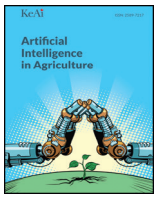
Artificial Intelligence in Agriculture 8 (2023) 60–76

Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/



GxENet: Novel fully connected neural network based approaches to incorporate GxE for predicting wheat yield

Sheikh Jubair [a,*], Olivier Tremblay-Savard [a], Mike Domaratzki [b]

[a] Department of Computer Science, University of Manitoba, 66 Chancellors Cir, Winnipeg, MB, R3T 2N2, Canada
[b] Department of Computer Science, University of Western Ontario, 1151 Richmond St, London, ON, N6A 3K7, Canada

ARTICLE INFO

Article history:
Received 27 November 2022
Received in revised form 30 April 2023
Accepted 15 May 2023
Available online 19 May 2023

Keywords:
Genomic prediction
Multi-environment trial
Deep learning
GxE
Enviromics

ABSTRACT

The expression of quantitative traits of a line of a crop depends on its genetics, the environment where it is sown and the interaction between the genetic information and the environment known as GxE. Thus to maximize food production, new varieties are developed by selecting superior lines of seeds suitable for a specific environment. Genomic selection is a computational technique for developing a new variety that uses whole genome molecular markers to identify top lines of a crop. A large number of statistical and machine learning models are employed for single environment trials, where it is assumed that the environment does not have any effect on the quantitative traits. However, it is essential to consider both genomic and environmental data to develop a new variety, as these strong assumptions may lead to failing to select top lines for an environment. Here we devised three novel deep learning frameworks incorporating GxE within the deep learning model and predicted line-specific yield for an environment. In the process, we also developed a new technique for identifying environment-specific markers that can be useful in many applications of environment-specific genomic selection. The result demonstrates that our best framework obtains 1.75 to 1.95 times better correlation coefficients than other deep learning models that incorporate environmental data depending on the test scenario. Furthermore, the feature importance analysis shows that environmental information, followed by genomic information, is the driving factor in predicting environment-specific yield for a line. We also demonstrate a way to extend our framework for new data types, such as text or soil data. The extended model also shows the potential to be useful in genomic selection.

1. Introduction

Food crop production faces impending challenges in feeding the global population, including increasing population, reduction in agricultural inputs, and global climate change. These challenges have led to global problems in food security. Around 85 million more people were facing a severe food crisis in 2021 compared to what was reported in 2016, which resulted in around 193 million people facing severe hunger across 36 countries (FAO, 2022). This crisis has been worsened recently both by human made and natural phenomena such as domestic food price inflation, war and pandemic (FAO, 2022; World Bank Group, 2022; UN World Food Programme, 2022; Kakaei et al., 2022). To address this problem, we need to produce more food with the limited resources available by creating improved varieties of crops that can perform well in different environments.

A genotype of a crop organism is its complete set of genetic makeup that influence its traits. However, a genotype often refers to genes that have different alleles. Thus to avoid confusion, we will refer genotypes as lines in the rest of the article. To create a new variety from lines with improved traits, we need to consider a line's genetics and its interaction with the environment where it is sown (Washburn et al., 2021; Lin et al., 2020). This phenomenon is known as genotype by environment interaction (GxE), which refers to the fact that even if a variety produces the desired values of quantitative traits in one environment, it may not provide us with the same outcome in another environment (Lenz et al., 2017). Thus, any tools that are developed to aid in crop breeding need to replicate the impact of GxE within the model by incorporating genetic and environmental information. In this work, we aim to build such computational tools that estimate a trait even before sowing the crop.

The advancement of the next-generation sequencing technology, such as genotyping by sequencing (GBS) and restriction-site associated DNA sequencing (RAD-seq), enables us to capture the genetic diversity among different lines of the same species (Nguyen et al., 2019; Esposito

* Corresponding author.
E-mail address: jubairs@myumanitoba.ca (S. Jubair).

https://doi.org/10.1016/j.aiia.2023.05.001

et al., 2020). These sequencing technologies usually provide us with many molecular markers, typically covering the whole genome of the species. These markers can be employed for studying genetic diversity, identification of quantitative trait loci for a specific trait (Boudhrioua et al., 2020; Zhang et al., 2019; Zegeye et al., 2018; Tang et al., 2018) and for estimating genomic breeding values for different traits (Tong and Nikoloski, 2021; Jubair et al., 2021; Khaki and Wang, 2019; Ma et al., 2018; Rachmatia et al., 2017; Crossa et al., 2016).

The environment of a crop is the combination of the weather, soil and field management information of where it is sown (Jubair and Domaratzki, 2023; Sharma et al., 2022; Washburn et al., 2021; Khaki and Wang, 2019). While the most frequently obtained weather and soil information includes precipitation, temperature, pressure, radiation, wind speed, humidity, day length, soil electrical conductivity, calcium carbonate, saturated hydraulic conductivity, gypsum content and pH, the most frequently used field management variables include management practices such as sowing pattern, number of pre-irrigations, and the amount of fertilizer and insecticide applied on the field (Washburn et al., 2021; Lin et al., 2020; Khaki and Wang, 2019; Montesinos-López et al., 2019; Shook et al., 2021; Khaki et al., 2020; Guo et al., 2020; Sandhu et al., 2021). However, based on the previous research, this is a minimal list of the weather, soil and field management variables. The potential list of variables can be arbitrarily large. The effect of environmental variables on crops differs from growing cycle to growing cycle, even for the same trait and line (Sonkar et al., 2019; Tadesse et al., 2019). For example, researchers developed different wheat varieties suitable for different traits and weather conditions such as heat-stressed (Ly et al., 2018), high rainfall (Tadesse et al., 2010) and favourable environments where crops are provided with optimum water and heat (Juliana et al., 2017).

Genomic selection (GS) is a computational technique of selecting top lines to create new varieties of a crop. GS takes genotyped data and, increasingly, environmental information as input and predicts quantitative traits as outputs (Khaki and Wang, 2019; Meuwissen et al., 2001). There have been many genomic selection applications that use machine learning with genomic data only (Zhang et al., 2019; Jubair et al., 2021; Ma et al., 2018; Gianola et al., 2011; Pérez-Rodríguez et al., 2012; González-Camacho et al., 2012, 2016; Rachmatia et al., 2017; Jubair and Domaratzki, 2019). This type of genomic selection is known as a single-environment trial as it is assumed that the environmental effect on plants remains constant; hence single environment trials are not able to capture the environmental effect on genotypes of a line. Then, some genomic selection models take environmental information only as the input and predict average quantitative trait of lines for that specific environment (Lin et al., 2020; Shook et al., 2021; Khaki et al., 2020). As no genetic information is given as the input to the models, they also do not capture the effect of the environment on genotypes. These models are known as multi-environment models as they are able to predict average quantitative traits for different environments. On the other hand, other more recent multi-environment models consider both environmental and genomic data and the interaction between them to predict line specific phenotype (Washburn et al., 2021; Khaki and Wang, 2019; Montesinos-López et al., 2019; Jarquín et al., 2014). However, the process of building multi-environment models incorporating environmental and genomic data is not well understood. In this work, we aim to build novel machine learning approaches for a wheat dataset that combines genomic and environmental information that is capable of making predictions in novel environments where previous crop performance data is not available.

As GS for multi-environment trials requires different types of data, such as weather and genomic data, incorporating these pieces of data together to capture GxE is a significant challenge. Deep learning (DL) methods that employ neural networks are known for their ability to handle heterogeneous data and have been successfully applied in some recent papers for GS with multi-environment trials (Washburn et al., 2021; Lin et al., 2020; Khaki and Wang, 2019; Montesinos-López et al., 2019). The main building blocks of deep learning models are artificial neural networks, such as fully connected neural networks (linear layers), recurrent neural networks and convolutional neural networks. Each deep learning model has at least one input layer, more than two hidden layers of neural networks and an output layer. The input and output of the neural networks are the neurons, where the input layer neurons are the input features such as genetic marker data or environment variables. The input to the hidden neural networks are the features from the previous layer and produce a learned feature representation as the output by applying some functions, based on the type of employed neural network. Finally, the output layer takes the output of the last hidden layer as the input and employs the neural network functions to make the final prediction. In this work, all our proposed frameworks employ fully connected neural networks where each neuron in a hidden layer is the linear function of all neurons of the previous layer. Thus each neuron of the current layer represents summarized information of all previous neurons.

In this work, we proposed three deep learning frameworks that combine genotyped data and environmental information, such as weather and field management data, to replicate GxE and predict wheat yield in multi-environment trials. These frameworks differ on how GxE is incorporated within the deep learning model or whether field management information is integrated. Overall, we have the following contributions:

1. We proposed a novel concept of global and local marker sets for feature selection where the global marker sets are the markers important for yield prediction irrespective of any environment. On the other hand, local markers are environment-specific important markers for a certain trait.
2. We devised two deep learning frameworks where we carefully modelled the interaction between weather variables and genotyped data and predicted line-specific yield value of wheat.
3. We extended one of the frameworks (third framework) to integrate unstructured text about field notes.
4. We employed DeepLift (Shrikumar et al., 2017), a method to identify which features contribute more towards prediction in a deep learning model, to understand how environmental data, genetic information and text data contribute to predicting yield in our models.

2. Materials and methods

2.1. Dataset

2.1.1. Genotyped and phenotyped data

Genotypic data for Spring Wheat *(Triticum aestivum)* is collected from the CIMMYT dataverse used in the Feed the Future Innovation Lab (Poland et al., 2021). The phenotypic data and the environmental information are also obtained from the CIMMYT dataverse for four different nurseries: International Bread Wheat Screening Nursery (IBWSN), High Rainfall Wheat Yield Trial (HRWYT), Elite Selection Wheat Yield Trial (ESWYT) and Wheat Yield Collaboration Yield Trial (WYCYT). Here, the meaning of trial and nursery is the same, and we will use trial to indicate both of them in the later part of this work. Each trial is located in many places. Trials are also categorized into different mega-environments based on weather conditions such as the amount of rainfall, soil acidity, the necessity of irrigation, and altitudes of the locations. Thus locations in a trial have similar weather conditions. Typically lines are sown in multiple cycles, and in a cycle, the same lines are sown in numerous locations of the same trial. The cycles are usually numbered, such as 45th IBWNSN and 1st WYCYT, where the first part indicates the cycle number and the second part is the trial. We collected the data of 1st WYCYT to 6th WYCYT, 11th HRWYT to 27th HRWYT, 29th ESWYT to 36th ESWYT and 36th IBWSN to 52nd IBWSN.

Although the locations of the CIMMYT wheat breeding program have eight mega-environments, our collected data mostly falls in two

mega-environments: mega-environment 1 and mega-environment 2. Locations in mega-environment 1 have favourable conditions for wheat breeding where rainfall is usually low and irrigation is optimal. On the other hand, locations in mega-environment 2 are high rainfall areas where precipitation occurs during the growing cycle, and irrigation is not needed. Table 1 shows the nursery information along with their mega-environment information.

Locations in mega-environment 1 have favourable conditions for wheat breeding where rainfall is usually low and irrigation is optimal. On the other hand, locations in mega-environment 2 are high rainfall areas where precipitation occurs during the growing cycle, and irrigation is not needed. Table 1 shows the nursery information along with their mega-environment information.

The lines were sown over multiple years in different locations which we are going to refer as a site-year (combination of locations and year). As the environment of a specific location is not constant and changes each year, each site-year is considered a different environment. Table 2 shows the number of unique locations and lines, number of cycles and total lines for each nursery type. From the table, we observe that IBWSN trials have the highest number of unique lines that are sown in 171 unique locations which creates 72,776 line-site-year combinations. The quantity of line-site-year combinations of IBWSN trials are followed by ESWYT, HRWYT and WYCYT respectively.

We performed the Anderson-Darling test to check whether the distributions of yield in any of the trials follow a normal distribution. Our result shows that none of the four trials are normally distributed as the statistics of the Anderson-Darling test range from 224 to 11, which is much higher than the critical values of all significance levels. Table 3 shows the detailed result of the test along with their significance levels.

