frontiers | Frontiers in Artificial Intelligence

TYPE Review
PUBLISHED 10 January 2023
DOI 10.3389/frai.2022.1040295

OPEN ACCESS

EDITED BY
Lyndon Estes,
Clark University, United States

REVIEWED BY
Jinran Wu,
Australian Catholic University, Australia
Aalt-Jan Van Dijk,
Wageningen University and Research, Netherlands

*CORRESPONDENCE
Sheikh Jubair
✉ jubairs@myumanitoba.ca

SPECIALTY SECTION
This article was submitted to AI in Food, Agriculture and Water, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 09 September 2022
ACCEPTED 22 December 2022
PUBLISHED 10 January 2023

CITATION
Jubair S and Domaratzki M (2023) Crop genomic selection with deep learning and environmental data: A survey. *Front. Artif. Intell.* 5:1040295. doi: 10.3389/frai.2022.1040295

Crop genomic selection with deep learning and environmental data: A survey

Sheikh Jubair[1]* and Mike Domaratzki[2]

[1]Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada, [2]Department of Computer Science, University of Western Ontario, London, ON, Canada

Machine learning techniques for crop genomic selections, especially for single-environment plants, are well-developed. These machine learning models, which use dense genome-wide markers to predict phenotype, routinely perform well on single-environment datasets, especially for complex traits affected by multiple markers. On the other hand, machine learning models for predicting crop phenotype, especially deep learning models, using datasets that span different environmental conditions, have only recently emerged. Models that can accept heterogeneous data sources, such as temperature, soil conditions and precipitation, are natural choices for modeling GxE in multi-environment prediction. Here, we review emerging deep learning techniques that incorporate environmental data directly into genomic selection models.

KEYWORDS

genomic selection, machine learning, GxE, environment, MET, deep learning

1. Introduction

Production of sufficient food for the increasing world population is a major concern. Industrialization and development of infrastructure in developing countries are causing a shortage of land for growing populations in urban areas, which leads to unplanned expansion of cities into agricultural land (Azadi et al., 2011). Soil erosion due to water, wind, or excessive use for cultivation affects the topsoil and fertility, thus reduces crop production. A large amount of surface and groundwater has already been used, causing a decrease in groundwater level (Van Meijl et al., 2018). Global temperature is increasing and heat waves have become more frequent, which leads a significant decrease in crop production (Bourgault et al., 2018; Nawaz and Chung, 2020). Though several regions will benefit from the effect of climate change, especially because of the increase in temperature, overall food production will decrease by 2050 (Van Meijl et al., 2018).

These problems will increase the price of the food and people, especially in developing countries, will suffer from hunger and deficiency in nutrition, causing low growth in children or low weight (Linehan et al., 2012; United Nations, 2019; Nawaz and Chung, 2020). It is projected by the UN that by 2050, the world population will reach 9.7 billion and to accommodate this vast number of people, a large amount of new agricultural land will be needed (Searchinger et al., 2019). This will lead to the "more people, less agricultural land" problem (United Nations, 2019; Nawaz and Chung, 2020).

To ensure food security and keep the food affordable to everyone, by 2050, we will need to increase our food production by 50% of our current production (Nawaz and Chung, 2020).

To face the challenge of food production in the future, selection of varieties with desired phenotypes from a collection of varieties of a crop is essential to breeders, as the right selection can lead to improvements such as drought resistance, biotic and abiotic stress resistance, yield improvement and disease resistance (Varshney et al., 2017). While the amount of water, fertilizer, pest control, and sound production practices contribute to the environment for the plant, the genotype of the plant defines the ability to produce a desired phenotypic value within that environment (Milton, 1979). Thus, as environmental factors and breeding practices are standardized and measured, it is vital to create improved varieties for that environment.

Genomic selection (GS), first defined by Meuwissen et al. (2001), is a marker-assisted selection method that uses dense whole-genome molecular markers to improve the quantitative traits of an organism such as a crop or livestock by identifying the top germplasms. That is, GS is a computational tool for choosing the most advantageous individuals from a set of varieties and has the potential to save money and time by accelerating improvements to crops or livestock (Acquaah, 2009; Varshney et al., 2017).

GS for single environment trials employs GS to identify top individuals to create a new variety for a specific environment (Meuwissen et al., 2001; Heffner et al., 2009; Crossa et al., 2017; Jubair et al., 2021). If the environment changes, single environment GS does not guarantee that the new variety will have the desired outcome in that new environment (Oakey et al., 2016). GS for multi-environment trial is a generalization that is able to identify top organisms even if the environment is new (Washburn et al., 2021). In this survey, we focus on applications of deep learning in both single and multi-environment trial and analyze the differences between single environment and multi-environment models. In particular, we are interested in those multi-environment models that incorporate data such as hourly temperature, rainfall or other time series data from environments into deep learning models to improve prediction. The reader may wish to consult existing reviews of genomic selections for material focused on statistical models of single environment (Wang et al., 2018; van Dijk et al., 2021; Anilkumar et al., 2022) and multi-environment trials (Tong and Nikoloski, 2021; van Dijk et al., 2021). Additionally, several reviews cover fully the use of machine learning models for single environment trials (Montesinos-López et al., 2021; Tong and Nikoloski, 2021; van Dijk et al., 2021; Anilkumar et al., 2022; Danilevicz et al., 2022). Xu et al. (2022) also review GS and describe the potential for the use of multiple sources of data beyond genomic data, including enviromental data. This includes the use of machine learning models. In contrast, multi-environment deep learning approaches are an emerging area that enable detailed weather data to be incorporated directly into the model (Khaki and Wang, 2019; Khaki et al., 2020; Lin et al., 2020; Shook et al., 2020). Our survey focuses specifically on recent works involving this latter class of models that employs genomic and weather data together to inform deep learning models and predict phenotypes.

Traditionally, we can identify two broad approaches to GS. Linear methods such as BLUP and variants (Burgueño et al., 2012; Bandeira e Sousa et al., 2017; Cuevas et al., 2017, 2019; Ferrão et al., 2017; Howard et al., 2019; Millet et al., 2019) explicitly model the phenotype in terms of contributions from different factors, including pedigree, individual markers or distinct site-years. Typically, these models perform well for additive traits due to the linear nature of the models. On the other hand, machine learning models, such as Random Forests (RFs) (Holliday et al., 2012; Ali et al., 2020; Sawitri et al., 2020), Support Vector Machines (SVMs) (Ogutu et al., 2011; Wang et al., 2019) and Neural Networks (NNs) (Jubair and Domaratzki, 2019; Pérez-Enciso and Zingaretti, 2019) can model traits in non-linear but typically opaque ways. For a complete introduction to machine learning and deep learning (DL), see Emmert-Streib et al. (2020) or Dey (2016). In this paper, our focus is on the deep learning methods in this area.

Crops respond differently in different environmental conditions (Millet et al., 2019), an effect known as genome by environment interaction (GxE). This leads to differences in production quantity or quality (Cuevas et al., 2017). In a single environment trial, it is typically assumed that the environment is constant, thus, there is no effect of environment on genotypes. A number of deep learning methods for single environment trials have been published (McDowell, 2016; Rachmatia et al., 2017; Ma et al., 2018; Jubair and Domaratzki, 2019; Zingaretti et al., 2020; Jubair et al., 2021; Montesinos-Lopez et al., 2021). These methods differ in their deep learning architectures and focus on how they capture the genetic information. Multi-environment models can be thought of an extension of single environment trial as the models consider the interaction between environment and genome. Though multi-environment trials are an extension of single-environment GS, there are very few deep learning methods that have been developed for this problem (Montesinos-López et al., 2018, 2019b; Khaki and Wang, 2019; Khaki et al., 2020; Lin et al., 2020; Shook et al., 2020) that take GxE interaction in crops into account because of the complexity in incorporating the environmental interaction into the model and lack of complete environmental data. In the past 3 years, new research has demonstrated the potential of incorporating environmental information into deep learning models for GS (Khaki and Wang, 2019; Khaki et al., 2020; Lin et al., 2020; Shook et al., 2020). This survey focuses specifically on deep learning methods for integrating weather

TABLE 1 An example of genotyped data.

Genotype	M_1	M_2	...	...	M_D
$Geno_1$	1	−1			−1
$Geno_2$	0	1			1
...	0	-1	...	...	1
...	0	0	...	...	1
$Geno_N$	−1	0	...	...	0

In the column header, M means markers. This dataset contains D markers and N genotypes. Thus, each line is represented by D markers. Each of these markers can have one of the three values: 1, 0, and −1.

data into GS. The ability to integrate heterogenous data into a model is a known strength of machine learning models in general, and deep learning models in particular. However, this research is one facet of a large, active research community that seeks to improve GS accuracy, using various models, through integration of types of environmental data (Costa-Neto et al., 2022; Montesinos-López et al., 2022; Putra et al., 2022; Song et al., 2022).

In this survey, our aim is to provide a comprehensive overview of genomic selection process with deep learning that starts from data and ends with creating a new variety for both single and multi-environment trial. To do this, (i) we provide an overview of different data of GS and how these data need to be processed, (ii) discuss popular components of deep learning models typically employed in GS and then (iii) review existing deep learning architectures and motivation behind them for both single and multi-environment trials.

2. Datasets for GS

Crop organisms are usually genotyped using high throughput sequencing technology that uses a large number of genomic markers to cover the whole genome of that organism (Goddard and Hayes, 2007; Heffner et al., 2009; Crossa et al., 2017). These markers are usually represented by categorical values based on their zygosity or sequencing technology. For example, a diploid organism is usually represented by 1, 0 and -1 where 1 and -1 represent homozygous allele and 0 represents heterozygous allele. If DArT assays are used for sequencing, SNPs are represented by binary values, indicating a gene's presence or absence (Crossa et al., 2016b; Jubair and Domaratzki, 2019). Table 1 shows an example dataset.

As the data may contain uninformative markers and missing values, the genotyped data often need pre-processing. The preprocessing steps may involve removing uninformative markers, imputation of missing values and representing the features in some other forms. If the minor allele frequency $\leq 5\%$ (Ma et al., 2018; Jubair et al., 2021) or more than 30% values are missing, then the marker is usually removed as those markers do not bear any relevant information. To replace the missing values, one popular imputation techniques is k-nearest neighbor. For example, at first, the k-nearest genotypes of the genotype of interest are identified. From those genotypes, the missing value is replaced by the most frequent value for the specific marker.

Most neural networks consist of a linear equation that multiplies a weight vector with a feature vector (LeCun et al., 2015; Dong et al., 2021). If a feature is represented with a zero, it means the feature will not have any influence on final outcome as the resulting multiplication between the weight and feature will also be zero. Thus, providing traditional marker data as input to the deep learning models may result in a loss of information. This may lead us to think that representing the allele with other categorical values such as 1, 2, and 3 will solve this issue. This leads to another problem as multiplying weights with a high value of a specific allele may mislead the deep learning model to give higher priority to that specific allele. To solve these problems, one-hot encoded vector (Liu et al., 2019b) or Hardy-Weinberg equlibrium can be used to represent markers (Jubair et al., 2021). A one-hot encoded vector is an n dimensional sparse vector where n is the number of alleles of a specific marker. Each allele of a marker is associated with a specific position in the vector. If that allele is present in the marker, the specific position for the allele is represented with 1 and other positions with 0. Sometimes, an extra position is also added to the one hot encoded vector to represent missing values (Liu et al., 2019b). As an alternative to categorical encoding and one hot encoded representation, markers can also represented by their allele frequency (Jubair et al., 2021), which can be obtained following the Hardy-Weinberg equilibrium formula. For example, suppose, in 10 genotypes, allele AA, Aa, and aa for a specific marker occurs 6, 3, and 1 times, respectively. Then the frequency of AA, Aa, and aa is 0.6, 0.3, and 0.1, respectively.

The environment of crops comprises weather, soil and field management data. Weather information, such as maximum and minimum temperature, precipitation, vapor pressure, wind speed and radiation, plays an essential part in GS for multi-environmental trials (Khaki and Wang, 2019; Gangopadhyay et al., 2020; Khaki et al., 2020; Shook et al., 2020). Weather information can be integrated as daily, weekly, monthly or yearly averages based on the architecture of the deep learning model (Khaki and Wang, 2019; Khaki et al., 2020; Washburn et al., 2021). In addition, soil information such as percentage of clay, silt and sand, water capacity, soil pH, number of irrigation, organic matter, and cation-exchange capacity also plays a vital role (Washburn et al., 2021). Sometimes, field management information such as the number of irrigations, sowing pattern of crops, amount of water used in irrigation, and amount of fertilizer or insecticide applied is also recorded. These can also be integrated with soil data as they carry valuable information (Washburn et al., 2021). As the variables from environmental information are in different ranges, these variables are usually scaled by zero-centering as a pre-processing

TABLE 2 An example of genotyped and environmental data after pre-processing in a tabular format.

