Deep Learning for Genomic Selection

by

Sheikh Jubair

A thesis submitted to
The Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements
of the degree of

Ph. D. in Computer Science

Department of Computer Science
The University of Manitoba
Winnipeg, Manitoba, Canada
February 2022

Thesis advisor

Mike Domaratzki and Olivier Tremblay-Savard

Author

Sheikh Jubair

Deep Learning for Genomic Selection

Abstract

One of the significant challenges in the world is to feed its population, as around 193 million people are facing severe hunger. The solution to this problem is to increase the quality and quantity of food while facing the challenge of decreasing or degrading agricultural land. Genomic selection is a predictive technique to identify the top genotypes and develop new cultivars. It uses the whole genome molecular markers of a genotype to predict crop traits even before growing them. There are two main categories of genomic selection: i) single environment trial, where it is assumed that the environment does not impact the crop's development, and ii) multi-environment trial, where the environment influences the crop development by interacting with the genetic component of crops known as GxE. Deep learning models can extract meaningful information from different data sources, such as weather and text data. However, they need to be better developed, especially for multi-environment trials. Here we devised one ensemble deep learning model and one transformer model for single environment trials, and three deep learning frameworks for multi-environment trials. While devising these models and frameworks, we introduced some new techniques for genomic selection, such as representing markers with genotype frequency, environment-specific markers and global markers that are not related to any environ-

ment but to a specific trait. The results demonstrate that our single environment models are competitive with or comparable to existing methods, while our multi-environment frameworks are better than some existing methods. We anticipate that this research will help future research build more complex deep learning frameworks for both categories. For example, future multi-environment frameworks may incorporate different data types, such as soil data and images of agricultural land. One of the proposed frameworks can be extended further to facilitate additional data and models.

Contents

List of Figures

List of Tables

Acknowledgments

I want to begin by thanking my advisors, my committee, my parents, my wife, and all the people who have supported me along the way. I am blessed to be supervised by Dr. Mike Domaratzki and Dr. Olivier Tremblay-Savard. Dr. Mike helps me develop as his student and in my professional career, always encourages me to do the best work, and guides me in the right direction when I am lost. Through his extraordinary supervision, he becomes my role model and I want to be a supervisor like him.

Although Dr. Olivier was not my supervisor from the beginning, he was one of the professors in our Bioinformatics Lab. Because of his effort, sharing knowledge among peers becomes easy, enjoyable and fun. After becoming my cosupervisor, he provided valuable feedback to complete the last part of the thesis.

I want to give special thanks to my committee, as their expertise in agriculture and machine learning helps me to earn more domain knowledge. Moreover, I applied some of their advice not only in my thesis but also in my workplace.

The impact of my parents during my Ph.D. journey is tremendous as my mother, Dr. Nasima Begum, motivated me a lot to pursue Ph.D. and supported every positive decision I took. My father, Dr. Modabber Hashmi, is the main person for whom I am here today. He always provided me with the best of everything and worked relentlessly without making me understand how hard he is working.

I married my beautiful wife, Tahsina Shuchi, on December 2021. However, she was my girlfriend for the last ten years before we married. We maintained a long-distance relationship since September 2015. During this long period of staying apart, she always tried to ensure that I did not feel the distance between us and supported me mentally.

Finally, I would like to thank my friends and roommates in Winnipeg. During these four years, they became next to my family and always cheered me up when needed.

This thesis is dedicated to my parents and my wife and to the people who are suffering from hunger.

Chapter 1

Introduction

Food insecurity refers to people's inability to meet the required amount of nutrition by consuming a sufficient diet. The inability comes from the shortage of agricultural production, the excessive price of agrifood or the seasonal lack of access to food [Government of Canada, 2020; FAO, 2006]. Due to food insecurity, around 85 million more people were facing a severe food crisis in 2021 compared to what was reported in 2016, which has resulted in around 193 million people facing severe hunger across 36 countries [FAO, 2022]. Although much effort is going on to increase agricultural food production, it is estimated that around 10% of the world population (828 million) in 2022 is facing hunger which shows no improvement in the percentage of the nutrition-deprived population from 2015 [WFP et al., 2022; WFP, 2022]. Figure 1.1 shows the food insecure areas worldwide. Most of the under-developed and developing countries in Africa, Asia, Latin America and the Caribbean suffer from moderate to high food insecurity due to pre and post-harvest losses caused by various plant diseases, unfavourable weather conditions (i.e., drought, heavy rainfall and

heat waves), climate change, increase in population, wars, inability to purchase agri-products from international markets and spike in food prices [Otekunrin et al., 2019, 2020; Anderson et al., 2021; Falkendal et al., 2021; Ben Hassen and El Bilali, 2022]. However, developed countries such as Canada also face food insecurity as 10.8% to 11.6% households faced moderate to severe food insecurities between 2018 and 2020 primarily because of financial constraints and prices of agrifood [Tarasuk et al., 2022].

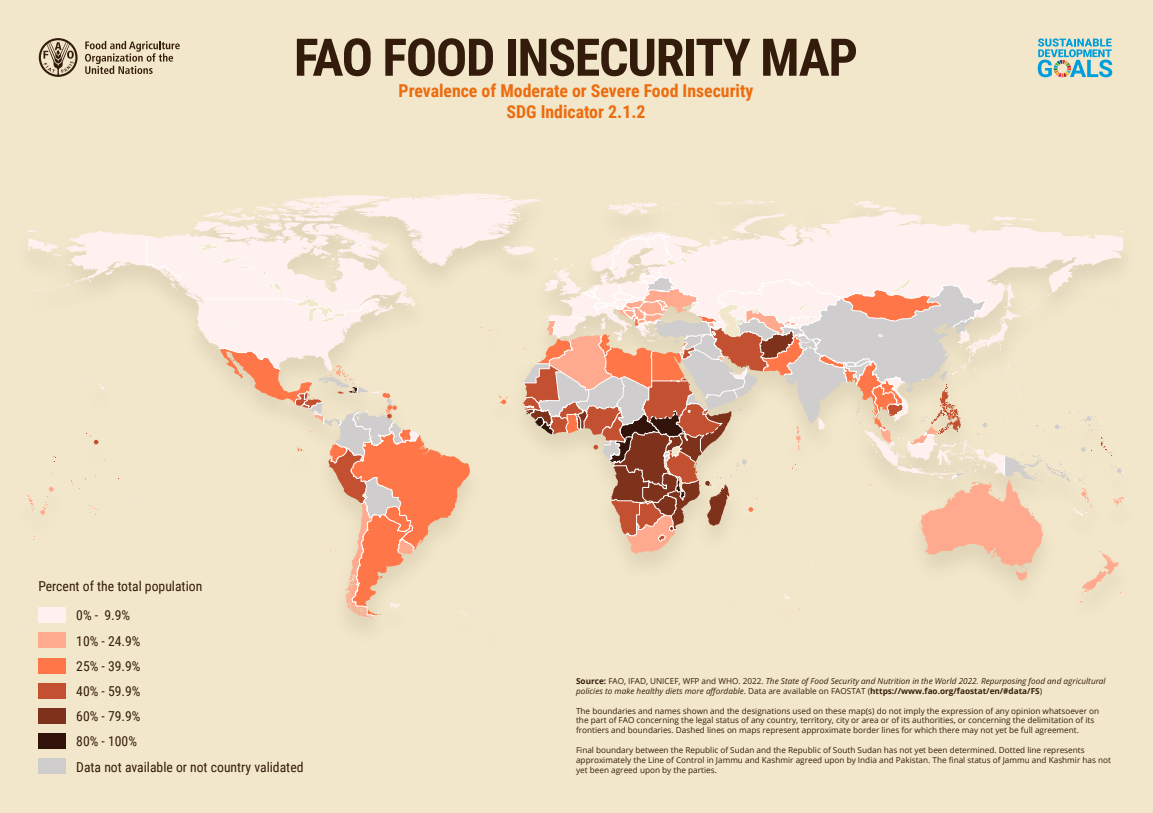


Figure 1.1: Percent of food insecured population [FAO, 2021][1].

To ensure food security, it is estimated that we need to increase our food production by 50% by 2050 while facing the challenge of climate change, soil erosion, decreasing amount of agri-land for the increasing population and extreme weather

[1]copyright FAO, reproduction permitted for academic purposes.

events such as floods and droughts [Van Meijl et al., 2018; Bourgault et al., 2018; Nawaz and Chung, 2020]. Different measures have been taken to face the challenge of climate change that also has negative impact on food production. For example, in Canada, the federal government looks to reduce GHG emissions by reducing fertilizer use by 2050 [Agriculture and Agri-Food Canada, 2022].

Plant breeding is a procedure for improving the traits of a crop, such as pest resistance, higher yield and higher nutritional value [Acquaah, 2009]. To improve the quality of the plants, it is essential to identify top lines or genotypes that possess the desired traits. A line or genotype is a crop organism that contains all its genetic materials. However, genotype also refers to the alleles of a gene of a specific locus. On the other hand, traits are the characteristics of an organism that are expressed by genes and influenced by the environment. The expression of the quantitative traits is the phenotype of an organism. In other words, phenotypes are the physical appearance or biochemical characteristics of an organism determined by the interaction of its genotype and the environment.

Genomic selection (GS), also referred as genomic prediction (GP), is a prediction tool that develops new cultivars by estimating phenotypes of a quantitative trait from genotypes. It makes the link between traits and the underlying genomic information by using whole-genome molecular markers. 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 crop [Le Nguyen et al., 2019; Esposito et al., 2019]. These sequencing and genotyping technologies usually provide us with many molecu-

lar markers, typically covering the whole genome of the crop. 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., 2021a; Khaki and Wang, 2019; Ma et al., 2018; Rachmatia et al., 2017a; Crossa et al., 2016b]. GS predicts traits by combining all DNA marker data to build one single model. On the other hand, traditional marker-assisted selection (MAS) works with only known genes that are important for the trait [Jannink et al., 2010]. However, many important genes for crops are still unknown. Moreover, statistical methods used in MAS performed poorly in improving complex traits [Heffner et al., 2009] due to the insufficient training population size that sometimes causes underfitting of the statistical models [Dekkers and Hospital, 2002; Schön et al., 2004; Wong and Bernardo, 2008].

In GS, the experimental environments can be of two types: i) single environment and ii) multiple environments. In single environment trials, cultivars are developed without considering the impact of the environment, assuming that they will be sown in the same location and time of the year. However, multi-environment trials (MET) are crop experiments where the crop is grown in multiple geographic locations, seasons, or years. The main challenge of considering MET is that quantitative traits are affected by genotype by environment interaction known as GxE. GxE can both negatively and positively affect the traits based on different environments [Lin et al., 2020; Bellairs et al., 1996; Shrestha et al., 2012]. Thus, considering GxE in a learning model can help breeders choose appropriate genotypes based on environments [Roorkiwal et al.,

2018]. In this research, we started our work on single environment trials and then extended it to multi-environment trials.

There are two types of models that are primarily employed for GS: i) statistical models and ii) machine learning models. Most of the statistical models are linear [Burgueño et al., 2012; Cuevas et al., 2017; Ferrão et al., 2017; Wang et al., 2018; Bandeira E Sousa et al., 2017; Cuevas et al., 2019; Howard et al., 2019; Millet et al., 2019; van Dijk et al., 2021; Anilkumar et al., 2022] and may not capture relationships between genotypes and complex traits. On the other hand, machine learning, especially deep learning models able to predict complex traits and perform either better or as good as the statistical models [Holliday et al., 2012a; Ogutu et al., 2011b; Rachmatia et al., 2017b]. This research focuses on devising deep learning methods for predicting traits from genetic markers to improve the breeding of plants for a) single environment trials and b) multi-environment trials. For the multi-environment trial, we incorporate environmental data, such as the amount of rainfall, maximum and minimum temperatures, altitudes and day length of the next breeding cycle with the genetic data to replicate GxE and predict traits.

It has been observed that deep learning methods perform equally or better than statistical methods, but there is no single method that is better than all other methods and the performance of the same method can differ for different traits of the same species [Holliday et al., 2012a; Ogutu et al., 2011b; Rachmatia et al., 2017b; Ma et al., 2018]. In addition, there are very few machine learning methods that take GxE interaction into account because of the complexity of incorporating the environmental interaction into the model and lack of environmental information [Khaki and Wang,

2019; Shook et al., 2020; Lin et al., 2020; Khaki et al., 2020; Gangopadhyay et al., 2020; McCormick et al., 2021; Washburn et al., 2021; Måløy et al., 2021; Zhong et al., 2022]. We reviewed the existing deep learning models with a particular focus on multi-environment trials in chapter 2.

We employed two deep learning approaches to solve genomic selection for a single environment: i) ensemble deep learning approach on a wheat dataset for predicting six different traits in chapter 3 [Jubair and Domaratzki, 2019] and ii) transformer-based deep learning model named GPTransformer to develop disease resistant cultivar of Barley in chapter 4 [Jubair et al., 2021a]. The ensemble technique in machine learning employs multiple weak machine learning algorithms to predict target values (phenotypes in GS). Although ensemble learning methods have been applied in animal breeding programs successfully [Li et al., 2018; Abdollahi-Arpanahi et al., 2020; Liang et al., 2021], they are not explored much in GS for plants [González-Camacho et al., 2018].

On the other hand, the transformer is a more recent deep learning approach [Vaswani et al., 2017] that has been initially applied to Natural Language Processing (NLP) tasks [Devlin et al., 2018] and later adopted to bioinformatics to perform downstream tasks, such as identifying important biological regions (i.e., promoters, transcription sites) and gene expression prediction [Zaheer et al., 2020; Ji et al., 2021; Avsec et al., 2021]. To the best of our knowledge, we employed the first transformer-based model for genomic prediction. Our results for both models is either better or equal to the state-of-the-art model for GS.

Finally, we devise two deep learning frameworks, GxENet, in chapter 5 that take

GxE into account within the model by incorporating environmental and genotype data along with field notes to predict environment specific yield for each line of wheat. While devising one of the frameworks, we utilized our observation from our previous two works that the deep learning models can predict traits for a single environment trial successfully, described in chapter 3 and chapter 4, to build one of the representation learning models of the framework. The results demonstrate that GxENet can identify environment-specific top lines of wheat successfully in two different test scenarios.

In this research, we have the following contributions:

- We provided a comprehensive overview of the genomic selection process with deep learning that starts from data and ends with creating a new cultivar for both single and multi-environment trials. To do it, i) we provided an overview of different data of GS and how these data need to be processed, ii) discussed popular components of deep learning models typically employed in GS and then iii) reviewed existing deep learning architectures and motivation behind them for both single and multi-environment trials.

- We devised an ensemble deep learning method to predict six traits of wheat in a single environment trial that obtained either better or state-of-the-art performance.

- We employed a Transformer-based deep learning method, GPTransformer, that uses genotypic and phenotypic data to predict FHB severity and DON levels in a two-row barley population. We investigated a new approach to represent markers using genotype frequency obtained from employing the Hardy-Weinberg

equilibrium rule. In addition, we investigate the effect of feature selection on genomic prediction using mutual information and examine the biological relevance of the top markers identified by the mutual information method. The results from this model also demonstrate state-of-the-art performance.

- We devised two multi-environment genomic selection frameworks that predict environment specific yield for each line. We proposed a novel concept of global and local marker sets for feature selection in one of the frameworks. The global marker sets are essential for yield prediction, irrespective of any environment. On the other hand, local markers are environment-specific important markers for a particular trait. In addition, 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 and genetic information contribute to predicting yield in our models. Finally, we also extended one of the frameworks (third framework) to integrate unstructured field notes.

Chapter 2

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

Sheikh Jubair[1], Mike Domaratzki[2]

1 – Department of Computer Science, University of Manitoba, 66 Chancellors Cir, Winnipeg, MB R3T 2N2

2 – Department of Computer Science, University of Western Ontario, 1151 Richmond St, London, ON N6A 3K7

This chapter is an adaptation of the article Jubair and Domaratzki [2023] published in Frontiers in Artificial Intelligence in 2023.

Author Contributions

SJ: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing- review and editing; MD: Writing - review and editing, Supervision.

Abstract

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, multi-environment, MET, deep learning

2.1 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 [Nawaz and Chung, 2020; Linehan et al., 2012; United Nations, 2019]. 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 [Nawaz and Chung, 2020; United Nations, 2019]. 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 [Varshney et al., 2017; Acquaah, 2009].

GS for single environment trials employs GS to identify top individuals to create

a new cultivar for a specific environment [Meuwissen et al., 2001; Heffner et al., 2009; Crossa et al., 2017; Jubair et al., 2021a]. If the environment changes, single environment GS does not guarantee that the new cultivar 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; Anilkumar et al., 2022; van Dijk et al., 2021] 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 [Danilevicz et al., 2022; Tong and Nikoloski, 2021; van Dijk et al., 2021; Anilkumar et al., 2022; Montesinos-López et al., 2021]. 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 [Shook et al., 2020; Khaki and Wang, 2019; Khaki et al., 2020; Lin 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; Cuevas et al., 2017; Ferrão et al., 2017; Cuevas et al., 2019; Bandeira E Sousa 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., 2012b; Ali et al., 2020; Sawitri et al., 2020], Support Vector Machines (SVMs) [Ogutu et al., 2011b; Wang et al., 2019] and Neural Networks (NNs) [Pérez-Enciso and Zingaretti, 2019; Jubair and Domaratzki, 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 [Jubair et al., 2021a; Montesinos-Lopez et al., 2021; Zingaretti et al., 2020; Jubair and Domaratzki, 2019; Ma et al., 2018; Rachmatia et al., 2017a; McDowell, 2016]. 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 [Shook et al., 2020; Montesinos-López et al., 2018a, 2019b; Khaki et al., 2020; Lin et al., 2020; Khaki and Wang, 2019] 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 three years, new research has demonstrated the potential of incorporating environmental information into deep learning models for GS [Shook et al., 2020; Khaki et al., 2020; Lin et al., 2020; Khaki and Wang, 2019]. This survey focuses specifically on deep learning methods for integrating weather 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 [Montesinos-López et al., 2022; Song et al., 2022; Putra et al., 2022; Costa-Neto 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 cultivar 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.

2.2 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 [Crossa et al., 2017; Heffner et al., 2009; Goddard and Hayes, 2007]. 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 [Jubair and Domaratzki, 2019; Crossa et al., 2016b]. Table 2.1 shows an example dataset.

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

Table 2.1: An Example of genotyped data. 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.

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., 2021a] 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 neighbour. 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 the 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., 2021a]. 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., 2021a], 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, vapour pressure, wind speed and radiation, plays an essential part in GS for multi-environmental trials [Shook et al., 2020; Khaki and Wang, 2019; Khaki et al., 2020; Gangopadhyay 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 step. Table 2.2 shows an example of genotyped and environmental data after pre-processing.

2.3 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 [Dong et al., 2021; Li et al., 2020; LeCun et al., 2015]. The dominant paradigm in DL is a network. A

		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

Table 2.2: An example of genotyped and environmental data after pre-processing in a tabular format. 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.

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 [Dong et al., 2021; LeCun et al., 2015]. 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 2.3.1). 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 2.3.2). Normalization and reg-

ularization layers such as L1, L2 and dropout are applied after the activation layer to generalize the model to avoid overfitting (discussed in section 2.3.3). Figure 2.1 shows the general architecture of a deep learning method.

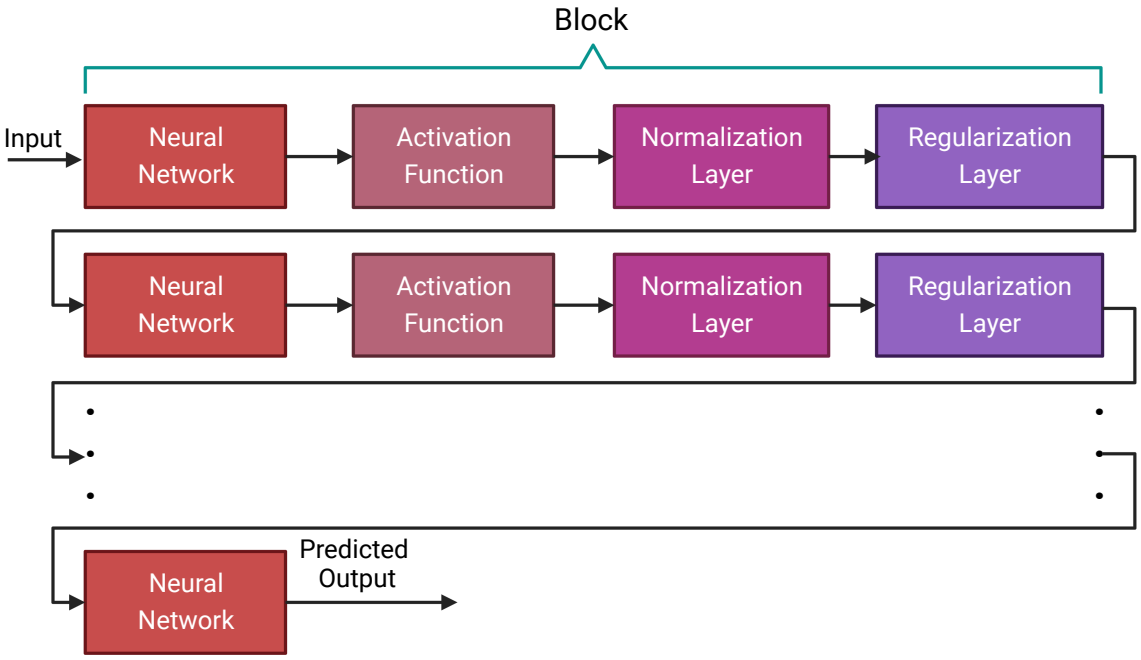


Figure 2.1: General architecture of a deep learning algorithm. All the layers between the first and last layer are called hidden layers. The first layer is the input layer and the last layer is the output layer. A neural network layer is typically followed by an activation function and then by normalization 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.

2.3.1 Neural Networks

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 2.2 shows a fully connected network.

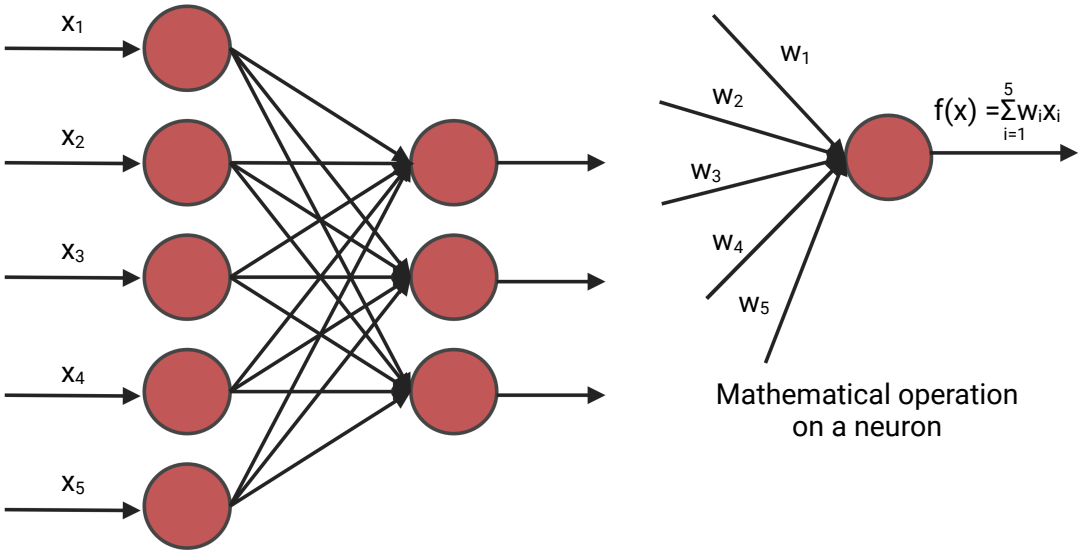


Figure 2.2: An example of fully connected layer. x_i is the input and w_i is the weight. The weights are initialized randomly and are optimized iteratively for prediction. In GS, x_i is the marker.