2.1.2. Weather data

The weather data for each site-year is collected from the CIMMYT dataverse from 1990 to 2018 containing all locations of International Wheat Improvement Network (IWIN). The weather data contains 769 locations along with nine weather variables of each location such as the hourly average amount of precipitation, maximum relative humidity, minimum relative humidity, shortwave radiation ($MJ/m^2/d$), maximum and minimum temperature (C), maximum vapour pressure deficit (kPa), 2 m wind speed (m/s) and 10 m wind speed (m/s). Although the sowing date of crops are recorded in entirety, there are many missing values for when the crops are harvested; hence, we consider nine months of environmental data as the input to the machine learning model. The nine months period starts at least two months before the sowing date to capture the environmental effect on the soil before sowing (as the sowing date is available), and the following seven months are considered as the growing season. A monthly average of all the weather variables for each of the nine months are calculated, which provides us with $9 \times 9 = 81$ weather variables for each environment.

2.1.3. Field notes

Field notes are obtained from the CIMMYT dataverse for each site-year. These notes are mostly unstructured text and contain a wide variety of information. Data also differs from trial cycle to trial cycle. This data contains information such as how much and which fertilizer is applied, disease development information, number of irrigations before sowing, moisture available before sowing, major weed species, soil aluminum toxicity and many more. Though our aim was to collect information before sowing, we are unable to verify that all the information in this data is taken before sowing.

Table 1
Environments of each nursery. ME refers to Mega-Environment. CIMMYT has 6 M-environments for Spring Wheat.

	IBWSN	HRWYT	ESWYT	WYCYT
Rainfall	Low rainfall	>500 mm	Low rainfall	Mixed
Mega-Environment	ME1	ME2	ME1	Mixed
Irrigation	Optimal	No	Optimal	No information
Overall Condition	Favourable	Rainfall during cropping cycle	Favourable	Mixed

Table 2
Nursery-type specific information of genotypes and locations.

	IBWSN	HRWYT	ESWYT	WYCYT
Unique Locations	171	122	216	77
Unique Lines	2619	305	369	109
Number of Cycles	17	14	8	6
Number of Line-Site-Year Combinations	72,776	6,984	25,712	3,012

2.2. Train-test Split

From the genotyped data, we created five different training, test and validation partitions where each set has 70%–15%–15% training, validation and test split. While dividing the data, we ensure that lines that are selected for the test set in a partition are not observed in the training set. Thus the training data contains the information of the environments where this 15% test will be sown as other lines are already sown in those environments. As the model already observed the environment of test lines in the training data, in later part of this work, we will refer this test case as environment observed scenario or test scenario one.

For each previously created partitions, we also randomly sample some locations with 85% probability that the location will be in the training set and 15% probability that the location is in the test set. In this scenario, if the location is in the test set, we did not include any of the site-year data for that location in the training set or validation set. Although most of the lines in the test set were sown in other site-years, the training set also does not contain the 15% lines separated for testing in the previous step. Thus there may be some lines in the test set that are not present in the training data. As the model did not observe the test locations in training data, we will refer this test case as environment unobserved scenario or test scenario two. Finally, the training-test partitioning strategy creates two test scenarios for each partition: i) no lines in the training set are also in the test set, but the test set and training set may contain different lines grown in the same environment, and ii) the training data and the test data do not contain any locations in common, but the training set may contain information on how some lines performed in other site-years.

2.3. Weather data clustering

In this work, we group site-years into clusters to identify statistically related locations and use these groups to find group-specific important markers. To obtain the grouping, we calculated the yearly average of each weather variable for each location from 1990 until 2019 of all IWIN locations. We then applied hierarchical agglomerative clustering with the number of clusters $c = 25$. The output of the clustering method

Table 3
Anderson-Darling test of yield distribution. Since the statistics are larger than the critical values of all significance levels, the hypothesis that the data comes from a normal distribution is rejected.

		Critical Values				
Trials	Test Statistics	15%	10%	5%	2.5%	1%
IBWSN	224.479	0.576	0.656	0.787	0.918	1.092
HRWYT	193.598	0.576	0.656	0.787	0.917	1.091
ESWYT	28.912	0.576	0.656	0.787	0.918	1.092
WYCYT	11.149	0.576	0.656	0.787	0.917	1.091

is the cluster assignment for each site-year. Finally, one of the cluster categories is assigned to a location in which the site-year combination of that location is the most frequent. Though we clustered all IWIN locations, there are 320 unique IWIN locations (1483 site-year) in four trials for which we identified the cluster category. The primary purpose of clustering in our work is to use the cluster to find environment-specific important markers. Thus we did not focus on finding the appropriate number of clusters. We use the cluster category information of a location to identify cluster-specific important markers of wheat for yield.

2.4. Feature selection

Each marker in our dataset is represented by three values: 1, 0, −1. We applied the Hardy-Weinberg equilibrium (HW) (Acquaah, 2009) to each marker in the training set to obtain the genotype frequency. Each genotype is then replaced by one of the three quantities obtained from applying HW. As we have five training sets, each marker of a line will have five frequency values. An average of these five frequency values is used as the genotype frequency. Fig. 1 shows an example of how the average frequency is calculated.

Research shows that some markers contribute towards yield irrespective of the environment (Lenz et al., 2017). Also, some specific markers contribute more in a specific condition. For instance, Lenz et al. (Lenz et al., 2017) demonstrated that selecting the top 250 previously known important markers of black spruce results in the same correlation coefficient score obtained by randomly selecting 4993 markers. They also observed that selecting fewer than 500 markers randomly decreases the correlation score between the predicted traits and the true traits. When the important markers are not known, previous research shows that feature selection methods were able to identify biologically and statistically significant markers for yield prediction (Jubair et al., 2021). Thus by applying feature selection, we aim to identify important markers irrespective of any environment (global marker set) as well as markers that play an essential role in a specific condition (local marker set).

To identify the global marker set, the first step is to calculate the average yield of each line over all environments. After finding the average yield of lines, mutual information (MI) regression (Ross, 2014) feature selection is applied to each marker. Again, as we have five different training sets, each marker will have five different MI scores. We calculated the average MI score for all the markers across five folds and then selected the top 2000 markers with the highest MI scores. Fig. 2 shows how MI is obtained for a specific training set.

To obtain the local marker set, we first exclude all markers that are in the global marker set. Then an average phenotypic value is calculated for all lines in each environmental cluster obtained previously in section 2.3. For each cluster, the average MI (averaged over all training sets) for each marker is measured and the top 100 markers were chosen for each cluster. As the global feature sets are excluded from these markers, the expectation is that the identified markers are more related to the environmental effect. This marker selection process selected another 2052 unique markers. After combining global and local marker sets, we have 4052 markers for our machine learning model. Fig. 3 shows the procedure for obtaining the local marker set.

2.5. Deep learning framework 1 (F1)

We now describe our first DL framework. This framework was designed to test whether genomic and environmental information can be treated equally as data in the model, and whether genomic markers can be scaled with environmental data to capture a marker-by-environment effect. In this framework, the first step is to represent markers of each line with their genotype frequency, as described in section 2.4. After obtaining the genotype frequency, global and local marker sets are obtained by applying the procedure described previously, again in section 2.4. As different environmental variables have different ranges of values and the genotype frequency obtained by applying HW ranges between 0 and 1, environmental variables are normalized by applying Min-Max scaler (Buitinck et al., 2013) to bring them in the same range of genotype frequency. The input to the deep learning model is lines represented by 4052 markers and the corresponding normalized site-year environmental data. The output of this model is the predicted yield. Fig. 4 shows the overall workflow of deep learning framework 1. We applied three different deep learning models in this framework. The major difference between the three models is their depth and when the environmental information is integrated. Details of the deep learning models are given below.

2.5.1. Deep learning model 1 (F1M1)

In this model, the assumption is that all the markers and weather variables may interact with each other at the same time. The input to the model is the concatenated vector of marker data and weather variables totalling 4133 input neurons. It then contains 15 blocks of linear and ReLU layers which are followed by an output regression layer. The output of the odd blocks are connected by a residual connection from the previous odd block to the current odd block. Fig. 5 shows the architecture of this model. This model is similar to the model of Khaki and

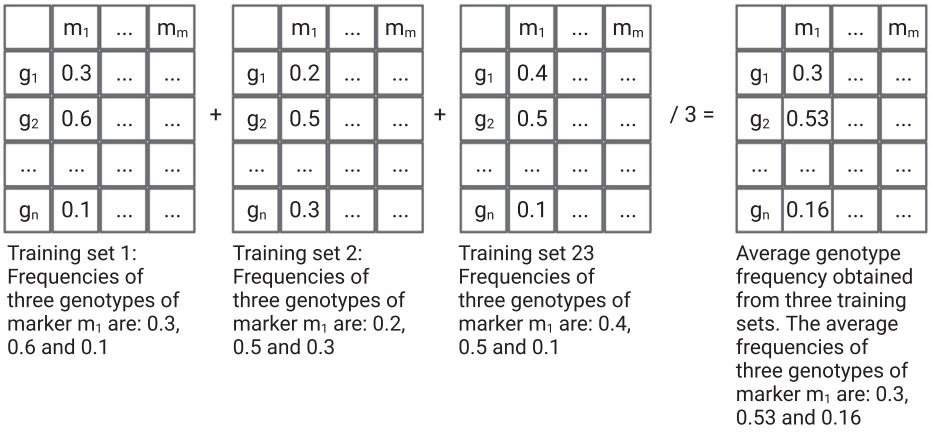


Fig. 1. Markers represented by the average genotype frequency. Here, the example is shown with three training sets. We used five training sets to calculate the genotypic frequency average of each marker.

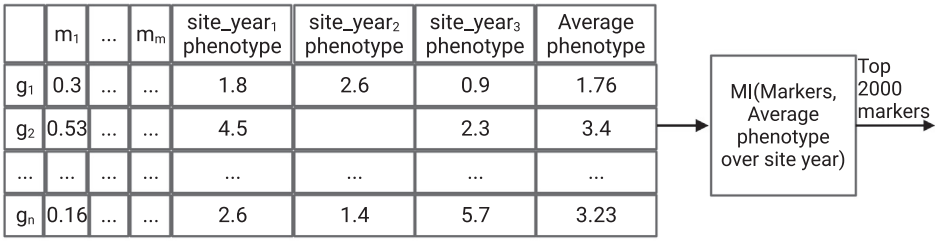


Fig. 2. Top global marker selection procedure. An average of phenotypes across all site-year is calculated for each line. Then, mutual information is applied.

Wang (Khaki and Wang, 2019) as both models employed linear layers and consider that all environmental variables and markers to interact with each other at the same time. We will refer to this variant of the F1 framework as the F1M1 framework.

2.5.2. Deep learning model 2 (F1M2)

To imitate the interaction between all environment variables and a marker, the first block of this model is 4052 parallel linear layers, where the input of each linear layer is all 81 environmental variables along with one marker. We chose 54 neurons as the hidden neurons for the linear layers by experimenting with various numbers of output neurons as they minimize the validation loss for some initial epochs quicker. A ReLU is applied on the stacked output, followed by a block of linear and ReLU layers. A linear layer is applied as the regression layer at the end. Fig. 6 shows the architecture of the deep learning model in the F1M2 framework. We will refer to this variant of the F1 framework as the F1M2 framework.

2.5.3. Deep learning model 3 (F1M3)

The intuition of the deep learning model in the F1M3 framework is that markers interact with each other even before they interact with the environment. It is the result or summary of the marker interaction that interacts with the environment. To imitate this, we constructed a deep learning model with fifteen blocks of neural networks. Each block contains a linear layer followed by a ReLU activation function. The first linear layer of the first block is the input layer, which takes all the selected markers as the input. The output of this layer is a 750-dimensional vector. The subsequent 11 blocks of neural networks take the input from the previous block and again produce an output of a 750 dimensional vector. A residual connection between the odd blocks of the first 12 blocks of neural networks is employed to make sure that none of the blocks suffer from the vanishing gradients problem. After the first 12 blocks of neural networks, 750 parallel linear layers were employed where the input of each linear layer was all 81 environment variables along with one of the 750 neurons from the previous block. The outputs of each of these linear layers are 54-dimensional vectors that are stacked together. After the parallel linear layers, another three blocks of neural networks are applied with a residual connection between the output of parallel neural networks and the output of block fourteen. After these three neural network blocks, a linear layer is applied to perform regression. Fig. 7 shows the neural network architecture. We will refer to this variant of the F1 framework as the F1M3 framework.