		Markers			Weather variables			Soil variables			Field management		
Envs	**Geno**	M_1	…	M_D	W_1 $t=1$	…	W_w $t=T$	S_1	…	S_s	F_1	…	F_f
Env_1	$Geno_1$	0.6	…	0.4	0.32	…	0.27	0.2	…	0.15	0.4	…	0.6
Env_1	$Geno_2$	0.2	…	0.4	0.32	…	0.27	0.2	…	0.15	0.2	…	0.21
Env_2	$Geno_3$	0.6	…	0.4	0	…	0.4	0.32	…	0.24	0.25	…	0.05
Env_3	$Geno_4$	0.6	…	0.2	0.65	…	0.1	0.3	…	0.31	0.4	…	0.1
…	…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…
Env_k	$Geno_n$	0.2	…	0.4	0.65	…	0.1	0.3	…	0.31	0.2	…	0.1

In this example, each genotype has D markers after removing minor alleles and imputing missing values. Marker values are represented by their allele frequency. There are w weather variables where each weather variables are divided in T time steps. Apart from the weather variables, there are s soil variables and f field management variables too. All the data are normalized.

step. Table 2 shows an example of genotyped and environmental data after pre-processing.

3. Deep learning

in recent years, Deep Learning has emerged as a leading paradigm for supervised machine learning tasks. Significant innovation has occurred in diverse areas like Natural Language Processing, Computer Vision, and Bioinformatics (LeCun et al., 2015; Li et al., 2020; Dong et al., 2021). The dominant paradigm in DL is a network. A deep learning network is made up of blocks and each block has several different types of layers. A block usually contains multiple layers of one or more neural networks, activation function, normalization layer and regularization layer (LeCun et al., 2015; Dong et al., 2021). In this section, we discuss each of the layers of neural network blocks and describe the function of the most common layers. It is worth mentioning that we chose these layers based on their usage in previous research conducted in GS.

In a deep learning model, the layers between the input and output are called hidden layers. Each layer consists of several nodes called neurons where we receive input and perform computation on the data from previous layers. Typically, the neural network layer contains one or more feed-forward (Bebis and Georgiopoulos, 1994), convolution (Kim, 2017; Kiranyaz et al., 2021) or Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997; Yu et al., 2019b) layers (discussed in Section 3.1). As these neural networks are generally linear functions, activation functions such as ReLU and sigmoid are applied to the output of the neural network layer to introduce non-linearity (discussed in Section 3.2). Normalization and regularization layers such as L1, L2 and dropout are applied after the activation layer to generalize the model to avoid overfitting (discussed in Section 3.3). Figure 1 shows the general architecture of a deep learning method.

3.1. Neural networks

3.1.1. Fully connected neural networks

A fully connected neural network (FNN), often referred to as a linear layer, is an Artificial Neural Network where all the neurons of the previous layer are connected to each neuron of the current layer. The mathematical operation of the fully connected neural network can be compared to n linear regression methods (Montgomery et al., 2021) where n is the number of hidden neurons of the current layer. A deep fully connected neural network is often called Multi-Layer Perceptron (MLP). Figure 2 shows a fully connected network.

MLPs have been applied to predict phenotypes both in single environment trial (Gianola et al., 2011; González-Camacho et al., 2016; Jubair and Domaratzki, 2019; Montesinos-López et al., 2019a; Jubair et al., 2021) and multi-environment trial Montesinos-López et al. (2018), Khaki and Wang (2019). In case of single environment trials, the input is the genotyped data of crops. When the prediction of phenotypes is for multi-environment trials, additional information such as environmental data are concatenated with the genotyped data. This concatenated vector is the input of the feed-forward networks and the output is the environment-specific predicted yield (Khaki and Wang, 2019).

3.1.2. Convolutional neural networks

Convolutional neural networks are a successful model of DL that employ convolution operations to incorporate targeted regions of input in decision making (Li et al., 2021). A convolution operation summarizes point-wise multiplication between a small kernel that slides over the input of the convolution layer. The weights of the kernels are shared across all the sliding windows. These kinds of neural networks are known for capturing local information within the data since, in

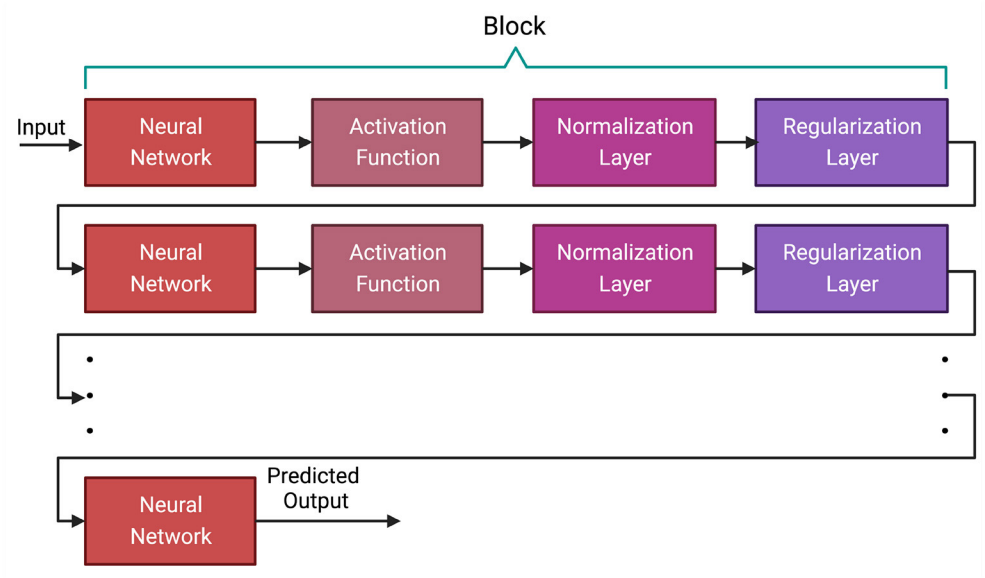

FIGURE 1
General architecture of a deep learning algorithm. All the layers between the first and last layer are called hidden layers. The first layer is the input layer and the last layer is the output layer. A neural network layer is typically followed by an activation function and then by normalization and regularization layers. Based on the architecture of the deep learning model, some of these layers, such as normalization and regularization layers, may not be present in a block.

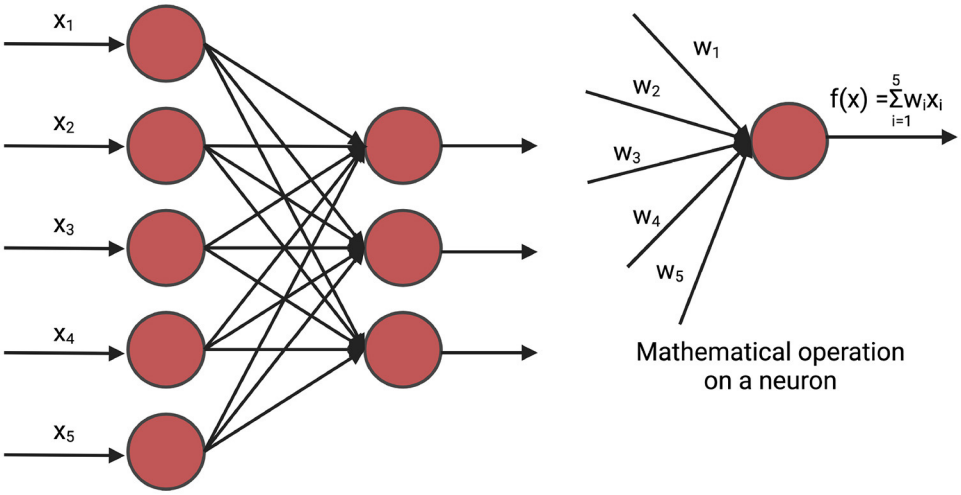

FIGURE 2
An example of fully connected layer. x_i is the input and w_i is the weight. The weights are initialized randomly and are optimized iteratively for prediction. In GS, x_i is the marker.

each sliding window, the network is on a small subset of the data (LeCun et al., 2015; Dong et al., 2021). Convolution operations were first developed in vision to help identify features of an image in a restricted window as the spatial information in the image plays a vital role in most vision applications (Dong et al., 2021; Li et al., 2021). The applications of convolutional neural networks have also been extended to other domains such as GS (Ma et al., 2018; Jubair and Domaratzki, 2019; Liu et al., 2019b; Zingaretti et al., 2020).

There are three types of convolution, conv1D, conv2D, and conv3D, available in different deep learning frameworks (Abadi, 2016; Paszke et al., 2017; Chollet, 2018). The choice of the

convolution layer depends on the dimension of the input to the convolution layer. In GS, as the data is generally one-dimensional, conv1D is typically used (Ma et al., 2018). As the genotyped data is often categorical (1, 0, and −1), the marker data can also be converted to a one-hot encoded vector which will be the input of a conv2D layer (Liu et al., 2019b; Washburn et al., 2019; Avsec et al., 2021; Ji et al., 2021). Figure 3 shows an example of how 1D convolution works. In this example, a sequence of length 5 is processed with a kernel of size 3 and stride 1. The weights of the kernel are randomly initialized. A point-wise multiplication operation between the input window (in this example, the input window = 3) and the kernel takes place and after that an aggregation operation is performed. As the stride = 1, the input window then shift one space and the same operation of point-wise multiplication and aggregation takes place. This continues till the total input space is covered. The result is a sequence of length 3 where each neuron bears spatial information of the sequence.

To apply convolutional neural network for multi-environment trials, the algorithm should be developed carefully as a concatenated input vector of environment, genetic and soil data may not properly represent relationships between different data sources. The reason is that since the sliding window of convolution operation captures local information, a convolution operation on the concatenated vector may not properly reflect the effect of environment on the genetic data, as these are represented in regions of the concatenated vector that are not adjacent. To solve this problem, different types of neural networks can be employed on different types of data (Khaki et al., 2020; Washburn et al., 2021; Sharma et al., 2022). The predictions from different networks can be combined to obtain an overall prediction.

3.1.3. Recurrent neural networks

Recurrent neural networks (RNNs) are distinct from both MLPs and CNNs as they are not feed-forward. Neurons in RNNs may have connections to themselves. RNNs are a family of neural networks, such as Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Cho et al., 2014), that typically work with time-series and sequence data (Hochreiter and Schmidhuber, 1997). These networks have been successfully applied in weather prediction (Qing and Niu, 2018; Salman et al., 2018; Yu et al., 2019a) and in GS (Shook et al., 2020). Particularly, LSTM has been applied in genomic selection task mostly with environmental information (Shook et al., 2020). LSTM either preserves or forgets past information for future prediction by applying a particular structure called gates. The input of LSTM is time-steps or sequences and the output depends on all the previous time-steps or sequences. As LSTM are applicable to time-series data, the use with environmental data in GS allows the networks to efficiently summarize large-scale data. We refer the readers to the review on LSTM by Yu et al. (2019b) to know more about LSTM.

Generally, in multi-environment GS, historical weather information is the input to the RNNs. Genetic information is incorporated in the later part of the network (Shook et al., 2020). As the genetic information is not a time series data in nature, this part of the network generally does not contain any LSTM layers. The outcome is the predicted phenotypes for a specific weather condition.

3.1.4. Transformers

Transformers are another type of neural networks that transform one sequence to another sequence. That is, the transformer architecture is designed to take a sequence as input but also produce a sequence as output (Vaswani et al., 2017; Ji et al., 2021; Jubair et al., 2021; Le et al., 2022), as opposed to a single value, which is the output of MLPs or and CNNs. The transformer architecture contains an encoder and a decoder. This encoder and decoder can be used separately or together. The transformer encoder has been applied in GS (Jubair et al., 2021) and other fields such as DNA representation learning (Ji et al., 2021; Le et al., 2022) and gene expression prediction of humans (Avsec et al., 2021). Here, we discuss only the transformer encoder to predict crop traits.

The main building block of a transformer encoder is the multi-head attention layer which applies self attention (Vaswani et al., 2017). In GS, self-attention measures how important a marker is with respect to other markers for the phenotype prediction. Thus, the self attention captures the relationship of distant markers that influence the final phenotypic outcome (Jubair et al., 2021). Usually, the importance of markers with respect to a specific marker m is represented in a vector called attention vector. If multiple attention vectors are generated per marker, the final attention vector is the weighted average of all the attention vectors. The multiple attention vector is called multi-head attention. Apart from the multi-head attention layer, a transformer also contains a feed-forward neural network and layer normalization. Figure 4 shows a transformer encoder. The input of the transformer can be a one hot encoded vector or the genotype frequency (Avsec et al., 2021; Jubair et al., 2021). The embedding layer then embeds each marker to a d dimensional expanded representation. Usually a feed-forward neural network or a convolutional neural network is applied to embed the input features. The embedded representation of the markers are the input of the attention layers of the transformer.