MLPs have been applied to predict phenotypes both in single environment trial [Jubair and Domaratzki, 2019; Jubair et al., 2021a; González-Camacho et al., 2016; Gianola et al., 2011; Montesinos-López et al., 2019a] and multi-environment trial Khaki and Wang [2019]; Montesinos-López et al. [2018b]. 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].

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 each sliding window, the network is on a small subset of the data [Dong et al., 2021; LeCun et al., 2015]. 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 [Li et al., 2021; Dong et al., 2021]. The applications of convolutional neural networks have also been extended to other domains such as GS [Zingaretti et al., 2020; Jubair and Domaratzki, 2019; Liu et al., 2019b; Ma et al., 2018].

There are three types of convolution, conv1D, conv2D and conv3D, available in different deep learning frameworks [Abadi, 2016; Chollet et al., 2018; Paszke et al., 2017]. 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, -1), the marker data can also be converted to a one-hot encoded vector which will be the input of a conv2D layer [Ji et al., 2021; Avsec et al., 2021; Washburn et al., 2019;

Liu et al., 2019b]. Figure 2.3 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 until 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 a convolutional neural network to 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 a 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 [Washburn et al., 2021; Khaki et al., 2020; Sharma et al., 2022]. The predictions from different networks can be combined to obtain an overall prediction.

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

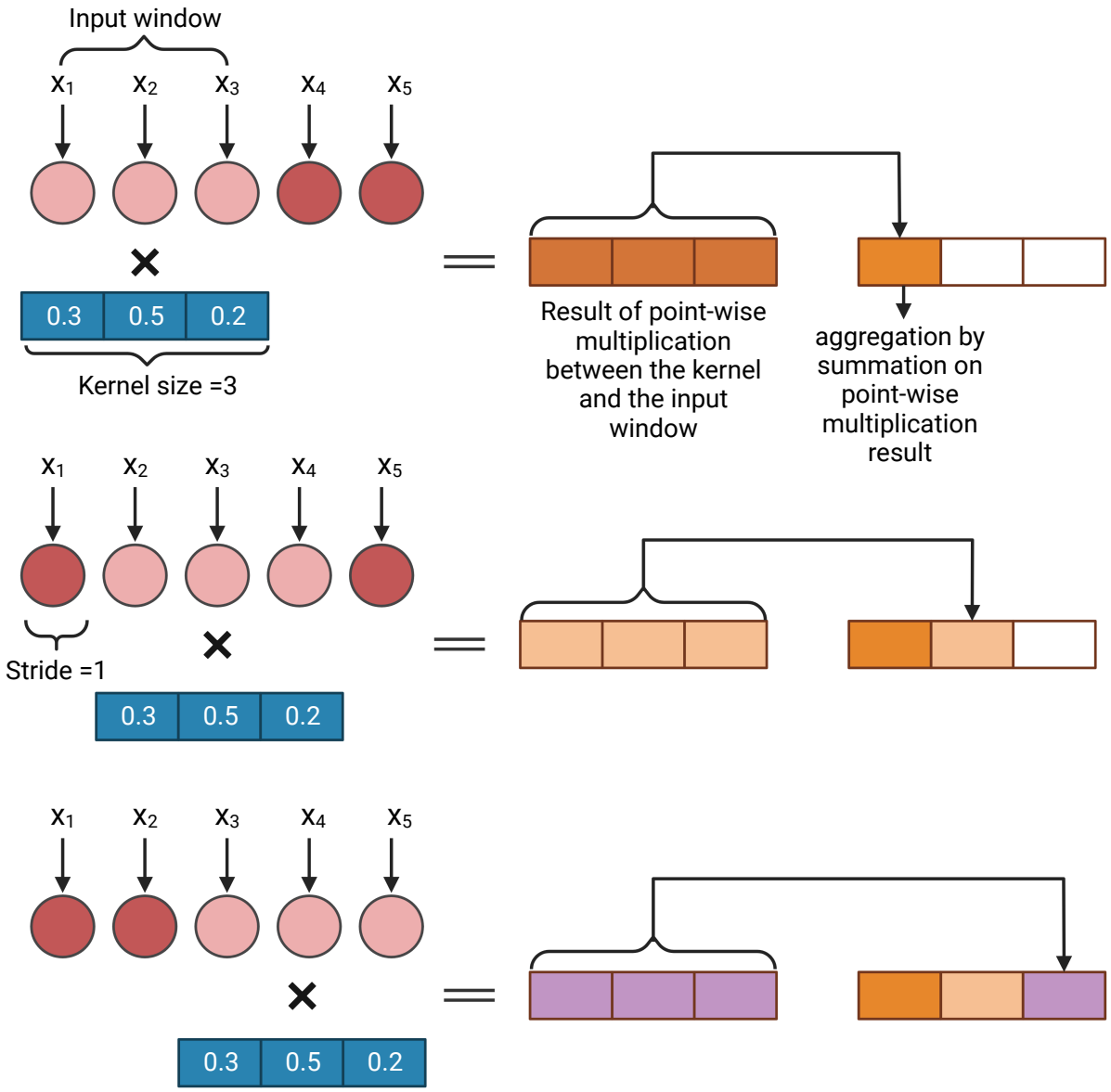


Figure 2.3: An example of convolution operation on a one dimensional input vector. In GS, x_i is a marker.

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.

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; Jubair et al., 2021a; Ji 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., 2021a] 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., 2021a]. 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 2.4 shows a transformer encoder. The input of the transformer can be a one hot encoded vector or the genotype frequency [Jubair et al., 2021a; Avsec 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.

2.3.2 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,

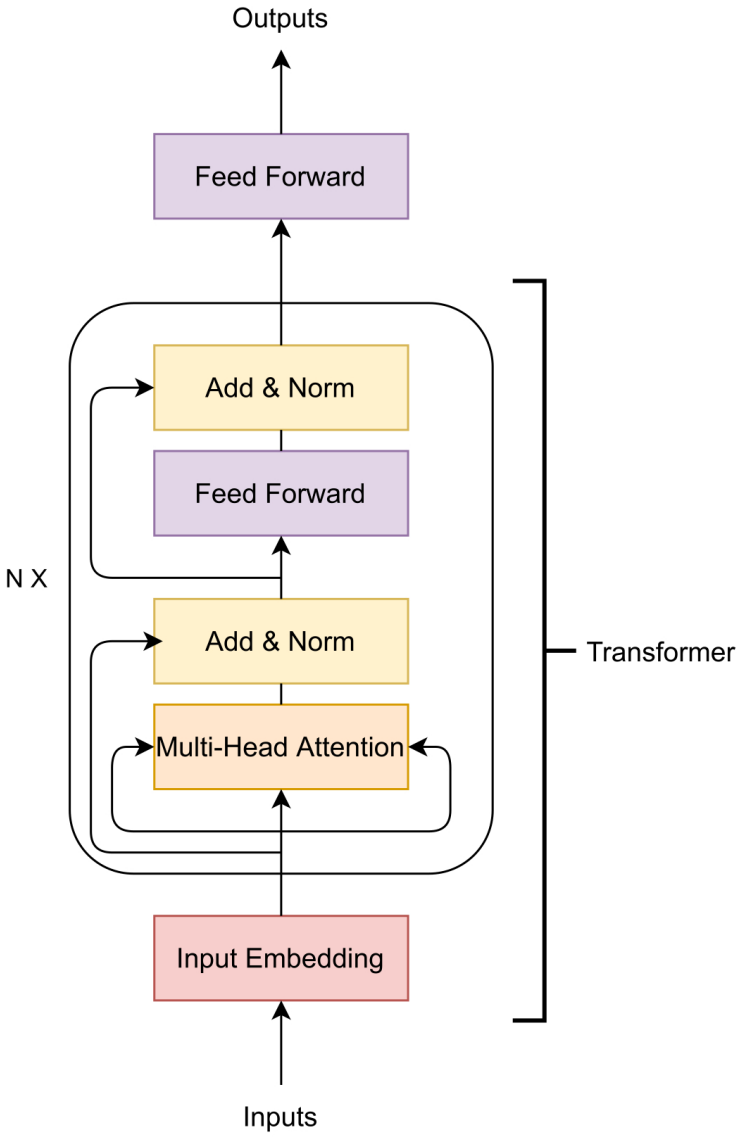


Figure 2.4: A transformer encoder. Here N represents N transformers can be stacked together.

we need to 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; Shook et al., 2020; Khaki et al., 2020; Washburn et al., 2021; Måløy et al., 2021]. Hence, we provide an overview of these activation functions below.

The sigmoid activation produces the output neuron between 0 to 1 by applying the sigmoid function [Szandała, 2021; Dubey et al., 2022]. Though the sigmoid function is one of the most used activation functions, 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 towards 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.

2.3.3 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 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.

2.3.4 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 [Shook et al., 2020; Khaki and Wang, 2019; Khaki et al., 2020; Rachmatia et al., 2017a; Ma et al., 2018; Jubair et al., 2021a]. 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].

2.3.5 Optimization

The objective of training is to optimize the DNN. For optimizing, after each iteration, the weights need to be adjusted to minimize the 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].

2.3.6 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 cultivar, 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 the linear relationship between 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].

2.3.7 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 set-

tings. 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 [Crossa et al., 2016a; Cuevas et al., 2016; de Los Campos et al., 2010; Endelman, 2011; Hassen et al., 2018; Lopez-Cruz et al., 2015; Pérez and de Los Campos, 2014; Pérez-Elizalde et al., 2015]. 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 2.5, 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.

2.4 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 develop a new cultivar 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 cultivar for crops [Tong and Nikoloski, 2021; Pérez-Enciso and Zingaretti, 2019]. 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 2.5 shows how a new cultivar 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; Pérez-Rodríguez et al., 2012; González-Camacho et al., 2012]. Below, we discuss the advancement and 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,

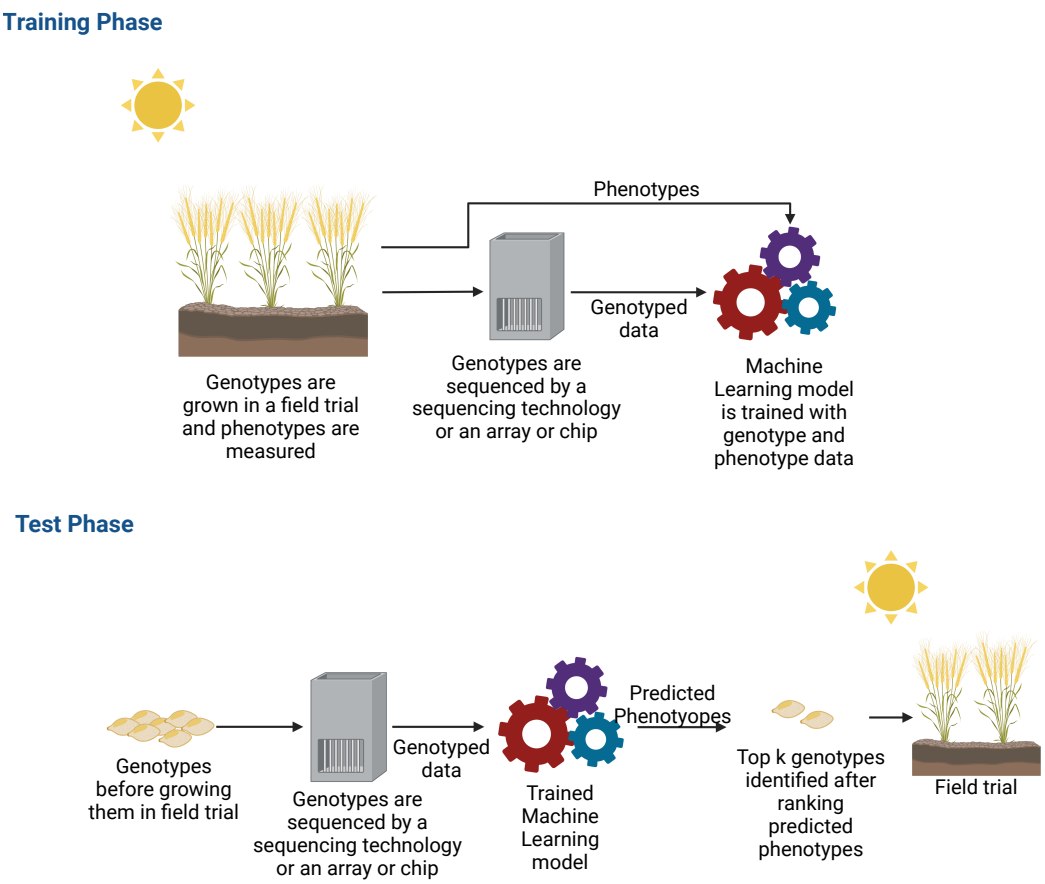


Figure 2.5: Workflow of genomic selection in single environment trial. For collecting training data, plants are grown in a field trial and phenotypes were measured. These plants are also genotyped. After obtaining both genotyped and phenotyped data, a machine learning model is trained during the training 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.

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 observed an improvement of

11% to 18% with their BRNN model 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 of yield, 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. [2017a] 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 labelled. 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. [2017a] 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% to 7.5% for different traits.

Though most of the research has found that machine learning performs better than the statistical methods [Ma et al., 2018; McDowell, 2016; Rachmatia et al., 2017a], 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% to 7% improvements in the ranking of traits such as grain length, grain width, grain hardness, thousand-kernel weight, test weight, sodium dodecyl sulphate 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 to 30% better than single machine learning methods and slightly better than RRBLUP in correlation coefficient 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 [Ji et al., 2021; Avsec et al., 2021; Le et al., 2022]. Jubair et al. [2021a] 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.

2.5 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 [Shook et al., 2020; Lin et al., 2020; Khaki 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 trialled 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 2.6 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 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

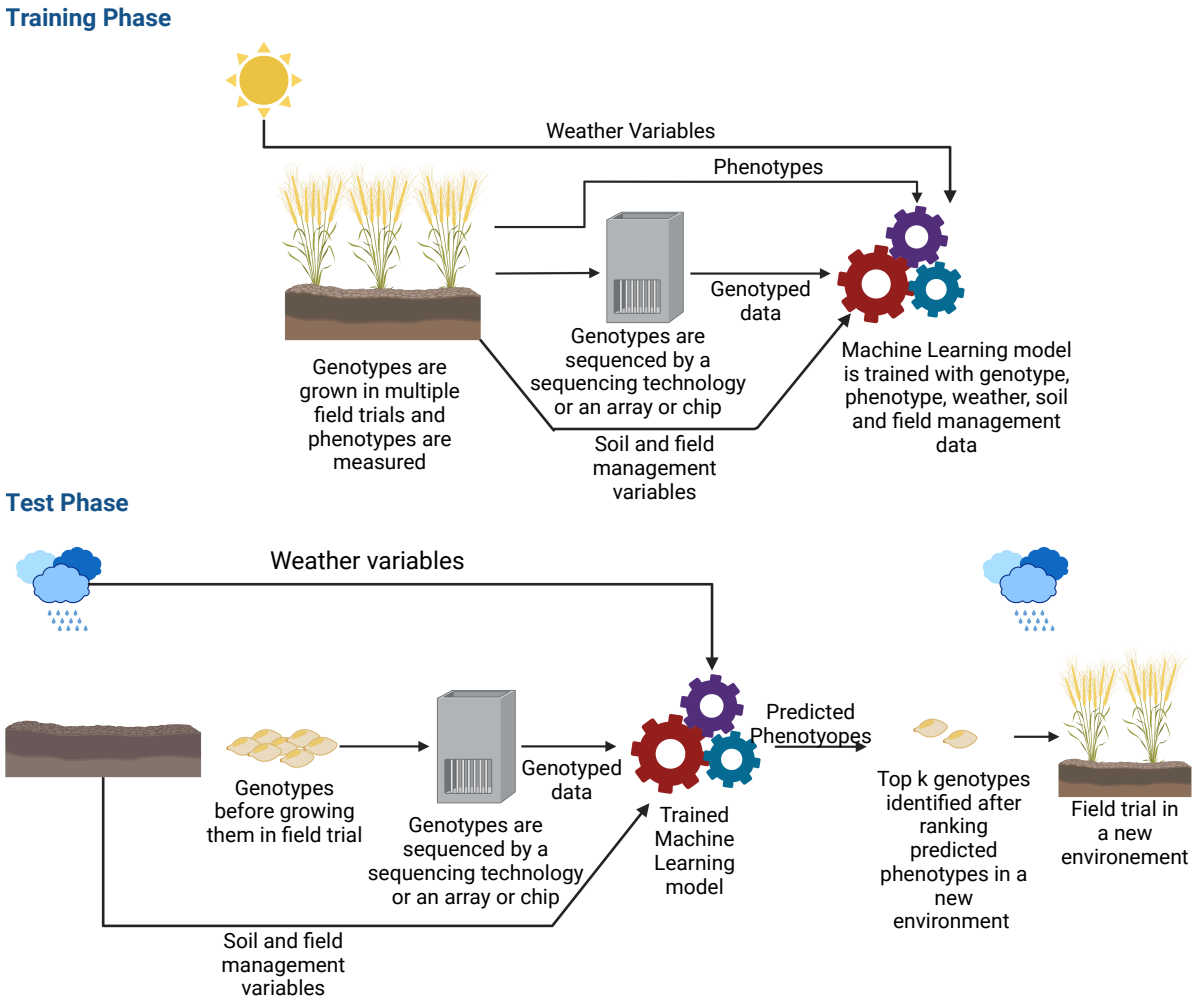


Figure 2.6: Workflow of genomic selection in a multi-environment trial. Before training the machine learning model, along with genotyped and phenotyped data, environmental information such as weather variables, soil and field management data are also collected. The genotypes are typically grown in multiple 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.

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 four 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 time 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 vari-

ables. 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% to 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, vapour 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 2.3, we list the deep learning-based academic papers that work with multi-environment trial and environmental data. Some single-environment models [Sandhu et al., 2021c,b, 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.

2.6 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

Year	Author	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

Table 2.3: Papers on Multi-Environment Deep Learning Models. In the table, MLP means Fully Connected Networks and Att means attention networks.

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 vapour pressure [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 vari-

ables 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; Shook et al., 2020; Khaki 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, the time window can be shorter, and in the later stage, the size of the 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 (Internet of Things) 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 considered 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., 2021a; Måløy et al., 2021], we believe attention-based architectures are the most promising approaches 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., 2021a; 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 [Xu et al., 2022; Spindel and McCouch, 2016]. 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.

Conflict of Interest Statement

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.

Funding

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

Chapter 3

Ensemble supervised learning for genomic selection

Sheikh Jubair[1], Mike Domaratzki[2]

1 – Department of Computer Science, University of Manitoba, 66 Chancellors Cir, Winnipeg, MB R3T 2N2

2 – Department of Computer Science, University of Western Ontario, 1151 Richmond St, London, ON N6A 3K7

This chapter is an adaptation of the article Jubair and Domaratzki [2019] published in IEEE International Conference on Bioinformatics and Biomedicine (BIBM) in 2019.

Author Contributions

SJ: Conceptualization, Formal analysis, Investigation, Writing - original draft, Writing-review and editing; MD: Writing - review and editing, Supervision.

Abstract

To meet the world's growing food and nutrition demands, agricultural breeders need to grow crops with improved phenotypes and create cultivars that allow increased production. Genomic selection enables the breeders to select individuals with improved phenotypes even before growing them. Existing genomic selection is made mostly through statistical methods that do not accurately predict complex non-linear traits. Deep learning and other machine learning methods have been applied, but most of the deep learning methods are not specifically designed for genomic selection. Also, there has been relatively little comparison between different machine learning methods. We propose three ensemble learning methods: i) ensemble support vector

regression; ii) ensemble deep convolutional neural networks and iii) random forests; to predict phenotype with high accuracy. We also propose a feature selection strategy that identifies important markers and contributes to improved phenotype prediction. The proposed marker selection strategy is independent of machine learning methods; thus, the markers that are selected remain the same when the machine learning model is changed. We employed our methods to Iranian wheat landraces. The result shows that ensemble learning methods are better than the single machine learning methods with the lowest PCC 0.339 for plant height and the highest PCC 0.747 for grain length. Our models are also robust as they rank both top twenty and bottom twenty individuals well with nDCG@20 ranges from 0.188 to 0.712.

Keywords– genomic selection, deep learning, machine learning, ensemble learning

3.1 Introduction

Selection of proper individuals with intended phenotypes from a collection of varieties of a crop is essential to breeders as the right selection can lead to improvements in the crop such as drought resistance, biotic and abiotic stress resistance, yield improvement and disease resistance. While the amount of water, fertilizer, pest control, and good production practices constitutes the environment for the plant, the cultivar of the plant defines the ability to produce desired phenotypic value within that environment [Milton, 1979]. Thus, if the environmental factors and breeding practices are standardized, it is also vital to create improved cultivars for that environment. Genomic selection (GS) is a marker-assisted selection method that uses whole-genome

molecular markers to improve the quantitative traits or phenotypes of an organism, such as a crop or livestock, by identifying the top lines. That is, GS is a computational tool for choosing the most advantageous individuals from varieties and has the potential to save money and time by accelerating improvements to crops or livestock. Thus GS can solve the two main objectives of the breeders: building variation and selection of leading individuals from the cultivar that fulfills the breeding objective [Acquaah, 2009]. Though GS has been successfully applied to livestock, GS for crops is not as well developed [Bhat et al., 2016] and therefore, there is a need for new computational tools of GS for plants. With proper GS software, it is possible to address the problem of feed quality, increased supply needs for food in a growing world population, and adaptation of crops for a specific environment such as drought stress and wet conditions.

GS links traits and the underlying genomic information. A trait is a characteristic or a feature of an organism, such as height, length, yield, and disease resistance. A phenotype is the expression of a particular trait that is determined by the interaction between an organism's genotype and the environment. A large number of small effect genes known as polygenes cumulatively contribute towards the final expression of the phenotype. Though many markers contribute to the complex phenotypes of plants, some markers mostly interact with the environment and are responsible for a specific phenotype, and other markers remain stable[Oakey et al., 2016]. Identification of markers that interact with the environment and respond to a phenotype is necessary to understand the crops and build a better cultivar.

In this paper, we propose three ensemble machine learning models: i) ensemble

support vector regression; ii) ensemble deep convolutional neural networks and iii) random forests to predict different phenotypes of Iranian wheat landraces. We also compare the performance of our ensemble models with single machine learning models, such as support vector machines and convolutional neural networks, and a statistical model RR-BLUP. We also combine the concept of binning the continuous values of labels [Trohidis et al., 2008] and apply a filter-based method feature selection algorithm [Doquire and Verleysen, 2013] to identify important markers for obtaining improved phenotypic values. As the filter-based feature selection method, we use chi-square feature selection. In general, for GS, the features are the genotyped markers and the labels are the traits. In the training phase, each of our models takes genotyped markers and a phenotyped trait as the input, performs feature selection on the marker data, creates several subsets of markers from the selected markers and then trains each subset employing a machine learning algorithm. In the testing phase, the inputs of the trained models are the same subset of markers that are used to train a specific model. The trained models predict the phenotypes and the final output is the average of all the predicted phenotypes. The training data is both genotyped and phenotyped, but the test data is only genotyped but not phenotyped.

Machine learning methods are known to perform better than statistical methods generally, but for GS, there is no single method that is better than all other methods and the performance of the same method can differ for different traits of the same species [Ogutu et al., 2011a; Holliday et al., 2012a; Rachmatia et al., 2017b]. Very recently, Pérez-Enciso and Zingaretti [Pérez-Enciso and Zingaretti, 2019] have reviewed deep learning techniques for GS, showing a limited amount of existing re-

search on genomic selection for both plant and animal breeding [Mcdowell, 2016; Liu and Wang, 2017; Ma et al., 2018; Khaki and Wang, 2019]. Many of these models show that multi-layer perceptron models and convolution neural networks are comparable to or exceed established statistical techniques such as GBLUP. Ma et al. [Ma et al., 2018] employed DeepGS, a deep learning model with convolution neural network to predict the phenotypes from the lines and to the best of our knowledge, this is the only model that is specifically designed for GS on wheat. This model obtained better performance than existing statistical models, and the authors demonstrated that decreasing the number of markers increases the performance of DeepGS. Though the number of markers plays a crucial role in the prediction of phenotype, the selection of the markers in DeepGS is done randomly, and there is a need to identify the important markers for each trait, which leads to improved phenotype prediction.