2.6. Deep learning framework 2 (F2)

Previous research shows that deep learning models can successfully predict traits for trials where the input is the genomic information (Zhang et al., 2019; Jubair et al., 2021; Ma et al., 2018; Gianola et al., 2011; Pérez-Rodríguez et al., 2012; González-Camacho et al., 2012, 2016; Rachmatia et al., 2017; Jubair and Domaratzki, 2019) or weather information (Lin et al., 2020; Shook et al., 2021; Khaki et al., 2020). In the former scenario, the predicted traits are the average of all locations for a specific line. In the later scenario, the predicted traits are the average over all genotypes grown in a particular environment. To estimate traits in multi-environment, statistical models such as BLUP and GBLUP try to capture the average effect of genetic information and then add variance due to environmental changes. Some of the statistical methods employ dimensionality reduction techniques (Rogers and Holland, 2021) and kernel tricks (Jarquín et al., 2014) to capture GxE. However, the memory requirements of these models increase rapidly with the increase of lines and environments in the training set.

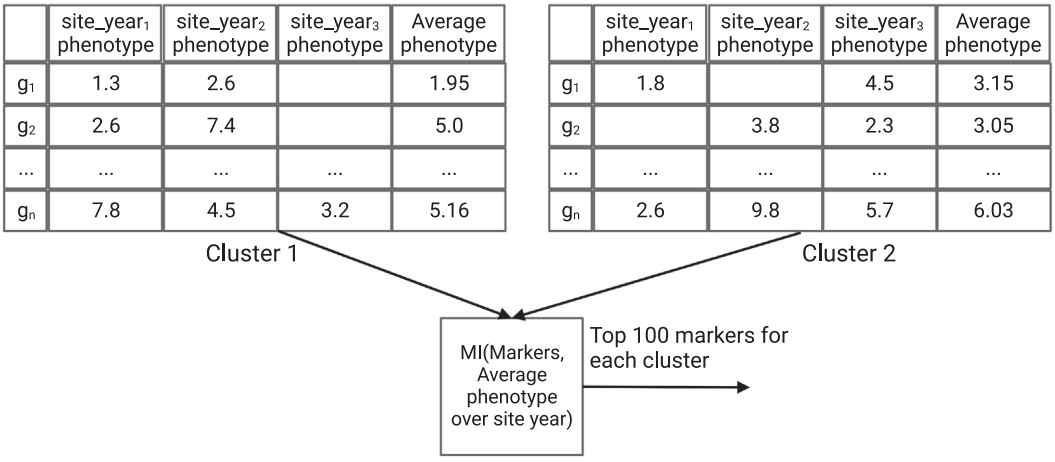


Fig. 3. Top local marker selection procedure (markers were not shown). Tables in the figure shows how average phenotype is calculated. 100 markers are selected from each individual cluster. There are 25 clusters in total. As some markers are common among the clusters, this leads to 2052 unique local markers.

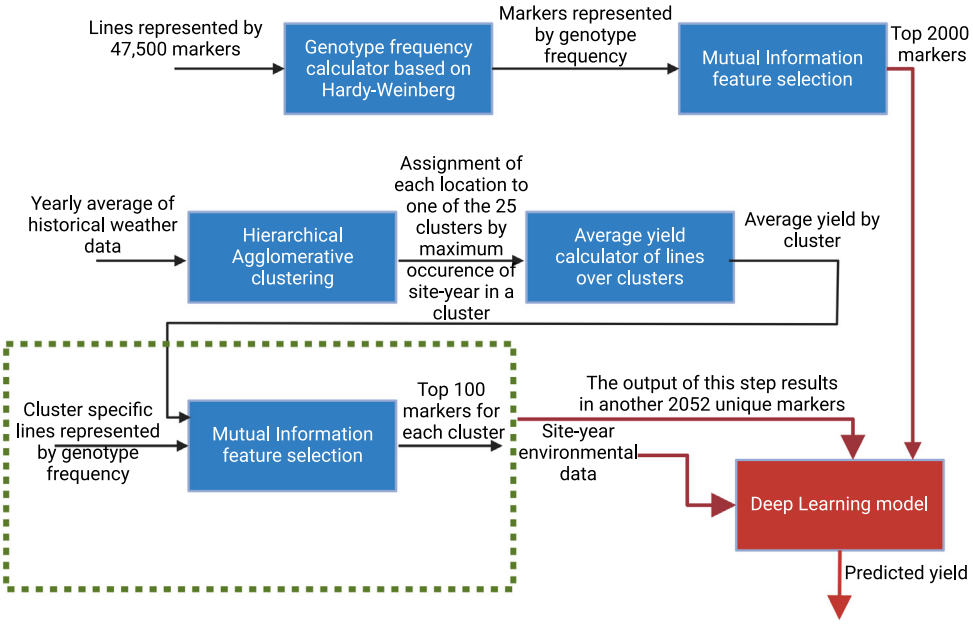


Fig. 4. Deep learning framework 1 workflow. In this workflow, we employed three different deep learning models.

In our proposed architecture, we first learn representations of genotyped data and environmental variables separately. These two models are optimized to predict the average yield over environments and genotypes, respectively. We then concatenate these two representations and predict the environment-specific yield for a specific line assuming that markers and environments work as two groups and one group has an effect on another group for environment specific yield prediction of each line.

Fig. 8 shows the proposed deep learning framework 2. This deep learning framework is based on one deep regression model and two deep representation learning models: i) optimized for an average yield of a line over all environments (line-specific average yield) and ii) optimized for predicting average yield over all lines for an environment (environment-specific average yield). The first step for predicting line-specific average yield is to identify the global marker set by applying the procedure described in section 2.4. After obtaining the global marker set, a neural network model is trained with this marker set as features to predict the line-specific average yield. The last layer before the regression layer of this neural network model produces a 256-dimensional representation vector for each line which is one of the inputs to the deep regression model.

The input to the second representation learning model optimized for predicting environment-specific average yield is nine months of environmental variables for each site-year. This model produces a representation of a 54-dimensional vector for each input for the environmental variables. After training and testing these two models, for each site-year, the representation vectors of these two models are concatenated, which serve as the input to the deep regression model that predicts yield for each site-year. This framework will be referred to as F2. Now, we describe the architecture of each of the models of F2 individually.

2.6.1. Representation learning model optimized for predicting line-specific average yield

Fig. 9 shows the representation learning model that predicts the average yield over environments. The input to this model is a line represented by marker frequency. Each block of neural networks contains a

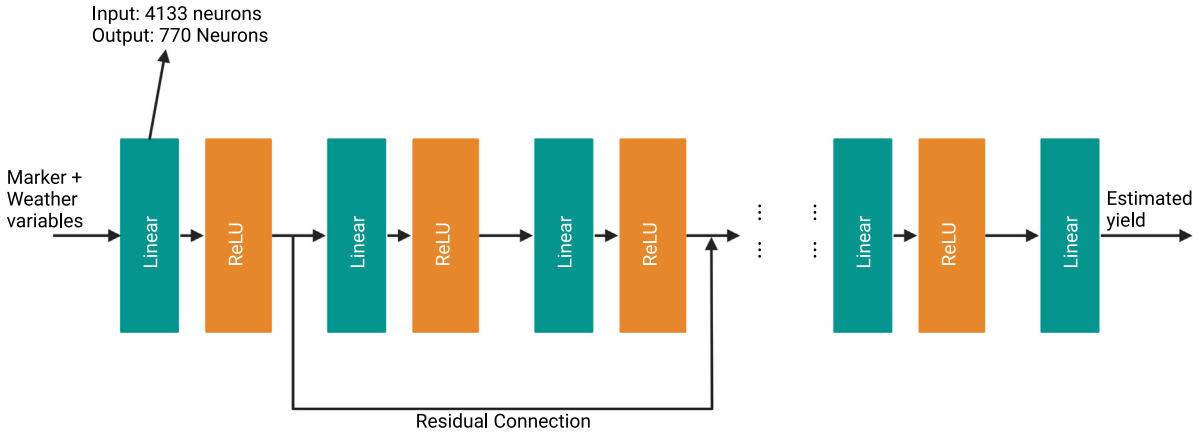


Fig. 5. Architecture of the deep learning model in the F1M1 framework. This model concatenates the marker and weather variables and passed to a linear neural network block that contains a linear and ReLU layer. The architecture of this model is similar to the model of (Khaki and Wang, 2019).

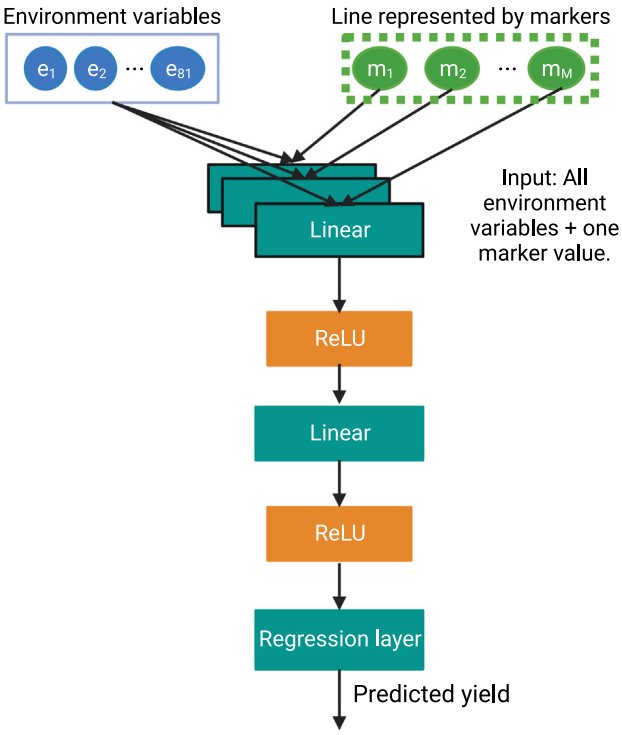


Fig. 6. Architecture of deep learning model in the F1M2 framework.

linear layer, a leaky ReLU activation function and a dropout layer. The hidden layer of the liner layer contains 2000 neurons in the first five blocks of the model. The last three blocks of neural networks have 666, 444 and 296 hidden nodes. All leaky ReLU layers have the same slope of 0.1 for the negative values. The first five dropout layers have the probability of 0.5 to drop a neuron. The next two dropout layers have the probability of 0.4, and the last dropout layer has the probability of 0.2 to drop a neuron. The last layer is the regression layer that predicts average yield across environments.

2.6.2. Representation learning model optimized for predicting environment-specific average yield

Fig. 10 shows the representation learning model that predicts average yield over lines. The input to this model is the environmental variables of nine months normalized by a min-max scaler. There are four blocks of neural networks where each block contains a linear layer, a ReLU activation function and a dropout layer. Each block produces an output of 54 hidden neurons with a dropout probability of 0.25. The last layer is the regression layer that predicts the average yield for an environment.

2.6.3. Yield prediction model

Fig. 11 shows the yield prediction model. This is a shallow model that contains three blocks of neural networks. Each block has a linear layer and a ReLU activation function. The last layer is the regression layer that predicts yield for a specific line in a specific environment.

2.7. Deep learning framework 3 (F3)

None of the previous two frameworks contain any information about the field management and soil information. In this framework, we integrated unstructured text data which may provide information about management and field conditions.