3.2. Activation functions

The previous discussion shows that neural networks typically compute a linear function. However, as it is known that complex traits such as yields are non-linear, we need to

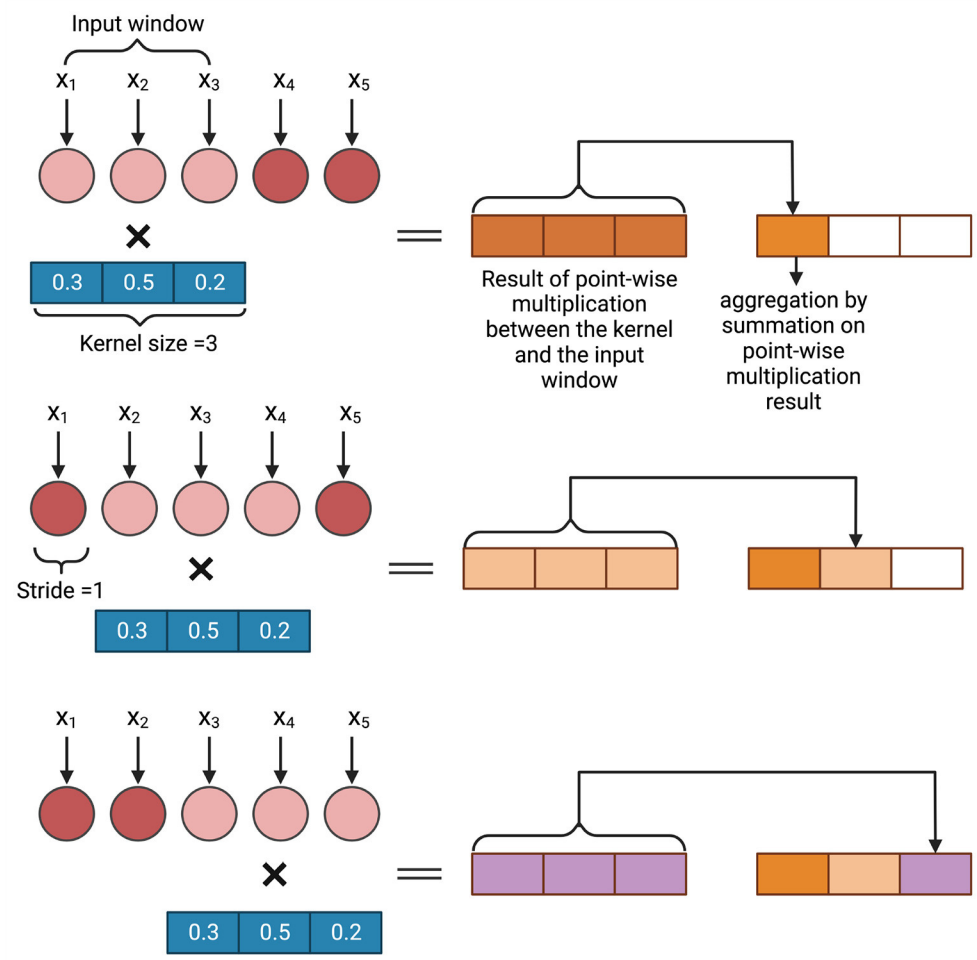

FIGURE 3
An example of convolution operation on a one dimensional input vector. In GS, x_i is a marker.

introduce non-linearity in the network. Activation functions introduce non-linearity to the network by deciding which neuron should be activated. Each activation function addresses different limitations; see the survey of Szandała (2021) for information on different activation functions used in the literature. However, sigmoid, ReLU and tanh are the most widely used activation functions for GS (McDowell, 2016; Ma et al., 2018; Jubair and Domaratzki, 2019; Khaki and Wang, 2019; Khaki et al., 2020; Shook et al., 2020; Måløy et al., 2021; Washburn et al., 2021). Hence, we provide an overview of these activation functions below.

The sigmoid activation produces the output neuron between 0 and 1 by applying the sigmoid function (Szandała, 2021; Dubey et al., 2022). Though sigmoid function is one of the most used activation function, it suffers from the vanishing gradient problem, that is, the gradient of the loss function approaches zero, which causes the model parameters of the DNN to not update or update very slowly. It is also not zero centered, causing difficulties during optimization.

The tanh activation function solves the zero centered problem as the output of this function ranges from −1 to 1 (Szandała, 2021; Dubey et al., 2022). However, it suffers from vanishing gradient problem, as very high value and very low value of the input neuron will be mapped to −1 and 1 and other values will be toward zero.

ReLU (Rectified Linear Unit) is the most popular activation function which ranges from 0 to ∞ (Szandała, 2021; Dubey et al., 2022). It solves the vanishing gradient problem and because of the simplicity of the function, it converges quicker than other activation functions.

3.3. Regularization layer

A regularization layer helps DL algorithms avoid overfitting and leads to better generalization by reducing the model complexity (Kukačka et al., 2017; Moradi et al., 2020). The most popular generalization techniques employed in GS are

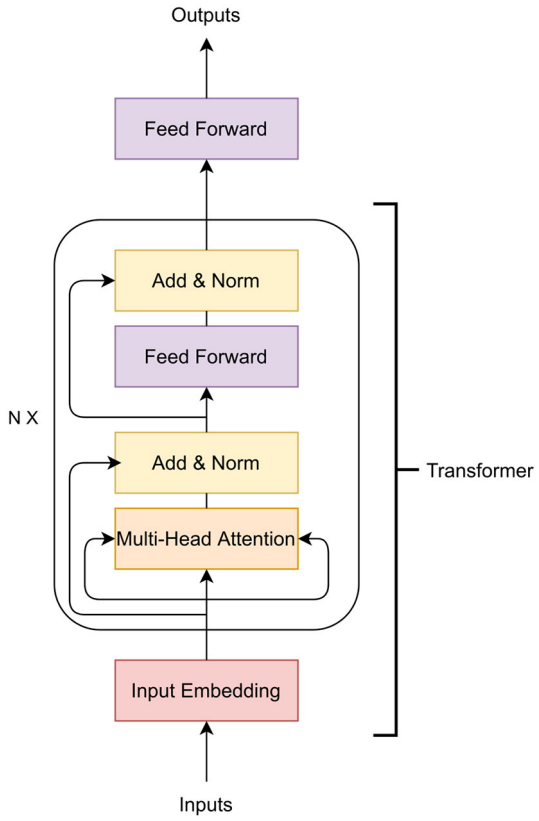

FIGURE 4
A transformer encoder. Here *N* represents *N* transformers can be stacked together.

L1, L2 and dropout regularizer. L1 regularization calculates the summation of the absolute value of the weight vectors while trying to estimate the median of the data. On the other hand, L2 regularization calculates the summation of the square of the weight vector that tries to estimate the mean of the data. Dropout (Srivastava et al., 2014) is the most popular regularization technique. Dropout regularization randomly drops a neuron with a probability p and thus reduces the complexity of the model.

3.4. Loss functions

A loss function calculates the loss between the observed phenotype and predicted phenotype during training. The most popular loss function for GS is mean squared error (MSE). MSE measures the average squared difference between the observed and predicted phenotypes (Rachmatia et al., 2017; Ma et al., 2018; Khaki and Wang, 2019; Khaki et al., 2020; Shook et al., 2020; Jubair et al., 2021). Categorical cross entropy has also been applied as the loss function where the prediction task is converted to a classification problem (González-Camacho et al., 2016).

3.5. Optimization

The objective of training is to optimize the DNN. For optimizing, after each iteration, the weights need to be adjusted to minimize loss function. An iteration over the whole training set is called an epoch. Optimizers adjust the weights by applying certain algorithms and optimizing the loss function (LeCun et al., 2015; Dong et al., 2021). Optimization functions typically apply gradient descent to optimize the weights of the neural networks. The gradient measured is in relation to the loss function, that is, between the true and predicted value of the network as it currently predicts at this point in training. Stochastic Gradient Descent (SGD) (Ruder, 2016) is an optimizer that uses a subset of the training data to calculate and update the gradient of each weight. It uses a hyper-parameter called the learning rate to control how much it will adjust the weights from each iteration. There are also some algorithms that employ an adaptive learning rate strategy such as Adagrad (Ruder, 2016) and Adam (Kingma and Ba, 2014). Instead of using a fixed learning rate for all the weights, they use different learning rates for each of them. Adam calculates the first and second moments of the gradients and updates weights based on this calculation. For more detail on Adam and other optimization methods, we refer the readers to the review by Sun (2020).

3.6. Performance metrics

Performance metrics measure the performance of a machine learning model on a test dataset, which indicates how well the model will perform in production. As the ultimate goal is to rank genotypes to create a new variety, most of the research applied a correlation based performance metric such as Pearson Correlation Coefficient (PCC), or a ranking based measure such as Normalized Distributed Cumulative Gain (nDCG) (Järvelin and Kekäläinen, 2017). Some research also applied MSE as the performance metric.

PCC measures how linear the predicted phenotypes and the true phenotypes. PCC values range from −1 to 1 where a perfect linear relationship is indicated by 1 and completely non-linear relationship is indicated by −1. The formula of PCC is given below:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

In the above equation, x_i is the observed phenotype, $\bar{x}$ is the mean of observed phenotype, y_i is the predicted phenotype and $\bar{y}$ is the mean of predicted phenotype.

nDCG@k is a key measure for GS because it measures the quality of the ranking of the predicted phenotypes for the top k individuals (Järvelin and Kekäläinen, 2017; Jubair and Domaratzki, 2019). The formula for calculating *nDCG@k* is given below:

$$nDCG@k = \frac{DCG@k}{IDCG@k}$$

In the above equation, *DCG@k* means the discounted cumulative gain for the top k individuals. *DCG@k* measures the graded relevance of top k predicted genotypes. On the other hand, *IDCG@k* is the ideal *DCG* for the top k genotypes. The value of nDCG@k ranges from 0 to 1 where nDCG@k is 1 for perfectly ranked genotypes. nDCG was previously employed for measuring performance in GS by Ma et al. (2018) and then adopted by Jubair and Domaratzki (2019).

3.7. Training, test, and validation set

Supervised machine learning algorithms learn from the training data and their corresponding labels. Validation data is used to optimize the parameter of a machine learning algorithm while the final performance is measured on the test data. During training, the input of the DL algorithm is both genotyped and phenotyped data, with phenotypes being our target value to predict. An iteration for training a DL algorithm is called an epoch. After each epoch, the DL algorithm is validated on validation data to decide on the necessity of further training. During the validation step, the input to the DL algorithm is genotyped data while the model predicts the phenotypes. A loss between actual and predicted phenotypes for the validation data is measured. The training stops if there is no improvement in validation loss in n consecutive epochs. The final performance of the DL model is measured on the test data with the model that is obtained from the last most succesful epoch.

For a single environment trial, k-fold cross validation can be applied to divide the data into training test and validation sets. Runcie and Cheng (2019) recommended separating the training data and test data first and then applying k-fold cross validation on training data to divide the data in k training and validation sets (Refaeilzadeh et al., 2009).

For a multi-environment trial, a deep learning model can be evaluated in four scenarios, as described by Gillberg et al. (2019). In the first scenario, the authors used the trained model to observe the test lines in some environments. As some environments did not contain the test lines, the objective is to estimate traits of unobserved lines in those environments. In the second scenario, some lines are observed in some environments, but a subset of lines in the test set were never observed in any environments. The second scenario is more complex than the first one as the machine learning model has no prior knowledge of the test lines from any environment. In the third scenario, the machine learning model did not observe the environment where we want to grow the genotypes; however, the genotypes may be observed in other settings. The goal here is to predict traits for this new environment. Finally, the fourth scenario is the most extreme case of all scenarios. In this scenario, machine learning models do not have any prior information about the test lines and environment. That is, both lines and environments are new to the model and the objective is to predict traits for these new lines in a new environment.

In classical linear models, such as extensions to Genomic Best Linear Unbiased Prediction (GBLUP), environments are treated as a discrete category or as a relationship matrix between environments (de Los Campos et al., 2010; Endelman, 2011; Pérez and de Los Campos, 2014; Lopez-Cruz et al., 2015; Pérez-Elizalde et al., 2015; Crossa et al., 2016a; Cuevas et al., 2016; Hassen et al., 2018). Because of this, only the first two scenarios can be simulated, as environments unknown to the training set cannot be modeled. This demonstrates power of using deep learning models that are capable of incorporating heterogenous weather data directly into predictive models. In the examples we see in Section 5, deep learning models that incorporate weather data directly are capable of being evaluated in all four scenarios. However, the extent to which all these evaluations are performed varies.

4. Deep learning methods for single environment trials

Single environment trials have been the subject of many approaches. The main objective of GS for a single environment trial is to build a new variety of crops for that specific environment. A variety of deep learning models have been demonstrated to be successful for single environment datasets and building a new variety for crops (Pérez-Enciso and Zingaretti, 2019; Tong and Nikoloski, 2021). During the training phase of a deep learning algorithm, the typical inputs to the neural networks are the genotyped data and phenotypes. The model learns from these observed data, and then, after learning, it predicts the phenotypes of unobserved genotypes. From the predicted phenotype values, the top k genotypes are chosen as potential candidates for new varieties. Figure 5 shows how a new variety is developed by applying machine learning.