As the traits we are going to predict are continuous variables, to select the features, one of the strategies is to make several bins within a range of values and consider each of those bins as a category [Trohidis et al., 2008]. After obtaining the categorical labels for each individual, any filter-based feature selection method can be applied to retain the markers that contribute to discriminate the groups [Doquire and Verleysen, 2013]. In this paper, we use the chi-square feature selection as the filter method. In general, the filter-based feature selection does not rely on any machine learning algorithm; instead, they consider each feature individually and calculate a value based on some criteria to identify the potential to separate the classes. The features are then ranked based on the calculated values and the top features are used as inputs in any machine learning algorithm [Tang et al., 2014].

Ensemble learning methods combine the prediction of multiple weak machine learning methods and produce a reliable prediction [Dietterich et al., 2002]. There are two types of ensemble learning methods: bagging and boosting [Quinlan et al., 1996]. In this paper, we employed bagging to build our models and make predictions of phenotypes. In bagging, each model is trained independently with different subsets of the data. When making the prediction, each model produces an independent outcome, and the final result is the average of all the guesses. As each model is trained with different subsets, they capture the information of only a small part of the data. After combining all the models, they produce an overall picture and make a stronger prediction.

We organize the rest of the papers as follows. In section 3.2 of this chapter, we described our materials and methods; section 3.3 contains the result, and section 3.4 is our conclusion.

3.2 Materials and Methods

3.2.1 Dataset

The Iranian bread wheat (Triticum aestivum) dataset was obtained from the wheat gene bank of CIMMYT [CIMMYT, accessed July, 2019]. The dataset contains 2000 individuals of Iranian bread wheat and genotyped with genotype by sequencing method (GBS) using $33,709$ DArT (Diversity Array Technology) markers where the values are either 0 or 1 for each marker. All the phenotypic traits were measured in a single standard environment. The traits are thousand-kernel weight (tkw), test

weight (tw), grain hardness (gh), grain length (gl), grain width (gw), and plant height (pht). More details can be found in [Crossa et al., 2016b].

3.2.2 Marker Selection

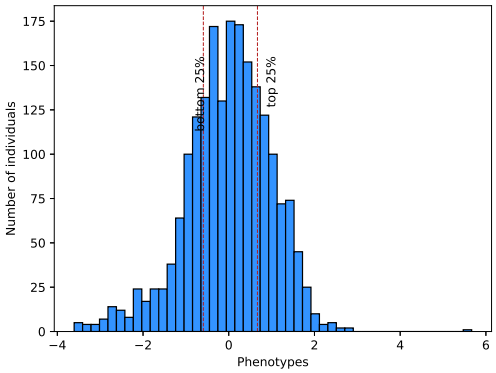
Previous work [Ma et al., 2018] on this dataset indicated that not all features are informative for genomic selection. This is consistent with other research that demonstrates that feature selection generally improves phenotype prediction [Oakey et al., 2016].

To select features, we consider each phenotype individually. We created three bins where the first bin contains the top 25% lines based on their phenotypic value; in the second bin, the middle 50% were placed, and in the last bin, the bottom 25% were kept. The distribution of all traits is shown in Figure 3.1. The left side of the left red line shows the bottom 25% and the right side of the right red line indicates the top 25% individuals.

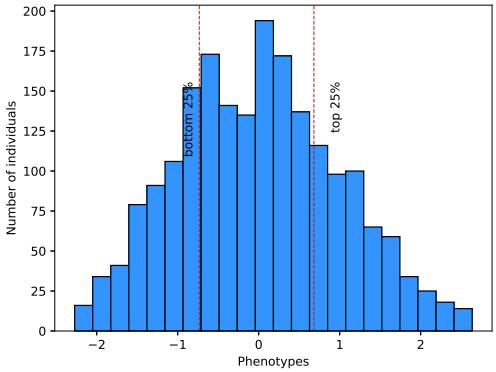
We labeled the bins with discrete values (1,2 and 3) and then applied chi-square feature selection [Saeys et al., 2007] to rank the markers where higher values of chi-square indicate a more prominent feature. We retained markers that have a chi-square p-value of less than 0.005.

3.2.3 Marker Ensemble

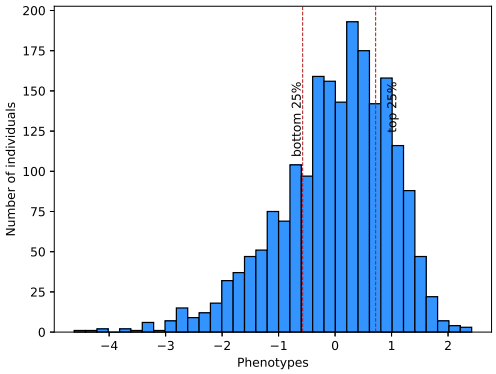
Our genomic selection is an ensemble model that trains N independent machine learning models on subsets of markers. From the training set of the data, we created N subsets of markers where each subset contains m markers. To ensure that all



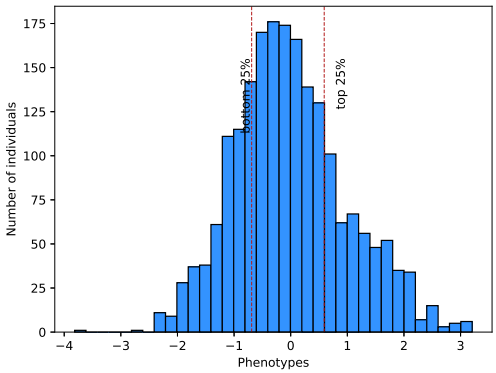
(a) Distribution of grain length.



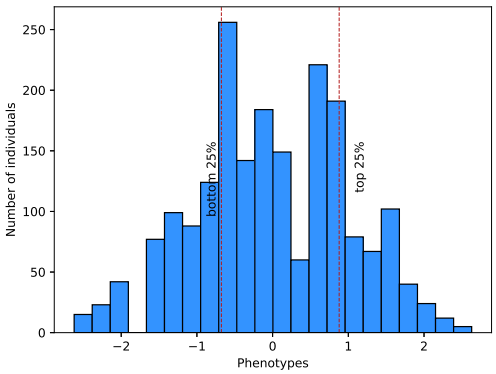
(b) Distribution of thousand kernel weight.



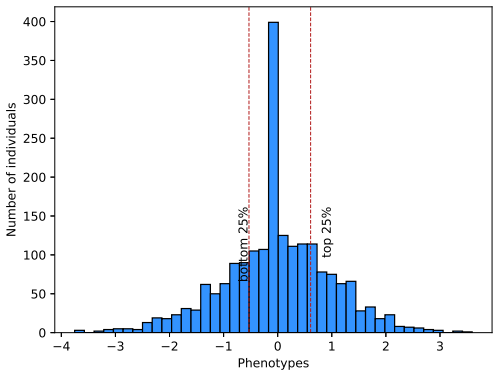
(c) Distribution of test weight.



(d) Distribution of grain width.



(e) Distribution of grain hardness.



(f) Distribution of plant height.

Figure 3.1: Distribution of phenotypes of different traits of Iranian Wheat.

the markers are at least selected once, we selected m markers per subset without replacement. This process was continued until all the markers are selected once. If all the markers are selected before obtaining N subsets, we restarted the process again from the beginning and continue until we get N subsets. Each of the N selected subsets of features is used to train a machine learning model described in section 3.2.4 and 3.2.5.

3.2.4 Deep Learning Model

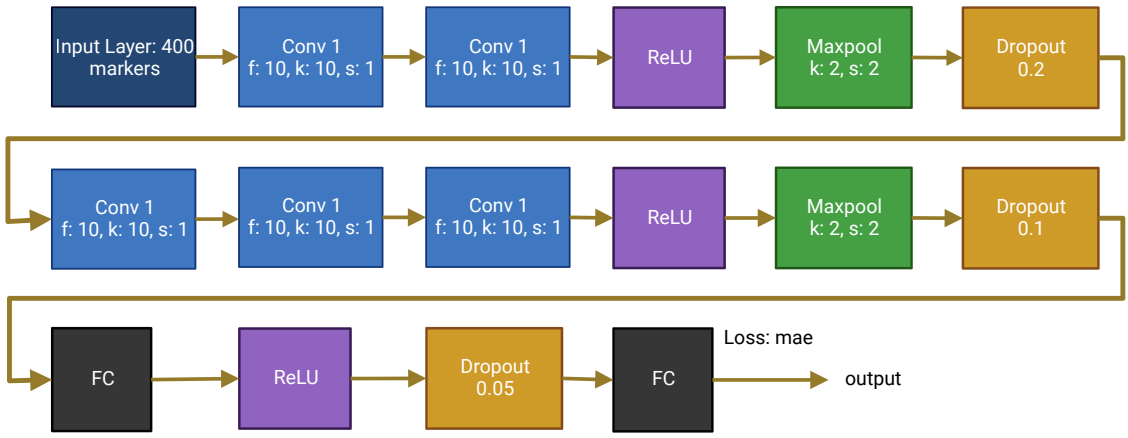


Figure 3.2: Architecture of one convolution neural network. In the layers, k indicates the kernel size, f refers to the filter size and s is the stride.

Figure 3.2 shows the architecture of our convolution neural network (CNN). Our CNN model consists of one input layer, five convolution 1D layers, three ReLU layers, two max-pooling layers, three dropout layers, two fully connected layers, and one output layer. The input layer takes an input of M neurons and passes the input neurons to blocks of convolution layers. Each convolution layer has ten filters with kernel size ten and a stride of one. There are two groups of convolution layers. In the first group, there are two convolution layers, and in the second group, there are

three convolution layers. Each convolution layer group is followed by a ReLU layer, a max-pool layer where both the kernel size and stride are 2 and a dropout layer. There are two fully connected layers where a ReLU and a dropout layer follow the first layer. The last fully connected layer is the output layer. We use Adam as the optimizer and Mean Absolute Error (mae) as the loss function. The batch size and learning rate are 128 and 0.001, respectively. The CNN is implemented using the Keras package in python [Chollet et al., 2018].

3.2.5 Support Vector Regression

Support vector regression (SVR) [Drucker et al., 1997] maps the data from one vector space to another vector space to find out better separability for prediction. We employ Support Vector Regression (SVR) as an alternative to CNNs for both individual and ensemble prediction. In our SVR model, we used the radial basis function (RBF) kernel as RBF can make a non-linear prediction and often performs better than other kernel functions. Two parameters can be optimized for the RBF kernel: cost and gamma. For the non-ensemble model, the cost is optimized in the range of $\{1, 2, 4, 16\}$. The gamma parameter is optimized from 2^{-4} to 2^4 in powers of 2. For the ensemble model, we use the default parameter values. Scikit-learn [Pedregosa et al., 2011] is used to implement SVR.

3.2.6 Overall Architecture

Figure 3.3 shows the overall architecture of our framework for ensemble CNN and ensemble SVR. The genotyped and phenotyped data are first binned based on their

phenotypic value, and then chi-square feature selection [Jović et al., 2015] [Saeys et al., 2007] is applied to find the most discriminating features, which are chosen as those that have chi-square p-value less than 0.005. After that, an ensemble of N subsets of features are created, and a machine learning model, either CNN or SVR, is independently trained on each subset. The final output is the average of all the predicted outputs of each of the models.

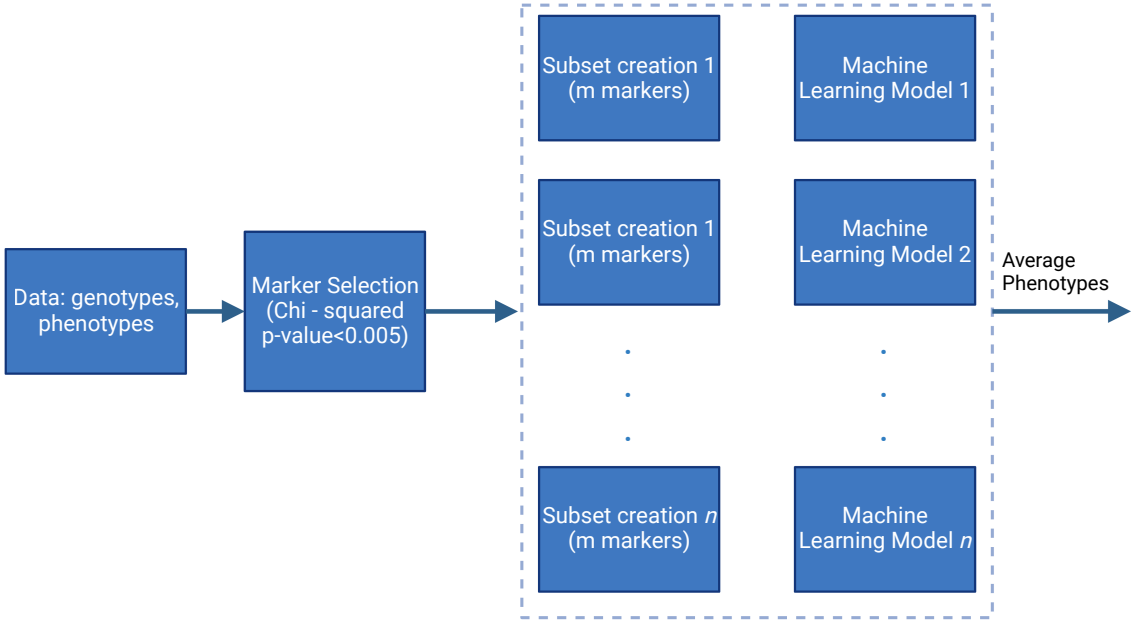


Figure 3.3: Full architecture of the framework.

3.2.7 Random Forest

Random forest (RF) [Breiman, 2001] is an ensemble machine learning method that uses a large number of individual decision trees. Each of the decision trees predicts the phenotypes separately, and the final output is the average of all the prediction of the decision trees. To make each tree different from the others, RF uses bagging,

and the markers of each tree in a random forest are picked from a subset of random markers. We optimize two parameters for RF: i) the number of trees in a forest from 50 to 1000 with an increase of 25 at each iteration and ii) the number of features to consider when looking for the best split from 10 to 200 with an increase of 10 at each iteration. We applied the grid search to optimize these two parameters. Scikit-learn [Pedregosa et al., 2011] is used to implement RF.

3.3 Results

For measuring the performance of our models, we used stratified 5-fold cross-validation. This means that the data is divided into five subsets without overlapping and the machine learning algorithms are trained with four subsets and evaluated with one subset. The training and testing were done five times, each time taking a different test-set. In each fold, 20% of the data from each bin are kept as the test-set and the rest of them are kept for training-set.

As the main objective of GS is to identify individuals that will harvest better phenotypes, it is more beneficial to obtain a linear relation between original phenotypes and predicted phenotypes than to predict the phenotypes accurately. If the relationship between the original phenotype and the predicted phenotype is linear or the order of predicted phenotypes and original phenotypes are the same, this means that the machine learning model performs well. We use two performance measures, i) Pearson correlation coefficient (PCC), and ii) normalized discounted cumulative gain at k (nDCG@k) [Al-Maskari et al., 2007] that either consider the linearity or the orders of the predicted phenotypes.

PCC measures the linear relation between the predicted phenotypes and the original phenotypes. Equation 3.1 shows the formula for calculating PCC. In this paper, x is the original phenotypes and y is the predicted phenotypes. If the original phenotypes and predicted phenotypes are perfectly linear, the PCC is 1. If the relationship is the opposite, the PCC is -1.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.1}$$

As the new cultivars are formed using the top individuals of an existing variety, 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. Equation 3.2 shows formula for calculating nDCG@k.

$$nDCG@k = \frac{DCG@k}{IDCG@k} \tag{3.2}$$

In the equation 3.2, $DCG@k$ means the discounted cumulative gain [Al-Maskari et al., 2007] for the top k individuals and $IDCG@k$ is the ideal DCG for the top k individuals. DCG measures the graded relevance, and in GS, the relevance is the ranking of predicted phenotypes compared to the original phenotypes. In $IDCG$, the relevance is the ranking of original phenotypes. If the ranking is perfect, the nDCG@k is 1 and if the ranking is exactly opposite of the expected ranking, nDCG@k is 0. nDCG was previously used by Ma et al. [Ma et al., 2018] for evaluating GS.

3.3.1 Marker Ensemble

After using chi-square feature selection for each trait, we obtained a reduced set of important markers. Table 3.1 shows the number of markers in each trait that have p-value for chi-square ≤ 0.005. Gw has the highest number of markers, and pht has the lowest. A higher value of chi-square indicates better separability and 0 means the marker does not have any effect to predict the phenotypes. Though we use the p-value for chi-square ≤ 0.005, from our result, we observe that there are very few markers that have zero chi-square value.

Table 3.1: Number of markers in each trait after using chi squared feature selection.

Traits	**# of selected markers.**
Grain Length	6,532
Grain Width	11,175
Grain Hardness	8,028
Thousand-Kernel Weight	6,958
Plant Height	1,023
Test Weight	8,887

Table 3.2 shows the percentage of common markers between different traits that are selected after feature selection. Most of the traits have $\approx 50\%$ common markers between them except pht. Pht has on average 14.598% markers in common with other phenotypic traits. One reason behind this is that pht has only 1023 markers that have chi-square p-value ≤ 0.005, which is approximately six times fewer markers than the second smallest trait.

Table 3.2: Percentage of common markers between two traits. The percentage is based on the average number of markers between two traits.

	Grain Length	**Thousand-Kernel Weight**	**Test Weight**	**Grain Width**	**Grain Hardness**	**Plant Height**
Grain Length	100	57.26	54.18	50.75	48.28	17.07
Thousand-Kernel Weight		100	49.21	65.05	56.16	13.74
Test Weight			100	63.74	57.36	15.71
Grain Width				100	69.28	12.65
Grain Hardness					100	13.82
Plant Height						100

3.3.2 Ensemble model vs single model

The single model of SVR and deep CNN are trained with a set that includes all the markers that are selected with feature selection. For the ensemble model, through experiments (results not shown), we selected an ensemble of size $N = 75$ with each machine learning model (SVR and deep CNN) considering $M = 400$ markers. The architecture of the deep CNN is the same for the single model except the input layer takes $11,176$ markers as the input. We chose $11,176$ markers as the input because it covers all the markers that have p-value for chi-square less than 0.005 for all the traits. Table 3.3 shows the comparison between the ensemble model and the single machine learning model. From the table, we observed that the ensemble model of SVR performs better than the single model of SVR. This means that the single model

is affected by a large number of markers and creates a "large p, small n problem". In the ensemble model of SVR, each model is trained with a small subset of markers from the marker set which solves the "large p, small n" problem. The final output is the average of all the predicted values. The PCC of both deep learning models (ensemble and single) are almost similar though in most of the traits, the ensemble models have slightly higher PCC than the single model.

Table 3.3: Comparison of PCC between actual and predicted traits, for both single model and ensemble model. Bold indicates the best performance obtained for that specific trait.

Traits	**Deep CNN**		**SVR**	
	Ensemble	**Single**	**Ensemble**	**Single**
Grain Length	**0.738**	0.728	0.732	0.488
Thousand-Kernel Weight	**0.663**	0.661	0.660	0.481
Test Weight	**0.618**	0.614	**0.618**	0.266
Grain Width	0.724	**0.731**	0.724	0.311
Grain Hardness	0.648	**0.661**	0.660	0.422
Plant Height	0.339	0.323	**0.379**	0.110

Figure 3.4 shows the comparison of nDCG@20 for deep CNN model, ensemble deep CNN and ensemble SVR model. In this figure, we did not consider the single SVR model as all the other models have very high PCC compared to the single SVR model. Single deep CNN has obtained the highest value of nDCG@20 for three traits. The ensemble SVR outperformed the other two models twice, while the ensemble deep CNN outperformed single CNN three times.

Both the performance measures we applied show that the single deep learning model and both the ensemble models provide similar results with minimal improvement in performance among different models. The difference in performance between the single SVR model and the other models are very high due to training with a large

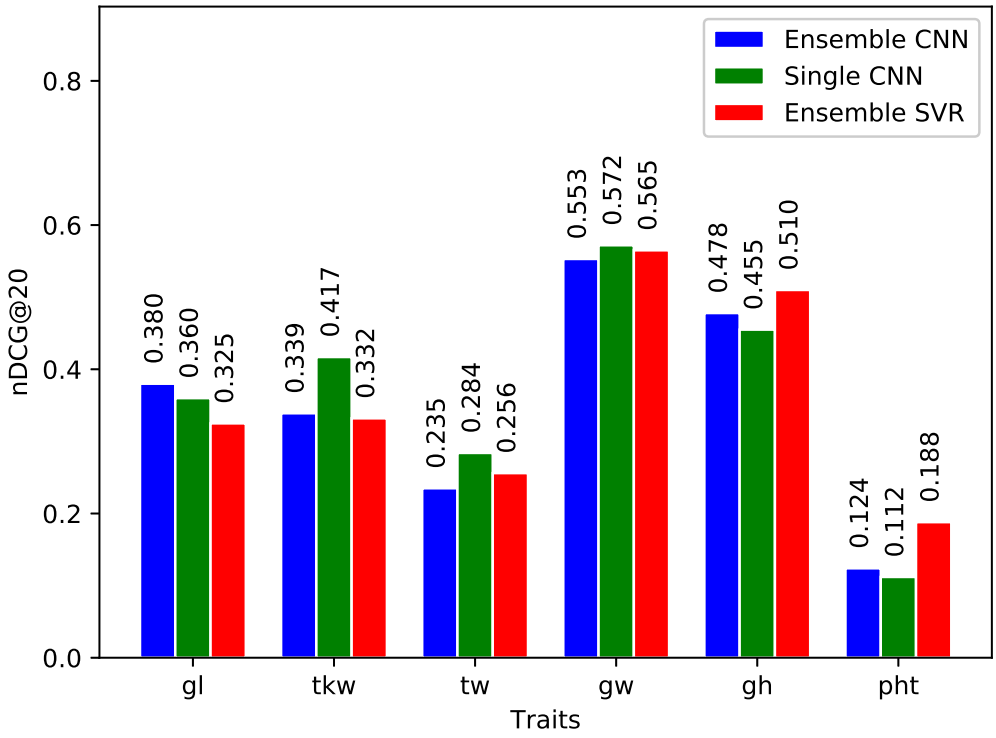


Figure 3.4: Comparison of nDCG@20 for ensemble and single models.

number of markers. Though the hyper-parameters of the ensemble models are not optimized, they can produce better or similar performance than the single model. Hyper-parameter optimization of the ensemble models can further improve the performance of the ensemble models.

3.3.3 Ensemble Models

We employed three ensemble models: i) ensemble deep CNN, ii) ensemble SVR and iii) RF. Each of these models creates ensembles with subsets of features. Table 3.4 shows the comparison of PCC among these models. From the table, we observe that

RF has better PCC in four traits, though the increase in PCC from both ensemble deep CNN and ensemble SVR is small. In two traits, ensemble SVR is better than the other two models.

Table 3.4: Comparison of PCC between actual and predicted traits. Bold indicates the best performance obtained for that specific trait.

Traits	Ensemble DL	RF	Ensemble SVR
Grain Length	0.738	**0.747**	0.732
Thousand-Kernel Weight	0.663	**0.672**	0.660
Test Weight	0.618	**0.624**	0.618
Grain Width	0.724	**0.738**	0.724
Grain Hardness	0.648	0.653	**0.660**
Plant Height	0.339	0.352	**0.379**

Though the PCC of the models on the pht trait is lower compared to other traits, all the traits have good PCC for all ensemble models. We compare the result of our best model with the performance of RR-BLUP that was reported in DeepGS [Ma et al., 2018]. We observe from Table 3.5 that we obtained an improvement of 1.013 times to 1.021 times than the RR-BLUP for all the traits except gh. Though the performance of Ensemble SVR on the gh trait is better in RR-BLUP, PCC of 0.660 is a good score.

Table 3.5: Comparison of PCC between RR-BLUP and the best model for each trait.