Fig. 12 shows framework 3. This framework is an extension of the F2. The major difference between the two frameworks is that an agri-bert

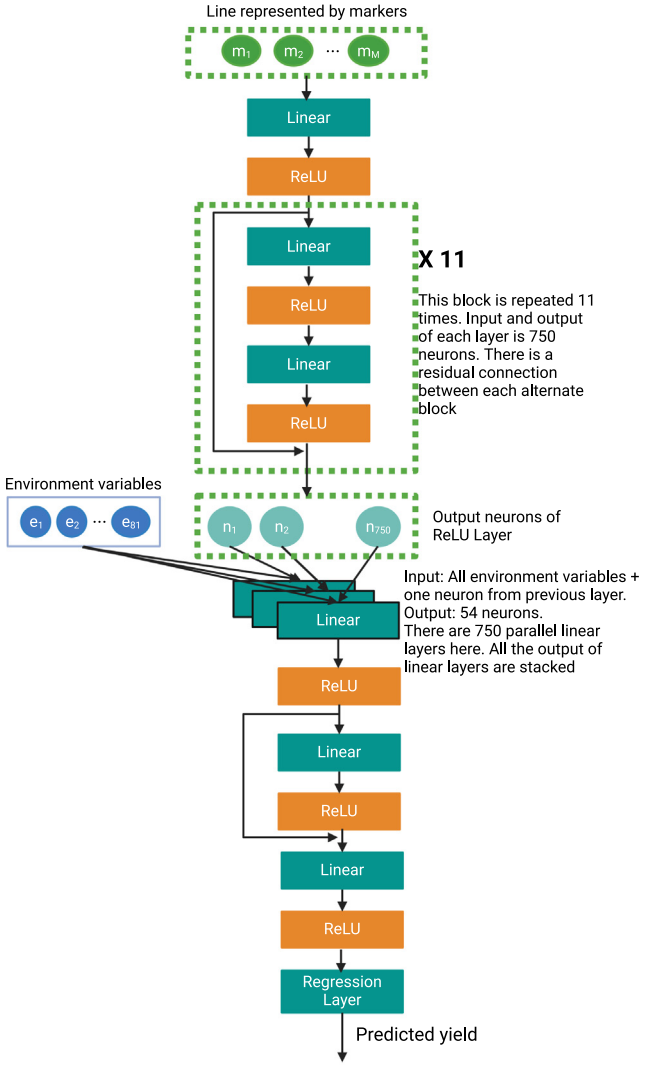


Fig. 7. Architecture of the deep learning model in the F1M3 framework. Output of each odd number of ReLU layer is connected through a residual connection.

model (Rezayi et al., 2022) is employed to obtain a representation of soil and environment-related text. A 768-dimensional representation is obtained for all texts and then an average representation is calculated for each site-year. The length of all individual notes are <256 tokens. As the maximum length of texts of the agribert model is 512 tokens, it can obtain the representation of the full note. This 768-dimensional vector is also concatenated with the output of two representation learning models and provided as the input to the shallow model.

2.8. General settings of deep learning models

All the models are optimized using the Adam optimizer with a learning rate of $1e^{-5}$. Mean square error is applied as the loss function. Training is stopped when there is no improvement in PCC for the validation set in at least 30 consecutive epochs.

3. Results

3.1. Environment data cluster

To understand how similar the locations are in the same cluster, we applied TSNE on the yearly average of weather variables to reduce the

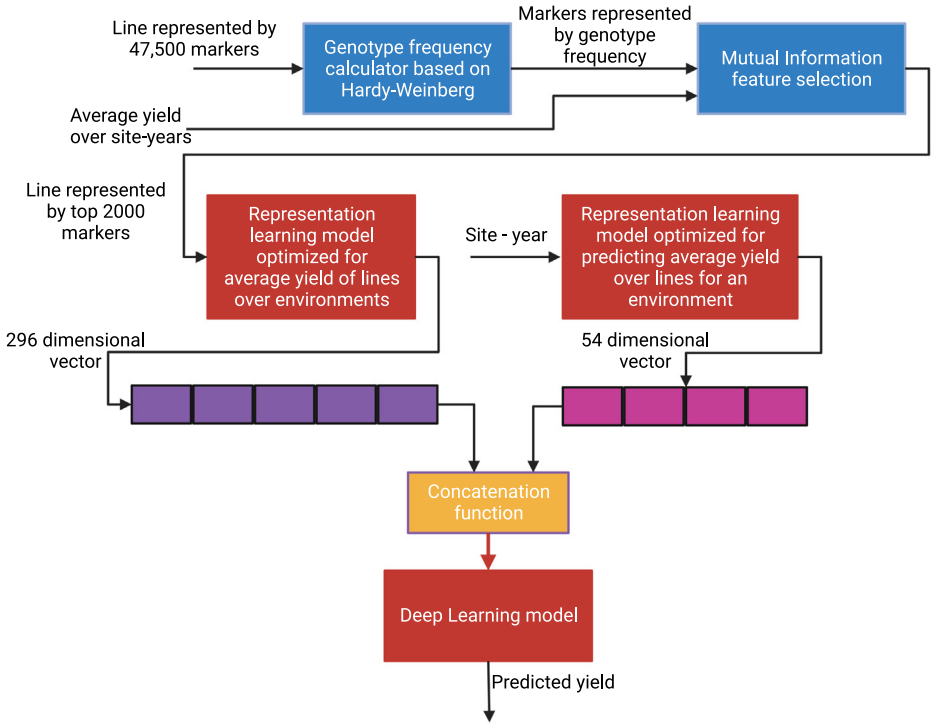


Fig. 8. Deep learning framework 2 (F2).

dimension to a three-dimensional space of each site-year. We then label each site-year by its nursery and cluster assignment to visualize how separable the site-years are in the three-dimensional space. Fig. 13 shows how different trials and clusters are mapped in a three-dimensional space. From the figure on the left (site-year weather data by nursery), we observe that the trials are not well separable from each other in a three-dimensional space though they are created based on similar environments. On the other hand, from figure on the right (site-year weather data by cluster group), we observe that the groupings created by the clustering algorithm have better separability than the nursery-based grouping. As it is very difficult to understand the figure with this large number of clusters, we also generated interactive figures for both.[1]

3.2. Effect of adding environmental variables

We experimented with three deep learning models in deep learning framework 1 (F1M1, F1M2 and F1M3), where these models differ primarily on how the environment and genotype interactions are captured within the model. In the F1M1, the assumption was that all genomic and environmental factors interact with each other at the same time. On the other hand, in the F1M2, the assumption is that each marker separately interacts with the environment first, and the resulting outcome then affects yield. Finally, in the F1M3, the relationship between markers is first taken into account, and then the environment interacts with the combined marker relationship to estimate yield.

During training, as we have five folds, we trained and tested the F1M1 and F1M3 for all folds. However, after training the first fold, we identified that the F1M2 significantly overfits as we obtained a low Pearson Correlation Coefficient (PCC) on test and validation data ($\approx$0.11) and a high PCC ($\approx$0.8) on training data. Furthermore, as 4052 parallel fully connected neural networks are employed just after the input layer, the trainable parameters and training time are also very high for this F1M2 model (around 1.5 h per epoch on NVIDIA GTX 1080). Thus, we did not train the F1M2 for the rest of the folds.

Table 4 shows the F1M1 and F1M3 comparison on the test sets. From the table, we observe that models in F1 that consider genetic interaction first and then capture GxE (F1M3) perform better on the following two test case scenarios: i) when environments are observed, but lines are not (test scenario one) and ii) when lines may be observed, but locations are not (test scenario two). Models in F1M3 are 1.62 to 2.05 times better than F1M1 counterparts on PCC score for different folds of test scenario one. Models in F1M3 also outperform models in F1M1 in the second test scenario, with PCC scores 1.35 to 1.82 times higher than F1M1 on different folds. However, the standard deviation of the F1M1 is lower than the F1M3, which indicates that models in F1M1 have less variations (more stable performance) across folds. This result is consistent with other research where it is observed that deep learning models that allow interactions between datatypes perform better than the models that do not consider these interactions (Sharma et al., 2022; Kick et al., 2023; Jarquin et al., 2021).

3.3. Effect of global markers vs global + local markers in F1M3

To understand the effect of the global and local marker sets, in this section, we present the results obtained when our F1M3 framework is trained with the global marker set only. Thus the input to this F1M3 framework is 2000 markers instead of the 4052 markers. The rest of the architecture is the same as the previous F1M3. Fig. 14 shows the performance of the two models in the F1M3 on two test scenarios. From the figure, we observe that although models in F1M3 trained only on the global marker set have higher PCC values in four folds out of five on

[1] See https://htmlpreview.github.io/?https://github.com/sheikhjubair/GxENet/blob/main/figures/cluster_label.html and https://htmlpreview.github.io/?https://github.com/sheikhjubair/GxENet/blob/main/figures/trial_cluster.html.

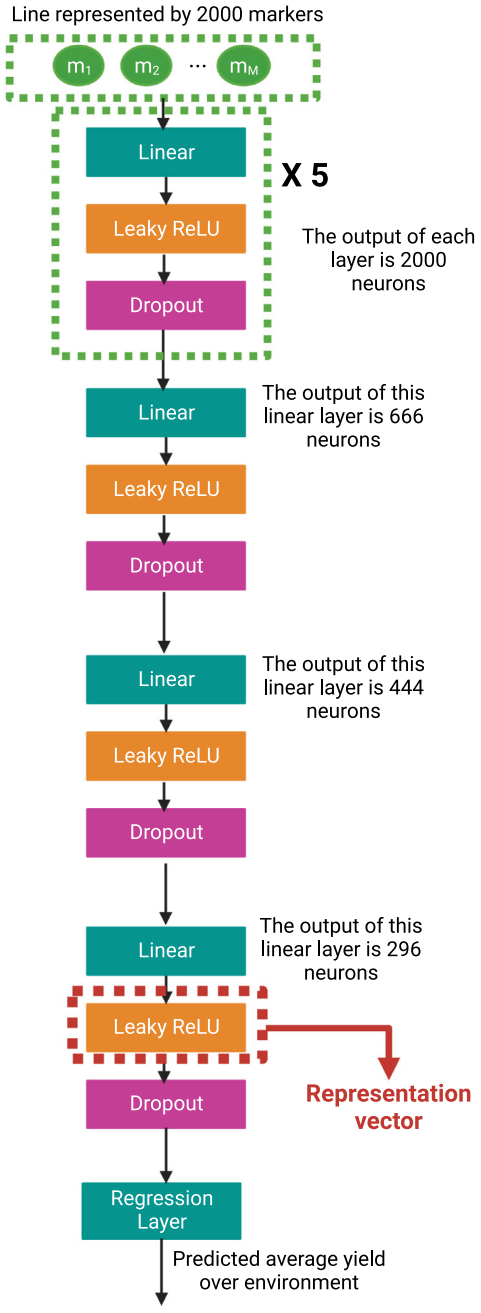


Fig. 9. Representation learning model for predicting line-specific average yield.

test scenario one, we observe the opposite outcome for test scenario two. The average PCC value of test scenario one trained with the global marker set is 0.729, which is 1.7% higher than the model trained on the global + local marker set. However, the average PCC value of test scenario two trained with global marker set is 0.381, which is 2.2% lower than the models trained on the global + local marker set. We also conducted a *t*-test on the PCC scores for both test scenarios. The *p*-value of test scenario one is 0.543 and test scenario two is 0.524 which indicates that the PCC score of the two models has identical variance and there is not a statistically significant difference between the two ways of training the model. Overall, the results showed that the effect of the environment-specific markers on the models is minimal for predicting yield. However, local markers improve PCC when the model does not observe the locations.

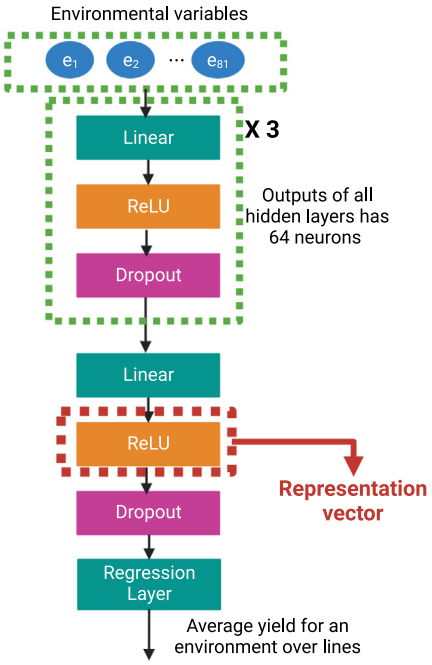


Fig. 10. Representation learning model for predicting environment-specific average yield.

3.4. Performance of the F1M3 vs the F2

Our F2 framework is the combination of three different deep learning models. The first representation learning model is a deep learning model that predicts line-specific average yield. Table 5 shows the PCC scores across five-fold for both test scenarios. The average PCC for test scenario one, where the model knows environments is 0.606. However, when the environments were not observed, the average PCC goes down to 0.167. As the input to this model is genotyped data and the output is average yield, this model supports the observation of other recent research which is models that do not incorporate environmental information are not suitable for predicting top lines for a new environment

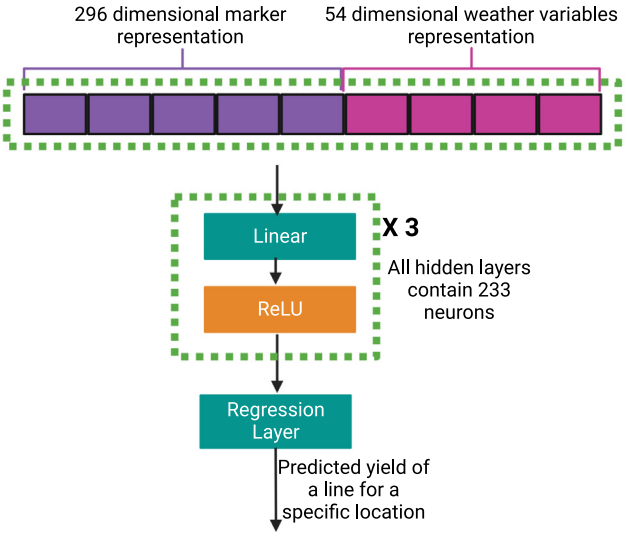


Fig. 11. Yield prediction model for predicting line specific yield for each site-year.

