data, such as genomic sequences, while RNNs are adept at modeling time-series data, such as environmental fluctuations. Deep learning, in particular, excels at processing time series data and integrating various environmental and soil parameters [24]. This capability has led to promising results in predicting crop yields [25]. Accurately anticipating crop yields is crucial for optimizing agricultural practices and ensuring food security. Autoencoders and Variational Autoencoders (VAEs) are now being employed to reduce dimensionality and generate new hypotheses about genetic and phenotypic traits [26]. Autoencoders are neural network architectures that learn to compress data into a lower-dimensional latent space and then reconstruct the original data. This ability to reduce dimensionality is particularly useful in genomics, where datasets often contain a vast number of features, such as single nucleotide polymorphisms [27]. VAEs extend the concept of autoencoders by introducing a probabilistic approach to encoding data. They learn a distribution over the latent space rather than a fixed representation, enabling the generation of new samples that resemble the training data. This capability is particularly advantageous for generating hypotheses about genetic traits and simulating potential phenotypic outcomes based on genetic variations [28]. VAEs have been shown to capture population structures and genetic diversity effectively. They can identify clusters of samples with similar genetic compositions, facilitating the exploration of genotype-phenotype relationships across diverse populations. For example, VAEs have been applied to analyze genomic data from various populations, providing insights into how genetic variations correlate with phenotypic traits [29]. The information on the main predictive algorithms has been compiled in Table 1.

Table 1. The features, principles, advantages, and limitations of key predictive algorithms.

Algorithm	Principle	Key Features	Advantages	Limitations
CART (Classification and Regression Trees)	Recursive binary splitting of data based on feature thresholds.	Decision tree structure with internal nodes representing feature splits and leaf nodes with predictions.	Simple, interpretable, and useful for categorical and continuous outcomes.	Prone to overfitting and sensitive to small changes in the data.
SVM (Support Vector Machine)	Maximizes the margin between data classes using a hyperplane in high-dimensional space.	Utilizes support vectors (critical data points) to define the optimal hyperplane.	Effective in high-dimensional spaces and works well for both classification and regression.	Computationally expensive for large datasets, sensitive to kernel choice, and challenging to interpret.
RF (Random Forest)	Ensemble method that builds multiple decision trees and aggregates their predictions.	Uses bootstrapping and random feature selection to reduce variance and overfitting.	Handles large datasets, robust to overfitting, and provides feature importance.	Slower prediction time compared to single trees, and reduced interpretability.
CNN (Convolutional Neural Network)	Extracts spatial hierarchies from grid-like data (e.g., images) using convolutional and pooling layers.	Employs convolutional, pooling, and fully connected layers for feature extraction.	Excellent for image, spatial, and sequential data. Automatic feature extraction.	Computationally intensive, requires large datasets, and often acts as a "black box".

Table 1. *Cont.*

Algorithm	Principle	Key Features	Advantages	Limitations
RNN (Recurrent Neural Network)	Processes sequential data with loops that allow information to persist through "hidden states".	Captures temporal dependencies and processes data with order dependencies.	Ideal for time-series, sequential, and text data. Effective in capturing time-dependent relationships.	Suffers from vanishing/exploding gradient issues, and training can be slow for long sequences.
VAE (Variational Autoencoder)	Encodes data into a probabilistic latent space and decodes to generate synthetic samples.	Probabilistic encoder-decoder model with a latent variable space.	Generates new data samples, useful for dimensionality reduction, and unsupervised learning.	Requires careful tuning of latent space size, and reconstructions may be blurry for image data.

2.2. Cross-Validation and Optimization

Cross-validation is a crucial technique in machine learning that enhances model evaluation and optimization by providing a robust framework for assessing how well a model generalizes to unseen data. While cross-validation and optimization are widely used in machine learning, they are not exclusive to this field. These techniques are also employed in other computational and statistical disciplines, such as bioinformatics, econometrics, and operations research. It involves partitioning the dataset into multiple subsets, allowing for iterative training and testing of the model. Using different subsets for training and testing provides a better understanding of how the model will perform on unseen data. It reduces the risk of overfitting, where a model performs well on training data but poorly on new data. Cross-validation ensures that the model is tested on various data points, leading to a more reliable estimate of its generalization capabilities. A common form of cross-validation is k-fold cross-validation. The dataset is divided into k equal-sized folds. The model is trained k times, each time using k − 1 folds for training and 1-fold for testing [30]. In a genomic prediction study on maize (*Zea mays* L.), researchers employed a 5-fold cross-validation approach to assess the performance of statistical learning methods that used genome-wide molecular marker data to predict genetic values of target traits [31]. Each model was trained on four folds of data (training set) and validated on the remaining one-fold (test set). This process was repeated five times, with each fold serving as the test set once. By comparing prediction accuracy across all folds, the study demonstrated the robustness of the methods and gained insights into how well these models generalized to unseen data. However, common challenges are encountered when applying k-fold cross-validation in machine learning, especially in the context of genomic research. Uneven distribution of phenotypic classes can affect model training and validation. Repeating the training and validation process k times increases computational demands, especially with large genomic datasets. If proper partitioning protocols are not followed, information may leak from the training set to the testing set, leading to overly optimistic performance estimates. Overfitting may still occur if hyperparameters are tuned using information from multiple folds, requiring careful separation of training, validation, and testing datasets.

3. AI in Genomic Research

ML is a field of computer science that aims to develop algorithms that allow computers to learn and draw conclusions from available data without being directly programmed [32]. In the context of genomic data analysis, ML makes it possible to automate complex computational processes, such as identifying patterns in DNA sequence data, significantly accelerating the development of plant genomics research.

3.1. Types of Machine Learning in Genomic Study

Machine learning algorithms are generally divided into two main types: supervised and unsupervised learning [33] (Figure 1). Supervised learning involves algorithms such as random forests, support vector machines, and k-nearest neighbors, which use labeled data to classify instances or predict numerical values (regression). In contrast, unsupervised learning algorithms like principal component analysis, k-means clustering, and self-organizing maps do not require labeled data and are mainly used for clustering and feature extraction. Examples of applications include the prediction of regulatory and non-regulatory regions in the maize genome [34], prediction of mRNA expression levels [35], polyadenylation sites identification in Arabidopsis [36], classification of macronutrient deficiencies on development in tomato [37].