DL models have received a significant amount of attention recently (Pérez-Enciso and Zingaretti, 2019) and can predict complex traits. DL methods have been mostly either based on fully connected networks or convolutional neural networks, with the exception of the early neural networks for genomic selection (Gianola et al., 2011; González-Camacho et al., 2012; Pérez-Rodríguez et al., 2012). Below, we discuss the advancement and

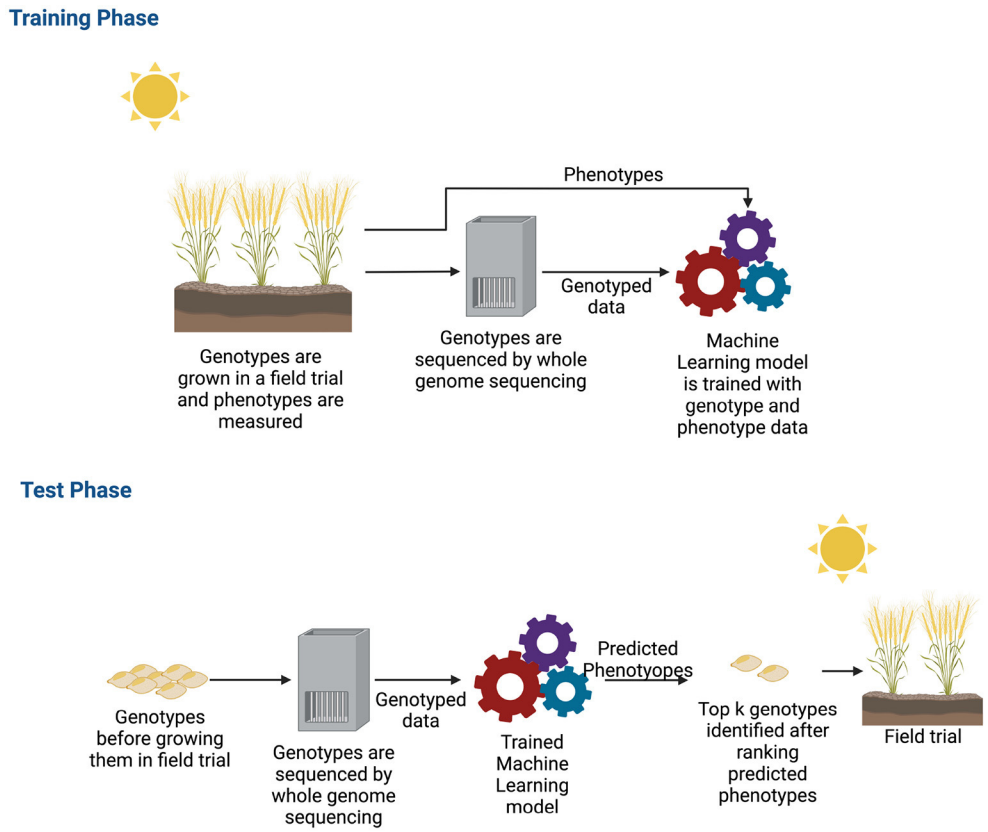

FIGURE 5
Workflow of genomic selection in single environment trial. For collecting training data, plants are grown in a field trial and phenotypes were measured. These plants are also genotyped. After obtaining both genotyped and phenotyped data, a machine learning model is trained during the training phase with both types of data. After the machine learning model is trained, potential genotypes that will be grown in the field are genotyped. These genotyped data are the input to the trained machine learning model. The machine learning model estimates the phenotypes. The estimated phenotypes are ranked and the top k phenotypes are chosen to select new varieties that will be grown in the field.

motivation of different neural networks for single environment trials.

Early implementations of neural networks in GS were mostly based on Bayesian Regularization, known as Bayesian Regularization Neural Network (BRNN) and Radial Basis Function Neural Network (RBFNN). Since some phenotypes follow a Gaussian distribution for some species, this works as the motivation to apply BRNN and RBFNN. Bayesian Regularization assumes the weights of the neural network come from a Gaussian distribution and calculates the loss between predicted phenotypes and true phenotypes by applying the Bayesian probabilistic approach. RBFNN, on the other hand, applies the radial basis function on each hidden neuron and thus, works as an activation function. These networks usually have one input layer, one hidden layer and an output layer. Gianola et al. (2011) proposed a BRNN for genomic selection and applied their framework to predict wheat yield. They compared the model with Bayesian Ridge Regression and showed that 11–18% improvements with their BRNN depending on the number of hidden neurons. Pérez-Rodríguez et al. (2012) compared two different shallow neural networks: BRNN and RBFNN with linear statistical models such as Bayesian ridge regression (BRR) (Bishop and Tipping, 2003), Bayesian LASSO (Hans, 2009), BayesA (Meuwissen et al., 2001), and BayesB (Meuwissen et al., 2001) on twelve different single environment trials and two phenotypes, grain yield and days to heading. Though there was no single winner for all traits and phenotypes, the research showed that non-linear models perform better than linear statistical models in general. Similar research is conducted by González-Camacho et al. (2012), which applied RBFNN on twenty-one traits of maize. The results showed that RBFNN performs similarly or better than statistical models.

After the moderate success of BRNN and RBFNN, researchers have applied shallow fully connected neural networks to GS. The shallow fully connected neural networks usually contain one or two hidden layers. González-Camacho et al. (2016) conducted a large study between a probabilistic ANN (PNN) and shallow MLP model on 33 datasets comprising

wheat and maize. The PNN model is the extension of RBFNN where a softmax activation function is applied to convert the output of the RBF kernel layer to a probability of c classes. The shallow MLP model consists of two hidden layers and also predicts a class as the output. As their model predicts a class, they transformed the regression problem into a classification problem by dividing the data into three categories, where the top category contains 30%, the middle category is 40% and the bottom category is the remaining 30%. The results showed that the PNN is better than the shallow MLP model for classification.

McDowell (2016)'s M.Sc. thesis also employed three shallow fully connected neural networks to GS consisting of one to three hidden layers. In their shallow models, they also employed different regularization techniques such as L2 and dropout regularization on some benchmark datasets, such as wheat and maize. Overall, the single hidden layer regularized neural networks performed better than the unregularized ones. The research showed that though increasing the number of hidden layers decreases the performance of their model, the neural networks are as good as the statistical models.

Rachmatia et al. (2017) proposed a different model than MLP known as Deep Belief Network (DBN). The motivation of applying DBN is to learn the genetic structure of the genomic data for a specific phenotype prediction. DBNs are usually applied in a semi-supervised setting where only a limited portion of the data is labeled. Thus, from all the available genomic data, it first tries to identify the pattern within the data by applying Restricted Boltzman Machine (RBM) (Zhang et al., 2018) blocks. Each RBM block in the DBN focuses on learning the probability distribution of its previous layer and, in the end, produces a feature vector for each input. This feature vector is the input to an output layer that predicts the phenotypes. Rachmatia et al. (2017) employed three block RBMs to predict both additive and non-additive effect phenotypes of maize, such as grain yield, female flowering, male flowering, and the anthesis-silking interval. The results showed that while the DBN is better than the existing statistical methods (BLUP and Bayesian LASSO) for predicting non-additive phenotypes, the performance for additive phenotypes drops significantly below BLUP by 3.5–7.5% for different traits.

Though most of the research has found that machine learning performs better than the statistical methods (McDowell, 2016; Rachmatia et al., 2017; Ma et al., 2018; Montesinos-López et al., 2019a) found that statistical methods are as good as machine learning methods and that SVMs (Hearst et al., 1998) are better than fully connected deep learning models. However, they also discussed the reason for the low performance of DL methods might be because of the small dataset they used, which only contained 270 wheat lines.

To the best of our knowledge, DeepGS (Ma et al., 2018) was the first method that applied CNN for GS. As GS data are high dimensional, DeepGS employed a combination of convolution, dropout and pooling layers. Conceptually, the adoption of CNN, with strides and window size, allows the possibility to integrate the effect of proximal markers and later when a linear layer is applied, capture the overall influence of markers on the phenotype. Ma et al. (2018) used a ranking procedure called Mean Normalized Cumulative Gain to rank the predicted individuals and obtained 2–7% improvements in the ranking of traits such as grain length, grain width, grain hardness, thousand-kernel weight, test weight, sodium dodecyl sulfate sedimentation, grain protein, and plant height, compared to RR-BLUP. They also showed that the selection of input markers and reducing the data dimension improved the performance of the deep learning model.

Jubair and Domaratzki (2019) proposed an ensemble CNN model to predict six traits of wheat. Each CNN model in the ensemble is created by a subset of randomly selected markers from the marker set. The final output is the average of the models in the ensemble. They compared their model with other non-ensemble and ensemble machine learning methods such as: support vector regression (SVR), CNN, ensemble SVR and Random Forests (Breiman, 2001) and RRBLUP. The work showed that overall ensemble machine learning methods are 20–30% better than single machine learning methods and slightly better than RRBLUP in correlation coeffcient and genotype ranking. The notable observation from this research is when CNNs are applied on a random marker set, the model still performs well, indicating little importance of the spatial relationship of GS for wheat. This observation also aligns with the observation of Ma et al. (2018).

Liu et al. (2019b) applied a dual-CNN architecture where after the input layer, they applied two separate streams of CNN that are not connected. The first stream has two CNN blocks and the second stream has one CNN block. The motivation behind employing two CNN streams is to use the second stream as a residual connection to the first CNN stream by aggregating two CNN streams together. The aggregated output is then passed to another CNN block, followed by a fully connected block for further processing and predicting phenotypes. Their model is trained and tested on a soybean dataset which performs better than DeepGS (Ma et al., 2018), MLP and statistical methods such as RRBLUP, BRR, BayesA, and Bayesian Lasso. The saliency map they applied also showed that the dual stream CNN model puts more importance on known biologically important markers for the specific traits.

There have been some other researches that employed CNN with limited success. Zingaretti et al. (2020) applied CNN in two polyploid species: strawberries and blueberries for predicting five different phenotypes. Their study showed that while CNN outperformed statistical models and Reproducing Kernel Hilbert Spaces (RKHS) for epistatic traits, it was not as successful for additive and mixed traits. Pook et al. (2020) showed the importance of dataset size while applying CNN in genomic selection. In an arabidopsis dataset, they showed that increasing training data could allow a CNN model to

outperform state-of-the-art models such as GBLUP and MLP. Sandhu et al. (2021c) applied MLP and CNN on multiple traits of spring wheat data. Their research showed that no unique MLP or CNN models worked well with all traits, since the number of hidden neurons, activation functions and the number of hidden layers differs from trait to trait. While there is 0 to 5% improvement in correlation score from RRBLUP with CNN and MLP, MLP performs consistently better than CNN by a very small margin.

Self-attention is a recent mechanism in DL which identifies the relationship among features and has been applied primarily to natural language processing (Devlin et al., 2018; Liu et al., 2019a; Raffel et al., 2020). One of the popular methods for incorporating self-attention is the transformer model. Though the transformer and attention have not been the subject of much research for GS, they have been applied successfully in similar research areas (Avsec et al., 2021; Ji et al., 2021; Le et al., 2022). Jubair et al. (2021) proposed a transformer-based DL method for genomic selection. The main motivation for employing the transformer in genomic selection was to capture and use the information on internal relationships between markers to predict phenotypes. To the best of our knowledge, this was the first transformer-based DL method for GS in a single environment trial. The model was trained on a barley dataset to predict Fusarium Head Blight (FHB) and Deoxynivalenol (DON) content in barley. Their work showed that even with a small amount of data (400 genotypes), the transformer-based DL method can be as good as or better than the state-of-the-art GS methods such as BLUP. It also outperformed other machine learning methods such as MLP, linear regression and decision trees. However, the authors also mentioned the limitation of the transformer in terms of memory and time complexity, as it needs a massive amount of memory and computation time and may not be feasible to consider all markers representing the whole genome.

Montesinos-Lopez et al. (2021) proposed an MLP model that applied negative log-likelihood of Poisson distribution as the loss function to predict counts of symptomatic spikelets of Fusarium Head Blight (FHB) in wheat in three different environments. The model was compared with the MLP model without the Poisson loss, Generalized Poisson Ridge regression, Generalized Poisson Lasso regression, Generalized Poisson Elastic net regression, Bayesian normal Ridge regression and Bayesian log normal Ridge regression. The MLP model with negative log-likelihood of Poisson distribution loss was better than the normal MLP model and performed similarly to Bayesian normal Ridge regression. The use of Poisson distributions in this research was motivated by the particular phenotype of FHB-affected spikelets: Poisson distributions are an accurate model for situations when counting of some quantity. The authors note that this extends beyond physical counts (as of spikelets) but to other situations as well, like laboratory test results and adverse drug events. Further attention is necessary for integrating Poisson models, as they are not commonly used in many datasets that fall into these categories.

Ubbens et al. (2021) also explained deep learning for GS. The work examined a kernel method for masking marker data while making prediction, to investigate the role that other factors, such as marker location, play on prediction. The authors concluded that deep learning models for GS may suffer from so-called shortcut-learning (Geirhos et al., 2020), where models learn from contextual information that is correlated with the outcome variable rather than the intended data, which in this case is the marker data. This suggests that further attention is necessary for using deep learning with GS. This also gives motivation for incorporating environmental data into models, as this yields larger data set and may mitigate overfitting.

5. Deep learning methods for multi-environment trials

The previous section shows that deep learning methods can predict complex traits in a single environment trial. However, extending models to multi-environmental datasets is challenging (Oakey et al., 2016; Crossa et al., 2017; Rincent et al., 2017). Here, a multi-environment deep learning model is defined as a deep learning architecture that takes environmental and/or genetic data as the input and predicts phenotype for a specific environment. Though the ideal scenario is training a model with genotyped data along with weather, soil and field management information (Khaki and Wang, 2019; Washburn et al., 2021), some of this data is sometimes not available and some of the multi-environment models are developed with environmental data only (Khaki et al., 2020; Lin et al., 2020; Shook et al., 2020; Zhong et al., 2022). Since in a multi-environment task, our goal is to estimate phenotypes of a crop in a new environment, the machine learning model typically needs field trialed data in many different environments. An environment is the growing cycle of a crop; for example, if a crop is grown multiple times of the year in the same field, each instance will be a different environment. As crops need to be grown numerous times in various locations, collecting these data may take years before it is possible to train a machine learning model (Spindel and McCouch, 2016). In addition, as the sources and types of data are different (genetic, weather, soil and field management data), the machine learning model can become very complex. Figure 6 shows the workflow of a multi-environment trial.