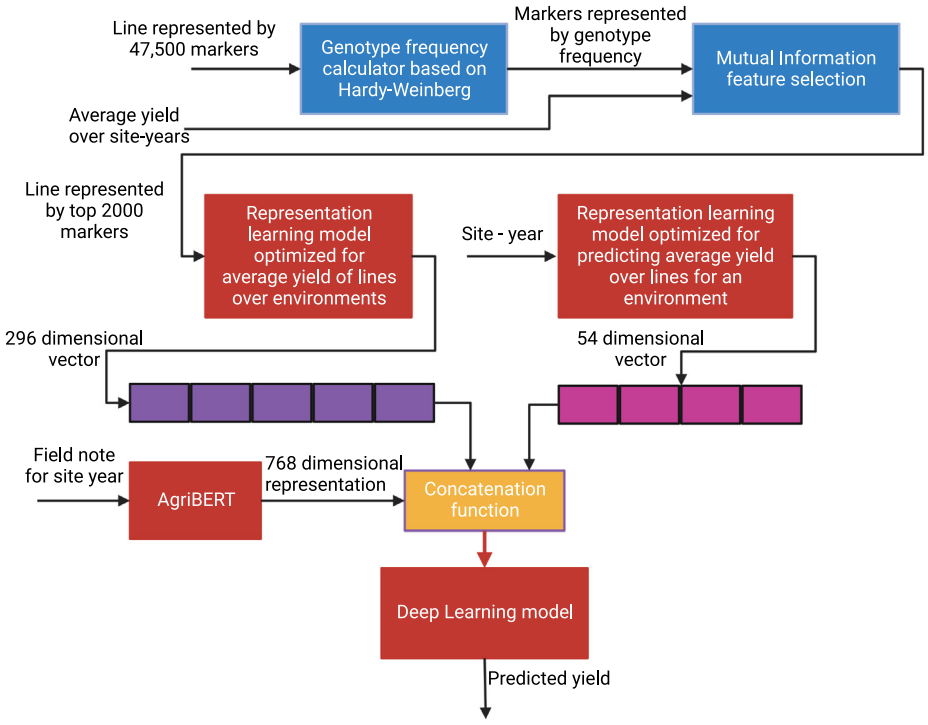


Fig. 12. Deep learning framework 3 (F3).

(Washburn et al., 2021; Lin et al., 2020; Khaki and Wang, 2019; Montesinos-López et al., 2019).

The objective of the second representation learning model is to capture the environmental effect on yield by estimating the average yield over genotypes for an environment. Table 6 shows PCC scores across five folds for both test scenarios. From the table, we observe that although the average PCC in the observed environment is higher than in the second test scenario, the performance in the second test scenario is also satisfactory as PCC score 0.518 indicates that there exists some linear relationship between the target and predicted average yield.

The yield prediction model of the F2 framework estimates the environment-specific yield of a specific line by combining the

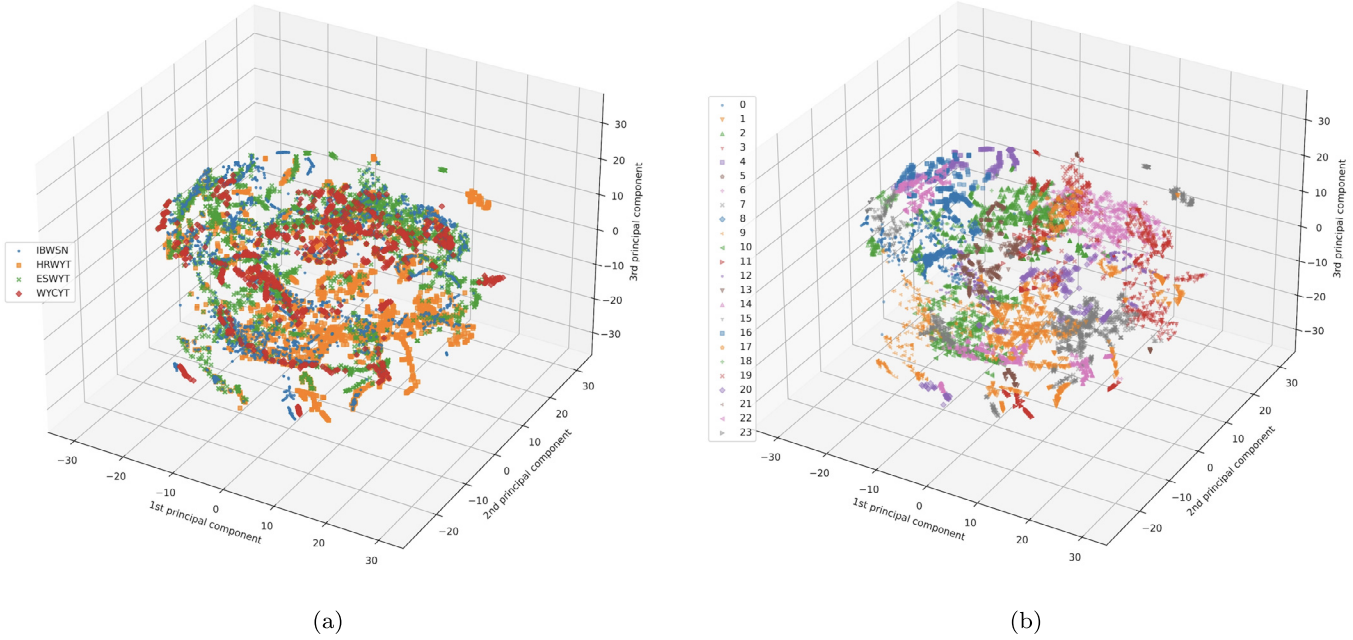


Fig. 13. Site-year weather data by nursery (left) and cluster group (right). Each point in the figure indicates a site-year. Different colors either indicate a nursery (left) or a cluster group (right).

Table 4
Comparison of PCC over five folds between F1M1 and F1M3 for test scenario one and test scenario two.

Folds	F1M3 (test scenario one)	F1M1 (test scenario one)	F1M3/F1M1	F1M3 (test scenario two)	F1M1 (test scenario two)	F1M3/F1M1
1	0.697	0.375	1.85	0.359	0.257	1.39
2	0.759	0.372	2.04	0.382	0.257	1.48
3	0.690	0.354	1.94	0.470	0.258	1.82
4	0.740	0.360	2.05	0.353	0.260	1.35
5	0.677	0.417	1.62	0.452	0.258	1.75
Average	0.712	0.375	1.89	0.403	0.258	1.56
Std	0.031	0.022		0.041	0.001	

representation learnt from the previous two models. Fig. 15 shows the PCC scores for five folds on both test scenarios. The average PCC score of the yield prediction model of the F2 framework in test scenario one is 0.734, while the average PCC score of F1M3 was 0.712, indicating 2.2% improvement over F1M3 framework. Although the average PCC score of the yield prediction model of the F2 framework for test scenario one is higher, we observe that the models in F1M3 have higher PCC scores in two folds out of five, indicating no clear advantage of using one framework over another. However, in test scenario two, yield prediction models of the F2 have higher PCC scores in all five folds, with an average PCC of 0.454, which is 5.18% improvement over the F1M3. The result demonstrates that F2 shows improvement when compared to all the variants of the F1.

In addition, we conducted a *t*-test on the results of the F1M3 and F2 frameworks. The *p*-value obtained from the t-test of test scenario one is 0.398 and test scenario two is 0.164, which indicates that the PCC scores of the two models have identical variance and there is no statistically significant difference between the two results of the two frameworks. However, a 5.18% improvement in the average PCC score of the F2 framework over the F1M3 framework on test scenario two is a notable improvement. Moreover, the F2 framework needs significantly less training time, as described in section 3.7.

In all the reported results above, we calculated a global PCC score where we considered all cycles of all trials in the test set of a specific fold. As PCC measures the linear relationship between actual yield and predicted yield, global PCC score and cycle-specific PCC score of a trial may vary. Thus, we calculated cycle-specific PCC scores to understand how our models perform for different cycles of a trial. As there are lots of cycles and trials combinations, to present the result, we divided the PCC score into three ranges: PCC ≤ 0 indicates the model did not learn anything, $0 < \text{PCC} \leq 0.4$ indicates the performance of the model is random and finally, PCC>0.4 is our desirable range which indicates there exist some linear relationship between predicted yield and true yield. Table 7 shows the result for our two test scenarios. From the table, we observe that yield prediction models of the F2 framework have a higher number of cycles with PCC scores >0.4 compared to the models of F1M3. On average, there are 30.2 cycles in the first test scenario across five-folds, and the F2 framework has 29.2 cycles with PCC > 0.4. In test scenario two, the average number of cycles in the F2 framework with PCC>0.4 is 5.4 while the average number of cycles in each fold is 6.6. We also observe that eight unique cycles have PCC < 0.4 in F2, with 51st IBWSN cycle having the most frequent appearance. While experimenting with the F1M3 framework, we identified that 11 unique cycles have low PCC scores (< 0.4), and again 51st IBWSN cycle is the most frequent across five-fold. Five cycles with low PCC in the F2 framework are also present in F1M3. There is no specific type of trials in the cycles that have low performance as the low-performance cycles occur in all trials except WYCYT.

Table 5
PCC scores across five folds of the representation learning model of the F2 framework for predicting line specific average yield.

Folds	Test Scenario One	Test Scenario Two
1	0.549	0.030
2	0.691	0.070
3	0.594	0.160
4	0.578	0.298
5	0.601	0.281
Average	0.606	0.167

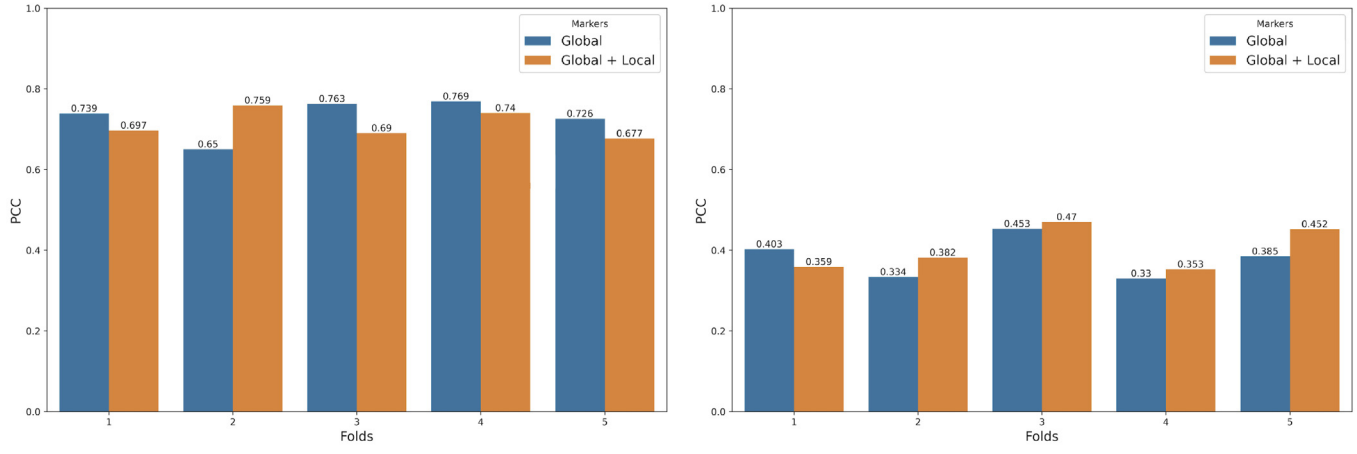


Fig. 14. Comparison of PCC scores of F1M3 models trained on global markers and global + environment specific markers and evaluated on the test scenario one (left) and test scenario two (right).