Traits	RR-BLUP	Best Ensemble Model	Improvement
Grain Length	0.735	0.747	1.016
Thousand-Kernel Weight	0.658	0.672	1.021
Test Weight	0.614	0.624	1.016
Grain Width	0.728	0.738	1.013
Grain Hardness	0.685	0.660	0.963
Plant Height	0.327	0.379	1.15

As it is essential to measure the accuracy of the ranking of individuals to build a better cultivar, we also measure the nDCG@20 for each trait. Table 3.6 shows the comparison of nDCG@20 for top individuals among different models. Though PCC is better in four traits with RF, only gw with RF has better nDCG@20 than PCC. As we are only considering the top 20, the scores are satisfactory for all the traits except pht. In addition, we calculated the p-value between three pairs of models and observed p-value > 0.9 for all pairs, indicating no statistically significant difference in these models.

Table 3.6: Comparison of nDCG@20 for the top individuals. Bold indicates the best performance obtained for that specific trait.

Traits	**Ensemble CNN**	**RF**	**Ensemble SVR**
Grain Length	**0.380**	0.355	0.325
Thousand-Kernel Weight	**0.339**	0.325	0.332
Test Weight	0.235	0.230	**0.256**
Grain Width	0.553	**0.612**	0.565
Grain Hardness	0.478	0.473	**0.510**
Plant Height	0.124	0.125	**0.188**

To find the robustness of our ensemble models, we also measure the nDCG@20 for bottom individuals. Table 3.7 shows the result. From the table, we observe that all the traits achieve nDCG@20 ≥ 0.3 except pht and tw.

When we perform feature selection, pht has very few features compared to other traits that can separate the top and bottom individuals. Figure 3.1f shows that for the pht trait, the dataset contains a large number of individuals in a specific range of phenotypic values and in other ranges, the number of individuals is very low. This distribution of phenotypes may cause machine learning models to overfit. Both PCC and nDCG@20 showed that pht is the hardest trait to predict in this

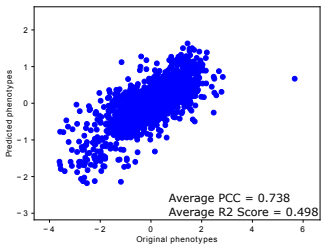
Table 3.7: Comparison of nDCG@20 for the bottom individuals. Bold indicates the best performance obtained for that specific trait.

Traits	Ensemble CNN	RF	Ensemble SVR
Grain Length	0.668	**0.712**	0.577
Thousand-Kernel Weight	0.314	0.319	**0.335**
Test Weight	0.236	0.276	**0.287**
Grain Width	0.370	**0.395**	0.323
Grain Hardness	**0.325**	0.290	0.265
Plant Height	0.198	0.233	**0.235**

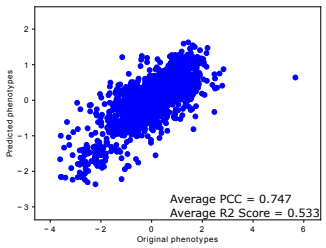
dataset. As in complex crops like wheat, a lot of markers contribute cumulatively to the final expression of phenotypes, the machine learning models do not have sufficient information from the markers from which it can predict the pht.

The distribution of phenotypes plays a role in selecting fewer markers and obtaining poor performance. Figure 3.5 shows the original vs. predicted phenotypes plot for gl, tw and pht. From the figure, we observe that the predicted phenotypes of gl are almost linear to the original phenotype; hence, it has high PCC and nDCG@20. The predicted phenotypes of trait tw are also close to linear to the original phenotypes though the nDCG@20 is low. The reason behind is there are many individuals with phenotypes that have a very close value between -2 to 1, making it difficult for the machine learning model to maintain the ranking while predicting, despite having a very high linear relationship with the original value. Thus we can consider that our model worked well for predicting tw. The predicted phenotypes of pht have a very little linear relation with the original phenotypes and thus results in low PCC. Gl has the best PCC and pht is the worst one we obtained with our models. The comparison of performance measure shows that there is no single method that outperforms the other models and there is no significant difference in performance with

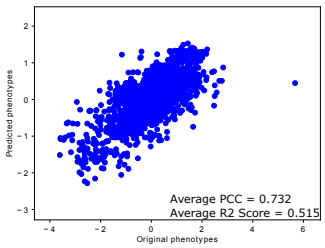
different ensemble models. Though RF obtained better PCC than others, RF models are optimized while other ensemble models are not.



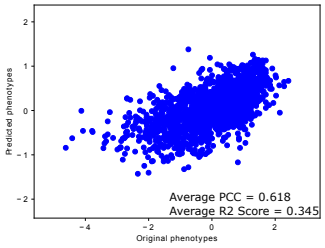
(a) Original vs predicted phenotypes of grain length with ensemble CNN.



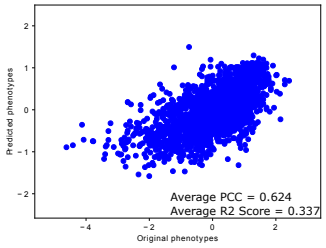
(b) Original vs predicted phenotypes of grain length with RF.



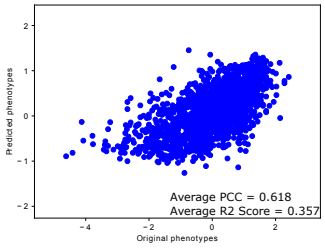
(c) Original vs predicted phenotypes of grain length with ensemble SVR.



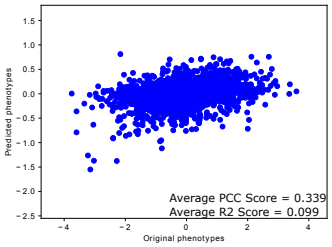
(d) Original vs predicted phenotypes of test weight with ensemble CNN.



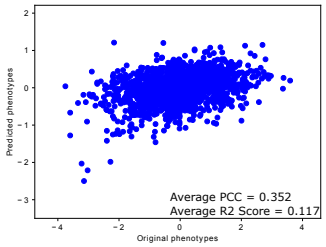
(e) Original vs predicted phenotypes of test weight with RF.



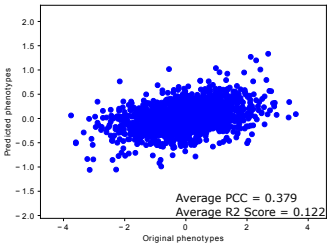
(f) Original vs predicted phenotypes of test weight with ensemble SVR.



(g) Original vs predicted phenotypes of plant height with ensemble CNN.



(h) Original vs predicted phenotypes of plant height with RF.



(i) Original vs predicted phenotypes of plant height with ensemble SVR.

Figure 3.5: Original vs. predicted phenotypes for length, test weight, and plant height.

The marker selection method of Ma et al. [Ma et al., 2018] is random. Thus the

performance of the model can vary when the subset of markers differs. In our model, we have a defined marker selection technique that gives the same subset of markers. Ma et al. reported a PCC of 0.742 for the DeepGS model for grain length only, where the model hyperparameters are appropriately tuned. In the random forest and ensemble deep learning model, we obtained PCC of 0.747 and 0.738, respectively, where the parameters of our ensemble deep learning model are not optimized. From this, we can observe that our proposed ensemble models are competitive with DeepGS, and that optimization of ensemble methods is a promising area to improve the results of the deep learning models.

3.4 Conclusion

In this paper, we proposed three ensemble learning methods: i) ensemble SVR ii) ensemble CNN and iii) random forest for GS in wheat. We also proposed a binning approach by applying chi-square feature selection to the identification of essential markers for a specific trait. We showed that the ensemble models on a wheat dataset are competitive with both DeepGS [Ma et al., 2018] and single models. The performance of different ensemble models are very similar to each other; thus, there is no definitive answer to which model is the best. In deep learning, the percentage of dropout neurons and the number of neurons in the fully connected layer plays a crucial part in tuning the model for better prediction. Cost and gamma are the two hyper-parameters that plays a similar role in SVR. In single models, we observe that the optimization of these parameters improves the accuracy of prediction dramatically. Thus, in the future, we will explore how the optimization of the hyper-parameters

influences the performance of ensemble models. Currently, our models predict a single trait. Deep learners are known for their ability to predict multiple outputs at the same time. We will investigate if the ensemble deep learning model can predict multiple traits with the same or better accuracy. We will also integrate environmental information with our models so that the models can predict phenotypic differences based on the environment.

Conflict of Interest Statement

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.

Funding

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

Chapter 4

GPTransformer: A Transformer-based deep learning method for predicting Fusarium related traits in Barley

Sheikh Jubair[1], James R. Tucker[2,3], Nathan Henderson[3], Colin W. Hiebert[4], Ana Badea[2], Michael Domaratzki[5], W. G. Dilantha Fernando[2]

1 – Department of Computer Science, University of Manitoba, 66 Chancellors Cir, Winnipeg, MB R3T 2N2

2 – University of Manitoba, Department of Plant Science, Winnipeg , Manitoba, Canada

3 – Brandon Research and Development Centre, Agriculture and Agri-Food Canada, Brandon, Manitoba, Canada

4 – Morden Research and Development Centre, Agriculture and Agri-Food Canada, Morden, Manitoba, Canada

5 – Department of Computer Science, University of Western Ontario, 1151 Richmond St, London, ON N6A 3K7

This chapter is an adaptation of the article Jubair et al. [2021b] published in Frontiers in Plant Science in 2021.

Author Contributions

SJ - Conceptualization, Methodology, Developing deep learning methods, Software and data analysis, Writing - original draft preparation, Writing - review and editing; JT - Conceptualization, Methodology, Micro-array genotyping, Software and data analysis, Writing - original draft preparation, Writing - review and editing; NH - Software and data analysis, Writing - review and editing; CH - Micro-array genotyping; AB - Conceptualization, Methodology, Writing-review and editing, Supervision, Project administration and funding acquisition; MD - Conceptualization, Writing - review and editing, Supervision; DF - Conceptualization, Methodology, Supervision,

Project administration and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Abstract

Fusarium head blight (FHB) incited by *Fusarium graminearum* Schwabe is a devastating disease of barley and other cereal crops worldwide. Fusarium head blight is associated with trichothecene mycotoxins such as deoxynivalenol (DON), which contaminates grains, making them unfit for malting or animal feed industries. While genetically resistant cultivars offer the best economic and environmentally responsible means to mitigate disease, parent lines with adequate resistance are limited in barley. Resistance breeding based upon quantitative genetic gains has been slow to date, due to intensive labour requirements of disease nurseries. The production of a high-throughput genome-wide molecular marker assembly for barley permits use in development of genomic prediction models for traits of economic importance to this crop. A diverse panel consisting of 400 two-row spring barley lines was assembled to focus on Canadian barley breeding programs. The panel was evaluated for FHB and DON content in three environments and over two years. Moreover, it was genotyped using an Illumina Infinium High-Throughput Screening (HTS) iSelect custom beadchip array of single nucleotide polymorphic molecular markers (50K SNP), where over 23K molecular markers were polymorphic. Genomic prediction has been demonstrated to successfully reduce FHB and DON content in cereals using various statistical models. Herein, we have studied an alternative method based on machine learning and compare it with a statistical approach. The bi-allelic SNPs represented pairs of alleles and were encoded in two ways: as categorical (-1, 0, 1) or using Hardy-

Weinberg probability frequencies. This was followed by selecting essential genomic markers for phenotype prediction. Subsequently, a Transformer-based deep learning algorithm was applied to predict FHB and DON. Apart from the Transformer method, a Residual Fully Connected Neural Network (RFCNN) was also applied. Pearson correlation coefficients were calculated to compare true vs. predicted outputs. Models which included all markers generally showed marginal improvement in prediction. Hardy-Weinberg encoding generally improved correlation for FHB (6.9%) and DON (9.6%) for the Transformer network. This study suggests the potential of the Transformer based method as an alternative to the popular BLUP model for genomic prediction of complex traits such as FHB or DON, having performed equally or better than existing machine learning and statistical methods.

Keywords– genomic prediction, deep learning, transformer, feature selection, quantitative traits, barley, fusarium head blight, deoxynivalenol

4.1 Introduction

Barley (*Hordeum vulgare* L.) is one of the most ancient grains, and is currently the fourth-most produced cereal globally measured both in area harvested and yield [FAO, 2019]. Barley is primarily grown as animal fodder, or used by the malting and brewing industries. As a cash crop, malting barley necessitates maximized yield performance, and requires strict management of numerous grain-quality characteristics with specific parameter ranges [Izydorczyk and Edney, 2017]. Barley achieving these superior standards can be sold into the lucrative malting barley market, where it returns a significant premium to the barley producer. Fusarium head blight (FHB), caused by

Fusarium graminearum Schwabe [teleomorph: *Gibberella zeae* (Schwein.) Petch], is a devastating disease of barley. The Primary concern of the disease is due to associated trichothecene mycotoxins such as deoxynivalenol (DON), which are potent inhibitors of protein synthesis [Pestka, 2010]. Due to potential adverse toxic effects, DON along with its alternative forms are highly regulated with maximum consumption limits set for humans and livestock (EFSA CONTAM Panel, 2017).

Breeding FHB resistant cultivars is a sustainable disease management solution, which has been achieved mainly through large disease nurseries. Several studies have demonstrated a positive association between visual symptoms of FHB infection and DON content in matured grains. However, this correlation is often moderate at best [Buerstmayr et al., 2004; Choo et al., 2004; Buerstmayr et al., 2004; He et al., 2015; Huang et al., 2018; Tucker et al., 2019]. Mycotoxin quantification is highly technical, where sampling protocols, quality controls and choice of analytical technologies are all implicated as important factors [Tittlemier et al., 2021]. Analytical chemistries are expensive and labour requirements for harvest and processing grains are substantial.

FHB and DON content resistances are both under quantitative genetic control in barley (affected by many genes, each with a small effect). Significant undertakings have been made in genetic studies of biparental populations to identify quantitative trait loci (QTL) for breeding resistant barley cultivars [Fernando et al., 2021]. While QTLs have been identified for FHB and DON, they are limited by the minimal level of genetic variance they explain, environmental specificity, and common association with negative agronomics such as extreme heading date and tall stature. Incorporating major QTLs from moderately resistant source 'Chevron' such as Qrgz-2H-8

into elite backgrounds, did not result in sufficient resistance levels [Linkmeyer et al., 2013]. Some commercial success has been achieved in developing moderately resistant cultivars such as six-row, malting barley 'Quest', through pyramiding of multiple resistances [Smith et al., 2013]. Association mapping was able to identify QTLs independent of negative agronomic traits, however these were small, only explaining 1-3% of the observed variance [Massman et al., 2011].

Cereal crops (Poaceae family, Triticeae tribe) are characterized by their large genomes, with frequent repetitive elements [Mascher et al., 2017]. With plummeting cost of genomic tools and availability of highly improved reference genomes, modern breeding approaches are now possible, which take advantage of genome-wide marker capabilities for the use in predicting complex traits [Jannink et al., 2010]. In the face of this challenge, genomic prediction of FHB has been possible using statistically based methodologies in hexaploid (bread) wheat [Rutkoski et al., 2012; Arruda et al., 2015; Jiang et al., 2015; Mirdita et al., 2015; Hoffstetter et al., 2016; Dong et al., 2018]; durum (pasta) wheat [Steiner et al., 2019; Moreno-Amores et al., 2020] and six-row barley [Sallam and Smith, 2016; Abed et al., 2018]. While cereal genomes are complex, initial results of genomic prediction for FHB and DON content are very promising, and demand further investigation.

Traditional statistical algorithms such as Best Linear Unbiased Prediction (BLUP) and variants [Burgueño et al., 2012; Cuevas et al., 2017; Ferrão et al., 2017; Cuevas et al., 2019; Howard et al., 2019] have been applied in many genomic prediction problems. These models are mostly linear in nature and perform well for additive traits. Machine learning methods have been applied in genomic prediction with mod-

erate success [Heslot et al., 2012; Ogutu et al., 2011b; Poland et al., 2012; González-Camacho et al., 2018]. The machine learning methods claim to capture non-additive effects better than the statistical methods [Heslot et al., 2012]. Deep learning is a subset of machine learning that is gaining popularity for genomic prediction [Rachmatia et al., 2017a; Ma et al., 2018; Jubair and Domaratzki, 2019; Khaki and Wang, 2019]. Deep learning differs from traditional machine learning by applying multiple networks along with non-linear functions that often imitate how the human brain learns and identifies patterns based on the learned representations. Under the training phase of genomic prediction, these deep learning algorithms take inputs of genotype data of different cultivars, and their corresponding phenotypes, to learn the parameters of the model. During the testing phase, only the genotype data of other cultivars is used as input and the trained model predicts the corresponding phenotypes of the test data. These deep learning methods have performed equally or better than existing statistical methods [Ma et al., 2018; Jubair and Domaratzki, 2019; Khaki and Wang, 2019]. For an overview of deep learning algorithms and their application in genomic prediction, we refer the readers to a recent review [Montesinos-López et al., 2021].

Neural networks such as feed-forward neural networks [Rachmatia et al., 2017a; Khaki and Wang, 2019] and Convolutional Neural Networks (CNNs) [Ma et al., 2018; Jubair and Domaratzki, 2019] have been applied in genomic prediction. The feed-forward network can be compared to n linear regressions where these n linear regressions are the hidden neurons of the feed-forward network. The output neurons of the CNN also represents multiple linear regression models where the linear combination is produced from a very small subset of markers. CNN uses a sliding window allowing

it to slide through the whole input space. Both feed-forward network and CNN do not reflect the polygenic interactive effects of markers as the relationship between markers are not considered in these algorithms.

Transformers are a family of deep learning algorithms that have been initially applied to Natural Language Processing (NLP) tasks such as classification, next sentence prediction and topic identification [Devlin et al., 2018; Raffel et al., 2019; Brown et al., 2020]. Historically these methods perform well when trained on a large amount of data [Devlin et al., 2018; Raffel et al., 2019; Brown et al., 2020] and can be used for transfer learning. Apart from NLP, Transformer architecture has been successfully applied to other fields such as image processing [Bazi et al., 2021; Dosovitskiy et al., 2020]. In this work, we proposed a Transformer-based genomic prediction model for predicting FHB and DON for barley. The Transformer consists of three main components: the self-attention, feed-forward networks, and layer normalization. The self-attention mechanism calculates the attention score for all genetic markers concerning a specific genetic marker [Vaswani et al., 2017] which helps to find the relation among markers. The layer normalization function converts each input marker to zero mean and unit variance. The Transformer network mainly identifies the inter-relation among markers.

Hardy-Weinberg equilibrium is a principle that states the allele frequency of a population will remain constant from generation to generation in the absence of disturbing factors [Acquaah, 2009]. Under random mating, a population can obtain the equilibrium even after a single generation if there are no selection pressures. The principle also applies for the marker frequency and provides additional information

about the population alongside genotype data [Acquaah, 2009]. In this paper, we apply Hardy-Weinberg equilibrium values as an input encoding for markers.

Hundreds of genes may contribute to a phenotype, such that identifying the top contributing genes and related markers is a challenging task. Feature selection algorithms identify essential features for a specific task [Saeys et al., 2007; Tang et al., 2014] and have been successfully applied in many classification problems of bioinformatics [Saeys et al., 2007]. Mutual information is a filter based feature selection algorithm that identifies top features based on a set of classes and features. Top features identified using mutual information may represent targets which bear biological value.

In this work, we evaluate a Transformer-based deep learning method, GPTransformer, that uses genotypic and phenotypic data to predict FHB severity and DON levels in a two-row barley population. Our specific objectives were to (i) compare the accuracy of the GPTransformer model to existing genomic prediction methods, (ii) study the outcomes of the model if categorical encoding was used or marker frequency-based encoding was used, and (iii) investigate the effect of feature selection on genomic prediction using mutual information and examine the biological relevance of the top markers identified by the mutual information method. The Transformer network is trained using a graphical processing unit (GPU). As the internal mechanism of the Transformer creates a four dimensional matrix of size batch size $\times$ number of heads$\times$ number of markers^2 at a certain point, which requires a large amount of GPU memory, the feature selection process also helps us to solve the GPU memory issue of the Transformer.

4.2 Methodology

4.2.1 Genotyping

The seed for a genetic panel was collected for a total of 400 spring habit two-row barley genotypes of mixed usage types of malt (171) and feed (229). Pure seed was provided by the Crop Development Centre, University of Saskatchewan, Saskatoon, Canada (CDC) for a diversity panel of barley (92) breeding lines tested in the Western Canadian Cooperative Two-Row Barley Registration Trials (WCTBRT) 1994-2006 [Beattie et al., 2010]. Additional elite lines (176) were selected from 2001-13 WCTBRT based on past performance, with the majority from three breeding programs: CDC; Field Crop Development Centre, Olds College; Agriculture and Agri-Food Canada (AAFC), Brandon Research and Development Centre. Moreover, breeding lines (105) targeting FHB resistance and involving crosses to exotic sources, were also selected from these programs. The remainder of lines were represented by American and exotic germplasm.

Two seeds were germinated on moist cotton balls for a week. At the two-leaf stage [Zadoks et al., 1974], leaves were cut from a single plant and flash-frozen in liquid nitrogen, then freeze-dried in a lyophilizer (Labconco Corporation, Kansas City, MO, USA). Genomic DNA was extracted from 100 mg of tissue using Qiagen, DNeasy 96 Plant Kit (Qiagen, Canada). The isolated DNA was evaluated by a NanodropTM 1000 spectrophotometer (Thermo Fisher Scientific Inc., Wilmington, DE, USA) for quality and concentration, then normalization to 50 ul mL^{-1}. Samples were assayed on an Illumina iScan (Illumina, San Diego, CA, USA) using a custom iSelect - 50K

SNP microarray [Bayer et al., 2017] at AAFC, Morden Research and Development Centre, Morden, MB. A custom cluster file provided by M. Ganal (TraitGenetics GmbH, Germany) was used to call SNP alleles using Illumina GemomeStudio V2.0.5 software (Please see data availability statement). Data were filtered for $\geq 5\%$ minor frequency alleles and $\leq 20\%$ missing data.

4.2.2 Field Studies

FHB nurseries were grown in 2014 and 2015 at 3 locations: Brandon, Manitoba (49°51'56.0"N 99°58'57.7"W); Carman, Manitoba (49°29'52.9"N 98°02'19.2"W) and Carberry, Manitoba (49°54'16.6"N 99°21'19.0"W). The experiments followed a randomized complete block design at all locations (n = 3) with two replications per site. Plots were sown with approximately 30-40 seeds and consisted of 0.9 m rows, 30 cm row spacing (Brandon, Carberry) or 1 m rows, 34 cm row spacing (Carman). Two inoculation methods were used. Brandon and Carberry experiments were inoculated by the grain spawn method, where maize kernels infected with 2 isolates each of 3ADON- and 15ADON-producing strains of *F. graminearum* were spread on the soil surface at $5\,\mathrm{g\,m^{-2}}$ at flag leaf then weekly for 3 total applications. Irrigation was applied after first inoculum application until all plots were rated. Experiments at Carman were sprayed with a macroconidia suspension of 3ADON and 15ADON isolates in equal proportions and standardized to 5×10^4 spores $\mathrm{ml^{-1}}$. Plots were misted and sprayed at 75% spike emergence and then again two days following.

Plots were rated at the soft dough (Zadoks - Z85) stage. A visual scale (0-5) was used to evaluate a composite measure of incidence and severity (A. Tekauz, personal

communication), where 0 = no infection. 1 = incidence low, up to 5% of spikes; severity low, 1 or 2 kernels per spike affected (up to 7% of head). 2 = incidence low to moderate, 5 to 15% of spikes infected; severity low to moderate, 1 to 4 kernels (up to 15% of head). 3 = incidence moderate, 15 to 30% of heads; severity moderate, 2 to 8 kernels (up to 25% of head). 4 = incidence moderate to high, 30 to 50% of spikes infected; severity moderate to high, 4 to 12 kernels (up to 40% of head). 5 = incidence high, 50% or more spikes affected; severity high, 5 to 15+ kernels (up to 50%+ of head diseased). Additional data were collected on days to heading (date 50% of row headed minus seeding date) at Brandon and Carberry and plant height (distance of soil surface to tip of spike excluding awns) at all locations.