We have discussed single trait trials, where the deep learning model estimated one phenotype. There have been studies that develop multi-trait deep learning models for multi-environment trials, to predict multiple phenotypes simultaneously. The intuition behind this approach is that deep learning models will capture the information of common factors as well as

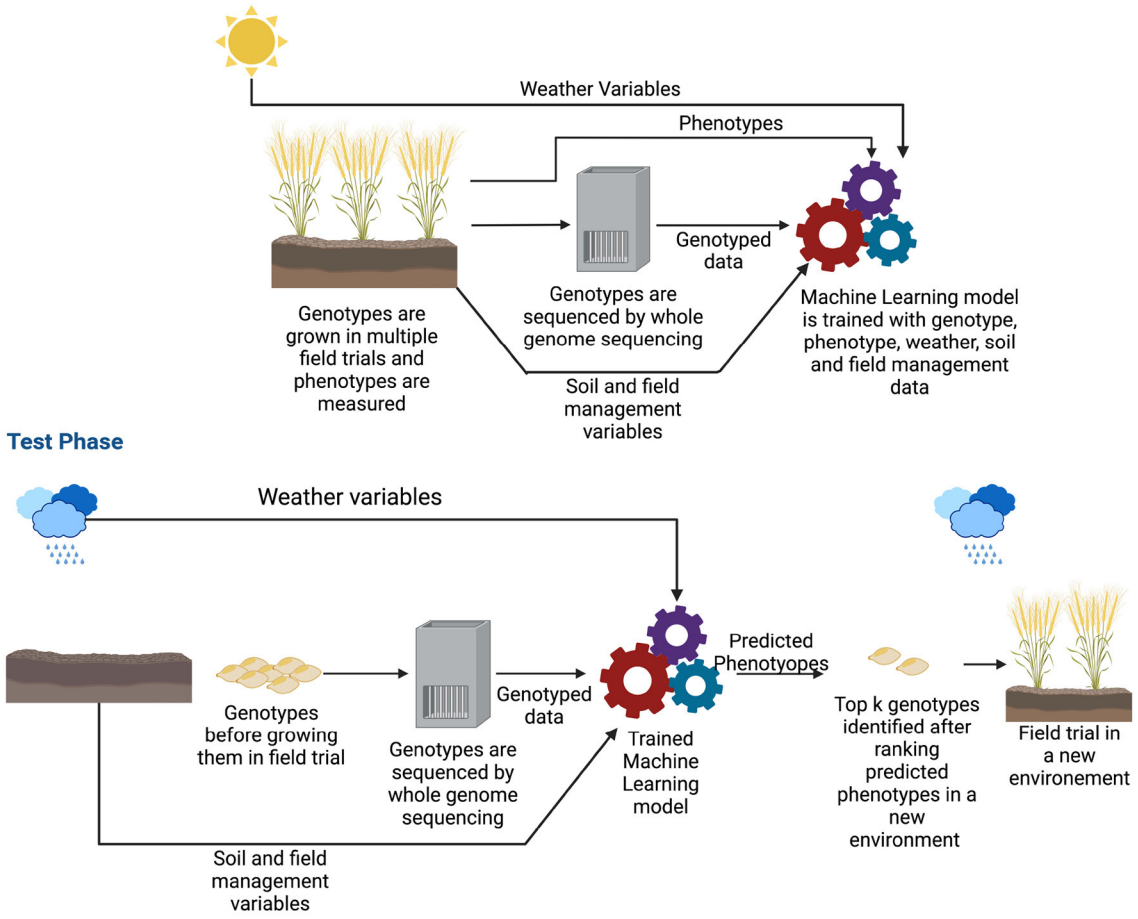

FIGURE 6
Workflow of genomic selection in a multi-environment trial. Before training the machine learning model, along with genotyped and phenotyped data, environmental information such as weather variables, soil and field management data are also collected. The genotypes are typically grown in multiple seasons/locations which provides a wide range of environmental data. During the training phase, the model is trained with all these data along with phenotypes. After the model is trained, in production, the model is given the genotyped data of crops along with environmental information of where the crops will be grown as the input. The model estimates the phenotype for that environment. Based on the estimated phenotypes, the top k genotypes are chosen and grown in the field.

phenotype-specific factors to predict phenotypes. Montesinos-López et al. (2019c) proposed an MLP containing three hidden layers and an output layer with three neurons to predict grain yield, days to heading and plant type of wheat. The input to this model is the concatenated matrix of environmental variables, a genomic relationship matrix obtained from genotypes, and a GxE term. The model was compared with GBLUP and MLP for the single phenotype. They observed that multi-trait MLP is better than the single trait MLP and overall, GBLUP model outperformed all of them with limited data (259 lines). Guo et al. (2020) also applied the same architecture of a multi-trait MLP model with a minimal wheat dataset (240 genotypes). Though their dataset was different than Montesinos-López et al. (2019c), as it consisted of genotyped data and environmental information, they also observed better performance. Sandhu et al. (2021a) applied the same MLP architecture on a wheat dataset comprised of spectral information of site-year and genetic information. These data were concatenated together to predict yield and protein content. The notable difference between this work and the previous two (Montesinos-López et al., 2019c; Guo et al., 2020) is the amount of data, as their dataset comprises 650 genotypes. The work showed that MLP performs similarly or better than GBLUP, BayesA, BayesB (Meuwissen et al., 2001), Random Forests (Breiman, 2001), CNN and Support Vector Machines (Hearst et al., 1998).

The model of Khaki and Wang (2019) was the first research to incorporate genetic information of corn and rich weather and soil data into a single deep learning framework. Their proposed method has two disjoint parts: (i) predicting weather variables for the growing cycle and (ii) predicting yield. In

the first part, they employed individual shallow MLP that take the previous 4 months' data of a specific weather variable as the input to predict the monthly weather variables of the growing cycle. In the second part, their deep learning model for predicting yield contained 21 fully connected neural network blocks where each block had 50 hidden neurons, an activation function and a regularization function. The input of this network was a concatenation of genetic information and weather variables obtained from the first part, along with soil information. The predicted output was the yield. As each hidden neuron combined environmental and genetic information, the motivation was to capture the GxE in each hidden neuron to predict yield. This model improved the correlation coefficient between predicted yield and original yield by 57% compared to the model that only had genomic data as the input.

Shook et al. (2020) proposed an LSTM-Fully Connected Neural Network based deep neural network that processed the inputs in two stages to predict soybean yield. In the first stage, LSTM blocks were employed on historical weather data. The weather data was divided into multiple time steps in the growing season where each time step is 30 days. An average of each weather variable was taken within the given time steps. LSTM blocks were applied on all the times steps to capture the temporal relationship and provide a context vector as an output optimized for yield prediction. After obtaining the context vector, maturity group information and a genotype cluster derived from applying k-means on the pedigree matrix were concatenated with the context vector. This concatenated vector was the input of the fully connected network that predicted yield. This model showed that when cluster and maturity group information are added, it leads to a lower root mean square error (RMSE).

Deep learning has also been successfully applied when no genetic or pedigree information is available. The deep learning model of Lin et al. (2020) had two parts: (i) attention-based LSTM network that captured the effect of environmental variables over time on yield, and (ii) multi-task learning (MTL) networks that predicted location-specific corn yield anomaly. The weather information was the weekly average of minimum and maximum temperature, precipitation, growing degree days and killing degree days. This model was compared to Random Forests and Lasso (Ranstam and Cook, 2018) and had the lowest RMSE among the three.

Khaki et al. (2020) employed a CNN-RNN based deep learning model on a dataset that contained historical yield and weather information and soil data for corn and soybean. In this work, CNNs were applied to yearly data to capture the spatial information of weather and soil information. Two separate CNN networks were employed that output two vectors to capture the spatial information of weather and soil variables. After obtaining the spatial information, LSTMs were applied to obtain the temporal relationship within the data. To employ LSTM, the distributed representations of soil and weather along with the corresponding yield of previous t years were concatenated and provided as the input to the LSTM, which predicted the yield of the current growing cycle. This model improved the correlation coefficient by 20–25% compared to LASSO (Ranstam and Cook, 2018) based on different years and crops.

Gangopadhyay et al. (2020) applied a dual attention neural network on a soybean dataset that comprised 13 years of data of 5,839 genotypes resulting in 103,365 observations. The attention networks are known for their ability to identify important features as it calculates an importance score (attention score) for each feature and aggregate all the features in a context vector by applying weighting based on the attention score. The dataset contained weekly weather variables such as average direct normal irradiance, average precipitation, average relative humidity, maximum direct normal irradiance, maximum surface temperature, minimum surface temperature and average surface temperature. A fully connected neural network followed by an attention layer was applied initially to the weather variables to capture the spatial information. Then, on the output of the spatial attention layer, multiple LSTM layers followed by another attention layer were applied to capture the temporal relation and predict the soybean yield. Though their model had comparable performance to the baseline model (LSTMs and LSTMs with temporal attention), they showed that the attention layer provided their model with more interpretability. They also observed that the attention mechanism identified average precipitation as the most influencing factor for soybean growth in most weeks.

McCormick et al. (2021) applied nine different architectures of LSTMs to predict the current growth stage of soybean. The architectures of LSTMs mostly differ in the number of layers and hidden neurons. These models were applied to a dataset consisting of 187 environments and 13,673 observations of soybean, based on different planting times and locations. Their weather variables included daily minimum and maximum temperature, solar radiation, night length, longitude and latitude. The task of these LSTM models was to predict, from seven growth stage variables, what the current stage of the plant is. In their LSTM model, they also included the output of a knowledge-based model named CROPGRO (Boote et al., 1998; Salmerón and Purcell, 2016) as features and showed that including the predicted output from CROPGRO as a feature improved the mean absolute error by 2.76 and 5.51% for different traits.

Washburn et al. (2021) applied a CNN-MLP based neural network on maize data. Their dataset is similar to Khaki and Wang (2019) as their data contains genetic, environmental, soil and field management information. Initially, this model processed the inputs in three parts: (i) fully connected blocks were applied to genetic data, (ii) CNN blocks were applied to environmental information and (iii) fully connected neural network blocks were employed on soil and field management data. Then the outputs of these three parts were concatenated and passed to fully

connected blocks to predict yields. They observed that soil and environmental factors play a bigger role than the genetic information for yield prediction as they comprised 35 and 22% of the importance score, respectively. From the feature perspective, precipitation, vapor pressure and plant density were the most influential features. They also observed that adding historical information for a specific location improved prediction and overall, the performance of the proposed CNN-MLP model was comparable to or better than GBLUP-based models.

Måløy et al. (2021) employed a variation of transformers named performers (Choromanski et al., 2020) on a barley dataset to predict yield. Performers were developed as attention-based models capable of capturing long-range interactions between features; this is appropriate for genomic data where attention related SNPs may be distant in the genome. In their work, the environment variables were of two types: (i) mean value of temperature and precipitation for the entire growing season and (ii) mean temperature and cumulative precipitation for each day of the growing season (historical data). Performers were applied to the genomic data to extract genomic features. An MLP was employed when the mean weather variables for the entire growing season were considered, or a performer was employed when historical weather data was considered as the input, to obtain the relevant features from the weather variables. Finally, both feature representations were concatenated and passed as the input to the regression layer to predict yield. Their results demonstrated that the model that considered historical weather information had the highest R^2 scores. Their model also outperformed a CNN + MLP model by 1.3% in R^2 score. In addition, as the historical weather data based model was better than average weather based models, the results showed that research needs to concentrate on integrating historical weather data and genomic data together in a meaningful way for different growth stages of crops to predict genotype-specific yield for a specific environment.

Zhong et al. (2022) proposed a multi-task learning model where each task-specific layer predicted the average yield of maize for a specific county. Their input variables contained weather, remote sensing and soil data. K-means clustering was applied to county-level yield and weather and soil data to obtain spatial features. In addition, an LSTM and a fully connected neural network were applied to the weather data and soil data, respectively, to extract temporal and soil features. Finally, these three outputs were combined and served as the input to the county-specific output layer that predicted yield for that specific county. The result of the proposed model showed that killing degree days was one of the major driving factors for yield loss in 2012. As this model predicted county-specific yield, it did not integrate genetic information. However, this model considered spatial-temporal relationships which can be integrated with genomic data and have the potential to play a vital role in capturing GxE.

Sharma et al. (2022) proposed a deep learning model that contains four modules: genome, weather, field management and soil module and predicted maize yield. For each of these modules, they obtained an embedded vector representing the feature set of that module by employing different types of neural networks. For example, two different CNNs were employed for weather and genomic data, while two separate MLPs were used for field management and soil data to obtain embeddings for each module. In addition, they applied an attention mechanism between the genome embedding and weather data embedding to learn an embedding that replicates GxE. Finally, the embeddings for GxE, weather, field management and soil were concatenated, and a fully connected layer was employed to predict the yield. The results demonstrated 1.45 times better correlation coefficient than GBLUP and CNN-based methods. This approach is unique compared to other methods as they used the attention mechanism to obtain GxE, which ideally puts more importance on the environmental variables that influence maize yield.