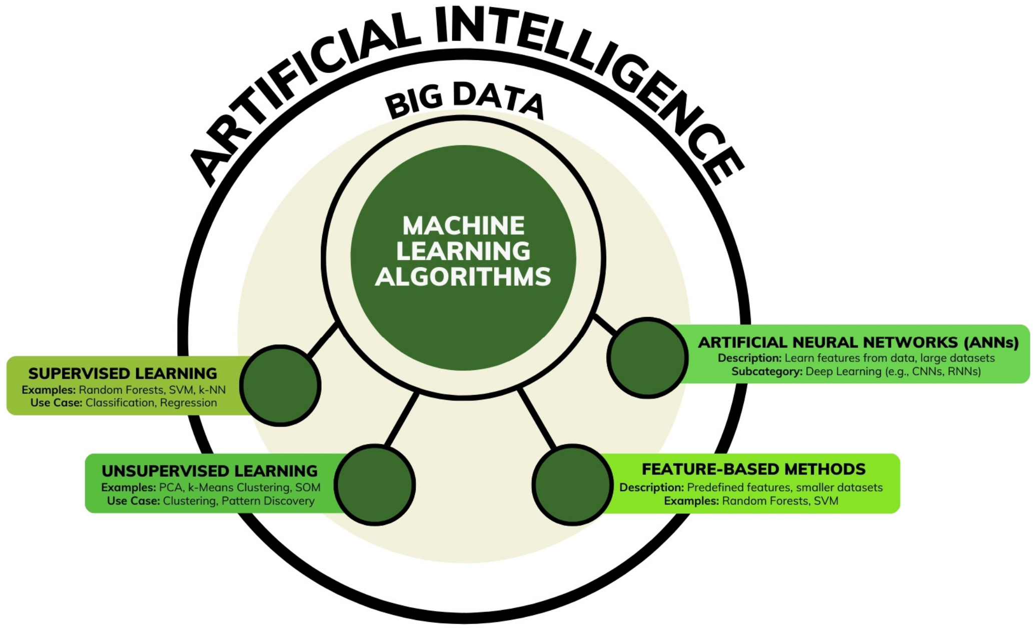
ARTIFICIAL INTELLIGENCE
BIG DATA
MACHINE LEARNING ALGORITHMS
SUPERVISED LEARNING
Examples: Random Forests, SVM, k-NN
Use Case: Classification, Regression
ARTIFICIAL NEURAL NETWORKS (ANNs)
Description: Learn features from data, large datasets
Subcategory: Deep Learning (e.g., CNNs, RNNs)
UNSUPERVISED LEARNING
Examples: PCA, k-Means Clustering, SOM
Use Case: Clustering, Pattern Discovery
FEATURE-BASED METHODS
Description: Predefined features, smaller datasets
Examples: Random Forests, SVM

Figure 1. Machine learning algorithm types and methods.

Additionally, machine learning methods can be categorized based on their approach into feature-based methods and artificial neural networks (ANNs). Feature-based methods, such as random forests and SVMs, require predefined features and are effective with smaller datasets. ANNs, including convolutional and recurrent neural networks, automatically learn features from large datasets, making them suitable for tasks requiring substantial amounts of data. When ANNs consist of multiple interconnected layers of neurons, they form a deep learning model, widely used in complex applications such as image and speech recognition [38]. Deep neural networks (DNNs), an extension of ANNs, contain multiple hidden layers, allowing them to analyze more complex patterns in the data. However, their greater predictive power requires more data resources and computing power. A subset of DNNs are CNNs, which automatically extract features from continuous data such as plant images or DNA/RNA sequences. By using spliced layers, CNNs are applied to tasks such as identifying phenotypic features of plants from images or analyzing motifs in genomic sequences. Examples include predicting regulatory sites in genomes with large numbers of repetitive sequences, such as maize [39,40]. The introduction of advanced machine learning techniques, such as CNNs and DNNs, opens up new possibilities in plant genomics (Table 2). Using these methods makes it possible to better understand the relationship between DNA sequences and phenotypes, which helps improve research into improving crop yield and resistance.

Table 2. The overview of the AI-powered tools in genomic research discussed in this review.

Task	AI Application
Genome assembly	AI improves accuracy in assembling complex genomes
Gap filling in genome assembly	AI better predicts missing genome fragments
Genome size estimation	AI defines the computational assessment of genome size
Structural variant detection	AI indicates large genomic variations
Functional annotation of plant genes	AI predicts and compiles coding domains and TFBS
Cis-regulatory element prediction	AI analyses of gene expression flanking regions
Prediction of TF binding sites	AI determines TFBS * prediction and cell-specific interactions
Annotation of regulatory regions	AI models long-range DNA sequence relationships
Genetic variation analysis	AI links phenotypic traits with genetic markers
CRISPR target site optimization	AI design of gRNAs for gene editing

* TFBS—Transcription Factor Binding Site.

3.2. Genome Assembly, Structure and Function

Next-generation sequencing technologies have significantly deepened our knowledge of genomes as it has become possible to sequence them. The sequencing results had to be assembled into a continuous genome. Genome assembly, the process of assembling short DNA fragments into complete genomic sequences, is crucial to understanding the genome structure of organisms [41]. It enables the identification of genes, regulatory elements, and genetic variants, with applications in medicine, agriculture, and biodiversity conservation [42]. AI greatly improves this process by providing more accurate and faster data analysis methods. AI algorithms can predict gene function, identify regulatory elements, and optimize genome assembly. For example, AI enables the better interpretation of sequencing data and integration of different types of biological data for more accurate genome mapping [43]. The use of AI in genome assembly increases the efficiency and accuracy of genetic element identification and enables the analysis of large and complex genomes, such as plant genomes and highly repetitive organisms. It is a groundbreaking tool for basic and applied research [44]. The use of neural networks and k-mers in plant genome assembly is an advanced method used in genomics to analyze large biological datasets. K-mers are commonly used in bioinformatics to analyze genomic data. They are obtained by sliding a window of length k over a sequence and recording all the overlapping subsequences. K-mers are essential for tasks such as sequence alignment, genome assembly, and error correction [45]. A k-mer is a substring of a given sequence of length k, where k refers to the number of characters (or nucleotides) in the substring. For example, in a DNA sequence, if k = 3, then a possible k-mer from the sequence "AGCT" would be "AGC" and "GCT" [45,46]. The size of k affects the specificity of the analysis, with smaller k-mers being more general and larger k-mers being more specific, helping to distinguish between different sequences or species [46].