Table 6
PCC scores across five fold of the representation learning model of the F2 framework that predicts environment specific average yield.

Folds	Test Scenario One	Test Scenario Two
1	0.786	0.503
2	0.844	0.564
3	0.797	0.480
4	0.855	0.536
5	0.837	0.507
Average	0.823	0.518

3.5. Feature importance

To find the feature importance of environmental variables, we employed DeepLift (Shrikumar et al., 2017) on the representation learning model optimized for predicting environment-specific average yield (one of the components of F2). DeepLift measures how the information is propagated along the network and assigns importance scores (referred to as the attribution scores) to input variables. We used the Captum library (Kokhlikyan et al., 2020) for DeepLift implementation and employed the default parameters to the DeepLift model. A positive attribution score for an environmental variable means it is positively influencing the prediction, while a negative attribution score means the opposite.

Fig. 16 shows the importance of each feature where each bar is an environmental variable. The variables are in the following order for each month: (1) precipitation, (2) maximum relative humidity, (3) minimum relative humidity, (4) shortwave radiation, (5) maximum temperature, (6) minimum temperature, (7) maximum vapour pressure deficit, (8) wind speed 2 m, (9) wind speed 10 m. The figure shows that maximum temperature has a positive effect in the first two months. It is worth mentioning that the first two months in our dataset are

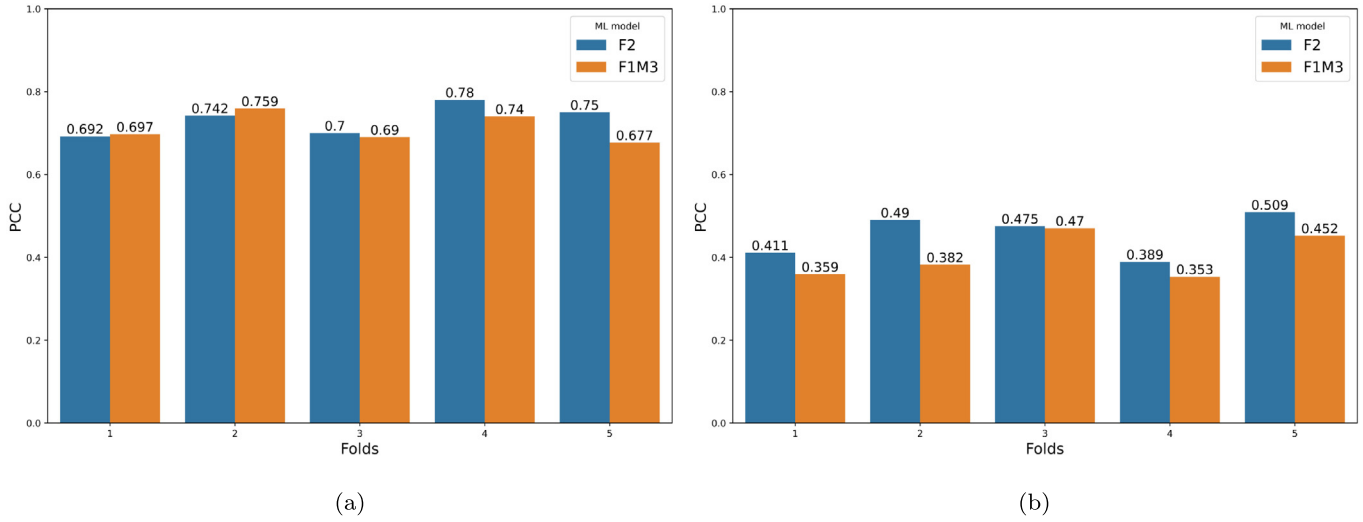


Fig. 15. Comparison of PCC scores between F2 and F1M3 on test scenario one (left) and test scenario two (right).

Table 7
Number of cycles in each ranges of PCC for both test scenarios.

Folds	PCC Range	Test Scenario One		Test Scenario Two	
		Number of Cycles (F1M3)	Number of Cycles (F2)	Number of Cycles (F1M3)	Number of Cycles (F2)
1	PCC ≤ 0	0	1	0	0
	0 < PCC ≤ 0.4	2	0	3	2
	PCC > 0.4	27	28	3	4
	Number of line and site-year combinations	12,543	22,996		
2	PCC ≤ 0	1	1	0	0
	0 < PCC ≤ 0.4	1	0	2	0
	PCC > 0.4	31	32	5	7
	Number of line and site-year combinations	14,915	15,884		
3	PCC ≤ 0	1	0	1	1
	0 < PCC ≤ 0.4	0	1	2	1
	PCC > 0.4	28	28	7	8
	Number of line and site-year combinations	14,094	15,884		
4	PCC ≤ 0	0	0	0	0
	0 < PCC ≤ 0.4	3	1	2	1
	PCC > 0.4	30	32	2	3
	Number of line and site-year combinations	12,792	24,312		
5	PCC ≤ 0	1	0	0	0
	0 < PCC ≤ 0.4	1	1	1	1
	PCC > 0.4	25	26	5	5
	Number of line and site-year combinations	12,684	20,400		
Average	PCC ≤ 0	0.6	0.4	0.2	0.2
	0 < PCC ≤ 0.4	1.4	0.6	2	1
	PCC > 0.4	28.2	29.2	4.4	5.4

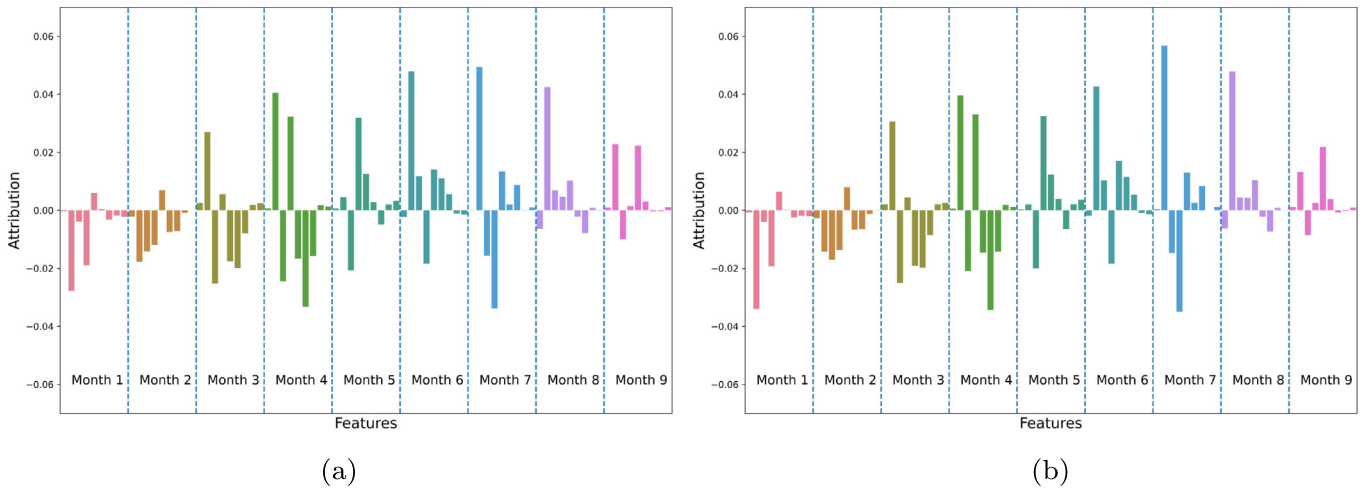


Fig. 16. Feature importance of weather variables obtained by employing DeepLift on environment specific average yield model of F2. The left figure is the test scenario one and the right figure is the test scenario two. Each bar in the figure represents an environmental variable. The variables are in the following order for each month: (1) precipitation, (2) maximum relative humidity, (3) minimum relative humidity, (4) shortwave radiation, (5) maximum temperature, (6) minimum temperature, (7) maximum vapour pressure deficit, (8) wind speed 2 m, (9) wind speed 10 m.

before sowing crops. In the third month, maximum vapour pressure contributes more than any other environmental variables for predicting average yield. We observe this trend for maximum vapour pressure for the rest of the months of the growing cycle except the fifth month, where shortwave radiation contributes more than the vapour pressure. The effect of shortwave radiation increases significantly in the fourth and fifth months, and then goes down gradually. In the third month, the importance of precipitation and maximum vapour pressure increases as we enter the months when seeds are sown. The maximum temperature has a continuous positive effect from the fifth month to the end of the growing cycle. Both wind speed variables have little to no impact from the third month to the end.

Previous research has shown that precipitation plays an important role, especially in the early month of the crop-growing season (Washburn et al., 2021; Måløy et al., 2021). However, in our work, we observe less impact of rainfall. The main reason is that about 90% of observations in our dataset (101,777 out of 111,836 observations combining training, test and validation set) are from IBWSN and ESWYT nurseries where rainfall is low and the land receives optimal irrigation. On the other hand, only 6.2% observations (6961 observations) are from HRWYT nurseries, where the crop gets rainfall during the cropping cycle. Since the environmental condition of the WYCYT nursery is mixed, we can not verify the amount of rainfall those nurseries received. However, WYCYT nurseries are only contributing to about 3.8% of the data.

To understand the effect of genomic information and environmental variables for predicting environment-specific yield of each line, we employed DeepLift in the final model of the F2 framework that combines the output of two representation learned models. Fig. 17 shows the attribution scores of representation learned features for both test scenarios. From the figure, we observe that although the number of learned features of the environmental representation is less than the marker representation, they contribute more towards environment-specific yield estimation for a line. However, many marker representation features have a small positive effect on the outcome. This observation aligns with other research that shows that yield is a complex trait and lots of small effect genes contribute to yield (Rogers et al., 2021).

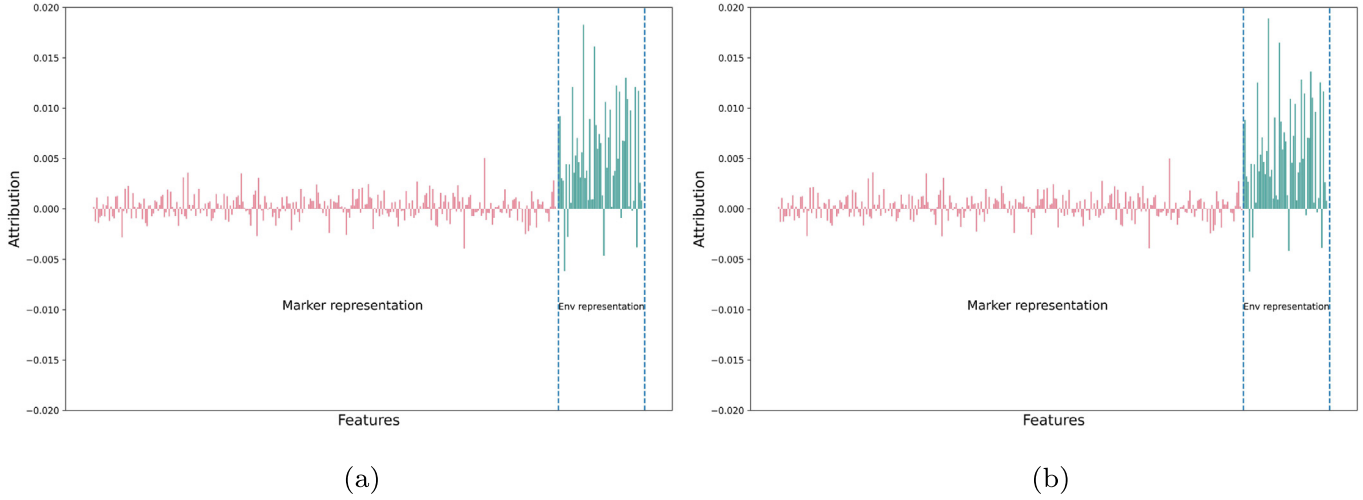


Fig. 17. Feature importance of representation learning features obtained by employing DeepLift on the final deep learning model of F2 that combines marker representation and environment variable representation to predict environment specific yield for a line.