In Table 3, we list the deep learning-based academic papers that work with multi-environment trial and environmental data. Some single-environment models (Sandhu et al., 2021b,c, 2022) employed an MLP, similar to the model of Montesinos-López et al. (2019c), to predict quantitative traits in another location or year. As these models did not incorporate environmental data into the model, we consider them single-environment models. Thus, this type of research, while important in demonstrating advances in prediction of traits in new situations, is not summarized in this survey. In addition, typically, environmental information is not readily available, and even if they are available, these models are complex in nature as different types of data need different types of ANNs to extract meaningful features. Thus, the development of new deep learning approaches in this new research area is comparatively slower than single environment trial models. We expect that, as data collection and integration continues in crop breeding programs, more detailed datasets containing rich genotypic, weather, soil and management data will be generally available. Models that incorporate this data will become more common as well, as the data becomes more reliable, standardized and available.

6. Discussion

Genomic selection is a well-established tool for crop breeding, and non-linear supervised deep learning models are increasingly being used to predict phenotypes for complex traits. As datasets become increasingly feature-rich and large enough to train complex models, the use of deep learning models becomes more feasible. This trend also enables incorporating heterogeneous weather, soil and field management data to be added to predict environmental effects on genotypes. Typically, weather variables such as precipitation and vapor pressure

TABLE 3 Papers on multi-environment deep learning models.

Year	References	DL model	Crops	Traits	Geno data	Weather data	Soil data	Other data
2019	Khaki and Wang (2019)	MLP	Corn	Yield	Yes	Yes	Yes	
2019	Montesinos-López et al. (2019c)	MLP	Wheat	Yield, Days to Heading	Yes	Yes	No	
2020	Shook et al. (2020)	LSTM-MLP	Soybean	Yield	No	Yes	No	Genotype Cluster
2020	Khaki et al. (2020)	CNN-RNN	Corn, Soybean	Yield	No	Yes	Yes	Historical Yield, Field Management
2020	Lin et al. (2020)	Att-LSTM	Corn	Yield	No	Yes	No	
2020	Gangopadhyay et al. (2020)	MLP LSTM Att	Soybean	Yield	No	Yes	No	
2020	Guo et al. (2020)	MLP	Wheat	Yield, Harvest Index, Spike Fertility, Thousand Grain Weight	Yes	Yes	No	
2021	Sandhu et al. (2021a)	MLP	Wheat	Yield, Protein Content	Yes	Yes	No	
2021	Washburn et al. (2021)	CNN MLP	Maize	Yield	yes	Yes	Yes	Field Management
2021	Måløy et al. (2021)	Transformers MLP	Barley	Yield	yes	Yes	No	
2022	Zhong et al. (2022)	LSTM MLP	Maize	Yield	No	Yes	Yes	
2022	Sharma et al. (2022)	CNN MLP Att	Maize	Yield	Yes	Yes	Yes	Field Management

In the table, MLP means Fully Connected Networks and Att means attention networks.

(Gangopadhyay et al., 2020; Washburn et al., 2021) are the most important. However, other environmental variables such as day length (Tacarindua et al., 2013; Rahman et al., 2018; Islam et al., 2019), and maximum and minimum temperature (Gul et al., 2020; Moore et al., 2021) may also become vital based on the crop species and environment. These weather variables are the most influential during the early stages of crop development (Washburn et al., 2021). As these weather variables are mostly available as hourly or daily data, determining how this information can be added to the deep learning models, especially during the early stages of development, is essential (Gangopadhyay et al., 2020). Most existing methods employed neural networks on monthly average data of weather variables for the whole growing season (Khaki and Wang, 2019; Khaki et al., 2020; Shook et al., 2020). To add more information in the early stage of development, a variable length time window approach can be adopted where in the beginning, time window can be shorter, and in the later stage, the size of time window can be increased. Additionally, the use of unsupervised learning techniques to learn appropriate representations of weather data is a potential area of additional exploration.

Some research (Khaki and Wang, 2019; Washburn et al., 2021) incorporated a wide range of soil and field management variables in their model, such as soil electrical conductivity, calcium carbonate content, saturated hydraulic conductivity, gypsum content, plant density, irrigation, and pH. Typically, water and nutrition-related soil variables are the most relevant (Washburn et al., 2021). Though it is observed that soil variables are more important than weather variables (Washburn et al., 2021), in most of the current research, these variables are not considered due to the lack of data. Recently, the use of IoT devices to collect soil and field data (for example, weather variables described above) is gaining popularity (Sharma et al., 2020). As IoT devices can collect data more accurately and frequently, it has become possible to estimate soil nutrients and moisture for the growing cycle (Sharma et al., 2020). These estimated values can be the input of the deep learning algorithm to estimate phenotypes. Another source of data that can work as the input of GS is high-quality image data of fields. Drones with high-quality cameras have been used recently to capture field images. These images can be fed into a deep learning model to add additional information about the field. Recent research has indicated that using early phenotypic data, including spectral data collected by drones, yields models that can be competitive with GS (Adak et al., 2021) in predicting phenotype at harvest. Since GS aims to estimate yield even before sowing, we need to ensure that the information added in the model is collected either before sowing the plants or is estimated for the growing season based on previously available data. Collecting phenotypic information during growing season to attempt to predict future phenotypes represents a different philosophy of approaching GS, either when this data is used alone or in conjunction with genomic data. This approach may be consider advantageous in forestry or perennial crops, where early phenotypic information may shape long-term field trials (Cros et al., 2015; Kwong et al., 2017; Faville et al., 2018; Crain et al., 2020; Lebedev et al., 2020; Archambeau et al., 2022).

Most of the multi-environment deep learning architecture we discussed so far sought to capture the spatial and/or temporal effect of environmental variables on traits and later incorporated genomic data into the model for estimating phenotypes. Though a few deep learning models were developed by employing attention for genomic selection (Gangopadhyay et al., 2020; Jubair et al., 2021; Måløy et al., 2021), we believe attention-based architectures are the most promising approach for genomic selection. Attention-based methods can capture both temporal and spatial information and summarize the input data by aggregating them based on importance scores. As a robust model needs to be trained on different types and data sources, attention may play a significant role by providing more importance to the critical parts of different data sources (Gangopadhyay et al., 2020; Jubair et al., 2021; Måløy et al., 2021).

As one of the major challenges of GS for multi-environment is the data, collaboration among breeders and a well-defined data collection strategy will be useful to take GS application into production (Spindel and McCouch, 2016; Xu et al., 2022). To the best of our knowledge, the only user-friendly software designed to integrate multiple data sources in genomic selection is learnMet (Westhues et al., 2022). This software allows the user to employ traditional machine learning methods, such as XGBoost and Random Forests, and MLP-based neural networks. However, complex models also need to be packaged as user-friendly software to make more accurate predictions and bring GS to breeders.

In summary, continued advances in deep learning, driven by disparate application areas such as vision and languages, will continue to be adapted to GS, especially in the context of large datasets incorporating environmental conditions. Future research should focus on extracting meaningful features from different data sources and leveraging their interactions to predict quantitative traits. To extract meaningful features, choosing an appropriate deep learning architecture that can capture different relationships within each type of data will be the first step. For example, weather and image data during the growing season contains a spatial-temporal relationship, whereas soil data before the growing season has a spatial relationship. There are also heterogeneous unstructured text data about field management, such as the sowing pattern of crops, the amount of water supplied during irrigation, and notes on the overall condition of fields. Deep learning architecture such as transformers may play a vital role as they have been successfully employed to extract meaningful features from genomic (Avsec et al., 2021; Ji et al., 2021; Monteiro et al., 2022), weather (Måløy et al., 2021), and unstructured text data (Devlin et al., 2018; Raffel et al.,

2020). However, GS for multi-environment model may need to employ different types of neural networks on different sources of data depending on the data property, such as the spatial, temporal and spatial-temporal relationship between variables. Future research also should focus on how to capture the interrelationship between genotypes and these features to predict quantitative traits.

Author contributions

SJ: formal analysis, investigation, and writing—original draft. MD: writing—review, editing, and supervision. All authors contributed to the article and approved the submitted version.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abadi, M. (2016). *TensorFlow: Learning Functions at Scale, Vol. 51*. New York, NY: Association for Computing Machinery. doi: 10.1145/3022670.2976746

Acquaah, G. (2009). *Principles of Plant Genetics and Breeding*. Chichester, UK: John Wiley & Sons.

Adak, A., Murray, S. C., and Anderson, S. L. (2021). Temporal phenomic predictions from unoccupied aerial systems can outperform genomic predictions. *BioRxiv*. doi: 10.1101/2021.10.06.463310

Ali, M., Zhang, Y., Rasheed, A., Wang, J., and Zhang, L. (2020). Genomic prediction for grain yield and yield-related traits in chinese winter wheat. *Int. J. Mol. Sci*. 21, 1342. doi: 10.3390/ijms21041342

Anilkumar, C., Sunitha, N., Devate, N. B., Ramesh, S., et al. (2022). Advances in integrated genomic selection for rapid genetic gain in crop improvement: a review. *Planta* 256, 1–20. doi: 10.1007/s00425-022-03996-y

Archambeau, J., Benito Garzón, M., Barraquand, F., de Miguel, M., Plomion, C., and González-Martínez, S. C. (2022). Combining climatic and genomic data improves range-wide tree height growth prediction in a forest tree. *Am. Nat*. 200, E141-E159. doi: 10.1086/720619

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* 18, 1196–1203. doi: 10.1038/s41592-021-01252-x

Azadi, H., Ho, P., and Hasfiati, L. (2011). Agricultural land conversion drivers: a comparison between less developed, developing and developed countries. *Land Degradat. Dev*. 22, 596–604. doi: 10.1002/ldr.1037

Bandeira e Sousa, M., Cuevas, J., de Oliveira Couto, E. G., Pérez-Rodríguez, P., Jarquín, D., Fritsche-Neto, R., et al. (2017). Genomic-enabled prediction in maize using kernel models with genotype× environment interaction. *G3* 7, 1995–2014. doi: 10.1534/g3.117.042341

Bebis, G., and Georgiopoulos, M. (1994). Feed-forward neural networks. *IEEE Potentials* 13, 27–31. doi: 10.1109/45.329294

Bishop, C. M., and Tipping, M. E. (2003). "Bayesian regression and classification," in *Advances in Learning Theory: Methods, Models and Applications*, eds J. A. K. Suykens, I. Horvath, S. Basu, C. Micchelli, and J. Vandewalle (Amsterdam: IOS Press), 267–285.

Boote, K. J., Jones, J. W., Hoogenboom, G., and Thornton, P. K. (1998). "The CROPGRO model for grain legumes," in *Understanding Options for Agricultural Production*, eds K. J. Boote, G. Hoogenbom, and P. K. Thornton (Dordrecht: Springer), 99–128. doi: 10.1007/978-94-017-3624-4_6

Bourgault, M., Löw, M., Tausz-Posch, S., Nuttall, J., Delahunty, A., Brand, J., et al. (2018). Effect of a heat wave on lentil grown under free-air CO_2 enrichment (face) in a semi-arid environment. *Crop Sci*. 58, 803–812. doi: 10.2135/cropsci2017.09.0565

Breiman, L. (2001). Random forests. *Mach. Learn*. 45, 5–32. doi: 10.1023/A:1010933404324

Burgueño, J., de los Campos, G., Weigel, K., and Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. *Crop Sci*. 52, 707–719. doi: 10.2135/cropsci2011.06.0299

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint* arXiv:1409.1259. doi: 10.3115/v1/W14-4012

Chollet, F., and Others. (2018). *Keras: The Python Deep Learning library*. Available online at: https://ui.adsabs.harvard.edu/abs/2018ascl.soft06022C

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., et al. (2020). Rethinking attention with performers. *arXiv preprint* arXiv:2009.14794. doi: 10.48550/arXiv.2009.14794

Costa-Neto, G., Crespo-Herrera, L., Fradgley, N., Gardner, K., Bentley, A. R., Dreisigacker, S., et al. (2022). Envirome-wide associations enhance multi-year genome-based prediction of historical wheat breeding data. *bioRxiv*. doi: 10.1093/g3journal/jkac313

Crain, J., Bajgain, P., Anderson, J., Zhang, X., DeHaan, L., and Poland, J. (2020). Enhancing crop domestication through genomic selection, a case study of intermediate wheatgrass. *Front. Plant Sci*. 11, 319. doi: 10.3389/fpls.2020.00319

Cros, D., Denis, M., Sánchez, L., Cochard, B., Flori, A., Durand-Gasselin, T., et al. (2015). Genomic selection prediction accuracy in a perennial crop: case study of oil palm (elaeis guineensis jacq.). *Theor. Appl. Genet*. 128, 397–410. doi: 10.1007/s00122-014-2439-z

Crossa, J., de los Campos, G., Maccaferri, M., Tuberosa, R., Burgue no, J., and Pérez-Rodríguez, P. (2016a). Extending the marker× environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. *Crop Sci*. 56, 2193–2209. doi: 10.2135/cropsci2015.04.0260

Crossa, J., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., Saint-Pierre, C., et al. (2016b). Genomic prediction of gene bank wheat landraces. *G3* 6, 1819–1834. doi: 10.1534/g3.116.029637

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci*. 22, 961–975. doi: 10.1016/j.tplants.2017.08.011

Cuevas, J., Crossa, J., Montesinos-López, O. A., Burgueño, J., Pérez-Rodríguez, P., and de los Campos, G. (2017). Bayesian genomic prediction with genotype× environment interaction kernel models. *G3* 7, 41–53. doi: 10.1534/g3.116.035584

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., de los Campos, G., et al. (2016). Genomic prediction of genotype× environment interaction kernel regression models. *Plant Genome* 9, 1–20. doi: 10.3835/plantgenome2016.03.0024

Cuevas, J., Montesinos-López, O., Juliana, P., Guzmán, C., Pérez-Rodríguez, P., González-Bucio, J., et al. (2019). Deep kernel for genomic and near infrared predictions in multi-environment breeding trials. *G3* 9, 2913–2924. doi: 10.1534/g3.119.400493

Danilevicz, M. F., Gill, M., Anderson, R., Batley, J., Bennamoun, M., Bayer, P. E., et al. (2022). Plant genotype to phenotype prediction using machine learning. *Front. Genet*. 13, 822173. doi: 10.3389/fgene.2022.822173

de Los Campos, G., Gianola, D., Rosa, G. J., Weigel, K. A., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel hilbert spaces methods. *Genet. Res*. 92, 295–308. doi: 10.1017/S0016672310000285

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805

Dey, A. (2016). Machine learning algorithms: a review. *Int. J. Comput. Sci. Inf. Technol*. 7, 1174–1179.