K-mers in assembly are used to construct de Bruijn graphs, which help in assembling DNA sequences by identifying overlapping genome fragments. This technique facilitates accurate mapping of the plant genome containing repetitive sequences and a large number of regulatory elements [47,48]. In cucumber B10 v3, the genome was first assembled from PacBio reads and then corrected using the short Illumina reads from P1. The BBnorm Ecc Linux from the BBTools v35.82 suite was used to correct the quality of the Illumina reads by k-mer distribution count-based modification [49]. K-mer-based GWAS has emerged as a method to assess genetic variation in plants without the need for a reference genome. For example, a study on *Aegilops tauschii*, which lacks a reference genome, used k-mers derived from sequencing data enriched for nucleotide-binding/leucine-rich repeat (NLR) genes. Significant k-mers were mapped to local assemblies of NLR genes, identifying candidate genes associated with resistance to wheat stem rust. This method demonstrated the efficiency of k-mers in linking genetic variations directly to phenotypic traits [46]. K-mer distributions have been employed to estimate genome sizes across various organisms, including cultivated potatoes and agricultural pests. In one study, researchers compared k-mer methods with traditional flow cytometry measurements and found that k-mer analysis could provide accurate estimates of genome size when applied correctly. They noted that using different values of k (e.g., 21) offered a balance between computational efficiency and accuracy in estimating genome sizes [50]. K-mers can also be used to detect genetic variants such as single nucleotide polymorphisms (SNPs), insertions, deletions, and structural variations without prior knowledge of these variants. This capability is particularly useful for species lacking comprehensive genomic information, enabling researchers to identify significant genetic markers associated with traits of interest [46]. Algorithms based on k-mers and ML enable the classification of genomic elements such as LTR retrotransposons. This technique has been successfully applied to the analysis of plant genomes, allowing for a better understanding of their structure [47]. DNN-based tools, such as GapFiller, use k-mer techniques to identify and fill missing fragments in the genome. The model analyzes sequence reads and predicts the most likely sequences in gap areas [51]. These technologies facilitate the analysis of highly complex plant genomes and open the way to more precise molecular studies. The combination of neural networks and K-mer analysis represents a significant advancement in genome assembly methodologies. Researchers can achieve more accurate and complete genome reconstructions by leveraging the strengths of both approaches—neural networks for complex pattern recognition and K-mers for efficient data representation. As these technologies continue to evolve, they promise to further transform our understanding of genomic structures and functions across diverse species. A comparison of CNN and k-mer methods shows that CNNs are more efficient at extracting sequence features, but their interpretation remains difficult, and the computational cost is high. Although powerful, CNNs are often criticized as 'black boxes', limiting their use when biological rules need to be elucidated. Alternatively, k-mer-based methods, which rely on frequency analysis of short sequence segments, are fast, accurate, and easy to interpret. They are particularly effective in identifying sequence signatures and allow quantitative comparison of sequence similarity. Examples of combining k-mer approaches with deep learning demonstrate the potential for synergy between predictive performance and interpretability of results [52]. ML and neural networks are widely used in the analysis of plant gene structures, contributing to the advancement of plant genomics and a better understanding of their biological mechanisms. Deep learning models predict plant gene expression patterns by analyzing the cis-regulatory code, e.g., flanking regions of genes. This allows the identification of key regulatory elements that affect gene expression, as it was conducted in *A. thaliana*, *S. lycopersicum*, *S. bicolor*, and *Z. mays* [53]. ML enables the identification of structural regions in the plant genome, such as coding domains or transcription regulators, supporting the functional annotation of genes [54]. ML supports the precise design of gene editing target sites in plants, which is useful in the design of gRNAs for CRISPR-Cas9 techniques [51]. Deep learning is used to analyze the structure of crop genomes to improve their phenotypic traits, such as resistance to environmental stresses and increased crop yield [22]. Neural networks analyze DNA sequencing data to detect and classify large structural variants in the plant genome, allowing for a better understanding of genetic diversity [55]. Neural networks combined with k-mer analysis, such as the kmerPMTF model, are used to analyze the interactions of miRNAs and their targets, which supports the interpretation of regulatory functions in the plant genome [56]. These technologies are crucial for studying plant genomes, predicting plant gene expression from regulatory sequences, characterizing gene function using ML, recognizing key genetic locations for gene editing, genomics modeling for crop improvement, and discovering structural genetic variants. AI will revolutionize breeding technologies, especially for those crop species with complex genomic structures [57,58].

3.3. Transcription Factor Binding Sites Studies

To analyze large datasets in plant genomics, several deep learning-based methods have been developed to model the specificity of transcription factor (TF) binding to DNA. For example, DeepBind allows the identification of multiple sequence motifs to predict the binding sites of DNA accurately- and RNA-binding proteins [59]. TFImpute optimizes the prediction of cell-specific TF interactions [60], while algorithms such as DeepSEA [61], DeFind [62], and DFIM [63] assess the impact of functional non-coding variants, which is crucial in the context of genome regulatory functions. In particular, DRNApred effectively distinguishes between DNA and RNA binding residues [64]. ML plays a key role in modeling TF binding sites in plant genomics. These models can be trained using different types of sequencing data, either alone or in combination with other sources, such as DNase I hypersensitivity data, significantly improving the prediction of TFBS binding in vivo [60]. Plant genomes have numerous repetitive elements and large intergenic regions, the identification of key regulatory regions is more challenging. To overcome these difficulties, approaches based on k-mer grammar and natural language processing have been used to accurately and cost-effectively annotate regulatory regions [34]. Further advances in this area include integrating methods that will, in the future, be able to model long-range relationships in DNA sequences, which could prove critical in studies of plants with large genomes, such as wheat or soybean. Initiatives such as PlantENCODE are attempting to address these issues by collecting extensive data on regulatory regions and TF binding, which will enable the training of more comprehensive AI models [65,66].

To sum up, next-generation sequencing has revolutionized genome analysis, enabling deeper insights into genome structure, gene identification, and regulatory elements. AI has further transformed genome assembly by enhancing the accuracy and efficiency of sequencing data interpretation, especially for complex genomes like plants. Deep learning models have advanced genome mapping, gene function prediction, and structural variant identification, transcription factor binding site prediction. This fusion is particularly impactful in agriculture, where it supports crop improvement, resilience, and phenotypic trait optimization. The AI-based methods are key to unraveling plant genome regulation, with continuous advancements enhancing precision, scalability, and applicability to large and complex genomes.

4. Data Challenges

As agricultural data's scale, complexity, and diversity continue to grow, integrating big data into AI model training becomes both a critical opportunity and a formidable challenge. In crop breeding programs, massive genomic, phenotypic, and environmental datasets can potentially transform how we develop high-yielding, stress-resilient varieties. By effectively harnessing these data, stakeholders can gain deeper insights, enhance predictive accuracy, and make more informed decisions. However, achieving this vision requires overcoming technical, infrastructural, and ethical hurdles. The following sections explore the power of big data in AI model training and the barriers that must be addressed to realize its full benefits in sustainable agriculture. This chapter underscores that while big data offers unparalleled opportunities for advancing AI-driven crop breeding programs, it also introduces complex challenges. From integrating heterogeneous datasets and ensuring data quality to establishing ethical guidelines and improving infrastructure, each step in harnessing big data demands careful attention. Researchers can unlock valuable genetic insights by adopting FAIR data principles, investing in high-throughput phenotyping platforms, and fostering open-access collaborations. Ultimately, overcoming these obstacles will pave the way for more robust, reliable, and responsive AI models—driving innovation in crop breeding, meeting global food demands, and contributing to the long-term sustainability of agriculture.