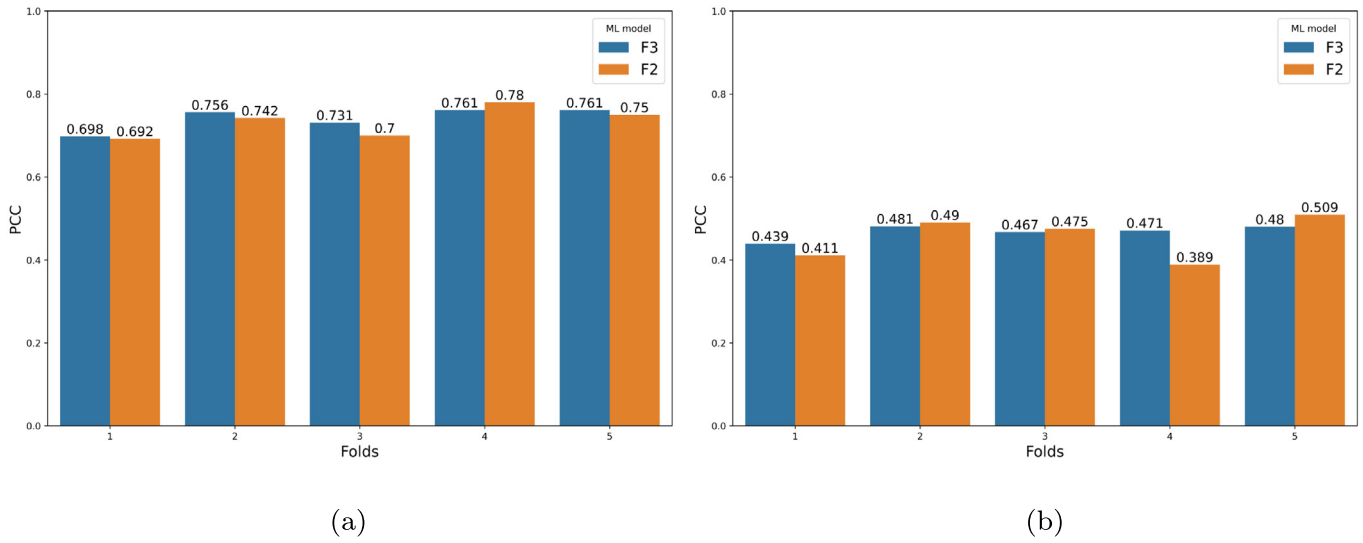


Fig. 18. Comparison of PCC scores between F3 and F2 frameworks.

In addition, we calculated the sum of the absolute attribution scores of all neurons for marker and weather feature representations of test scenario one. The results are 0.261 and 0.318 for marker and weather representations, respectively. However, when the sum of positive attribution scores is calculated, the scores drop to 0.156 and 0.299 for marker and weather representations, respectively. Moreover, the overall sum of attribution scores of all neurons for each feature representation is positive (0.05 and 0.28 for marker and weather representations, respectively). Furthermore, 57.09% neurons from marker representation have a positive impact, while 88.88% neurons from environmental representation have a positive impact for predicting yield. Overall, these observations and the Fig. 17 demonstrate the importance of adding environmental features for the genomic prediction task.

3.6. Performance of the F3 architecture

Since the F3 model is the extension of the F2 and we could not verify that all the text information for a specific environment can be obtained or predicted before sowing crops, we did not make any comparison of the F3 with models other than F2; rather, we presented it to demonstrate how other information such as text data or soil data can be incorporated in the F2 architecture. However, as these text data are mostly field management data collected before and during the growing season, this architecture may play a vital role in selecting superior lines for the next growing cycle if there are many similarities in field management among the growing seasons of a specific location.

Fig. 18 shows the PCC scores of the F3 architecture. The average PCC of F3 is 0.741 for test scenario one and 0.467 for test scenario two. The F3 model performs slightly better than the F2 in four folds in test scenario one and two folds in test scenario two. We also conducted a *t*-test on the PCC scores for both test scenarios. The *p*-value of test scenario one is 0.684 and test scenario two is 0.615 which indicates that the PCC score of the two models has identical variance and there is not a statistically significant difference between the two results of the two frameworks.

We also employed DeepLift on the F3 architecture to find out the text feature importance. Fig. 19 shows the attribution score of each input neuron to the final deep learning model of the F3 architecture. The figure shows that environment representation has the most significant impact on yield prediction. While 94.25% of the marker

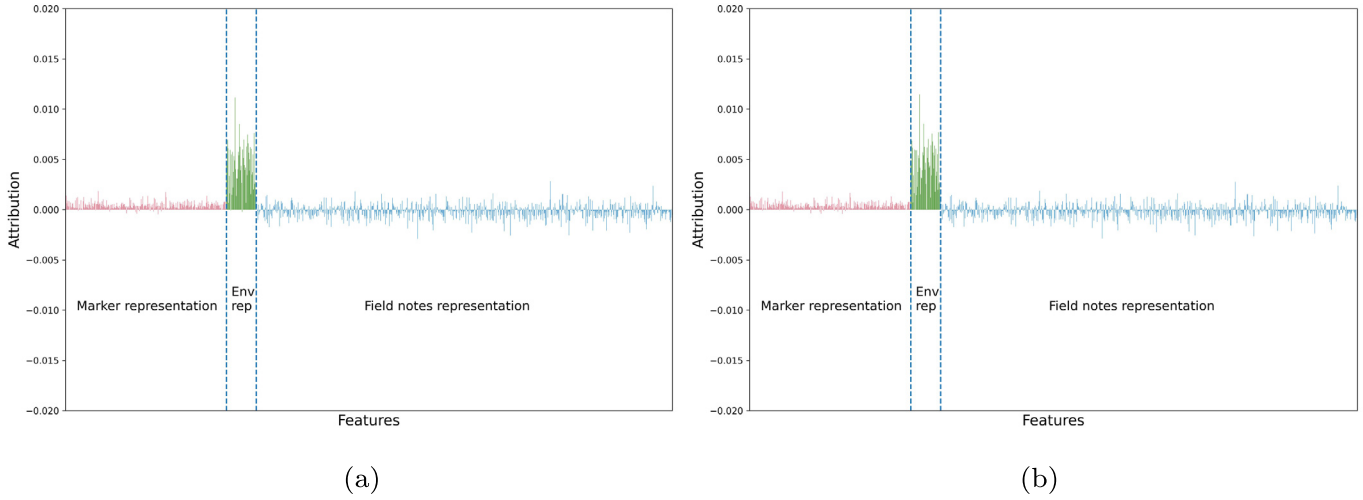


Fig. 19. Feature importance of representation learning features obtained by employing DeepLift on the yield prediction model of the F3 that combines marker representation, environment variable representation and field notes representation to predict environment specific yield for a line.

Table 8
Time required to train different models on NVIDIA RTX A6000 GPU.

Model	Average Time for One Epoch (seconds)	Approximate Number of Epochs for Training
F1M1	7.8	170
F1M2	3262.36	40
F1M3	280.26	100
Line Specific Average Yield (F2 and F3)	0.4	500
Environment Specific Average Yield (F2 and F3)	0.08	6000
Yield Prediction Model (F2)	3.05	350
Yield Prediction Model (F3)	5.51	300

representation neurons and 98.14% of the environmental neurons have a positive impact, only 26.11% of the text representation neurons have a positive impact on environment-specific yield estimation. Moreover, the sum of all attribution scores for each representation type shows that the environment and genotype have a positive sum (0.237 and 0.119, respectively) and the text representation has a negative sum (−0.144). However, when only the positive attribution scores are calculated for each of the representations, the text representation's sum of attribution scores is higher than the sum of marker representation's atrribution scores (0.141 and 0.122, respectively).

These observations show that the text data we had access to have a limited impact on our prediction task. However, since these text data are very short text and heterogeneous, more detailed and informative text data potentially could improve the model performance or have more positive influence in the prediction. In addition, the text representation is learnt from a BERT-based model known as agriBERT, which is primarily trained on agricultural journal papers. Since our texts are field notes, a BERT model trained on field notes would be more suitable. These demonstrate the challenges in incorporating highly heterogenous text into deep learning models, especially where missing data may be present and differences in data collection can be expected. We expect that future work may improve the ability to use management data in prediction. Moreover, since we can not verify that all the unstructured texts were before sowing season, this F3 model is a demonstration of how this textual information can be integrated with genomic and environmental data when an appropriate dataset is available.

3.7. GPU utilization and training time

For training these models, two GPUs are used: i) NVIDIA RTX A6000 and ii) NVIDIA GeForce GTX 1080. Training time depends on which GPU we are using. Overall, training time is much less in NVIDIA RTX A6000 GPU as this is faster than the NVIDIA GTX 1080. Overall, the training times of models in F1M2 are higher than any other models as they are much larger in terms of parameters. Deep learning models in F2 and F3 frameworks are faster to train than all other models. To compare the training time of each model, we run them on NVIDIA RTX A6000 GPU. Table 8 shows the time required to train different models for one epoch.

Since the line specific average yield and environment specific average yield models of the F2 and F3 frameworks are independent of each other, they can be trained in parallel, which reduces the overall training time of the F2 and F3 frameworks. Moreover, each epoch of these two models requires significantly less time compared to other models. Furthermore, each epoch of the yield prediction models of the F2 and F3 frameworks needs approximately 3.05 s and 5.51 s, which is less than the simple F1M1 framework. Overall, the simplicity in the architecture of the models in the F2 and F3 frameworks helps them train faster than the other frameworks.

4. Conclusion

In this work, we proposed three novel deep learning frameworks where the neural network models vary mostly on how environmental information is incorporated into the model. These models are curated to incorporate GxE. Among three models, we identified that the framework which employs two representation learning models optimized for predicting line specific average yield and environment specific average yield and then combines these two representations to estimate environment-specific yield for a specific line, is slightly better than the others. This framework shows 1.95 to 1.75 times better performance, depending on the test scenario, than some existing deep learning models. Later, we extend the F2 framework by integrating text data from field notes.

In this dataset, we do not have any information on the soil. As some research shows that soil plays a vital role in yield (Washburn et al., 2021), adding soil information to the model may help estimate yield more accurately. We showed that our F2 framework could easily be extended by adding new information as we extended the F2 framework by adding field notes. While devising these frameworks, we assumed that the weather of the growing season can be predicted ahead of time. Thus our models focus on only estimating the traits of the crops by using a representation of the weather during the growing season. Future work should incorporate weather prediction for the growing season from historical weather. While obtaining the representation of the weather variable, the input to the model was the monthly average of weather variables. Future work should also try to determine the effect of incorporating weekly or daily weather variables as the input to the model.

Finally, our F2 framework performs well in two test scenarios where the first test scenario is more straightforward to predict than the second one. In the first test scenario, we predicted yield in a scenario where environments are observed, but lines are not observed in any of the environments during the training of the model. In the second scenario, locations in test sets are not observed but the model may observe lines during training. All the models have better performance in the first test scenario compared to the second one. The result is understandable as the attribution score obtained by employing DeepLift shows that weather variables play a significant role in estimating yield.

Data and code availability

All data used in this work can be downloaded from the CIMMYT website (Poland et al., 2021; CIMMYT, 2022). The codes can be downloaded from GitHub: https://github.com/sheikhjubair/GxENet.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Sheikh Jubair: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Olivier Tremblay-Savard:** Writing – review & editing. **Mike Domaratzki:** Writing – review & editing, Supervision, Conceptualization.

Declaration of Competing 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.

References

Acquaah, G., 2009. Principles of Plant Genetics and Breeding. John Wiley & Sons.

Boudhrioua, C., Bastien, M., Torkamaneh, D., Belzile, F., 2020. Genome-wide association mapping of sclerotinia sclerotiorum resistance in soybean using whole-genome resequencing data. BMC Plant Biol. 20, 195. https://doi.org/10.1186/s12870-020-02401-8.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G., 2013. API design for machine learning software: experiences from the scikit-learn project. ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pp. 108–122.

CIMMYT, 2022. CIMMYT research data and software repository network. https://data.cimmyt.org/ accessed 2022-08-04.

Crossa, J., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., Saint-Pierre, C., Vikram, P., Sansaloni, C., Petroli, C., Akdemir, D., Sneller, C., Reynolds, M., Tattaris, M., Payne, T., Guzman, C., Peña, R.J., Wenzl, P., Singh, S., 2016. Genomic prediction of gene bank wheat landraces. G3 Genes|Genomes|Genetics 6, 1819–1834. https://doi.org/10.1534/g3.116.029637.

Esposito, S., Carputo, D., Cardi, T., Tripodi, P., 2020. Applications and trends of machine learning in genomics and phenomics for next-generation breeding. Plants 9. https://doi.org/10.3390/plants9010034.