Dong, S., Wang, P., and Abbas, K. (2021). A survey on deep learning and its applications. *Comput. Sci. Rev*. 40, 100379. doi: 10.1016/j.cosrev.2021.100379

Dubey, S. R., Singh, S. K., and Chaudhuri, B. B. (2022). Activation functions in deep learning: a comprehensive survey and benchmark. *Neurocomputing* 503, 92–108. doi: 10.1016/j.neucom.2022.06.111

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., and Dehmer, M. (2020). An introductory review of deep learning for prediction models with big data. *Front. Artif. Intell*. 3, 4. doi: 10.3389/frai.2020.00004

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrblup. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Faville, M. J., Ganesh, S., Cao, M., Jahufer, M., Bilton, T. P., Easton, H., et al. (2018). Predictive ability of genomic selection models in a multi-population perennial ryegrass training set using genotyping-by-sequencing. *Theor. Appl. Genet*. 131, 703–720. doi: 10.1007/s00122-017-3030-1

Ferrão, L. F. V., Ferrao, R. G., Ferrão, M. A. G., Francisco, A., and Garcia, A. A. F. (2017). A mixed model to multiple harvest-location trials applied to genomic prediction in coffea canephora. *Tree Genet. Genomes* 13, 95. doi: 10.1007/s11295-017-1171-7

Gangopadhyay, T., Shiik, J., Singh, A. K., and Sarkar, S. (2020). *Interpreting the Impact of Weather on Crop Yield Using Attention*. Available online at: https://www.researchgate.net/profile/Tryambak-Gangopadhyay/publication/346972032_Interpreting_the_Impact_of_Weather_on_Crop_Yield_Using_Attention/links/5fd66db0a6fdccdcb8c47b58/Interpreting-the-Impact-of-Weather-on-Crop-Yield-Using-Attention.pdf (accessed December 12, 2022).

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., et al. (2020). Shortcut learning in deep neural networks. *Nat. Mach. Intell*. 2, 665–673. doi: 10.1038/s42256-020-00257-z

Gianola, D., Okut, H., Weigel, K. A., and Rosa, G. J. (2011). Predicting complex quantitative traits with bayesian neural networks: a case study with jersey cows and wheat. *BMC Genet*. 12, 87. doi: 10.1186/1471-2156-12-87

Gillberg, J., Marttinen, P., Mamitsuka, H., and Kaski, S. (2019). Modelling G×E with historical weather information improves genomic prediction in new environments. *Bioinformatics* 35, 4045–4052. doi: 10.1093/bioinformatics/btz197

Goddard, M., and Hayes, B. (2007). Genomic selection. *J. Anim. Breed. Genet*. 124, 323–330. doi: 10.1111/j.1439-0388.2007.00702.x

González-Camacho, J. M., Crossa, J., Pérez-Rodríguez, P., Ornella, L., and Gianola, D. (2016). Genome-enabled prediction using probabilistic neural network classifiers. *BMC Genomics* 17, 208. doi: 10.1186/s12864-016-2553-1

González-Camacho, J. M., de Los Campos, G., Pérez, P., Gianola, D., Cairns, J., Mahuku, G., et al. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet*. 125, 759–771. doi: 10.1007/s00122-012-1868-9

Gul, F., Ahmed, I., Ashfaq, M., Jan, D., Fahad, S., Li, X., et al. (2020). Use of crop growth model to simulate the impact of climate change on yield of various wheat cultivars under different agro-environmental conditions in khyber pakhtunkhwa, Pakistan. *Arabian J. Geosci*. 13, 1–14. doi: 10.1007/s12517-020-5118-1

Guo, J., Khan, J., Pradhan, S., Shahi, D., Khan, N., Avci, M., et al. (2020). Multi-trait genomic prediction of yield-related traits in us soft wheat under variable water regimes. *Genes* 11, 1270. doi: 10.3390/genes11111270

Hans, C. (2009). Bayesian lasso regression. *Biometrika* 96, 835–845. doi: 10.1093/biomet/asp047

Hassen, M. B., Bartholomé, J., Valè, G., Cao, T.-V., and Ahmadi, N. (2018). Genomic prediction accounting for genotype by environment interaction offers an effective framework for breeding simultaneously for adaptation to an abiotic stress and performance under normal cropping conditions in rice. *G3* 8, 2319–2332. doi: 10.1534/g3.118.200098

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intell. Syst. Appl*. 13, 18–28. doi: 10.1109/5254.708428

Heffner, E. L., Sorrells, M. E., and Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Sci*. 49, 1–12. doi: 10.2135/cropsci2008.08.0512

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*. 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Holliday, J. A., Wang, T., and Aitken, S. (2012). Predicting adaptive phenotypes from multilocus genotypes in sitka spruce (picea sitchensis) using random forest. *G3* 2, 1085–1093. doi: 10.1534/g3.112.002733

Howard, R., Gianola, D., Montesinos-López, O., Juliana, P., Singh, R., Poland, J., et al. (2019). Joint use of genome, pedigree and their interaction with environment for predicting the performance of wheat lines in new environments. *G3* 9, 2925–2934. doi: 10.1534/g3.119.400508

Islam, M. R., Fujita, D., Watanabe, S., and Zheng, S.-H. (2019). Variation in photosensitivity of flowering in the world soybean mini-core collections (gmwmc). *Plant Prod Sci*. 22, 220–226. doi: 10.1080/1343943X.2018.1561197

Järvelin, K., and Kekäläinen, J. (2017). "Ir evaluation methods for retrieving highly relevant documents," in *ACM SIGIR Forum, Vol. 51* (New York, NY: ACM), 243–250.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. (2021). DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 37, 2112–2120. doi: 10.1093/bioinformatics/btab083

Jubair, S., and Domaratzki, M. (2019). "Ensemble supervised learning for genomic selection," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (San Diego, CA: IEEE), 1993–2000.

Jubair, S., Tucker, J. R., Henderson, N., Hiebert, C. W., Badea, A., Domaratzki, M., et al. (2021). Gptransformer: a transformer-based deep learning method for predicting fusarium related traits in barley. *Front. Plant Sci*. 12, 761402. doi: 10.3389/fpls.2021.761402

Khaki, S., and Wang, L. (2019). Crop yield prediction using deep neural networks. *Front. Plant Sci*. 10, 621. doi: 10.3389/fpls.2019.00621

Khaki, S., Wang, L., and Archontoulis, S. V. (2020). A cnn-rnn framework for crop yield prediction. *Front. Plant Sci*. 10, 1750. doi: 10.3389/fpls.2019.01750

Kim, P. (2017). "Convolutional neural network," in *MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence* (Berkeley, CA: Apress), 121–147. doi: 10.1007/978-1-4842-2845-6_6

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint* arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., and Inman, D. J. (2021). 1d convolutional neural networks and applications: a survey. *Mech. Syst. Signal Process*. 151, 107398. doi: 10.1016/j.ymssp.2020.107398

Kukačka, J., Golkov, V., and Cremers, D. (2017). Regularization for deep learning: a taxonomy. *arXiv preprint* arXiv:1710.10686. doi: 10.48550/arXiv.1710.10686

Kwong, Q. B., Ong, A. L., Teh, C. K., Chew, F. T., Tammi, M., Mayes, S., et al. (2017). Genomic selection in commercial perennial crops: applicability and improvement in oil palm (elaeis guineensis JACQ.). *Sci. Rep*. 7, 1–9. doi: 10.1038/s41598-017-02602-6

Le, N. Q. K., Ho, Q.-T., Nguyen, V.-N., and Chang, J.-S. (2022). Bert-promoter: an improved sequence-based predictor of dna promoter using bert pre-trained model and shap feature selection. *Comput. Biol. Chem*. 99, 107732. doi: 10.1016/j.compbiolchem.2022.107732

Lebedev, V. G., Lebedeva, T. N., Chernodubov, A. I., and Shestibratov, K. A. (2020). Genomic selection for forest tree improvement: methods, achievements and perspectives. *Forests* 11, 1190. doi: 10.3390/f11111190

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, H., Tian, S., Li, Y., Fang, Q., Tan, R., Pan, Y., et al. (2020). Modern deep learning in bioinformatics. *J. Mol. Cell. Biol*. 12, 823–827. doi: 10.1093/jmcb/mjaa030

Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst*. 33, 6999–7019. doi: 10.1109/TNNLS.2021.3084827

Lin, T., Zhong, R., Wang, Y., Xu, J., Jiang, H., Xu, J., et al. (2020). Deepcropnet: a deep spatial-temporal learning framework for county-level corn yield estimation. *Environ. Res. Lett*. 15, 034016. doi: 10.1088/1748-9326/ab66cb

Linehan, V., Thorpe, S., Andrews, N., and Beaini, F. (2012). "Food demand to 2050: opportunities for australian agriculture-algebraic description of agrifood model," in *ABARES Research Report, May 2012* (Canberra, NW).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019a). Roberta: a robustly optimized bert pretraining approach. *arXiv preprint* arXiv:1907.11692. doi: 10.48550/arXiv.1907.11692

Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., and Xu, D. (2019b). Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Front. Genet*. 10, 1091. doi: 10.3389/fgene.2019.01091

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J.-L., et al. (2015). Increased prediction accuracy in wheat breeding trials using a marker× environment interaction genomic selection model. *G3* 5, 569–582. doi: 10.1534/g3.114.016097

Måløy, H., Windju, S., Bergersen, S., Alsheikh, M., and Downing, K. L. (2021). Multimodal performers for genomic selection and crop yield prediction. *Smart Agric. Technol*. 1, 100017. doi: 10.1016/j.atech.2021.100017

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., et al. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* 248, 1307–1318. doi: 10.1007/s00425-018-2976-9

McCormick, R. F., Truong, S. K., Rotundo, J., Gaspar, A. P., Kyle, D., Van Eeuwijk, F., et al. (2021). Intercontinental prediction of soybean phenology via hybrid ensemble of knowledge-based and data-driven models. *In Silico Plants* 3, diab004. doi: 10.1093/insilicoplants/diab004

McDowell, R. (2016). *Genomic selection with deep neural networks* (Ph.D. thesis). Iowa State University.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Millet, E. J., Kruijer, W., Coupel-Ledru, A., Prado, S. A., Cabrera-Bosquet, L., Lacube, S., et al. (2019). Genomic prediction of maize yield across european environmental conditions. *Nat. Genet*. 51, 952–956. doi: 10.1038/s41588-019-0414-y

Monteiro, N. R., Oliveira, J. L., and Arrais, J. P. (2022). Dtitr: end-to-end drug-target binding affinity prediction with transformers. *Comput. Biol. Med*. 147, 105772. doi: 10.1016/j.compbiomed.2022.105772

Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J., and Hernández-Suárez, C. M. (2018). Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3* 8, 3813–3828. doi: 10.1534/g3.118.200740

Montesinos-López, O. A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C. M., Montesinos-López, A., et al. (2019a). A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding. *G3* 9, 601–618. doi: 10.1534/g3.118.200998

Montesinos-López, O. A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C. M., Montesinos-López, A., et al. (2019b). New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes. *G3* 9, 1545–1556. doi: 10.1534/g3.119.300585

Montesinos-López, O. A., Montesinos-López, A., Kismiantini, Roman-Gallardo, A., Gardner, K., Lillemo, M., Fritsche-Neto, R., et al. (2022). Partial least squares enhances genomic prediction of new environments. *Front. Genet*. 13, 920689. doi: 10.3389/fgene.2022.920689

Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., Barrón-López, J. A., Martini, J. W., Fajardo-Flores, S. B., et al. (2021). A review of deep learning applications for genomic selection. *BMC Genomics* 22, 1–23. doi: 10.1186/s12864-020-07319-x

Montesinos-López, O. A., Montesinos-López, A., Tuberosa, R., Maccaferri, M., Sciara, G., Ammar, K., et al. (2019c). Multi-trait, multi-environment genomic prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods. *Front. Plant Sci*. 10, 1311. doi: 10.3389/fpls.2019.01311

Montesinos-Lopez, O. A., Montesinos-Lopez, J. C., Salazar, E., Barron, J. A., Montesinos-Lopez, A., Buenrostro-Mariscal, R., et al. (2021). Application of a poisson deep neural network model for the prediction of count data in genome-based prediction. *Plant Genome* 14, e20118. doi: 10.1002/tpg2.20118

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to Linear Regression Analysis*. Hoboken, NJ: Wiley.