Grains were harvested at maturity, using a stationary research combine with low wind speed setting, then dried in a high capacity drier for a few days. Grains were cleaned using an SLN3 sample cleaner (Pfeuffer GmbH, Kitzingen, Germany). A 20 g subsample was removed, cleaned free of debris and/or chaff then ground using a Perten 3610 lab mill with fine particle disc set (PertenElmer Inc. Waltham, MA, USA). Deoxynivalenol content was analyzed by enzyme-linked immunosorbent assay (ELISA) technique using Veratox ® 5/5 (Neogen Corporation, Lansing, MI, USA) as per kit protocol (limit of detection = 0.1 mg kg^{-1}). Samples were each tested in sub-sample pairs, where samples deviating by $> 10\%$ were repeated.

4.2.3 Allele Frequency Based Encoding

As the barley population is diploid, a locus A can have two alleles, A and a and three genotypes AA, Aa and aa. We replaced the classical representation of genotype of

a marker (1, 0, -1) with genotypic frequency by applying Hardy-Weinberg equilibrium [Acquaah, 2009]. The frequency of alleles is calculated for each genetic marker. Suppose, the genotype AA and Aa appear D and H times respectively for a specific genetic marker. If the population size is N, the total number of alleles will be $2N$. Then, the frequency of allele A for a specific genetic marker is calculated by applying equation 4.1.

$$p = \frac{2D + H}{2N} \tag{4.1}$$

The frequency of the allele a is $q = 1 - p$. After calculating the allele frequency of a genetic marker, the expected genotypic frequencies of genotype AA, Aa and aa for a specific marker is obtained from p^2, $2pq$ and q^2. These expected genotypic frequencies are used as marker values for the machine learning algorithms.

4.2.4 Transformer

In this paper, we examine the application of Transformer, a variety of neural networks. Deep neural networks have been applied to GS in the past. See the survey of [Montesinos-López et al., 2021] for details on previous applications. The Transformer is a family of deep learning algorithms successfully applied to various NLP tasks such as classification, sequence to sequence modeling, and next sentence prediction. The Transformer architecture has two major components: i) encoder and ii) decoder [Vaswani et al., 2017]. In this work, we use the encoder part of the Transformer along with an additional feed-forward network 4.1. The encoder of the Transformer architecture contains an embedding layer, a multi-head self-attention layer, and finally a

feed-forward neural network.

The purpose of the embedding layer is to obtain an $n = 8$-dimensional expanded representations of markers. In this work, a feed-forward neural network is applied as the embedding layer. The input of the embedding layer is m markers and the output is an $m \times n$-dimensional vector. The vector is then reshaped to an (m, n)-matrix to obtain m expanded representation of the markers. The output of the embedding layer is then passed to a multi-head attention network.

The multi-head attention network is based on the self-attention mechanism. The input of this layer is the expanded representation of the markers obtained in the embedding layer. The main building block of the multi-head-attention is the self-attention mechanism that calculates the attention score for all other expanded representations of markers with respect to a specific expanded representation. To calculate the self-attention, at first, each embedded marker creates three vectors: a query vector q, a key vector k and a value vector v by applying a linear transformation on the embedding. The query vector is the candidate expanded representation with respect to which the attention is measured while the keys are the set of expanded representations where the importance scores are assigned. The attention score is calculated as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{4.2}$$

In equation 4.2, Q is a matrix of all the queries, K is a matrix of all the keys and V is a matrix of all the values and d_k is the dimension of keys. The last step is to generate a summation of the previous step, which produces the self-attention layer's

output. In a multi-head attention setting, the Transformer model creates h independent linear representation from queries, keys and values. These h representations are then concatenated and passed through a linear projection layer to obtain the final output.

In Figure 4.1, a residual connection from the output of the input embedding layer is added to the output of multi-head attention. A layer-normalization is applied to the output of the residual connection. The Transformer block contains another feed-forward network and a layer-normalization after the feed-forward layer. The N in Figure 4.1 indicates that this Transformer block can be stacked N times and the output of the Nth encoder block will be the input of the feed-forward layer that predicts the phenotypes.

4.2.5 Residual Fully Connected Neural Network

We now describe our second deep learning model which is based on a feed-forward network with residual connection. In general terms, a residual neural network is a neural network that has one or more residual connections. Residual connections allow skipping of layers in a neural network. In our implementation, the first layer of the Residual Fully Connected Neural Network (RFCNN) is a feed-forward neural network that takes M markers as the input and performs a linear transformation to produce an n-dimensional hidden representation of the markers. This hidden representation is the input of the batch-normalization layer. The batch-normalization layer normalizes the current batch by its mean and standard deviation. As the feed-forward network and batch normalization performs a linear operation on the input data, we apply

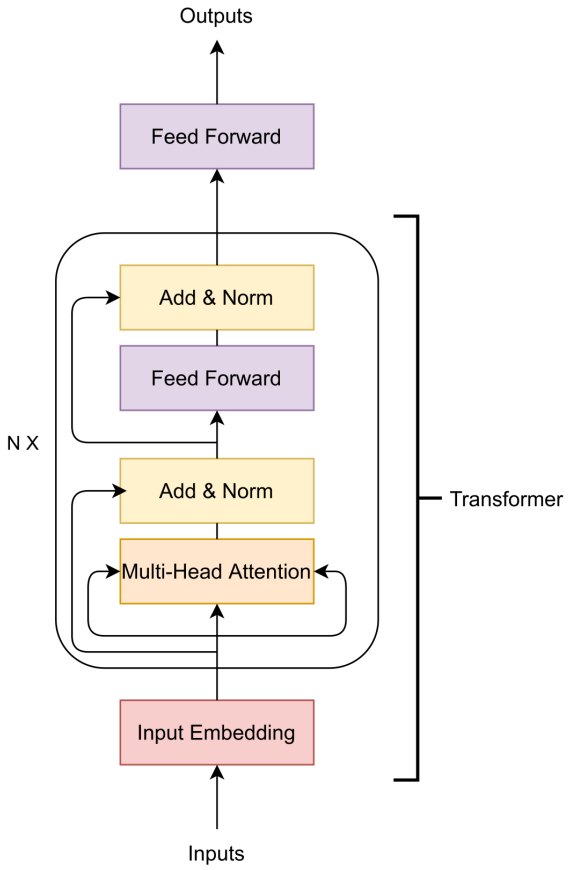


Figure 4.1: Transformer architecture.

the activation function ReLU to the batch-normalization layer's output. ReLU will return 0 if the input is ≤ 0; otherwise, it will return x where x is the input and thus, introduces the non-linearity. The output of ReLU is going to be the input of the next feed-forward network. In figure 4.2, the residual connection is shown as the arrow on the left-side of the figure skipping over the intermediary layers. A residual connection is added from the output of each odd ReLU layer to each odd batch-normalization layer (except the first batch-normalization). The residual block can be stacked N times. The output layer is the feed-forward network that predicts the phenotypes.

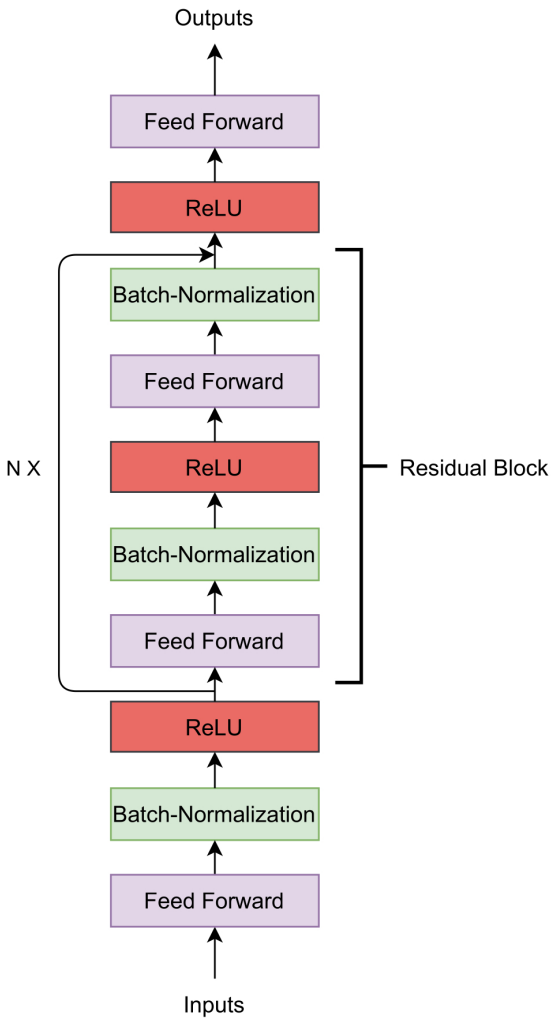


Figure 4.2: Residual Fully Connected Network architecture.

4.2.6 Other Statistical and Machine Learning Models for Baseline Comparison

Decision Tree

Decision Trees (DTs) are common statistical and machine learning methods used for predictive modelling. The baseline DT regressor models used in this analysis were

built with the Scikit-learn [Pedregosa et al., 2011] Python module under default settings. To train the models, an input dataset of an $n \times p$ matrix of encoded genotypes and a n-dimensional vector of known corresponding phenotype responses were supplied. At each node, splits based on the SNP markers are considered and the highest quality split is chosen; in this case, the quality measure of the split is the mean squared error. This process continues recursively, splitting the data into subsets of instances at each internal node, until the branches terminate with leaf node and produce a response value. The values at the leaf nodes are the arithmetic mean of the known associated response variables to the instances that are present in the adjoining edge. With a fitted model, genotypes with an unknown phenotype response can be predicted by iteratively testing the SNP marker values against the trees decision procedure until a leaf node is reached.

Linear Regression

Like the DT models, the linear regression (LR) models used in this analysis were built with the Scikit learn Python module using default settings. LR fits a linear model by calculating coefficients on the independent terms that minimize the sum squared error between the known observations responses and the approximated predictions. The following is the form of the linear models:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{4.3}$$

where $\mathbf{y}$ is the n-dimensional vector containing response variables (phenotypes) for each of n input genotypes. $\mathbf{X}$ is the $n \times p$ matrix (n genotypes and p markers in each

genotype), β is a p-dimensional column vector of unknown coefficient parameters, and ε is the n-dimensional unknown random error column vector. The linear regression model tries to learn β to make phenotype prediction.

Ridge Regression Best Linear Unbiased Prediction

The ridge regression (RR) models were built using JMP Genomics 9 [JMP Genomics, 1989-2021]. This process computes Best Linear Unbiased Predictions (BLUPs) that linearly correlate the genotypes, based on the input marker encodings, to a trait variable of interest. RR-BLUPs are linear mixed-models of the following from:

$$\mathbf{y} = \mathbf{F}\delta + \mathbf{Z}\gamma + \varepsilon \tag{4.4}$$

where $\mathbf{y}$ is the n-dimensional column vector containing response variables (phenotype) for each of n genotypes, $\mathbf{F}$ is the $n \times q$ matrix of known fixed-effects, δ is a q-dimensional column vector of unknown fixed-effects parameters, $\mathbf{Z}$ is the $n \times p$ matrix of known random-effects (n encoded genotypes), γ is the p-dimensional column vector of unknown random-effects parameters, finally, ε is the n-dimensional unknown random error column vector. It is assumed that the residuals ε and random-effects γ are normally distributed, $\varepsilon \sim N(0, \mathbf{I}\sigma^2_\varepsilon)$ and $\gamma_i \sim N(0, \sigma^2_{\gamma_i})$ where $\mathbf{I}$ is the identity matrix and $\sigma^2_{\gamma_i}$ are assumed equal for all SNP markers. The unknown model parameters are estimated from the solution of the mixed-model equations [Henderson, 1984]. A scoring file is produced that contains an equation of a linear combination of SNP markers for predicting phenotype response from the testing set data.

4.2.7 Train-Test-Validation Split

To divide the data into train, test and validation sets, we follow the recommendation of Runcie and Cheng [2019]. We randomly split the data in 85% - 15%, where 85% is the training data and the remaining 15% is the test data. From the training data, we again perform three random splits of 85% - 15%. The first 85% is the training set and the rest 15% is the validation set. Thus from the data, we created three training sets, three validation sets and one test set.

To further investigate the reliability of the proposed GPTransformer model, we again perform k random splits of the dataset into 70% - 15% - 15% train-test-validation sets ($k = 3$). This time, in each split, the test set also changes along with the training and validation set.

4.2.8 Feature Selection

Feature selection methods identify features that contribute to a specific expression. In our work, the purpose of the feature selection is to identify those genetic markers that contribute towards low FHB or low DON levels. We applied mutual information feature selection, a filter based method, where the input is the genetic markers and the phenotypes and the output is a mutual information score (ranging 0-1). Discretization was performed where we divide the genotypes into three bins based on phenotypes (lowest 25%, middle 50%, highest 25%). The categories are the labels and the genetic markers are the features of mutual information algorithm which produces a mutual information score for each marker. The final mutual information score for each marker

is the average mutual information score over the three training sets. Markers with average mutual information of ≥ 0.02 were selected.

4.2.9 Training Transformer

The input embedding of the Transformer network converts each marker to an eight-dimensional vector (hidden dimension, $n = 8$). Thus, if each genotype contains m genetic markers ($m = 25,000$ before performing feature selection), the input embedding layer's output will be an (m, n) matrix. This (m, n) matrix is the input of the multi-head attention. As the multi-head attention computes pairwise attention between each marker, the operation will result in an (h, m, m, n) matrix where h is the number of heads. This operation has significant memory requirements for the GPU. For instance, on this dataset, it requires over 48 GB of memory.

To circumvent memory limitations, only selected features of mutual information are taken as the Transformer network input. We also pass only one genotype at each batch for training. Thus, the input of the embedding layer is all the markers of a genotype selected by the mutual information algorithm. The output is an (f, n) dimensional matrix where f is the number of markers. This (f, n) matrix will be the input of the Transformer encoder. Our Transformer neural network contains two Transformer encoder blocks ($N = 2$). We use two heads ($h = 2$) for each multi-head attention layer and each feed-forward block inside the Transformer encoder contains 256 hidden neurons. The final Transformer encoder's output is also an (f, n) matrix which is flattened to create a vector that contains $f \times n$ elements. This vector is the input of the last feed-forward network which contains one output neuron. We

use the mean square error (MSE) loss function along with the Adam optimizer. The learning rate of the optimizer is $1e-5$. If there is no improvement in MSE loss in the validation set for ten consecutive epochs, we stop the training.

4.2.10 Training Residual Fully Connected Neural Network

The first feed-forward layer takes m markers (all the markers) or f markers (feature selected) as an input and produces a 512-dimensional hidden representation. Each subsequent feed-forward layer takes the previous layer's input and produces a 512-dimensional hidden representation. The last feed-forward layer contains only one output neuron. We stack five residual blocks ($N = 5$) one after another. We also use MSE as the loss function and Adam as the optimizer with a learning rate of $1e-5$.

4.3 Results

The Pearson Correlation Coefficient (PCC) was calculated to measure the performance. The PCC calculates the linear relation between the true output and predicted output. The PCC value ranges from -1 to 1 where 1 indicates a perfect linear relation between the predicted phenotypes and the true phenotypes whereas -1 indicates the opposite relationship between the true and predicted phenotypes.

In the remainder of the paper, especially in figures, we will denote Hardy-Weinberg genotype based encoding as HW, categorical encoding as CAT, Decision Tree algorithm as DT, Linear Regression as LR and Residual Fully Connected Neural Network as RFCNN.

4.3.1 Phenotype Assessment

The distribution of FHB and DON is shown in Figure 4.3 for all locations and years. It is observed that the FHB values are distributed over a 0.3 to 4.8 range (1.75±0.04), while the DON values range from 4.9 to 36.9 mg kg^{-1} (13.96±0.21) (Supplementary table S1). For both phenotypes, the distribution curve is similar to a normal distribution, however, a degree of positive skewness was observed in FHB (0.881) and DON (0.976). Shapiro-Wilk W tests conducted on FHB (W=0.947, Prob <0.0001) and DON (W=0.963, Prob <0.0001) indicated a degree departure from normality. Departures from normality were most obvious in tail regions of the distributions.

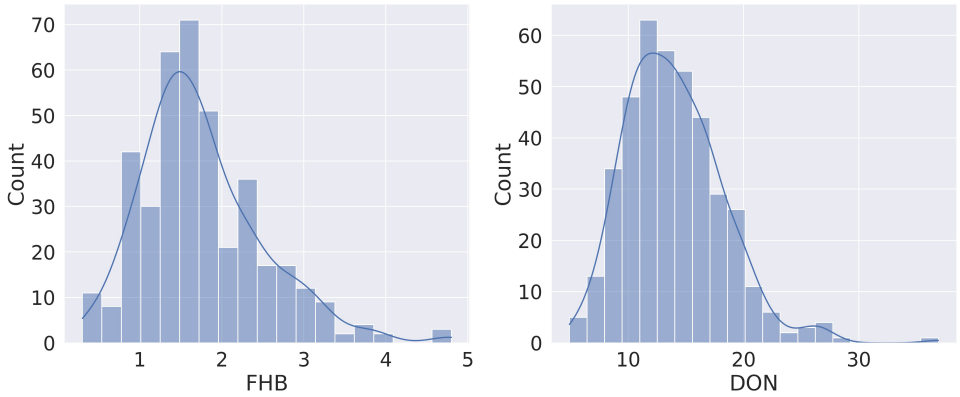


Figure 4.3: Distribution of phenotypes for Fusarium head blight (FHB, $0-5$) and deoxynivalenol content (DON, $mg\ kg^{-1}$) for all locations and years.

Though from Figure 4.3 it seems there may be a linear relation between FHB and DON, it is found that the Pearson Correlation Coefficient (PCC) between the two phenotypes is 0.381 ($p < 0.0001$). Figure 4.4 shows a scatterplot of FHB vs. DON for each genotype. From Figure 4.4, it is observed that there is a very little correlation between the two phenotypes, which is reflected by the PCC score. This

leads us to expect that similar genomic selection models may not immediately perform similarly when predicting the two phenotypes. FHB and DON were also examined for relationships with days to heading ($r = -0.18, P < 0.004; 0.18, P < 0.0003$) and height ($r = -0.60, P < 0.0001; -0.21, P < 0.0001$).

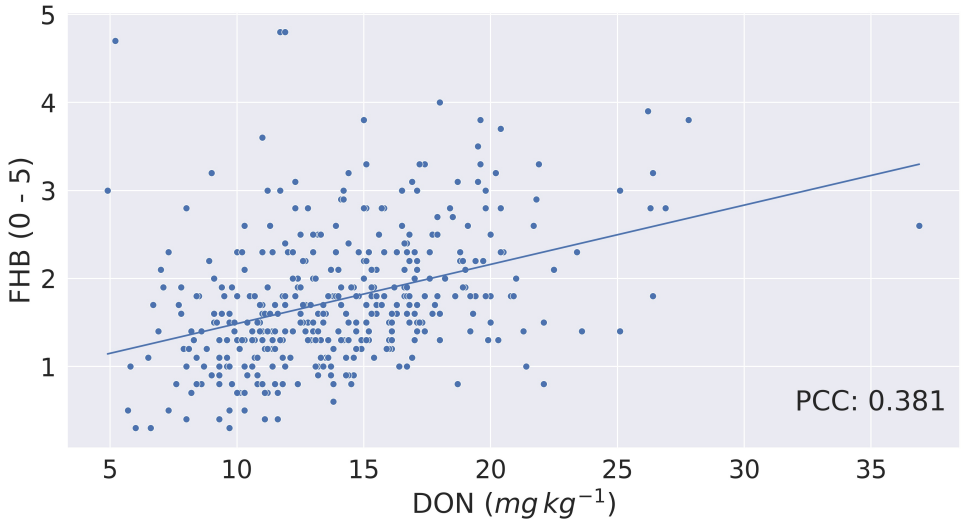


Figure 4.4: Fusarium head blight (FHB, $0-5$) vs. deoxynivalenol content (DON, $mg\ kg^{-1}$) for each of the barley genotype tested. Correlation between FHB and DON is 0.381.

4.3.2 Effect of Encoding Technique

Two encoding techniques were implemented: i) categorical encoding (-1, 0, 1) and ii) genotype frequency-based encoding that follows Hardy-Weinberg equilibrium. Figure 4.5 shows the comparison of PCC between two encoding schemes for various models. In most of the models, the categorical encoding outperforms the genotype frequency-based encoding. In the BLUP, correlation is very close to each other for both traits as the correlation score varies from 0.001 to 0.003 based on different traits.

For Transformer, Hardy-Weinberg encoding improves the correlation by 6.9% for FHB and 9.6% for DON. This can be explained by noting that, with the categorical encoding, the values of genotypes are 1, 0 and -1, and the heterozygous alleles will be considered neurons that do not have any effects. In particular, Figure 4.6 shows the effect of categorical encoding in the embedding layer. As the embedding layer's output is the input of the multi-head attention, multi-head attention also ignores any effect of heterozygous alleles. When applying genotype frequency-based encoding, different alleles of a specific gene have different values and these values even differ from gene to gene. For example, allele AA for gene X and allele AA for gene Y may appear in different frequencies and will have different Hardy-Weinberg values. Thus, the embedding layer does not suffer from multiplying-by-zero problems and improves the performance of the Transformer.

4.3.3 Effect of Feature Selection

No single molecular marker dominantly explained significant portions of the variation for FHB or DON. However, specific markers could be identified as 'top features' which may be associated with genes of interest which may operate closer to oligogenic vs. polygenic fashion (Supplementary Data Tables S2A & S2B). Biological endorsement of genomic features for FHB and DON were investigated through analysis of SNP effect annotations of markers on the 50K SNP chip [Bayer et al., 2017]. The top molecular markers with the highest mutual information are displayed in the supplementary file. Gene annotations generally concurred with resistance patterns.

In most of the experiments, when the machine learning or statistical methods are

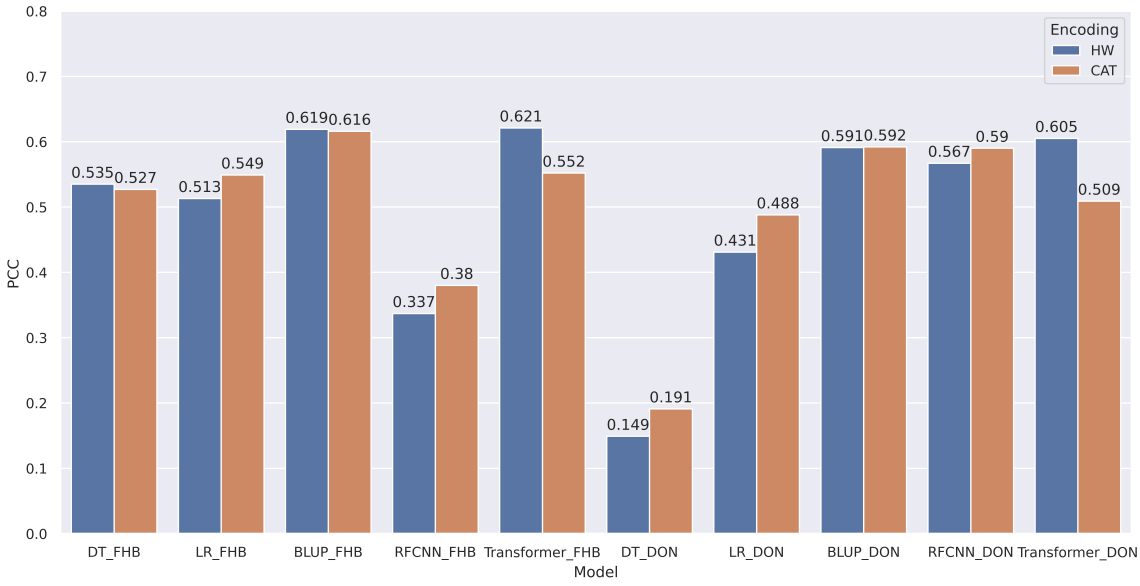


Figure 4.5: Comparison of Pearson Correlation Coefficient based on encoding techniques. HW and CAT represents Hardy-Weinberg and categorical encoding. The correlation is measured between the target and predicted phenotypes. Decision Tree, Linear Regression, BLUP, Residual fully connected neural network and Transformer are applied for each encoding technique.