4.1. The Role of Big Data in AI Model Training

The integration of big data into artificial intelligence model training is essential for advancing the capabilities of AI systems. Big data's unique characteristics enable AI models to achieve greater performance, accuracy, and applicability across diverse fields. Big data encompasses vast amounts of information generated from diverse sources such as genomic sequences, phenotypic records, environmental data, and agricultural management practices. For instance, genomic datasets such as those from the International Wheat Genome Sequencing Consortium (IWGSC) [67] allow researchers to train AI models for identifying yield-enhancing genotypes across different environments [68]. The integration of heterogeneous datasets helps AI systems uncover complex gene-environment interactions critical for breeding climate-resilient crops. These extensive datasets allow AI systems to train on a wide array of scenarios, capturing nuances in patterns that improve generalizability and predictive performance. For instance, the use of multi-omics data in AI-driven genomics has enabled more accurate identification of gene-disease associations and trait prediction [69]. The availability of big data allows AI algorithms to recognize intricate and non-linear relationships that may be overlooked in smaller datasets. AI algorithms trained on big data can identify intricate patterns and correlations often missed in smaller datasets. This is particularly valuable in genomics, where understanding subtle genetic relationships can lead to breakthroughs in disease resistance and stress tolerance. Deep learning models, such as CNNs, have been used to analyze genomic sequences for motif detection and structural variations that influence crop traits [61]. Big data improves the robustness of AI models by exposing them to diverse scenarios, including rare and extreme conditions. This exposure is particularly important in agriculture, where environmental variability and uncertainty demand resilient models. Multi-environment trials (METs), which capture data from diverse agro-climatic zones, provide the foundation for training robust AI systems to predict crop performance in different conditions [70]. The accuracy of AI models is highly dependent on the quality and quantity of training data. In genomics, high-throughput phenotyping and genotyping generate large, high-quality datasets that enable precise predictions of complex traits such as yield, flowering time, and disease resistance. AI models, such as Random Forest and GBM, have demonstrated significant improvements in predicting polygenic traits in crop breeding programs [7]. Big data technologies, combined with AI, enable real-time analysis and decision-making. For example, AI-driven platforms analyze sensor and satellite data to monitor crop health, predict pest outbreaks, and recommend timely interventions. Real-time insights are vital in applications such as irrigation scheduling and pest control, where immediate actions can prevent significant yield losses [71]. By improving big data analytics, agricultural stakeholders can make data-driven decisions that optimize resource use and enhance productivity. For example, AI models that integrate climate and soil data with genomic information can recommend site-specific crop varieties and management practices, maximizing yield and sustainability [22]. The combination of big data and machine learning allows continuous model improvement and adaptability. Machine learning models thrive on large datasets, learning to handle variability and complexity without explicit programming. In genomic selection, AI models trained on big datasets can predict breeding values more effectively than traditional statistical methods [3]. The integration of AI with big data enables advanced analytics, moving beyond descriptive insights to predictive and prescriptive solutions. Predictive analytics helps forecast future agricultural outcomes, such as yield potential and disease risk, while prescriptive analytics provides actionable recommendations for achieving optimal results.

4.2. Addressing Data Challenges in Crop Breeding Programs

Data collection and quality are foundational for success in crop breeding programs, significantly influencing the ability to develop high-yielding, stress-resilient crop varieties. However, several limitations impede the effective utilization of data, presenting challenges to discovering the full potential of big data and AI-powered genomic research for sustainable crop improvement. Crop breeding programs generate vast datasets from diverse sources, including genomic sequencing, field trials, and environmental monitoring. However, these datasets are often siloed across disparate platforms, limiting the ability to integrate and analyze them comprehensively [70]. The lack of unified data management systems hinders decision-making processes and restricts insights that could otherwise inform breeding strategies. Adopting FAIR (Findable, Accessible, Interoperable, and

Reusable) data principles to improve interoperability and streamline data integration [72]. Cloud-based platforms like CyVerse provide scalable solutions for integrating and analyzing fragmented datasets, enabling breeders to access comprehensive information in real time [73]. Accurate phenotypic data are critical for genomic selection and advanced breeding techniques, yet many crops lack high-quality phenotypic datasets. Variability in phenotyping protocols, inconsistent methodologies, and high labor costs impede the standardization and scalability of phenotyping processes [74]. High-throughput phenotyping platforms, such as drones and automated imaging systems, are revolutionizing data collection by enabling precise, scalable, and consistent measurements [75]. Machine learning models are being employed to analyze phenotypic data, reducing the reliance on manual assessment and increasing reproducibility. Many crops, particularly underutilized species and minor crops, lack comprehensive genomic resources. This scarcity limits breeders' ability to perform genomic selection or identify valuable traits [76]. Initiatives such as the African Orphan Crops Consortium (AOCC) and the CGIAR Excellence in Breeding Platform are addressing these gaps by developing genomic databases for neglected crops, expanding the scope of breeding programs [10,76]. Advances in next-generation sequencing and genotyping-by-sequencing have reduced the cost and time required to generate genomic data, accelerating the development of these resources [77]. The rise in digital agriculture has heightened concerns about data ownership, intellectual property rights, and accessibility. Breeders often face legal and ethical challenges when sharing or utilizing data collected from multiple sources, creating barriers to collaboration [78]. Establishing clear guidelines and agreements for data sharing to promote cooperation while respecting intellectual property rights. Open-access initiatives like the International Maize and Wheat Improvement Center (CIMMYT) [79] provide freely accessible genomic and phenotypic datasets, fostering global collaboration [7]. Modern breeding programs generate enormous volumes of data, requiring sophisticated analytical tools for integration and interpretation. Many programs lack the necessary infrastructure, computational power, or expertise to handle big data effectively. AI-powered platforms like DeepTools and TensorFlow enable the analysis of large-scale datasets, improve the accuracy of predictions, and reduce time-to-insight [80,81]. Investments in training programs for researchers and breeders are critical to building capacity for handling and interpreting big data. Environmental factors play a significant role in phenotypic expression, yet capturing this variability accurately in datasets is challenging. Variations in climate, soil, and management practices across trials can obscure genotype-phenotype relationships, complicating analysis [82]. Incorporating multi-environment trials (METs) into breeding programs to account for diverse agro-climatic conditions. AI models trained on integrated datasets that combine genomic, phenotypic, and environmental data can better account for variability, leading to more reliable predictions [7].

To summarize the above, big data integration is crucial for advancing AI-driven crop breeding and ensuring agricultural sustainability. Leveraging FAIR data principles, high-throughput phenotyping platforms, and collaborative initiatives enable researchers to overcome challenges like data fragmentation, limited genomic resources, and ethical concerns. By improving data quality and infrastructure, researchers can unlock complex gene-environment interactions and develop robust AI models for precise predictions and adaptive solutions. This synergy between big data and AI accelerates innovation, driving productivity, resilience, and sustainability in agriculture.