Moore, C. E., Meacham-Hensold, K., Lemonnier, P., Slattery, R. A., Benjamin, C., Bernacchi, C. J., et al. (2021). The effect of increasing temperature on crop photosynthesis: from enzymes to ecosystems. *J. Exp. Bot*. 72, 2822–2844. doi: 10.1093/jxb/erab090

Moradi, R., Berangi, R., and Minaei, B. (2020). A survey of regularization strategies for deep models. *Artif. Intell. Rev*. 53, 3947–3986. doi: 10.1007/s10462-019-09784-7

Nawaz, M. A., and Chung, G. (2020). Genetic improvement of cereals and grain legumes. *Genes* 11, 255. doi: 10.3390/genes111 11255

Oakey, H., Cullis, B., Thompson, R., Comadran, J., Halpin, C., and Waugh, R. (2016). Genomic selection in multi-environment crop trials. *G3* 6, 1313–1326. doi: 10.1534/g3.116.027524

Ogutu, J. O., Piepho, H.-P., and Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proc*. 5, S11. doi: 10.1186/1753-6561-5-S3-S11

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., Devito, Z., et al. (2017). "Automatic differentiation in PyTorch," in *NIPS 2017 Workshop on Autodiff* (Long Beach, CA). Available online at: https://openreview.net/forum?id=BJJsrmfCZ

Pérez, P., and de Los Campos, G. (2014). Genome-wide regression and prediction with the bglr statistical package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442

Pérez-Elizalde, S., Cuevas, J., Pérez-Rodríguez, P., and Crossa, J. (2015). Selection of the bandwidth parameter in a Bayesian kernel regression model for genomic-enabled prediction. *J. Agric. Biol. Environ. Stat*. 20, 512–532. doi: 10.1007/s13253-015-0229-y

Pérez-Enciso, M., and Zingaretti, L. M. (2019). A guide for using deep learning for complex trait genomic prediction. *Genes* 10, 553. doi: 10.3390/genes10070553

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manès, Y., and Dreisigacker, S. (2012). Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3* 2, 1595–1605. doi: 10.1534/g3.112.003665

Poehlman, J. M. (1979). *Breeding Field Crops*. Avi Publishing Company.

Pook, T., Freudenthal, J., Korte, A., and Simianer, H. (2020). Using local convolutional neural networks for genomic prediction. *Front. Genet*. 11, 561497. doi: 10.3389/fgene.2020.561497

Putra, A. R., Yen, J. D., and Fournier-Level, A. (2022). Forecasting trait responses in novel environments to aid seed provenancing under climate change. *Mol. Ecol. Resour*. 2022, 13728. doi: 10.1111/1755-0998.13728

Qing, X., and Niu, Y. (2018). Hourly day-ahead solar irradiance prediction using weather forecasts by lstm. *Energy* 148, 461–468. doi: 10.1016/j.energy.2018.01.177

Rachmatia, H., Kusuma, W., and Hasibuan, L. (2017). Prediction of maize phenotype based on whole-genome single nucleotide polymorphisms using deep belief networks. *J. Phys*. 835, 012003. doi: 10.1088/1742-6596/835/1/012003

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res*. 21, 1–67.

Rahman, H., Bennett, R. A., and Kebede, B. (2018). Molecular mapping of qtl alleles of brassica oleracea affecting days to flowering and photosensitivity in spring brassica napus. *PLoS ONE* 13, e0189723. doi: 10.1371/journal.pone.0189723

Ranstam, J., and Cook, J. (2018). Lasso regression. *J. Br. Surgery* 105, 1348–1348. doi: 10.1002/bjs.10895

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-validation. *Encyclopedia Database Syst*. 5, 532–538. doi: 10.1007/978-0-387-39940-9_565

Rincent, R., Kuhn, E., Monod, H., Oury, F.-X., Rousset, M., Allard, V., et al. (2017). Optimization of multi-environment trials for genomic selection based on crop models. *Theor. Appl. Genet*. 130, 1735–1752. doi: 10.1007/s00122-017-2922-4

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint* arXiv:1609.04747. doi: 10.48550/arXiv.1609.04747

Runcie, D., and Cheng, H. (2019). Pitfalls and remedies for cross validation with multi-trait genomic prediction methods. *G3* 9, 3727–3741. doi: 10.1534/g3.119.400598

Salman, A. G., Heryadi, Y., Abdurahman, E., and Suparta, W. (2018). Single layer multi-layer long short-term memory (lstm) model with intermediate variables for weather forecasting. *Procedia Comput. Sci*. 135, 89–98. doi: 10.1016/j.procs.2018.08.153

Salmerón, M., and Purcell, L. C. (2016). Simplifying the prediction of phenology with the dssat-cropgro-soybean model based on relative maturity group and determinacy. *Agric. Syst*. 148, 178–187. doi: 10.1016/j.agsy.2016.07.016

Sandhu, K. S., Aoun, M., Morris, C. F., and Carter, A. H. (2021b). Genomic selection for end-use quality and processing traits in soft white winter wheat breeding program with machine and deep learning models. *Biology* 10, 689. doi: 10.3390/biology10070689

Sandhu, K. S., Lozada, D. N., Zhang, Z., Pumphrey, M. O., and Carter, A. H. (2021c). Deep learning for predicting complex traits in spring wheat breeding program. *Front. Plant Sci*. 11, 613325. doi: 10.3389/fpls.2020.613325

Sandhu, K. S., Patil, S. S., Aoun, M., and Carter, A. H. (2022). Multi-trait multi-environment genomic prediction for end-use quality traits in winter wheat. *Front. Genet*. 13, 831020. doi: 10.3389/fgene.2022.831020

Sandhu, K. S., Patil, S. S., Pumphrey, M., and Carter, A. (2021a). Multitrait machine-and deep-learning models for genomic selection using spectral information in a wheat breeding program. *Plant Genome* 14, e20119. doi: 10.1002/tpg2.20119

Sawitri, S., Tani, N., Na'iem, M., Widiyatno, W., Indrioko, S., Uchiyama, K., et al. (2020). Potential of genome-wide association studies and genomic selection to improve productivity and quality of commercial timber species in tropical rainforest, a case study of shorea platyclados. *Forests* 11, 239. doi: 10.3390/f11020239

Searchinger, T., Waite, R., Hanson, C., Ranganathan, J., Dumas, P., Matthews, E., et al. (2019). *Creating a sustainable food future: A menu of solutions to feed nearly 10 billion people by 2050*. Final report. Availbale online at: https://agritrop.cirad.fr/593176/1/WRR_Food_Full_Report_0.pdf (accessed December 12, 2022).

Sharma, A., Jain, A., Gupta, P., and Chowdary, V. (2020). Machine learning applications for precision agriculture: a comprehensive review. *IEEE Access* 9, 4843–4873. doi: 10.1109/ACCESS.2020.3048415

Sharma, S., Partap, A., Balaguer, M. A., d,. L., Malvar, S., and Chandra, R. (2022). Deepg2p: fusing multi-modal data to improve crop production. *arXiv preprint* arXiv:2211.05986. doi: 10.48550/arXiv.2211.05986

Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S., and Singh, A. K. (2020). Crop yield prediction integrating genotype and weather variables using deep learning. *arXiv preprint* arXiv:2006.13847. doi: 10.1371/journal.pone.0252402

Song, H., Wang, X., Guo, Y., and Ding, X. (2022). G × EBLUP: A novel method for exploring genotype by environment interactions and genomic prediction. *Front. Genet*. 13, 972557. doi: 10.3389/fgene.2022.972557

Spindel, J. E., and McCouch, S. R. (2016). When more is better: how data sharing would accelerate genomic selection of crop plants. *New Phytol*. 212, 814–826. doi: 10.1111/nph.14174

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res*. 15, 1929–1958.

Sun, R.-Y. (2020). Optimization for deep learning: an overview. *J. Operat. Res. Soc. China* 8, 249–294. doi: 10.1007/s40305-020-00309-6

Szandała, T. (2021). "Review and comparison of commonly used activation functions for deep neural networks," in *Bio-inspired Neurocomputing*, eds A. K. Bhaoi, P. K. Mallick, C. -M. Liu, and V. E. Balas (Singapore: Springer), 203–224. doi: 10.1007/978-981-15-5495-7_11

Tacarindua, C. R., Shiraiwa, T., Homma, K., Kumagai, E., and Sameshima, R. (2013). The effects of increased temperature on crop growth and yield of soybean grown in a temperature gradient chamber. *Field Crops Res*. 154, 74–81. doi: 10.1016/j.fcr.2013.07.021

Tong, H., and Nikoloski, Z. (2021). Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *J. Plant Physiol*. 257, 153354. doi: 10.1016/j.jplph.2020.153354

Ubbens, J., Parkin, I., Eynck, C., Stavness, I., and Sharpe, A. G. (2021). Deep neural networks for genomic prediction do not estimate marker effects. *Plant Genome* 14, e20147. doi: 10.1002/tpg2.20147

United Nations (2019). *Department of Economic and Social Affairs|Food Security and Nutrition and Sustainable Agriculture*. Available online at: https://sdgs.un.org/topics/food-security-and-nutrition-and-sustainable-agriculture (accessed February 04, 2021).

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

Van Meijl, H., Havlik, P., Lotze-Campen, H., Stehfest, E., Witzke, P., Domínguez, I. P., et al. (2018). Comparing impacts of climate change and mitigation on global agriculture by 2050. *Environ. Res. Lett*. 13, 064021. doi: 10.1088/1748-9326/aabdc4

Varshney, R., Roorkiwal, M., and Sorrells, M. E. (2017). "Genomic selection for crop improvement: An introduction," in *Genomic Selection for Crop Improvement: New Molecular Breeding Strategies for Crop Improvement*, eds R. K. Varshney, M. Roorkiwal, and M. E. Sorrels (Cham: Springer), 1–6. doi: 10.1007/978-3-319-63170-7_1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural information Processing Systems* (Long Beach, CA), 5998–6008.

Wang, S., Wei, J., Li, R., Qu, H., Chater, J. M., Ma, R., et al. (2019). Identification of optimal prediction models using multi-omic data for selecting hybrid rice. *Heredity* 123, 395–406. doi: 10.1038/s41437-019-0210-6

Wang, X., Xu, Y., Hu, Z., and Xu, C. (2018). Genomic selection methods for crop improvement: current status and prospects. *Crop J*. 6, 330–340. doi: 10.1016/j.cj.2018.03.001

Washburn, J. D., Cimen, E., Ramstein, G., Reeves, T., O'Briant, P., McLean, G., et al. (2021). Predicting phenotypes from genetic, environment, management, and historical data using cnns. *Theor. Appl. Genet*. 134, 3997–4011. doi: 10.1007/s00122-021-03943-7

Washburn, J. D., Mejia-Guerra, M. K., Ramstein, G., Kremling, K. A., Valluru, R., Buckler, E. S., et al. (2019). Evolutionarily informed deep learning methods for predicting relative transcript abundance from dna sequence. *Proc. Natl. Acad. Sci. U.S.A*. 116, 5542–5549. doi: 10.1073/pnas.1814551116

Westhues, C. C., Simianer, H., and Beissinger, T. M. (2022). learnmet: an r package to apply machine learning methods for genomic prediction using multi-environment trial data. *G3* 12, jkac226. doi: 10.1093/g3journal/jkac226

Xu, Y., Zhang, X., Li, H., Zheng, H., Zhang, J., Olsen, M. S., et al. (2022). Smart breeding driven by big data, artificial intelligence and integrated genomic-enviromic prediction. *Mol. Plant* 15, 1664–1695. doi: 10.1016/j.molp.2022.09.001

Yu, Y., Cao, J., and Zhu, J. (2019a). An lstm short-term solar irradiance forecasting under complicated weather conditions. *IEEE Access* 7, 145651–145666. doi: 10.1109/ACCESS.2019.2946057

Yu, Y., Si, X., Hu, C., and Zhang, J. (2019b). A review of recurrent neural networks: Lstm cells and network architectures. *Neural Comput*. 31, 1235–1270. doi: 10.1162/neco_a_01199

Zhang, N., Ding, S., Zhang, J., and Xue, Y. (2018). An overview on restricted boltzmann machines. *Neurocomputing* 275, 1186–1199. doi: 10.1016/j.neucom.2017.09.065

Zhong, R., Zhu, Y., Wang, X., Li, H., Wang, B., You, F., et al. (2022). Detect and attribute the extreme maize yield losses based on spatio-temporal deep learning. *Fundament. Res*. doi: 10.1016/j.fmre.2022.05.006

Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P. R., et al. (2020). Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Front. Plant Sci*. 11, 25. doi: 10.3389/fpls.2020.00025