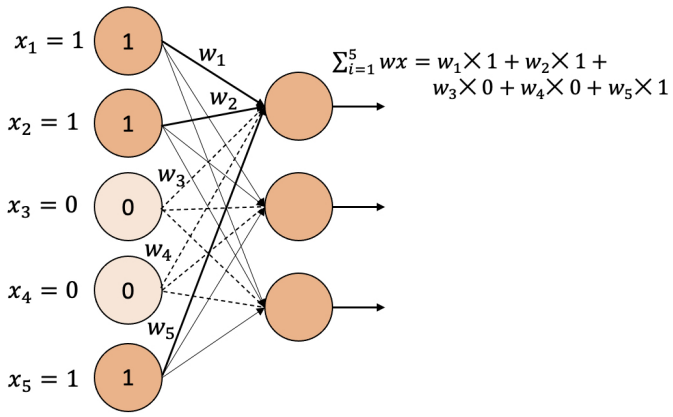


Figure 4.6: Categorical encoding when applied to a fully connected layer.

trained using all the markers, they performed better. Figure 4.7 shows the comparison of correlation scores between models trained on all markers and models trained on selected markers for DON. Using marker frequency as features for the DON, only

the BLUP method with reduced markers shows minor improvement (0.4%) over the BLUP method that uses all the markers. When the models were applied on categorical features for DON, only Decision Tree shows a 2.1% improvement when reduced marker sets were used as features. Recall that due to memory issues, the Transformer model with the full set of markers was not executed, and thus is absent from Figure 4.7.

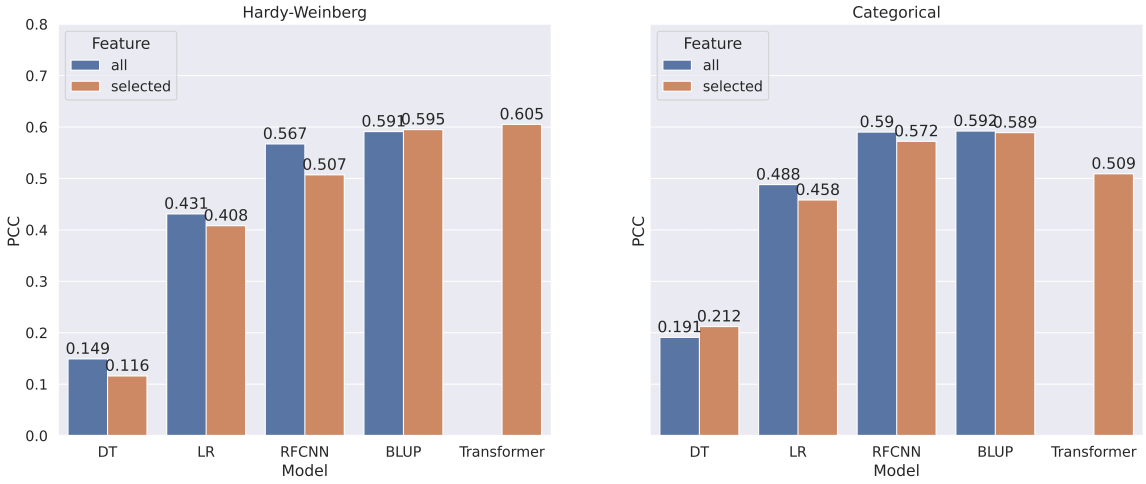


Figure 4.7: Comparison of Pearson Correlation Coefficient when taking all markers as features vs. selected markers for DON. The PCC is measured between target and predicted DON. Decision Tree, Linear Regression, BLUP, Residual fully connected neural network and Transformer are applied for each encoding technique.

In Figure 4.8, we show the comparison of correlation scores between models trained on all markers and models trained on selected markers for FHB. From the figure, we observe a similar pattern for FHB that we observed for DON. In contrast to categorical encoding, when using marker frequency as a feature value, selected markers improve the performance of the FHB RFCNN model by 27.6%. For categorical encoding of features, Linear Regression with selected features shows 4.2% improvement overall features.

Though there is a significant increase in correlation score when all the markers

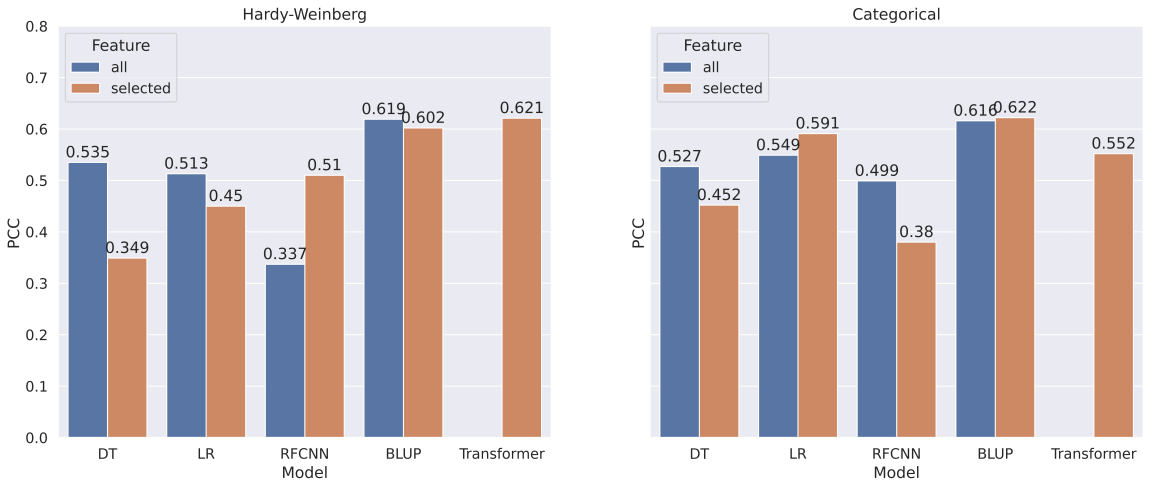


Figure 4.8: Comparison of Pearson Correlation Coefficient when taking all markers vs selected markers as features for FHB. The PCC is measured between target and predicted FHB.

are used for machine learning models, none of the BLUP models show any significant difference in performance when all markers or selected markers are used. Due to GPU memory limits, it was not possible to use all the features for Transformer architectures.

4.3.4 Best Performing Models

Overall, BLUP and Transformer models that use genotype frequency as features obtained better correlation scores than other models. Figure 4.9 shows the comparison among the best models for FHB and DON. BLUP and Transformer's performance are competitive as we observe only 1% improvement over BLUP for DON and the same correlation score for FHB.

Figure 4.10 shows the true versus predicted phenotype score using Transformer. From the figure, we observed a linear relationship between the target and predicted score, which also shows that the Transformer architecture performs well to predict

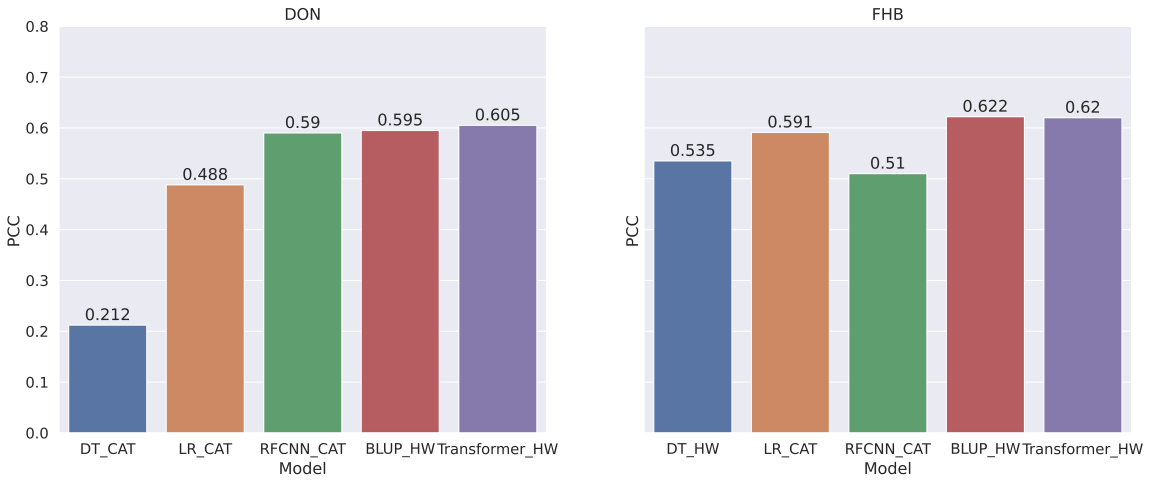


Figure 4.9: Comparison among the best models for each machine learning or statistical methods for DON and FHB.The PCC is measured between the target and predicted values of the phenotype.

phenotypes.

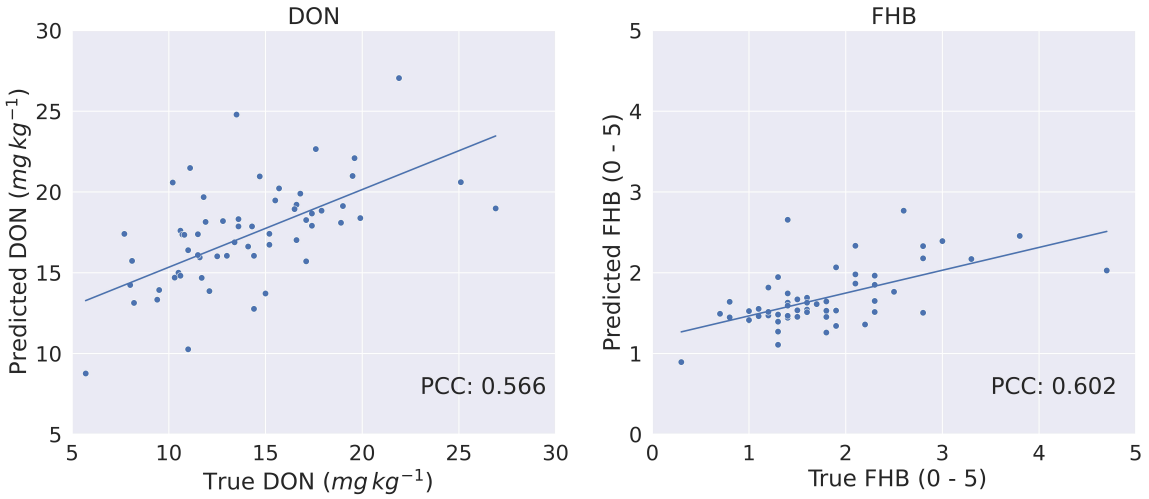


Figure 4.10: True vs. predicted phenotypes for DON ($\mathrm{mg\,kg^{-1}}$) and FHB ($0-5$) on the test set of 60 genotypes. As the PCC scores are 0.566 and 0.602 between the actual and predicted values of DON and FHB, respectively, this shows there exists some linear relationship between the actual and predicted values.

4.3.5 Reliability of the GPTransformer Model

To understand how much reliable our GPTransformer model, we train and test the model in three separate traning-validation-test splits. With the GPTransformer model, the average PCC between the target and predicted phenotype for DON is 0.748 and for FHB is 0.703. The BLUP model obtained PCC of 0.789 and 0.681 for DON and FHB respectively. The PCC we obtained with different splits using GPTransformer and BLUP models are not statisticallly significant as the p-value for DON is 0.576 and for FHB is 0.639. In addition, we observe an improvement in the average performance than the performance reported in the previous section when the model is trained and tested on different subsets of data.

4.4 Discussion

Transgressive segregation is common for FHB and DON in barley, where it is typical for offspring to deviate from mid-parent value. A greater degree of more susceptible genotypes was observed, which suggests that unique configuration of alleles over multiple loci may be responsible for resistance [Zhu et al., 1999]. Such complexities in multiple resistance genes which form near-continuous distributions may not be accounted for under assumptions of statistical models examined herein. While the machine learning approach did not substantially surpass the statistical models in prediction, the ML approaches we have used here are capable of capturing non-additive genetic components. Thus, the predictions given may be incorporating some of these interactions in their prediction. However, we aren't able to tell what effects are being

modeled in the ML models because of their opacity.

The current study confirms previous association mapping analysis of FHB and DON in barley [Massman et al., 2011], where quantitative trait loci (QTL) effects were small. While minor in nature, genes may additively contribute to resistance thereby lowering FHB and DON content. The advantage of such genes is that they do not typically carry issues seen when incorporating larger QTLs from exotics which tend to have tall stature or extreme heading date [Rudd et al., 2001]. In the context of the current study, the number of days to heading was not strongly associated with either character. Height was also weakly associated with DON content, however it did demonstrate a moderate, negative relationship with FHB. Top feature molecular markers and genes identified for FHB and DON did not overlap, as one might predict based on their moderate-to-low trait correlation. Within barley, the relationship of FHB disease and DON content is not as robust as seen in other cereals such as wheat. Application of GEBVs based on DON content may offer a better target for developing resistance, since it is the primary factor monitored by the industry.

Markers identified in the top features associated with FHB resistance were found on chromosomes 1H, 3H and 7H and all chromosomes for DON content, excluding 1H. Annotations of associated genes generally displayed direct biological function of resistance mechanisms. For instance FHB was associated with auxin transporter (HORVU1Hr1G073490) and response factor (HORVU7Hr1G033820), where this plant hormone has been associated with FHB severity and yield loss in barley (Petti et a. 2012). Also identified were genes involved in -glucan synthesis cell (HORVU7Hr1G003460, HORVU7Hr1G003460), which may contribute to resistance

via wall reinforcements or anti-oxidant properties [Martin et al., 2018]. The molecular marker BOPA2-12-31203 in the top features group in this study previously identified by Huang et al. [Huang et al., 2018] is a flanking marker for a QTL for FHB severity in the centromeric region of 7H. As a result of this toxic function, DON may induce programmed cell death (PCD, i.e. apoptosis). Development and Cell Death (DCD) domain protein (HORVU3Hr1G017930) and autophagy-related protein 18 (HORVU3Hr1G017150) underling removal of damaged cells may be involved in this process. Such top genomic features only explain a small percentage of total phenotypic variation for FHB and DON and could not be individually implemented under a marker-assisted selection program. However, biological functions associated with genes and markers highlighted above amongst others, may help explain why feature selections of a reduced marker subset may facilitate predictions with similar proficiency as when using all markers.

The proposed GPTransformer model takes the relationship among genetic markers into account within the model. The self-attention mechanism of the Transformer assigns a high weight to those markers that are associated with another specific marker. After applying the self-attention module, each obtained neuron is a combined representation of the genetic markers that are related to a specific marker. As many markers contribute towards a specific phenotype, GPTransformer has a unique attribute compared to other machine learning and statistical methods that takes marker relationships into account.

The frequency-based marker representation technique we applied for representing each allele carries more information as it indicates the zygosity and the frequency of

the allele. The traditional categorical encoding (1, 0, -1) only indicates the zygosity and remains the same for all the genetic markers. As each allele of a genetic marker is represented by its fixed frequency value, the frequency-based encoding provides us the information of the frequency as well as the zygosity. Though the frequency value remains the same within the same genetic marker for a specific allele, it may differ between different genetic markers. Thus, when the GPTransformer is combined with the frequency-based encoding, it performs better than the traditional categorical encoding-based model. The frequency-based representation is in the range of 0-1 and minimizes issues of vanishing gradient that may occur when training the GPTransformer or other neural network models.

The stability of the proposed model is tested with three different training and test data and the result shows the standard deviation of PCC on the test data for FHB is 0.04 and for DON is 0.09. The standard deviation of the BLUP model for FHB and DON for the same three different training and test is 0.008 and 0.04 respectively. The deviation from the average PCC for three different runs is higher for other machine learning methods we experimented with. This shows that the GPTransformer model is stable compared to other machine learning methods and as good as the popular BLUP model.

While the time commitment is higher and taking up to an hour to train the GPTransformer, it only took 24 epochs to complete training. The time complexity of the RFCNN is much lower than the Transformer though it took approximately 200 epochs to train. The most expensive task in the Transformer is the self-attention that requires substantial time and memory to complete. We ran both machine learning

models on an Intel Xeon E5-2690 v4 processor and an NVIDIA Tesla P40 GPU, which contains 24 GB memory. With our Transformer architecture, we were able to fit all the markers that have mutual information ≥ 0.02. To fit all the genetic markers, a larger memory or multi-GPU instance is needed.

Though the performance of most of the machine learning methods improves when all the markers are used for prediction, the Transformer architecture outperforms other methods with selected markers. To the best of our knowledge, this is the first method that uses Transformer architecture for genomic prediction. This work showed that this method could outperform existing machine learning methods with fewer data and obtain state-of-the-art performance. Based on the performance in the language model domain, it is expected that with an increased amount of data, the performance of the Transformer model will also increase.

Our work shows the potential of the Transformer-based method for genomic prediction. Though Transformer generally performed well with a large amount of data in other fields, in this work, we showed that when trained on a small dataset, the Transformer encoder performs equally or better compared to the existing machine learning and statistical methods. As the genotype data generally contains many markers, calculating self-attention in a GPU will require a large amount of GPU memory that may not be available. Our feature selection step in the model addresses the memory issue of the Transformer method. This step reduces the number of markers and identifies the biologically relevant markers for a specific phenotype. We also applied genotype frequency-based encoding for each genotype. This encoding performs better when combined with the Transformer. If a large amount of data is available,

the number of Transformer encoder blocks can be increased which may increase the overall performance.

Conflict of Interest Statement

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.

Funding

This research was funded by the Brewing and Malting Barley Research Institutes (BMBRI), Manitoba Crop Alliance, Growing Innovation: Agri-Food Research and Development Initiative (ARDI) project to D. Fernando and A. Badea. Support was also delivered through part of the Barley Cluster project led by Alberta Barley with funding from the Western Grains Research Foundation and Agriculture & Agri-Food Canada (AAFC) under the Growing Forward 2 program.

Acknowledgments

We gratefully thank Kevin Moore, Roger Larios and AAFC Brandon wheat breeding crew for Pathology technical assistance in FHB nurseries. We thank Bill Legge for project management of research in AAFC Brandon FHB nursery. We acknowledge Aaron Beattie, CDC, University of Saskatchewan and Pat Juskiw, FCDC, Olds College for providing breeder seed. We thank Barinder Bajwa, Shuzhen Zhang and

Mira Popovic for assistance in genotype assay. We thank Aria Dolatabadian, Abbot Oghenekaro, and Liang Zhao for reviewing the manuscript.

Data Availability Statement

The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus [Edgar et al., 2002] and are accessible through GEO Series accession number GSE188791 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE188791`)

Chapter 5

GxENet: Novel Fully Connected Neural Network Based Approaches to Incorporate GxE for Predicting Wheat Yield

Sheikh Jubair[1], Olivier Tremblay-Savard[1], Mike Domaratzki[2]

1 – Department of Computer Science, University of Manitoba, 66 Chancellors Cir, Winnipeg, MB R3T 2N2

2 – Department of Computer Science, University of Western Ontario, 1151 Richmond St, London, ON N6A 3K7

This chapter is an adaptation of the article Jubair et al. [2023] published in Artificial Intelligence in Agriculture in 2023.

Author Contributions

SJ: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing- review and editing; OTS: Writing - review and editing; MD: Writing - review and editing, Supervision.

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 cultivars are developed by selecting superior genotypes of seeds suitable for a specific environment. Genomic selection is a computational technique for developing a new cultivar 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 environment does not have any effect on the quantitative traits, and for multi-environment trials, where the assumption

is that crop genomics does not affect quantitative traits. However, it is essential to consider both genomic and environmental data to develop a new cultivar, 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 potential to be useful in genomic selection.

Keywords– genomic prediction, multi-environment trial, deep learning, GxE, enviromics

5.1 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 cultivars that can perform well in different environments.

A line is a crop organism from which we obtain genetic information after genotyping. To create a new cultivar 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 cultivar 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 [Le Nguyen et al., 2019; Esposito et al., 2019]. 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., 2021a; Khaki and Wang, 2019; Ma et al., 2018; Rachmatia et al., 2017a; Crossa et al., 2016b].

The environment of a crop is the combination of the weather, soil and field management information of where it is sown [Washburn et al., 2021; Khaki and Wang, 2019]. While the 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, 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., 2019c; Shook et al., 2020; Khaki et al., 2020; Guo et al., 2020; Sandhu et al., 2021a]. The effect of environment 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 cultivars. GS takes genotyped data and, increasingly, environmental information as input and predicts quantitative traits as outputs [Khaki and Wang, 2019; Hayes 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., 2021a;

Ma et al., 2018; Gianola et al., 2011; Pérez-Rodríguez et al., 2012; González-Camacho et al., 2012, 2016; Rachmatia et al., 2017b; 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 they 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., 2020; 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., 2019c]. 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., 2019c]. 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:

- 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.

- 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. The proposed frameworks perform better than a framework similar to an existing multi-environment framework [Khaki and Wang, 2019]. In addition, 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 and genetic information contribute to predicting yield in our models. We observed that while some environmental variables make a bigger positive contribution, a large number of genetic markers makes smaller contributions to estimating the yield.

- We extended one of the frameworks (third framework) to integrate unstructured text about field notes. This shows that the proposed models can be extended when new sources of data emerge.

5.2 Materials and Methods

5.2.1 Dataset

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 5.1 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 5.1 shows the nursery information along with their mega-environment information.

Table 5.1: Environments of each nursery. ME refers to Mega-Environment. CIMMYT has 6 mega-environments for Spring Wheat.

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

The lines were sown over multiple years in different location 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 5.2 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.

Table 5.2: 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	72776	6984	25712	3012

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 5.3 shows the detailed result of the test along with their significance levels.

Table 5.3: Anderson-Darling test of yields distribution. As the statistics are larger than the critical values of all significance levels, the hypothesis that the data comes from a normal is 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

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), 2m wind speed (m/s) and 10m 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.

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.

5.2.2 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.

5.2.3 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 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.

5.2.4 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. Figure 5.1 shows an example of how the average frequency is calculated.

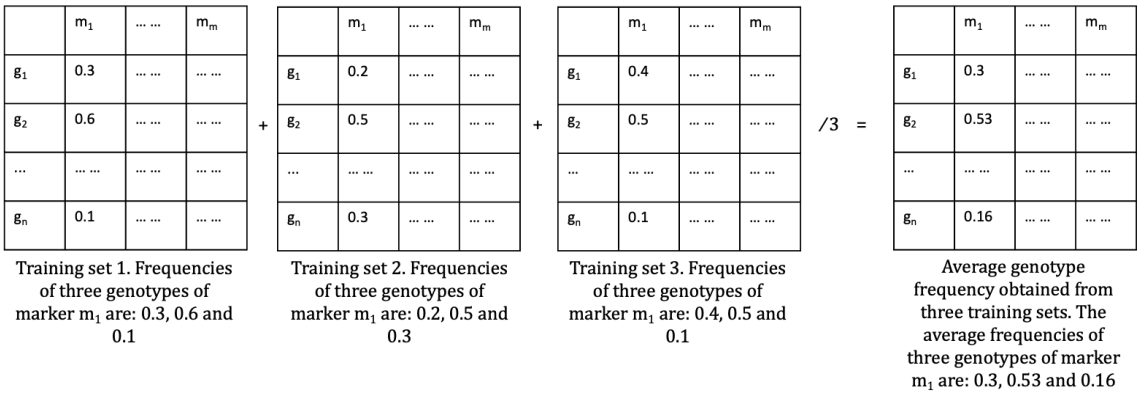


Figure 5.1: Markers represented by the average genotype frequency. Here, the example is shown with three training sets. We used five training sets to calculate the genotypic frequency average of each marker.

Research shows that some markers contribute toward 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 4,993 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., 2021a]. 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. Figure 5.2 shows how MI is obtained for a specific training set.

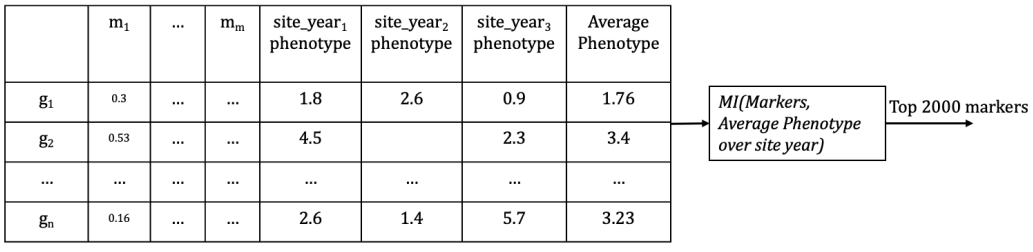


Figure 5.2: Top global marker selection procedure. An average of phenotypes across all site-year is calculated for each line. Then, mutual information is applied.