5. Future Perspectives

5.1. Addressing Ethical Concerns and the Role of AI in Sustainable Agriculture

AI represents a transformative force in agriculture, integrating various techniques such as DL, reinforcement learning, and ML. This evolution, particularly pronounced from the 1950s to the present, has significantly impacted agricultural practices, enhancing productivity and sustainability [83]. However, the integration of AI also brings forth critical ethical, social, and economic concerns that necessitate careful consideration to ensure sustainable agricultural practices. Jobin et al. [84] highlighted the significant debates that persist regarding the definition of ethical AI and the specific standards and practices required for its implementation. The five key ethical principles are privacy, transparency, justice and fairness, non-maleficence, and responsibility. The ongoing discourse around ethical AI aims to establish universal principles while also developing practical frameworks to ensure the effective implementation of these principles. Aldoseri et al. [85] draw attention to the vast volume and diversity of data sources, making it difficult to ensure that relevant and representative samples are gathered and, moreover, may encompass personal and sensitive information. Additionally, robust data management practices are necessary to maintain data integrity and security, including implementing version control and efficient storage systems [85]. Moreover, farmers must have full control over their data, including how it is collected, who can access it, and how it is utilized. This control is essential for protecting their privacy and ensuring their information is used responsibly. In the United States, farmers have formed cooperatives to collectively own and control their data from various applications. The Farmers Business Network (FBN) allows farmers to share data while retaining ownership rights, prioritizing privacy, and ensuring farmers control how their data are used. The Ag Data Transparent certification outlines FBN's principles on data ownership and privacy, stating that it will not sell or disclose non-aggregated farm data without a legally binding commitment to protect farmers' rights [86]. Transparency is another fundamental principle of ethical AI that fosters trust among stakeholders by promoting openness about policies, actions, and laws. In agriculture, a lack of transparency can hinder farmers' willingness to adopt AI solutions or share their data with technology providers. There is a need for clear communication regarding how AI models operate and make decisions. This transparency is crucial for building trust among farmers and consumers alike [87]. IBM's Watson Decision Platform enhances transparency by showing farmers how AI models make decisions based on data inputs. For example, when recommending crop management practices, it explains the algorithms and data used, helping farmers understand the reasoning behind these recommendations. This fosters trust and encourages greater adoption of AI technologies [88]. Fairness, justice, and equity issues in AI-driven agriculture are critical to ensuring equitable access to technology and preventing bias that can disadvantage marginalized farmers. As AI applications expand, it is essential to monitor and mitigate biases in data and algorithms, which can lead to unequal opportunities and exacerbate existing inequalities. Implementing diverse training data and inclusive design practices can help create fairer AI systems that cater to the unique needs of smallholders and underrepresented communities [89]. Non-maleficence, which means "do no harm", is a vital principle in the use of AI technologies in agriculture. It emphasizes the importance of minimizing any potential negative impacts on individuals, communities, and ecosystems. This requires conducting thorough risk assessments, addressing biases in AI models, and ensuring that those implementing these technologies have the necessary expertise. By prioritizing non-maleficence, AI applications can improve agricultural practices while protecting environmental integrity and food safety. Ultimately, this principle helps build trust among farmers and consumers and supports sustainable agricultural development [90]. In terms of the application of AI in agriculture responsibility, (accountability) is a term used for establishing frameworks that clarify who is responsible for decisions made by AI systems, which is vital. Without clear legal agreements outlining the responsibilities of developers, users, and stakeholders, it becomes difficult to hold anyone accountable for financial or reputational losses that may arise from errors. Additionally, oversight by legal entities is essential to ensure that the terms of use and data agreements are ethical and protect all parties involved [87]. Analyzing human and social issues, the use of AI in agriculture can automate (and already does) certain tasks, which may impact the labor force by displacing jobs. It is essential to recognize the social consequences of this displacement and to implement reskilling and upskilling initiatives to help workers transition smoothly and reduce negative effects [91]. In response to job displacement caused by automation, initiatives such as the Farmworker Jobs Program in

California aim to reskill agricultural workers for new roles arising from AI technologies. These programs offer training in areas like data analytics and machine operation, enabling workers to transition smoothly into new positions and helping to mitigate the negative social impacts of job loss [92]. AI can greatly improve agriculture by enhancing efficiency and sustainability. However, it is critical to address the ethical implications associated with its use. By focusing on privacy, transparency, fairness, and responsibility, stakeholders can ensure that AI contributes positively to agriculture while upholding ethical standards and social equity.

5.2. Prospects for Integrating AI Across Large-Scale Breeding Programs Worldwide

The integration of AI into large-scale breeding programs represents a transformative opportunity to enhance agricultural productivity and sustainability. This advancement is particularly relevant in light of the increasing global food demand, human population, and the challenges posed by climate change. By streamlining processes, enhancing precision, and accelerating the development of superior cultivars, AI holds immense potential to revolutionize the agricultural industry. AI is revolutionizing modern crop breeding, providing significant opportunities for advancements in plant science. AI technologies are remarkably useful because they enable the analysis of vast datasets generated from genomic, phenomic, and environmental sources. For instance, AI can facilitate high-throughput phenotyping, which allows breeders to rapidly assess plant traits and link them to genetic information [11]. This capability is essential for overcoming longstanding obstacles in breeding, particularly the challenge of connecting genotype to phenotype. By leveraging AI to automate data collection and analysis, breeders gain the ability to make informed decisions rapidly [93–95]. This not only accelerates the development of new crop varieties but also ensures these varieties are robust and resilient to environmental stresses, meeting the demands of a changing world [94]. Interestingly, the combination of speed breeding techniques with AI can dramatically shorten breeding cycles. Speed breeding involves growing plants under controlled conditions to accelerate their life cycles. AI enhances this process by efficiently managing complex datasets from various omics disciplines (genomics, transcriptomics, etc.), understanding the biological mechanisms that influence plant functions, and facilitating the successful implementation of speed breeding protocols aimed at improving crop yield and adaptability [94]. The prospects for integrating AI into large-scale breeding programs appear promising. Continued advancements in machine learning algorithms and computational power will likely lead to even more sophisticated applications in crop improvement. Additionally, as consumer demand shifts towards transparency and sustainability in food production, AI-driven approaches will be critical in developing breeding strategies that meet these expectations while addressing environmental concerns [96]. The integration of AI into agricultural breeding programs not only enhances productivity but also supports sustainability efforts in response to global challenges. By embracing these technological advancements, the agricultural sector can better equip itself to meet future food demands while ensuring environmental stewardship.

6. Conclusions

In conclusion, incorporating AI into genomic research and agricultural practices is not merely an enhancement of existing methods. It is a revolutionary shift that holds the potential to redefine food security and sustainability in agriculture. Continued research and development in this domain will be crucial for maximizing the benefits of AI technologies, ensuring that they are effectively harnessed to meet future agricultural challenges while promoting ecological balance and resource conservation.

As AI technology continues to evolve, its role in sustainable agriculture will likely expand, bridging the gap between food production demands and sustainable practices. Combining big data with AI enables robust model training that captures intricate gene-environment interactions essential for breeding climate-resilient crops. This synergy improves crop yield predictions, optimizes resource use, and minimizes waste.

Author Contributions: Conceptualization, M.P.; writing—original draft preparation, E.W.-G., B.Z. and M.P.; supervision, M.P.; project administration, M.P.; funding acquisition, M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a project from the National Science Center UMO-2020/37/B/NZ9/00586.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing does not apply to this article.