FAO, 2022. 2022 Global Report on Food Crises. https://www.fao.org/documents/card/en/c/cb9997en/ accessed 2022-08-04.

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

González-Camacho, J., de Los Campos, G., Pérez, P., Gianola, D., Cairns, J., Mahuku, G., Babu, R., Crossa, J., 2012. Genome-enabled prediction of genetic values using radial basis function neural networks. Theor. Appl. Genet. 125, 759–771. URL: https://doi.org/10.1007/s00122-012-1868-9.

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

Guo, J., Khan, J., Pradhan, S., Shahi, D., Khan, N., Avci, M., Mcbreen, J., Harrison, S., Brown-Guedira, G., Murphy, J.P., Johnson, J., Mergoum, M., Esten Mason, R., Ibrahim, A.M.H., Sutton, R., Griffey, C., Babar, M.A., 2020. Multi-trait genomic prediction of yield-related traits in us soft wheat under variable water regimes. Genes 11. https://doi.org/10.3390/genes11111270.

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., Burgueño, J., de los Campos, G., 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor. Appl. Genet. 127, 595–607. URL: https://doi.org/10.1007/s00122-013-2243-1.

Jarquin, D., de Leon, N., Romay, C., Bohn, M., Buckler, E.S., Ciampitti, I., Edwards, J., Ertl, D., Flint-Garcia, S., Gore, M.A., Graham, C., Hirsch, C.N., Holland, J.B., Hooker, D., Kaeppler, S.M., Knoll, J., Lee, E.C., Lawrence-Dill, C.J., Lynch, J.P., Moose, S.P., Murray, S.C., Nelson, R., Rocheford, T., Schnable, J.C., Schnable, P.S., Smith, M., Springer, N., Thomison, P., Tuinstra, M., Wisser, R.J., Xu, W., Yu, J., Lorenz, A., 2021. Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. Front. Genet. 11. https://doi.org/10.3389/fgene.2020.592769.

Jubair, S., Domaratzki, M., 2019. Ensemble supervised learning for genomic selection. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1993–2000 https://doi.org/10.1109/BIBM47256.2019.8982998.

Jubair, S., Domaratzki, M., 2023. Crop genomic selection with deep learning and environmental data: a survey. Front. Artif. Intell. 5. https://doi.org/10.3389/frai.2022.1040295.

Jubair, S., Tucker, J.R., Henderson, N., Hiebert, C.W., Badea, A., Domaratzki, M., Fernando, W.G.D., 2021. Gptransformer: a transformer-based deep learning method for predicting fusarium related traits in barley. Front. Plant Sci. 12. https://doi.org/10.3389/fpls.2021.761402.

Juliana, P., Singh, R.P., Singh, P.K., Crossa, J., Huerta-Espino, J., Lan, C., Bhavani, S., Rutkoski, J.E., Poland, J.A., Bergstrom, G.C., et al., 2017. Genomic and pedigree-based prediction for leaf, stem, and stripe rust resistance in wheat. Theor. Appl. Genet. 130, 1415–1430. URL: https://doi.org/10.1007/s00122-017-2897-1.

Kakaei, H., Nourmoradi, H., Bakhtiyari, S., Jalilian, M., Mirzaei, A., 2022. Effect of covid-19 on food security, hunger, and food crisis. COVID-19 and the Sustainable Development Goals. Elsevier, pp. 3–29.

Khaki, S., Wang, L., 2019. Crop yield prediction using deep neural networks. Front. Plant Sci. 10. https://doi.org/10.3389/fpls.2019.00621.

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

Kick, D.R., Wallace, J.G., Schnable, J.C., Kolkman, J.M., Alaca, B., Beissinger, T.M., Edwards, J., Ertl, D., Flint-Garcia, S., Gage, J.L., Hirsch, C.N., Knoll, J.E., de Leon, N., Lima, D.C., Moreta, D.E., Singh, M.P., Thompson, A., Weldekidan, T., Washburn, J.D., 2023. Yield prediction through integration of genetic, environment, and management data through deep learning. G3 Genes|Genomes|Genetics 13. https://doi.org/10.1093/g3journal/jkad006. URL: https://academic.oup.com/g3journal/article-pdf/13/4/jkad006/49809187/jkad006.pdf.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O., 2020. Captum: a unified and generic model interpretability library for pytorch arXiv:2009.07896.

Lenz, P., Beaulieu, J., Mansfield, S.D., Clément, S., Desponts, M., Bousquet, J., 2017. Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (picea mariana). BMC Genomics 18, 335. URL: https://doi.org/10.1186/s12864-017-3715-5.

Lin, T., Zhong, R., Wang, Y., Xu, J., Jiang, H., Xu, J., Ying, Y., Rodriguez, L., Ting, K.C., Li, H., 2020. Deepcropnet: a deep spatial-temporal learning framework for county-level corn yield estimation. Environ. Res. Lett. 15, 034016. URL: https://doi.org/10.1088/1748-9326/ab66cb.

Ly, D., Huet, S., Gauffreteau, A., Rincent, R., Touzy, G., Mini, A., Jannink, J.L., Cormier, F., Paux, E., Lafarge, S., Le Gouis, J., Charmet, G., 2018. Whole-genome prediction of reaction norms to environmental stress in bread wheat (triticum aestivum l.) by genomic random regression. Field Crop Res. 216, 32–41. URL: https://www.sciencedirect.com/science/article/pii/S0378429016308401. https://doi.org/10.1016/j.fcr.2017.08.020.

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

Måløy, H., Windju, S., Bergersen, S., Alsheikh, M., Downing, K.L., 2021. Multimodal performers for genomic selection and crop yield prediction. Smart Agricult. Technol. 1, 100017. URL: https://www.sciencedirect.com/science/article/pii/S2772375521000174. https://doi.org/10.1016/j.atech.2021.100017.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of Total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829. URL: https://doi.org/10.1093/genetics/157.4.1819. https://academic.oup.com/genetics/article-pdf/157/4/1819/42032331/genetics1819.pdf.

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

Nguyen, K.L., Grondin, A., Courtois, B., Gantet, P., 2019. Next-generation sequencing accelerates crop gene discovery. Trends Plant Sci. 24, 263–274. https://doi.org/10.1016/j.tplants.2018.11.008.

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J.M., Crossa, J., Manès, Y., Dreisigacker, S., 2012. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 Genes|Genomes|Genetics 2, 1595–1605. URL: https://doi.org/10.1534/g3.112.003665. https://academic.oup.com/g3journal/article-pdf/2/12/1595/40570733/g3journal1595.pdf.

Poland, J., Dreisigacker, S., Shrestha, S., Wu, S., Singh, R., Mondal, S., Juliana, P., Crossa, J., Basnet, B.R., Crespo, L., et al., 2021. Genotypic Data from Cimmyt Bread Wheat Breeding Lines Used in the Feed the Future Innovation Lab for Applied Wheat Genomics. URL: https://hdl.handle.net/11529/10695.

Rachmatia, H., Kusuma, W.A., Hasibuan, L.S., 2017. Prediction of maize phenotype based on whole-genome single nucleotide polymorphisms using deep belief networks. J. Phys. Conf. Ser. 835, 012003. URL: https://doi.org/10.1088/1742-6596/835/1/012003.

Rezayi, S., Liu, Z., Wu, Z., Dhakal, C., Ge, B., Zhen, C., Liu, T., Li, S., 2022. Agribert: knowledge-infused agricultural language models for matching food and nutrition. In: Raedt, L.D. (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, International Joint Conferences on Artificial Intelligence Organization , pp. 5150–5156. URL: 10.24963/ijcai.2022/715.

Rogers, A.R., Holland, J.B., 2021. Environment-specific genomic prediction ability in maize using environmental covariates depends on environmental similarity to training data. G3 Genes|Genomes|Genetics 12. https://doi.org/10.1093/g3journal/jkab440. https://academic.oup.com/g3journal/article-pdf/12/2/jkab440/45932746/jkab440.pdf.

Rogers, A.R., Dunne, J.C., Romay, C., Bohn, M., Buckler, E.S., Ciampitti, I.A., Edwards, J., Ertl, D., Flint-Garcia, S., Gore, M.A., Graham, C., Hirsch, C.N., Hood, E., Hooker, D.C., Knoll, J., Lee, E.C., Lorenz, A., Lynch, J.P., McKay, J., Moose, S.P., Murray, S.C., Nelson, R., Rocheford, T., Schnable, J.C., Schnable, P.S., Sekhon, R., Singh, M., Smith, M., Springer, N., Thelen, K., Thomison, P., Thompson, A., Tuinstra, M., Wallace, J., Wisser, R.J., Xu, W., Gilmour, A.R., Kaeppler, S.M., De Leon, N., Holland, J.B., 2021. The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. G3 Genes|Genomes|Genetics 11. https://academic.oup.com/g3journal/article-pdf/11/2/jkaa050/37042012/jkaa050.pdf.

Ross, B.C., 2014. Mutual information between discrete and continuous data sets. PLoS One 9, 1–5. URL: https://doi.org/10.1371/journal.pone.0087357.

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

Sharma, S., Partap, A., de Luis Balaguer, M.A., Malvar, S., Chandra, R., 2022. Deepg2p: fusing multi-modal data to improve crop production arXiv:2211.05986.

Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S., Singh, A.K., 2021. Crop yield prediction integrating genotype and weather variables using deep learning. PLoS One 16, 1–19. URL: https://doi.org/10.1371/journal.pone.0252402.

Shrikumar, A., Greenside, P., Kundaje, A., 2017. Learning important features through propagating activation differences. In: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning, PMLR, pp. 3145–3153. URL: https://proceedings.mlr.press/v70/shrikumar17a.html.

Sonkar, G., Mall, R., Banerjee, T., Singh, N., Kumar, T.L., Chand, R., 2019. Vulnerability of indian wheat against rising temperature and aerosols. Environ. Pollut. 254, 112946. https://doi.org/10.1016/j.envpol.2019.07.114.

Tadesse, W., Manes, Y., Singh, R.P., Payne, T., Braun, H.J., 2010. Adaptation and performance of cimmyt spring wheat genotypes targeted to high rainfall areas of the world. Crop Sci. 50, 2240–2248. https://doi.org/10.2135/cropsci2010.02.0102.

Tadesse, W., Sanchez-Garcia, M., Assefa, S.G., Amri, A., Bishaw, Z., Ogbonnaya, F.C., Baum, M., 2019. Genetic gains in wheat breeding and its role in feeding the world. Crop Breed. Genet. Genom. 1, e190005.

Tang, X., Liu, G., Zhou, J., Ren, Q., You, Q., Tian, L., Xin, X., Zhong, Z., Liu, B., Zheng, X., et al., 2018. A large-scale whole-genome sequencing analysis reveals highly specific genome editing by both cas9 and cpf1 (cas12a) nucleases in rice. Genome Biol. 19, 84. URL: https://doi.org/10.1186/s13059-018-1458-5.

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

UN World Food Programme, 2022. Update: Global Food Crisis 2022. https://www.wfp.org/publications/update-global-food-crisis-2022. accessed 2022-08-04.

Washburn, J.D., Cimen, E., Ramstein, G., Reeves, T., O'Briant, P., McLean, G., Cooper, M., Hammer, G., Buckler, E.S., 2021. Predicting phenotypes from genetic, environment, management, and historical data using cnns. Theor. Appl. Genet. 134, 3997–4011. URL: https://doi.org/10.1007/s00122-021-03943-7.

World Bank Group, 2022. Food security update. https://www.worldbank.org/en/topic/agriculture/brief/food-security-update. accessed 2022-08-04.

Zegeye, W.A., Zhang, Y., Cao, L., Cheng, S., 2018. Whole genome resequencing from bulked populations as a rapid qtl and gene identification method in rice. Int. J. Mol. Sci. 19. URL: https://www.mdpi.com/1422-0067/19/12/4000.

Zhang, H., Wang, X., Pan, Q., Li, P., Liu, Y., Lu, X., Zhong, W., Li, M., Han, L., Li, J., Wang, P., Li, D., Liu, Y., Li, Q., Yang, F., Zhang, Y.M., Wang, G., Li, L., 2019. Qtg-seq accelerates qtl fine mapping through qtl partitioning and whole-genome sequencing of bulked segregant samples. Mol. Plant 12, 426–437. https://doi.org/10.1016/j.molp.2018.12.018.