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 5.2.3. 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. Figure 5.3 shows the procedure for obtaining the local marker set.

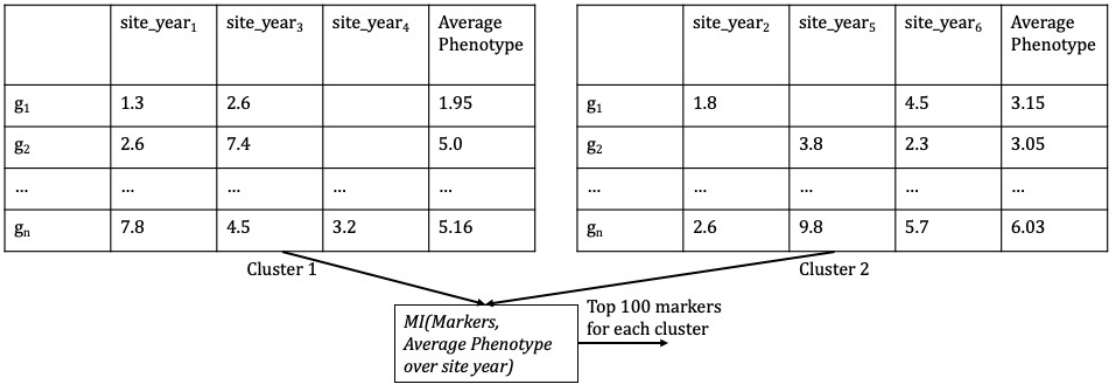


Figure 5.3: Top local marker selection procedure (markers were not shown). Tables in the figure shows how average phenotype is calculated. 100 markers are selected from each individual cluster. There are 25 clusters in total. As some markers are common among the clusters, this leads to 2052 unique local markers.

5.2.5 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 5.2.4. After obtaining the genotype frequency, global and local marker sets are obtained by applying the procedure described previously, again in section 5.2.4. 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. Figure 5.4 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.

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. Figure 5.5 shows the architecture

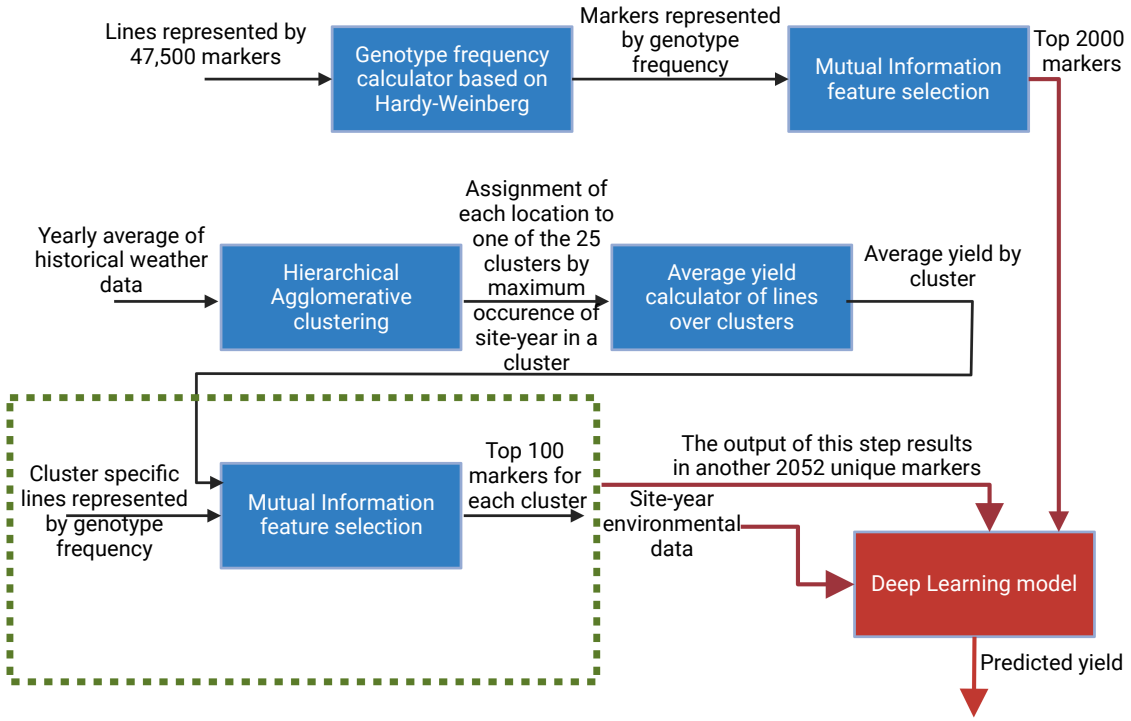


Figure 5.4: Deep learning framework 1 workflow. In this workflow, we employed three different deep learning models.

of this model. This model is similar to the model of Khaki and 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.

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

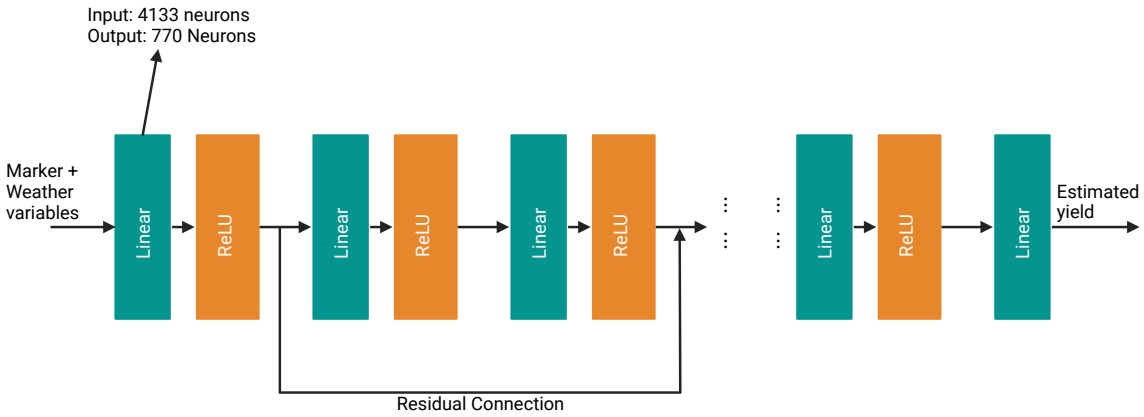


Figure 5.5: Architecture of the deep learning model in the F1M1 framework. This model concatenates the marker and weather variables and passed to a linear neural network block that contains a linear and ReLU layer. The architecture of this model is similar to the model of [Khaki and Wang, 2019]

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. Figure 5.6 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.

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

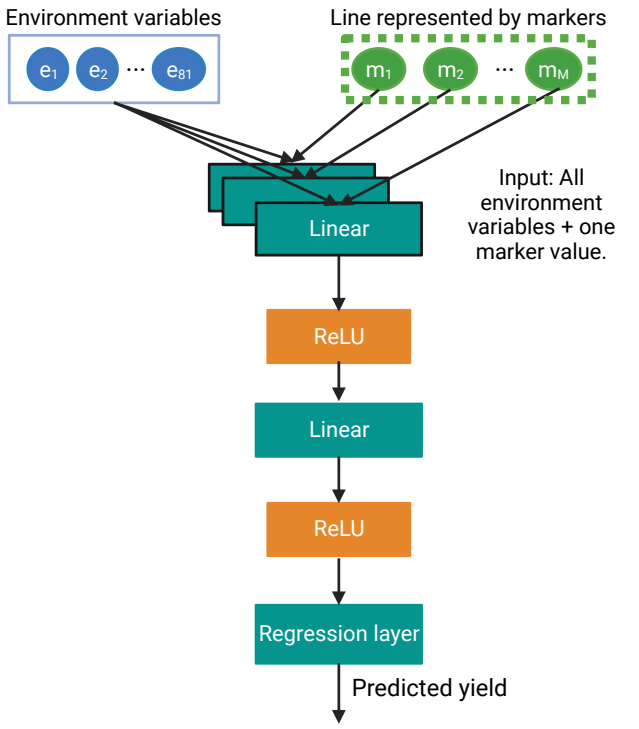


Figure 5.6: Architecture of deep learning model in the F1M2 framework.

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. Figure 5.7 shows

the neural network architecture. We will refer to this variant of the F1 framework as the F1M3 framework.

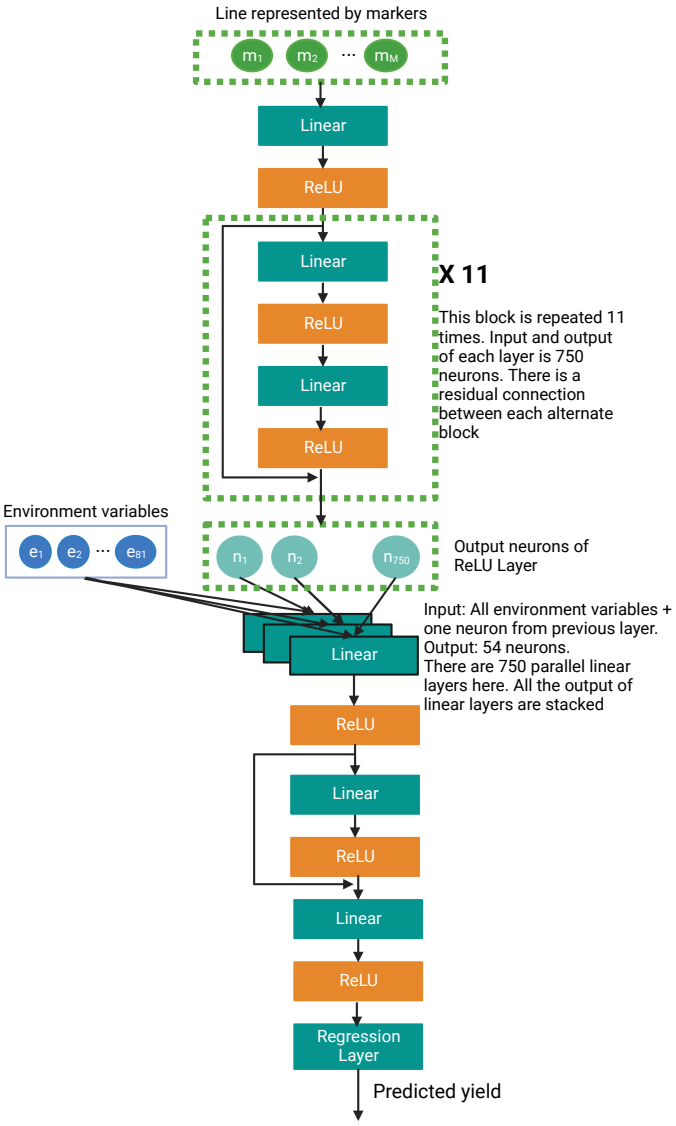


Figure 5.7: Architecture of the deep learning model in the F1M3 framework. Output of each odd number of ReLU layer is connected through a residual connection.

5.2.6 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., 2021a; Ma et al., 2018; Gianola et al., 2011; Pérez-Rodríguez et al., 2012; González-Camacho et al., 2012, 2016; Rachmatia et al., 2017b; Jubair and Domaratzki, 2019] or weather information [Lin et al., 2020; Shook et al., 2020; 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 specific 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. Inspired by their architecture, we first learn representations of genotyped data and environmental variables separately. These two models are optimized to predict the average yield over environment and over 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.

Figure 5.8 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 5.2.4. 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.

Representation learning model optimized for predicting line-specific average yield

Figure 5.9 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 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

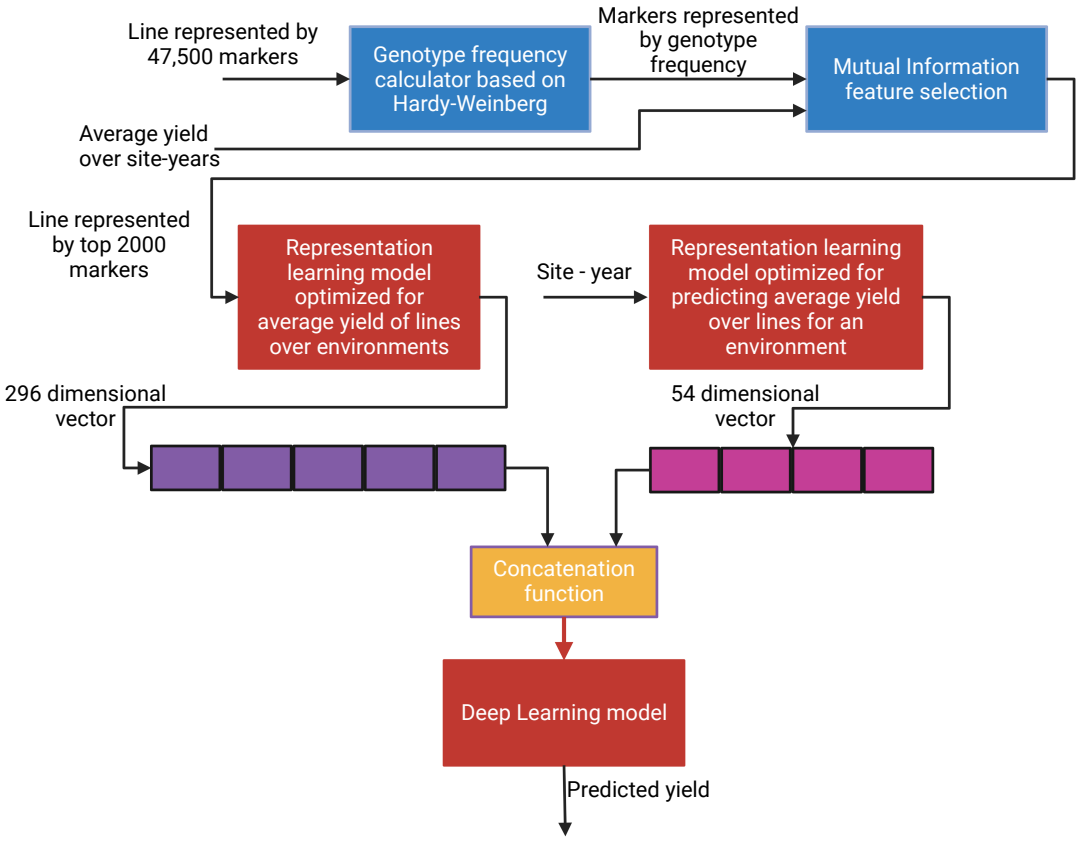


Figure 5.8: Deep learning framework 2 (F2).

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.

Representation learning model optimized for predicting environment-specific average yield

Figure 5.10 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

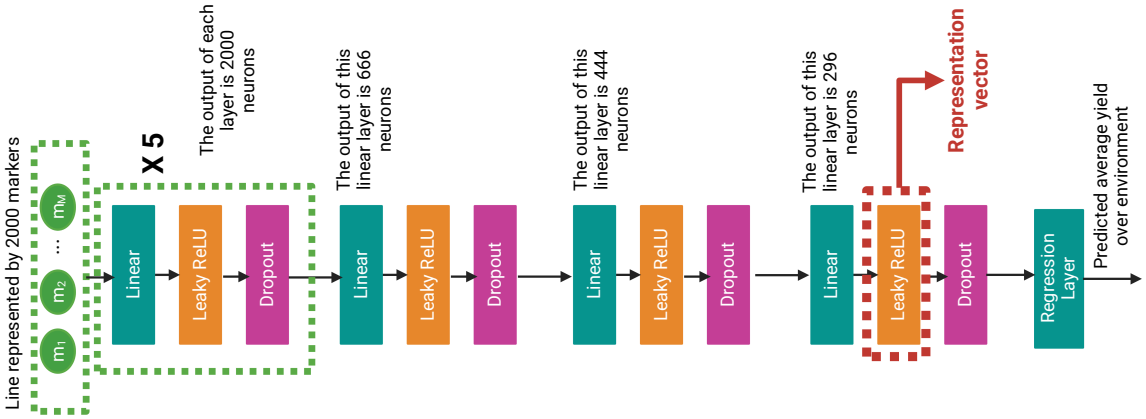


Figure 5.9: Representation learning model for predicting line-specific average yield.

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.

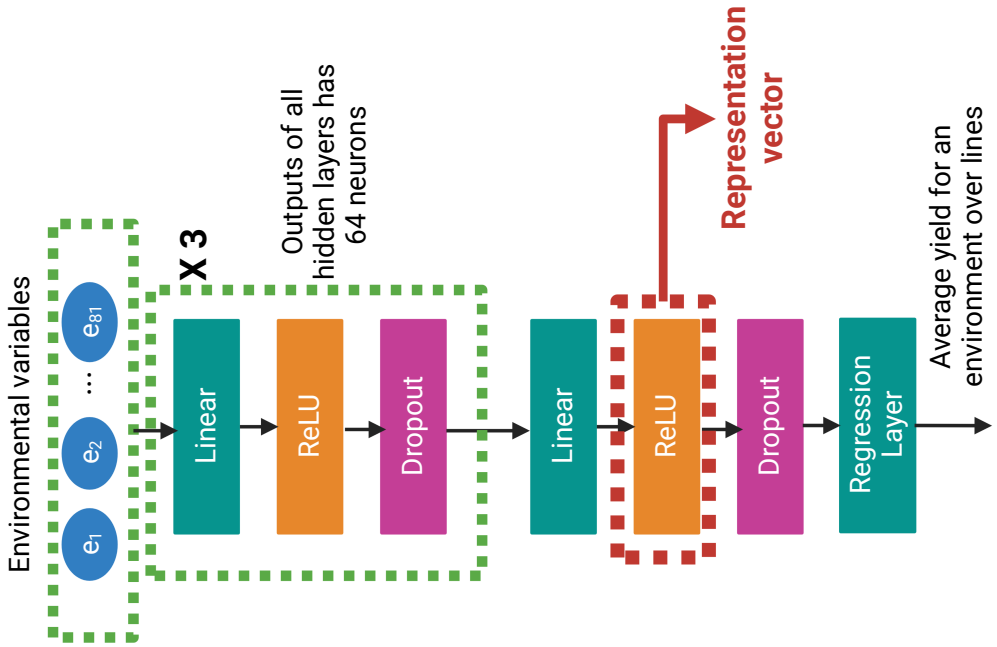


Figure 5.10: Representation learning model for predicting environment-specific average yield.

Yield prediction model

Figure 5.11 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.

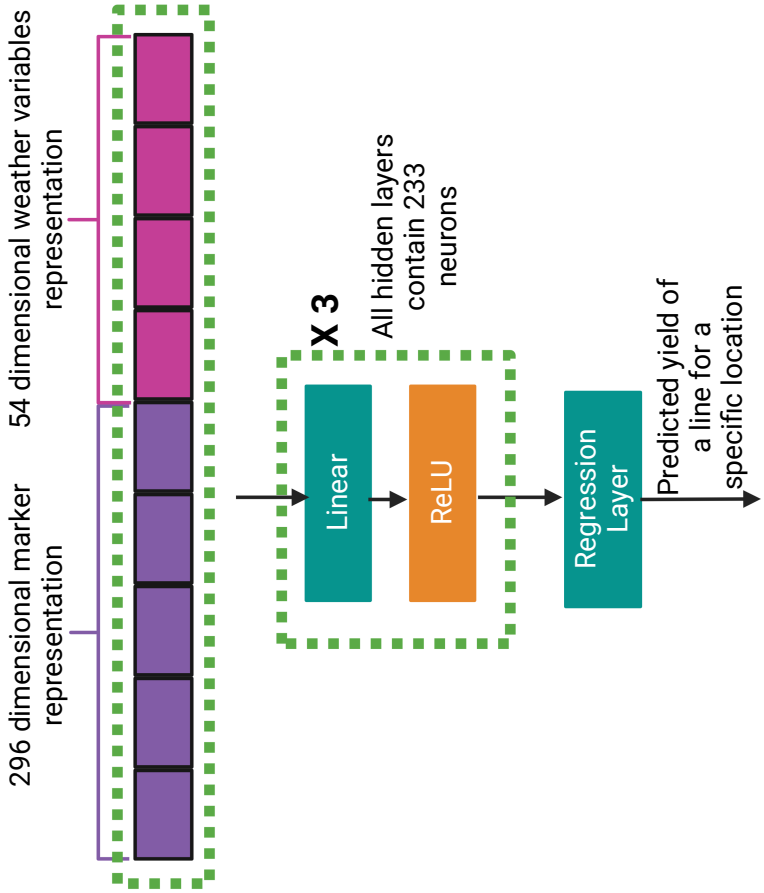


Figure 5.11: Yield prediction model for predicting line specific yield for each site-year.

5.2.7 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.

Figure 5.12 shows framework 3. This framework is an extension of the F2. The major difference between the two frameworks is that an agri-bert 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 less than 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.

5.2.8 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. 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

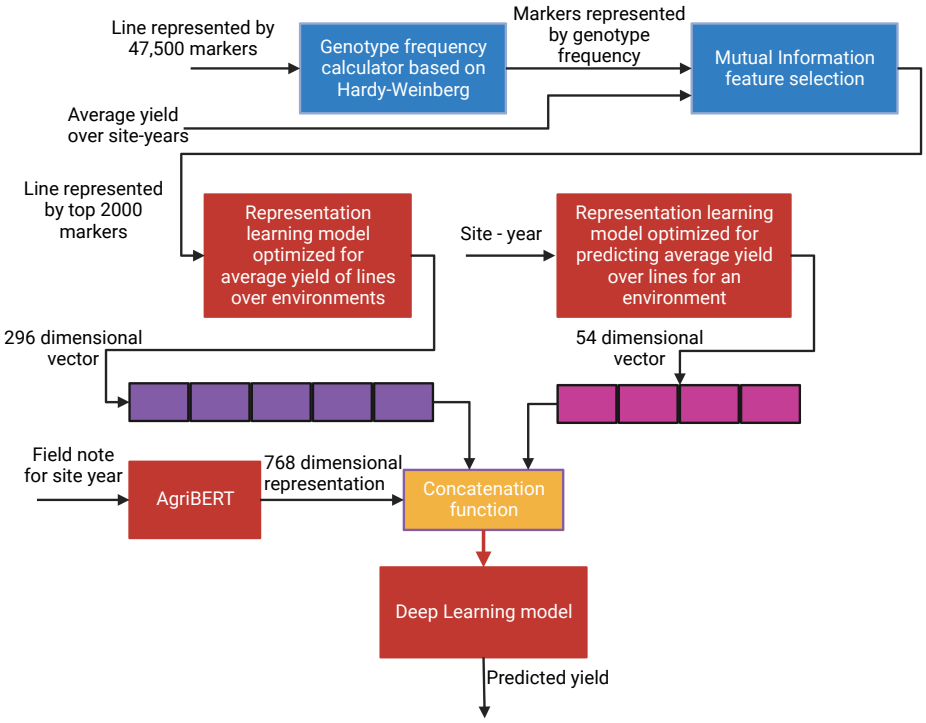


Figure 5.12: Deep learning framework 3 (F3).

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.

5.3 Results

5.3.1 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 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.

Figure 5.13 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].