Acknowledgments: We are very grateful to Marcin Ollik from Warsaw University of Life Sciences-SGGW for the critical review of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bose, S.; Banerjee, S.; Kumar, S.; Saha, A.; Nandy, D.; Hazra, S. Review of Applications of Artificial Intelligence (AI) Methods in Crop Research. *J. Appl. Genet.* **2024**, *65*, 225–240. [CrossRef] [PubMed]
2. Jubair, S.; Domaratzki, M. Crop Genomic Selection with Deep Learning and Environmental Data: A Survey. *Front. Artif. Intell.* **2023**, *5*, 1040295. [CrossRef] [PubMed]
3. Meuwissen, T.H.; Hayes, B.J.; Goddard, M.E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* **2001**, *157*, 1819–1829. [CrossRef] [PubMed]
4. Varshney, R.K.; Roorkiwal, M.; Sorrells, M.E. Genomic Selection for Crop Improvement: An Introduction. In *Genomic Selection for Crop Improvement*; Springer International Publishing: Cham, Switzerland, 2017; pp. 1–6, ISBN 9783319631684.
5. Montesinos-López, O.A.; Montesinos-López, A.; Pérez-Rodríguez, P.; Barrón-López, J.A.; Martini, J.W.R.; Fajardo-Flores, S.B.; Gaytan-Lugo, L.S.; Santana-Mancilla, P.C.; Crossa, J. A Review of Deep Learning Applications for Genomic Selection. *BMC Genomics* **2021**, *22*, 19. [CrossRef]
6. Voss-Fels, K.P.; Stahl, A.; Wittkop, B.; Lichthardt, C.; Nagler, S.; Rose, T.; Chen, T.-W.; Zetzsche, H.; Seddig, S.; Majid Baig, M.; et al. Breeding Improves Wheat Productivity under Contrasting Agrochemical Input Levels. *Nat. Plants* **2019**, *5*, 706–714. [CrossRef]
7. Crossa, J.; Pérez-Rodríguez, P.; Cuevas, J.; Montesinos-López, O.; Jarquín, D.; de Los Campos, G.; Burgueño, J.; González-Camacho, J.M.; Pérez-Elizalde, S.; Beyene, Y.; et al. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci.* **2017**, *22*, 961–975. [CrossRef]
8. Ismail, A.M.; Heuer, S.; Thomson, M.J.; Wissuwa, M. Genetic and Genomic Approaches to Develop Rice Germplasm for Problem Soils. *Plant Mol. Biol.* **2007**, *65*, 547–570. [CrossRef]
9. Wang, X.; Zeng, H.; Lin, L.; Huang, Y.; Lin, H.; Que, Y. Deep Learning-Empowered Crop Breeding: Intelligent, Efficient and Promising. *Front. Plant Sci.* **2023**, *14*, 1260089. [CrossRef]
10. Varshney, R.K.; Thudi, M.; Pandey, M.K.; Tardieu, F.; Ojiewo, C.; Vadez, V.; Whitbread, A.M.; Siddique, K.H.M.; Nguyen, H.T.; Carberry, P.S.; et al. Accelerating Genetic Gains in Legumes for the Development of Prosperous Smallholder Agriculture: Integrating Genomics, Phenotyping, Systems Modelling and Agronomy. *J. Exp. Bot.* **2018**, *69*, 3293–3312. [CrossRef]
11. Khan, M.H.U.; Wang, S.; Wang, J.; Ahmar, S.; Saeed, S.; Khan, S.U.; Xu, X.; Chen, H.; Bhat, J.A.; Feng, X. Applications of Artificial Intelligence in Climate-Resilient Smart-Crop Breeding. *Int. J. Mol. Sci.* **2022**, *23*, 11156. [CrossRef]
12. Pourkheirandish, M.; Hensel, G.; Kilian, B.; Senthil, N.; Chen, G.; Sameri, M.; Azhaguvel, P.; Sakuma, S.; Dhanagond, S.; Sharma, R.; et al. Evolution of the Grain Dispersal System in Barley. *Cell* **2015**, *162*, 527–539. [CrossRef] [PubMed]
13. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*; Springer US: Boston, MA, USA, 1989; ISBN 9780412317606.
14. History of CART. Available online: https://usdd-dev.thinkbluedata.com/history-cart (accessed on 28 November 2024).
15. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Routledge: London, UK, 2017; ISBN 9781315139470.
16. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
17. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
18. Madhukar, A.; Dashora, K.; Kumar, V. Climate Trends in Temperature and Water Variables during Wheat Growing Season and Impact on Yield. *Environ. Process.* **2021**, *8*, 1047–1072. [CrossRef]
19. Wójcik-Gront, E. Variables Influencing Yield-Scaled Global Warming Potential and Yield of Winter Wheat Production. *Field Crops Res.* **2018**, *227*, 19–29. [CrossRef]
20. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE Inst. Electr. Electron. Eng.* **1998**, *86*, 2278–2324. [CrossRef]
21. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
22. Wang, H.; Cimen, E.; Singh, N.; Buckler, E. Deep Learning for Plant Genomics and Crop Improvement. *Curr. Opin. Plant Biol.* **2020**, *54*, 34–41. [CrossRef]

23. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016.
24. Shanmugam, I.; Rethnaraj, J.; Rajendran, S.; Manickam, S. Prediction on Field Crops Yield Based on Analysis of Deep Learning Model. *Indones. J. Electr. Eng. Comput. Sci.* **2023**, *30*, 518. [CrossRef]
25. Khaki, S.; Wang, L. Crop Yield Prediction Using Deep Neural Networks. *Front. Plant Sci.* **2019**, *10*, 621. [CrossRef]
26. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
27. Ozdemir, O.B.; Chen, R.; Li, R. A Deep Ensemble Encoder Network Method for Improved Polygenic Risk Score Prediction. *medRxiv* **2024**.
28. Geleta, M.; Mas Montserrat, D.; Giro-i-Nieto, X.; Ioannidis, A.G. Deep Variational Autoencoders for Population Genetics. *bioRxiv* **2023**.
29. Battey, C.J.; Coffing, G.C.; Kern, A.D. Visualizing Population Structure with Variational Autoencoders. *G3 Genes|Genomes|Genet.* **2021**, *11*, jkaa036. [CrossRef]
30. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. In *Encyclopedia of Database Systems*; Springer US: Boston, MA, USA, 2009; pp. 532–538, ISBN 9780387355443.
31. Crossa, J.; Pérez, P.; Hickey, J.; Burgueño, J.; Ornella, L.; Cerón-Rojas, J.; Zhang, X.; Dreisigacker, S.; Babu, R.; Li, Y.; et al. Genomic Prediction in CIMMYT Maize and Wheat Breeding Programs. *Heredity* **2014**, *112*, 48–60. [CrossRef]
32. Geron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed.; O'Reilly Media: Newton, MA, USA, 2017.
33. Wu, B.; Zhang, H.; Lin, L.; Wang, H.; Gao, Y.; Zhao, L.; Chen, Y.-P.P.; Chen, R.; Gu, L. A Similarity Searching System for Biological Phenotype Images Using Deep Convolutional Encoder-Decoder Architecture. *Curr. Bioinform.* **2019**, *14*, 628–639. [CrossRef]
34. Mejía-Guerra, M.K.; Buckler, E.S. A K-Mer Grammar Analysis to Uncover Maize Regulatory Architecture. *BMC Plant Biol.* **2019**, *19*, 103. [CrossRef]
35. Washburn, J.D.; Mejia-Guerra, M.K.; Ramstein, G.; Kremling, K.A.; Valluru, R.; Buckler, E.S.; Wang, H. Evolutionarily Informed Deep Learning Methods for Predicting Relative Transcript Abundance from DNA Sequence. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 5542–5549. [CrossRef]
36. Gao, X.; Zhang, J.; Wei, Z.; Hakonarson, H. DeepPolyA: A Convolutional Neural Network Approach for Polyadenylation Site Prediction. *IEEE Access* **2018**, *6*, 24340–24349. [CrossRef]
37. Tran, T.-T.; Choi, J.-W.; Le, T.-T.; Kim, J.-W. A Comparative Study of Deep CNN in Forecasting and Classifying the Macronutrient Deficiencies on Development of Tomato Plant. *Appl. Sci.* **2019**, *9*, 1601. [CrossRef]
38. Wu, Y.-X.; Wu, Q.-B.; Zhu, J.-Q. Data-driven Wind Speed Forecasting Using Deep Feature Extraction and LSTM. *IET Renew. Power Gener.* **2019**, *13*, 2062–2069. [CrossRef]
39. Monaco, A.; Pantaleo, E.; Amoroso, N.; Lacalamita, A.; Lo Giudice, C.; Fonzino, A.; Fosso, B.; Picardi, E.; Tangaro, S.; Pesole, G.; et al. A Primer on Machine Learning Techniques for Genomic Applications. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4345–4359. [CrossRef] [PubMed]
40. Libbrecht, M.W.; Noble, W.S. Machine Learning Applications in Genetics and Genomics. *Nat. Rev. Genet.* **2015**, *16*, 321–332. [CrossRef] [PubMed]
41. Pawełkowicz, M.; Zieniuk, B.; Staszek, P.; Przybysz, A. From Sequencing to Genome Editing in Cucurbitaceae: Application of Modern Genomic Techniques to Enhance Plant Traits. *Agriculture* **2024**, *14*, 90. [CrossRef]
42. Bastani, O.; Kim, C.; Bastani, H. Interpretability via Model Extraction. *arXiv* **2017**, arXiv:1706.09773.
43. De La Vega, F.M.; Chowdhury, S.; Moore, B.; Frise, E.; McCarthy, J.; Hernandez, E.J.; Wong, T.; James, K.; Guidugli, L.; Agrawal, P.B.; et al. Artificial Intelligence Enables Comprehensive Genome Interpretation and Nomination of Candidate Diagnoses for Rare Genetic Diseases. *Genome Med.* **2021**, *13*, 153. [CrossRef]
44. Vilhekar, R.S.; Rawekar, A. Artificial Intelligence in Genetics. *Cureus* **2024**, *16*, e52035. [CrossRef]
45. Manekar, S.C.; Sathe, S.R. A Benchmark Study of K-Mer Counting Methods for High-Throughput Sequencing. *Gigascience* **2018**, *7*, giy125. [CrossRef]
46. Karikari, B.; Lemay, M.-A.; Belzile, F. K-Mer-Based Genome-Wide Association Studies in Plants: Advances, Challenges, and Perspectives. *Genes* **2023**, *14*, 1439. [CrossRef]
47. Orozco-Arias, S.; Candamil-Cortés, M.S.; Jaimes, P.A.; Piña, J.S.; Tabares-Soto, R.; Guyot, R.; Isaza, G. K-Mer-Based Machine Learning Method to Classify LTR-Retrotransposons in Plant Genomes. *PeerJ* **2021**, *9*, e11456. [CrossRef]
48. Moeckel, C.; Mareboina, M.; Konnaris, M.A.; Chan, C.S.Y.; Mouratidis, I.; Montgomery, A.; Chantzi, N.; Pavlopoulos, G.A.; Georgakopoulos-Soares, I. A Survey of K-Mer Methods and Applications in Bioinformatics. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 2289–2303. [CrossRef] [PubMed]
49. Osipowski, P.; Pawełkowicz, M.; Wojcieszek, M.; Skarzyńska, A.; Przybecki, Z.; Pląder, W. A High-Quality Cucumber Genome Assembly Enhances Computational Comparative Genomics. *Mol. Genet. Genomics* **2020**, *295*, 177–193. [CrossRef] [PubMed]
50. Pflug, J.M.; Holmes, V.R.; Burrus, C.; Johnston, J.S.; Maddison, D.R. Measuring Genome Sizes Using Read-Depth, k-Mers, and Flow Cytometry: Methodological Comparisons in Beetles (Coleoptera). *G3* **2020**, *10*, 3047–3060. [CrossRef]
51. Chen, L.; Liu, G.; Zhang, T. Integrating Machine Learning and Genome Editing for Crop Improvement. *Abiotech* **2024**, *5*, 262–277. [CrossRef]

52. Shen, Z.; Bao, W.; Huang, D.-S. Recurrent Neural Network for Predicting Transcription Factor Binding Sites. *Sci. Rep.* **2018**, *8*, 15270. [CrossRef]
53. Peleke, F.F.; Zumkeller, S.M.; Gültas, M.; Schmitt, A.; Szymański, J. Deep Learning the Cis-Regulatory Code for Gene Expression in Selected Model Plants. *Nat. Commun.* **2024**, *15*, 3488. [CrossRef]
54. Mahood, E.H.; Kruse, L.H.; Moghe, G.D. Machine Learning: A Powerful Tool for Gene Function Prediction in Plants. *Appl. Plant Sci.* **2020**, *8*, e11376. [CrossRef]
55. van Dijk, A.D.J.; Kootstra, G.; Kruijer, W.; de Ridder, D. Machine Learning in Plant Science and Plant Breeding. *iScience* **2021**, *24*, 101890. [CrossRef]
56. Zhang, W.; Zhang, P.; Sun, W.; Xu, J.; Liao, L.; Cao, Y.; Han, Y. Improving Plant miRNA-Target Prediction with Self-Supervised k-Mer Embedding and Spectral Graph Convolutional Neural Network. *PeerJ* **2024**, *12*, e17396. [CrossRef]
57. Niazian, M.; Niedbała, G. Machine Learning for Plant Breeding and Biotechnology. *Agriculture* **2020**, *10*, 436. [CrossRef]
58. Yan, J.; Wang, X. Machine Learning Bridges Omics Sciences and Plant Breeding. *Trends Plant Sci.* **2023**, *28*, 199–210. [CrossRef] [PubMed]
59. Alipanahi, B.; Delong, A.; Weirauch, M.T.; Frey, B.J. Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning. *Nat. Biotechnol.* **2015**, *33*, 831–838. [CrossRef] [PubMed]
60. Qin, Q.; Feng, J. Imputation for Transcription Factor Binding Predictions Based on Deep Learning. *PLoS Comput. Biol.* **2017**, *13*, e1005403. [CrossRef] [PubMed]
61. Zhou, J.; Troyanskaya, O.G. Predicting Effects of Noncoding Variants with Deep Learning-Based Sequence Model. *Nat. Methods* **2015**, *12*, 931–934. [CrossRef]
62. Wang, M.; Tai, C.; E, W.; Wei, L. DeFine: Deep Convolutional Neural Networks Accurately Quantify Intensities of Transcription Factor-DNA Binding and Facilitate Evaluation of Functional Non-Coding Variants. *Nucleic Acids Res.* **2018**, *46*, e69. [CrossRef]
63. Greenside, P.; Shimko, T.; Fordyce, P.; Kundaje, A. Discovering Epistatic Feature Interactions from Neural Network Models of Regulatory DNA Sequences. *Bioinformatics* **2018**, *34*, i629–i637. [CrossRef]
64. Yan, J.; Kurgan, L. DRNApred, Fast Sequence-Based Method That Accurately Predicts and Discriminates DNA- and RNA-Binding Residues. *Nucleic Acids Res.* **2017**, *45*, e84. [CrossRef]
65. Lane, S.W.; Williams, D.A.; Watt, F.M. Modulating the Stem Cell Niche for Tissue Regeneration. *Nat. Biotechnol.* **2014**, *32*, 795–803. [CrossRef]
66. Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.F.G.; Qin, C.; Zidek, A.; Nelson, A.; Bridgland, A.; Penedones, H.; et al. De novo structure prediction with deep-learning based scoring. *Annu Rev Biochem.* **2018**, *77*, 6.
67. International Wheat Genome Sequencing Consortium. Available online: https://www.wheatgenome.org/ (accessed on 28 November 2024).
68. International Wheat Genome Sequencing Consortium (IWGSC); Appels, R.; Eversole, K.; Stein, N.; Feuillet, C.; Keller, B.; Rogers, J.; Pozniak, C.J.; Choulet, F.; Distelfeld, A.; et al. Shifting the Limits in Wheat Research and Breeding Using a Fully Annotated Reference Genome. *Science* **2018**, *361*, eaar7191. [CrossRef]
69. Zhang, R.; Zhang, C.; Yu, C.; Dong, J.; Hu, J. Integration of Multi-Omics Technologies for Crop Improvement: Status and Prospects. *Front. Bioinform.* **2022**, *2*, 1027457. [CrossRef] [PubMed]
70. van Eeuwijk, F.A.; Bustos-Korts, D.; Millet, E.J.; Boer, M.P.; Kruijer, W.; Thompson, A.; Malosetti, M.; Iwata, H.; Quiroz, R.; Kuppe, C.; et al. Modelling Strategies for Assessing and Increasing the Effectiveness of New Phenotyping Techniques in Plant Breeding. *Plant Sci.* **2019**, *282*, 23–39. [CrossRef] [PubMed]
71. Wolfert, S.; Ge, L.; Verdouw, C.; Bogaardt, M.-J. Big Data in Smart Farming—A Review. *Agric. Syst.* **2017**, *153*, 69–80. [CrossRef]
72. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef]
73. Merchant, N.; Lyons, E.; Goff, S.; Vaughn, M.; Ware, D.; Micklos, D.; Antin, P. The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol.* **2016**, *14*, e1002342. [CrossRef]
74. Araus, J.L.; Cairns, J.E. Field High-Throughput Phenotyping: The New Crop Breeding Frontier. *Trends Plant Sci.* **2014**, *19*, 52–61. [CrossRef]
75. Yang, W.; Feng, H.; Zhang, X.; Zhang, J.; Doonan, J.H.; Batchelor, W.D.; Xiong, L.; Yan, J. Crop Phenomics and High-Throughput Phenotyping: Past Decades, Current Challenges, and Future Perspectives. *Mol. Plant* **2020**, *13*, 187–214. [CrossRef]
76. Hendre, P.S.; Muthemba, S.; Kariba, R.; Muchugi, A.; Fu, Y.; Chang, Y.; Song, B.; Liu, H.; Liu, M.; Liao, X.; et al. African Orphan Crops Consortium (AOCC): Status of Developing Genomic Resources for African Orphan Crops. *Planta* **2019**, *250*, 989–1003. [CrossRef]
77. Poland, J.A.; Rife, T.W. Genotyping-by-sequencing for Plant Breeding and Genetics. *Plant Genome* **2012**, *5*, 92–102. [CrossRef]
78. Jasanoff, S.; Hurlbut, J.B. A Global Observatory for Gene Editing. *Nature* **2018**, *555*, 435–437. [CrossRef]
79. International Maize and Wheat Improvement Center. Available online: https://www.cimmyt.org/ (accessed on 28 November 2024).
80. Ramírez, F.; Ryan, D.P.; Grüning, B.; Bhardwaj, V.; Kilpert, F.; Richter, A.S.; Heyne, S.; Dündar, F.; Manke, T. deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis. *Nucleic Acids Res.* **2016**, *44*, W160–W165. [CrossRef] [PubMed]

81. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A System for Large-Scale Machine Learning. *arXiv* **2016**, arXiv:1605.08695.
82. Voss-Fels, K.P.; Cooper, M.; Hayes, B.J. Accelerating Crop Genetic Gains with Genomic Selection. *Züchter Genet. Breed. Res.* **2019**, *132*, 669–686. [CrossRef]
83. El Jarroudi, M.; Kouadio, L.; Delfosse, P.; Bock, C.H.; Mahlein, A.-K.; Fettweis, X.; Mercatoris, B.; Adams, F.; Lenné, J.M.; Hamdioui, S. Leveraging Edge Artificial Intelligence for Sustainable Agriculture. *Nat. Sustain.* **2024**, *7*, 846–854. [CrossRef]
84. Jobin, A.; Ienca, M.; Vayena, E. The Global Landscape of AI Ethics Guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
85. Aldoseri, A.; Al-Khalifa, K.N.; Hamouda, A.M. Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. *Appl. Sci.* **2023**, *13*, 7082. [CrossRef]
86. Janzen, T. The Farmer's Business Network, Inc. Available online: https://www.agdatatransparent.com/certified/fbn (accessed on 12 December 2024).
87. Dara, R.; Hazrati Fard, S.M.; Kaur, J. Recommendations for Ethical and Responsible Use of Artificial Intelligence in Digital Agriculture. *Front. Artif. Intell.* **2022**, *5*, 884192. [CrossRef]
88. IBM Largest Ever AI Toolset Release Is Tailor Made for 9 Industries and Professions. Available online: https://www.agritechtomorrow.com/news/2018/09/25/ibm-largest-ever-ai-toolset-release-is-tailor-made-for-9-industries-and-professions/11028/ (accessed on 12 December 2024).
89. Abramov, M. Data Bias in AI Agriculture: Ensuring Fairness & Sustainability. Available online: https://keymakr.com/blog/data-bias-in-ai-agriculture-ensuring-fairness-and-sustainability (accessed on 28 November 2024).
90. Ryan, M. The Social and Ethical Impacts of Artificial Intelligence in Agriculture: Mapping the Agricultural AI Literature. *AI Soc.* **2022**, *38*, 2473–24485. [CrossRef]
91. Pandey, D.K.; Mishra, R. Towards Sustainable Agriculture: Harnessing AI for Global Food Security. *Artif. Intell. Agric.* **2024**, *12*, 72–84. [CrossRef]
92. Plevin, R. Central Valley Effort Aims to Train Farmworkers to Master the Technology Replacing Fieldwork. *Los Angeles Times*, 2024. Available online: https://www.latimes.com/california/story/2024-09-21/central-valley-effort-trains-farmworkers-to-master-technology-replacing-fieldwork (accessed on 12 December 2024).
93. Mushtaq, M.A.; Ahmed, H.G.M.-D.; Zeng, Y. Applications of Artificial Intelligence in Wheat Breeding for Sustainable Food Security. *Sustainability* **2024**, *16*, 5688. [CrossRef]
94. Rai, K.K. Integrating Speed Breeding with Artificial Intelligence for Developing Climate-Smart Crops. *Mol. Biol. Rep.* **2022**, *49*, 11385–11402. [CrossRef] [PubMed]
95. Xu, Y.; Zhang, X.; Li, H.; Zheng, H.; Zhang, J.; Olsen, M.S.; Varshney, R.K.; Prasanna, B.M.; Qian, Q. Smart Breeding Driven by Big Data, Artificial Intelligence, and Integrated Genomic-Enviromic Prediction. *Mol. Plant* **2022**, *15*, 1664–1695. [CrossRef] [PubMed]
96. Gupta, D.K.; Pagani, A.; Zamboni, P.; Singh, A.K. AI-Powered Revolution in Plant Sciences: Advancements, Applications, and Challenges for Sustainable Agriculture and Food Security. *Explor. Foods Foodomics* **2024**, *2*, 443–459. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.