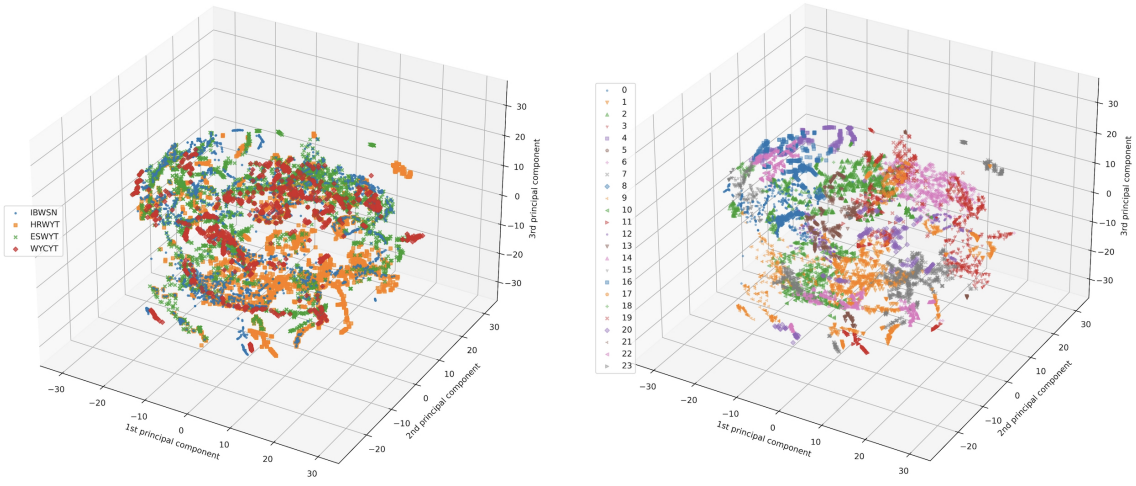


Figure 5.13: Site-year weather data by nursery (left) and cluster group (right). Each point in the figure indicates a site-year. Different colors either indicate a nursery (left) or a cluster group (right).

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

5.3.2 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 hours per epoch on NVIDIA GTX 1080). Thus, we did not train the F1M2 for the rest of the folds.

Table 5.4 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.

Table 5.4: Comparison of PCC over five folds between F1M1 and F1M3 for test scenario one and test scenrio two. Green colour indicates better performance.

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**	

5.3.3 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. Figure 5.14 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 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.

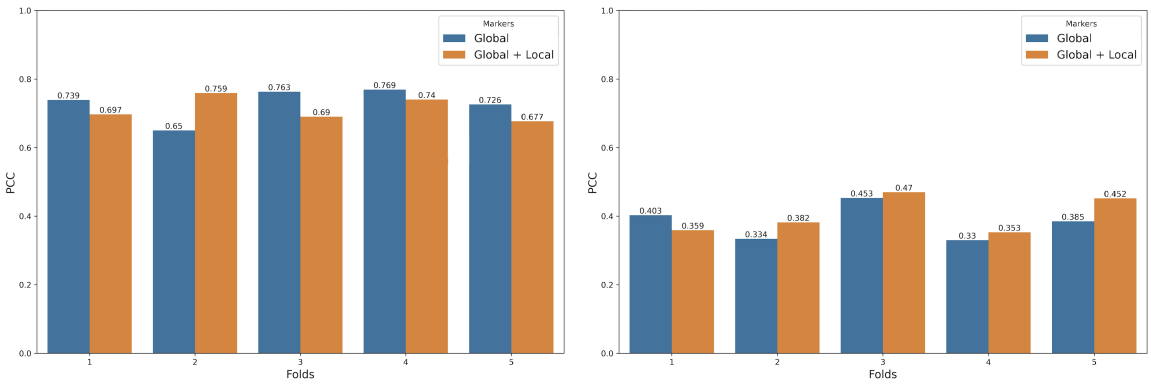


Figure 5.14: Comparison of PCC scores of F1M3 models trained on global markers and global + environment specific markers and evaluated on the test scenario one (left) and test scenario two (right).

5.3.4 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 5.5 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 [Washburn et al., 2021; Lin et al., 2020; Khaki and Wang, 2019; Montesinos-López et al., 2019c].

Table 5.5: 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**

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 5.6 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 representation learnt from the previous two models. Figure 5.15 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

Table 5.6: 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**

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% improvements over the F1M3. The result demonstrates that the F2 is the best architecture compared to all the variants of the F1.

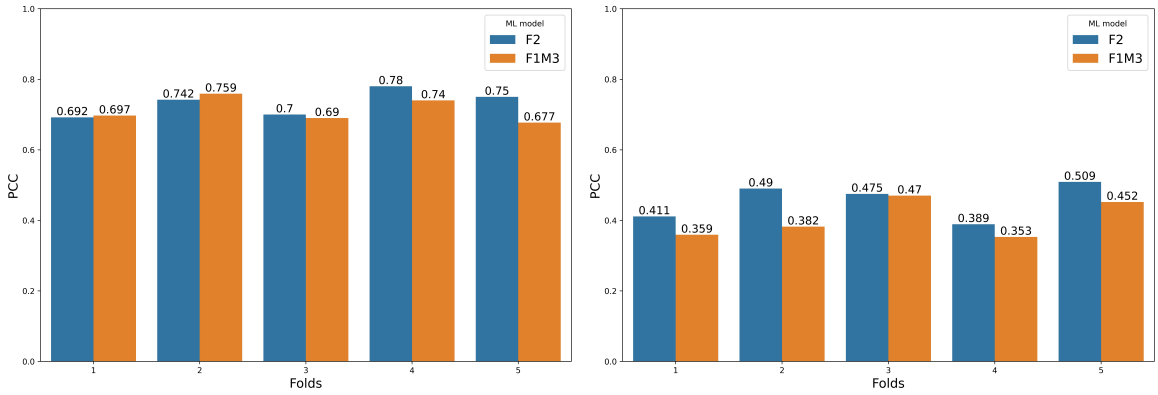


Figure 5.15: Comparison of PCC scores between F2 and F1M3 on test scenario one (left) and test scenario two (right).

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, $0 <$ PCC ≤ 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 5.7 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 5.7: Number of cycles in each ranges of PCC for both test scenarios. The green colour indicates the best framework.

Fold	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 $\leq$ 0**	0	1	0	0
	0 < PCC $\leq$ 0.4	2	0	3	2
	PCC > 0.4	27	28	3	4
	Number of line and site-year combination	**12,543**		**22,996**	
2	**PCC $\leq$ 0**	1	1	0	0
	0 < PCC $\leq$ 0.4	1	0	2	0
	PCC > 0.4	31	32	5	7
	Number of line and site-year combination	**14,915**		**15,884**	
3	**PCC $\leq$ 0**	1	0	1	1
	0 < PCC $\leq$ 0.4	0	1	2	1
	PCC > 0.4	28	28	7	8
	Number of line and site-year combination	**14,094**		**17,068**	
4	**PCC $\leq$ 0**	0	0	0	0
	0 < PCC $\leq$ 0.4	3	1	2	1
	PCC > 0.4	30	32	2	3
	Number of line and site-year combination	**12,792**		**24,312**	
5	**PCC $\leq$ 0**	1	0	0	0
	0 < PCC $\leq$ 0.4	1	1	1	1
	PCC > 0.4	25	26	5	5
	Number of line and site-year combination	**12,684**		**20,400**	
Average	PCC $\leq$ 0	0.6	0.4	0.2	0.2
	0 < PCC $\leq$ 0.4	1.4	0.6	2	1
	PCC > 0.4	28.2	29.2	4.4	5.4

5.3.5 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.

Figure 5.16 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 2m, (9) wind speed 10m. 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 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.

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. Figure 5.17 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 represen-

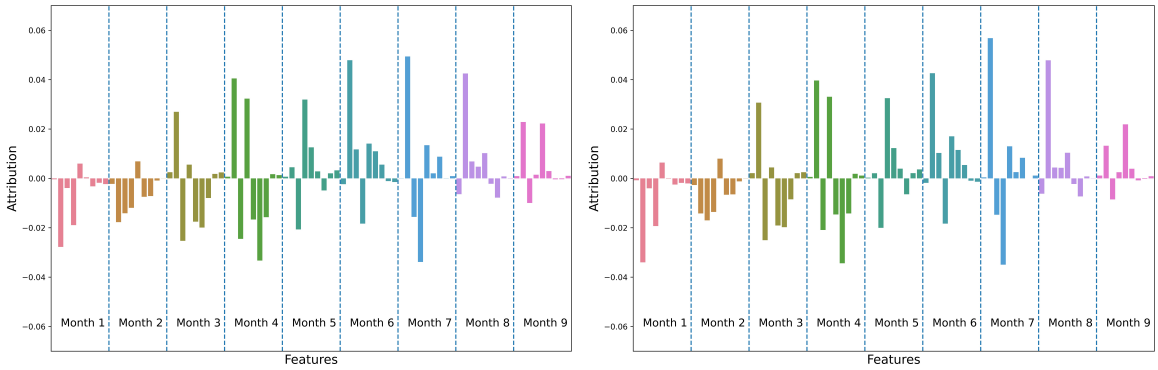


Figure 5.16: Feature importance of weather variables obtained by employing DeepLift on environment specific average yield model of F2. The left figure is the test scenario one and the right figure is the test scenario two. Each bar in the figure represents an environmental variable.

tation, they contribute more towards environment-specific yield estimation for a line. However, many marker representation features have a small positive effect on the outcome. Overall, this figure demonstrates the importance of adding more specific environmental features for the genomic prediction task.

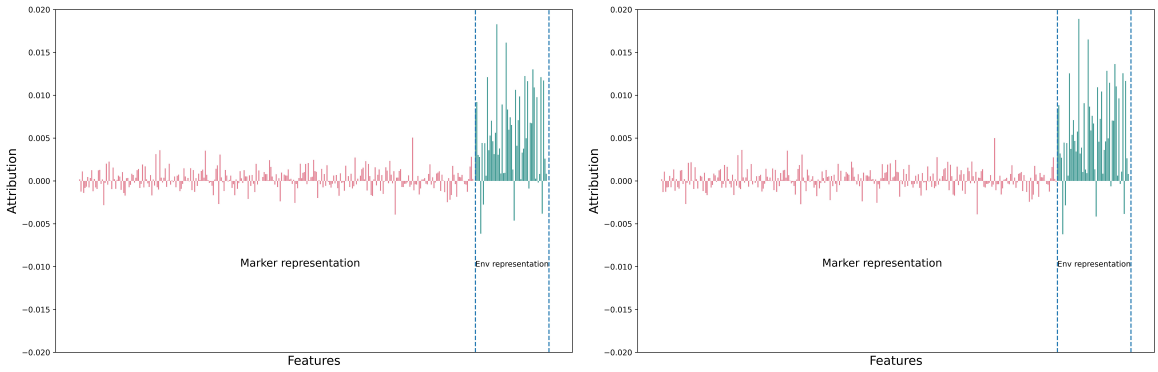


Figure 5.17: Feature importance of representation learning features obtained by employing DeepLift on the final deep learning model of F2 that combines marker representation and environment variable representation to predict environment specific yield for a line.

5.3.6 Performance of the F3 Architecture

As 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.

Figure 5.18 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. Figure 5.19 shows the attribution score of each input neuron to the final deep learning model of the F3 architecture. The figure shows that environment and marker representation have the most significant impact on yield prediction. While 94.25% of the marker representation neurons and 98.14% of the environmental neu-

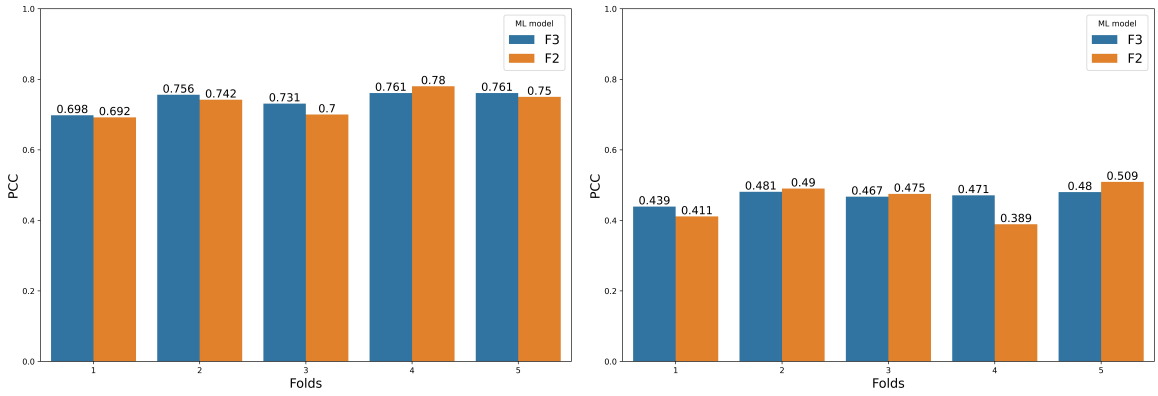


Figure 5.18: Comparison of PCC scores between F3 and F2 frameworks.

rons have a positive impact, only 26.11% of the text representation neurons have a positive impact on environment-specific yield estimation of a specific line. This result shows that the text data we had access to was not making an impact on our prediction task. However, as 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.

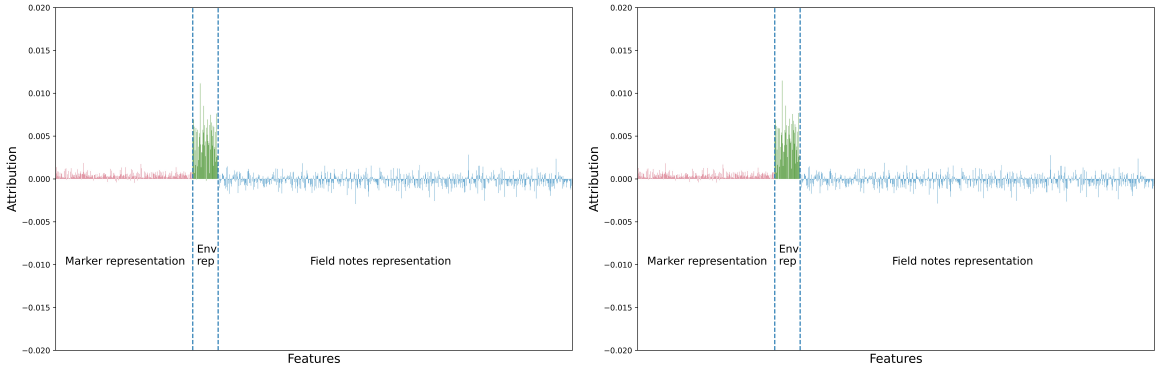


Figure 5.19: Feature importance of representation learning features obtained by employing DeepLift on the yield prediction model of the F3 that combines marker representation, environment variable representation and field notes representation to predict environment specific yield for a line.

5.4 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. Our evaluation shows that environmental information and genomic data positively affect yield estimation, while most textual representation has a negative effect. However, the text representation is learnt from a BERT-based model known as agriBERT, which is primarily trained on agricultural journal papers. As our texts are field notes, a BERT model trained on field notes would be more suitable. Furthermore, text data is heterogeneous, and there are lots of variations even within the same trial. Therefore, a more systematic approach for collecting field notes may give meaningful insight into the machine learning model.

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.

Conflict of Interest Statement

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.

Funding

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

Chapter 6

Conclusion

In this research, we devised two novel deep learning methods for predicting traits of a line for single environment trial and three deep learning frameworks to predict environment specific yield of each line. In the ensemble method for single environment trial, the ensembles are created using different subsets of markers while ensuring that all markers in the marker set are present in the ensemble method at least once. The results demonstrate that the performance of deep ensemble methods is as good as the state-of-the-art DeepGS model.

The proposed GPTransformer model for predicting the severity of FHB and DON content takes the relationship among genetic markers into account within the model. The self-attention mechanism of the transformer assigns high weight to those markers associated with another specific marker and uses these relationships to predict phenotypes. The transformer model typically provides better results when a large amount of data is available. However, the results of GPTransformer demonstrate either better or equal performance on PCC scores compared to the statistical BLUP model when

trained on 320 lines of barley. In addition, we proposed a frequency-based marker representation technique that carries more information than the traditional categorical marker representation. The major limitation of the transformer model is the memory requirement for a large number of features. We addressed this problem by selecting important markers identified by mutual information feature selection.

We proposed two novel deep learning frameworks for multi-environment trials where the models in the frameworks vary how GxE is incorporated. The results demonstrate that the performances of these frameworks are identical and better than some of the existing works. In addition, we also extended one of the frameworks to incorporate field notes. Our evaluation demonstrates that environmental information and genomic data positively affect yield estimation, while most textual representation has a negative effect. However, as these text data are very short text and heterogeneous, more detailed and informative text data could potentially improve the model performance or have more positive influence on the prediction.

Future works need to focus on incorporating more data, such as more robust environment and soil information, conditions of fields before sowing and sowing pattern, both for single and multi-environment trials. Since different types of data have different types of inter-relationship within them, they may need to be processed by different deep learning models and, finally, combine the outcome of those models to predict desired traits. Although we followed the same strategy in our representation learning based multi-environment framework (GxENet F2 and F3), we optimized three deep learning models of different tasks separately. Optimizing all the models simultaneously with a common loss function may further improve the framework. However, this

may require more memory and time to train the model. In addition, more research needs to be done on how environmental information can be added more meaningfully. For example, environmental information summarized based on the different known growth stages of the plants may reflect the effect of the environment on genotypes more meaningfully. Data scientists and plant breeders may collaborate closely to provide expert opinions on their respective domains so that more meaningful data can be collected at the proper interval and in an organized manner.

Bibliography

M. Abadi. Tensorflow: learning functions at scale. In *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming*, pages 1–1, 2016.

R. Abdollahi-Arpanahi, D. Gianola, and F. Peñagaricano. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution*, 52(1):1–15, 2020.

A. Abed, P. Pérez-Rodríguez, J. Crossa, and F. Belzile. When less can be better: how can we make genomic selection more cost-effective and accurate in barley? *Theoretical and Applied Genetics*, 131(9):1873–1890, 2018.

G. Acquaah. *Principles of plant genetics and breeding*. John Wiley & Sons, 2009.

A. Adak, S. C. Murray, and S. L. Anderson. Temporal phenomic predictions from unoccupied aerial systems can outperform genomic predictions. *BioRxiv*, 2021.

Agriculture and Agri-Food Canada. Discussion document: Reducing emissions arising from the application of fertilizer in canada's agriculture sector. *https://agriculture.canada.ca/en/about-our-department/transparency-and-corporate-reporting/public-opinion-research-and-consu*

lt at io ns /s ha re -i de as -f er ti li ze r-e mi ss io ns -r ed uc ti on -t ar ge t/ di sc us si on -d oc um en t-r ed uc in g-e mi ss io ns -a ri si ng -a pp li ca ti on -f er ti li ze r-c an ad as -a gr ic ul tu re -s ec to r, accessed 2022-11-20, 2022.

A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between ir effectiveness measures and user satisfaction. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 773–774. ACM, 2007.

M. Ali, Y. Zhang, A. Rasheed, J. Wang, and L. Zhang. Genomic prediction for grain yield and yield-related traits in chinese winter wheat. *International journal of molecular sciences*, 21(4):1342, 2020.

W. Anderson, C. Taylor, S. McDermid, E. Ilboudo-Nébié, R. Seager, W. Schlenker, F. Cottier, A. de Sherbinin, D. Mendeloff, and K. Markey. Violent conflict exacerbated drought-related food insecurity between 2009 and 2019 in sub-saharan africa. *Nature Food*, 2(8):603–615, 2021.

C. Anilkumar, N. Sunitha, N. B. Devate, S. Ramesh, et al. Advances in integrated genomic selection for rapid genetic gain in crop improvement: a review. *Planta*, 256(5):1–20, 2022.

J. Archambeau, M. Benito Garzón, F. Barraquand, M. de Miguel, C. Plomion, and S. C. González-Martínez. Combining climatic and genomic data improves range-wide tree height growth prediction in a forest tree. *The American Naturalist*, 200 (4):E141–E159, 2022.

M. P. Arruda, P. J. Brown, A. E. Lipka, A. M. Krill, C. Thurber, and F. L. Kolb. Genomic selection for predicting fusarium head blight resistance in a wheat breeding program. *The Plant Genome*, 8(3):plantgenome2015–01, 2015.

Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

H. Azadi, P. Ho, and L. Hasfiati. Agricultural land conversion drivers: A comparison between less developed, developing and developed countries. *Land Degradation & Development*, 22(6):596–604, 2011.

M. Bandeira E Sousa, J. Cuevas, E. G. de Oliveira Couto, P. Pérez-Rodríguez, D. Jarquín, R. Fritsche-Neto, J. Burgueño, and J. Crossa. Genomic-enabled prediction in maize using kernel models with genotype× environment interaction. *G3: Genes, Genomes, Genetics*, 7(6):1995–2014, 2017.

M. M. Bayer, P. Rapazote-Flores, M. Ganal, P. E. Hedley, M. Macaulay, J. Plieske, L. Ramsay, J. Russell, P. D. Shaw, W. Thomas, et al. Development and evaluation of a barley 50k iselect snp array. *Frontiers in plant science*, 8:1792, 2017.

Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan. Vision transformers for remote sensing image classification. *Remote Sensing*, 13(3):516, 2021.

A. D. Beattie, M. J. Edney, G. J. Scoles, and B. G. Rossnagel. Association mapping

of malting quality data from western canadian two-row barley cooperative trials. *Crop science*, 50(5):1649–1663, 2010.

G. Bebis and M. Georgiopoulos. Feed-forward neural networks. *IEEE Potentials*, 13 (4):27–31, 1994.

S. Bellairs, N. Turner, P. Hick, and R. Smith. Plant and soil influences on estimating biomass of wheat in plant breeding plots using field spectral radiometers. *Australian Journal of Agricultural Research*, 47(7):1017–1034, 1996.

T. Ben Hassen and H. El Bilali. Impacts of the russia-ukraine war on global food security: towards more sustainable and resilient food systems? *Foods*, 11(15):2301, 2022.

J. A. Bhat, S. Ali, R. K. Salgotra, Z. A. Mir, S. Dutta, V. Jadon, A. Tyagi, M. Mushtaq, N. Jain, P. K. Singh, et al. Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Frontiers in genetics*, 7:221, 2016.

C. M. Bishop and M. E. Tipping. Bayesian regression and classification. *Nato Science Series sub Series III Computer And Systems Sciences*, 190:267–288, 2003.

K. Boote, J. Jones, G. Hoogenboom, and N. Pickering. The cropgro model for grain legumes. In *Understanding options for agricultural production*, pages 99–128. Springer, 1998.

C. Boudhrioua, M. Bastien, D. Torkamaneh, and F. Belzile. Genome-wide association mapping of sclerotinia sclerotiorum resistance in soybean using whole-genome resequencing data. *BMC plant biology*, 20(1):1–9, 2020.

M. Bourgault, M. Löw, S. Tausz-Posch, J. Nuttall, A. Delahunty, J. Brand, J. Panozzo, L. McDonald, G. O'Leary, R. Armstrong, et al. Effect of a heat wave on lentil grown under free-air CO_2 enrichment (face) in a semi-arid environment. *Crop Science*, 58(2):803–812, 2018.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

H. Buerstmayr, L. Legzdina, B. Steiner, and M. Lemmens. Variation for resistance to fusarium head blight in spring barley. *Euphytica*, 137(3):279–290, 2004.

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

J. Burgueño, G. de los Campos, K. Weigel, and J. Crossa. Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. *Crop Science*, 52(2):707–719, 2012.

K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

F. Chollet et al. Keras: The python deep learning library. *Astrophysics source code library*, pages ascl–1806, 2018.

T. M. Choo, B. Vigier, Q. Q. Shen, R. A. Martin, K. M. Ho, and M. Savard. Barley traits associated with resistance to fusarium head blight and deoxynivalenol accumulation. *Phytopathology*, 94(10):1145–1150, 2004.

K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

CIMMYT. `http://genomics.cimmyt.org/mexican_iranian/traverse/iranian/`, accessed July, 2019.

G. Costa-Neto, L. Crespo-Herrera, N. Fradgley, K. Gardner, A. R. Bentley, S. Dreisigacker, R. Fritsche-Neto, O. Montesinos-Lopez, and J. Crossa. Envirome-wide associations enhance multi-year genome-based prediction of historical wheat breeding data. *bioRxiv*, 2022.

J. Crain, P. Bajgain, J. Anderson, X. Zhang, L. DeHaan, and J. Poland. Enhancing crop domestication through genomic selection, a case study of intermediate wheatgrass. *Frontiers in Plant Science*, 11:319, 2020.

D. Cros, M. Denis, L. Sánchez, B. Cochard, A. Flori, T. Durand-Gasselin, B. Nouy, A. Omoré, V. Pomiès, V. Riou, et al. Genomic selection prediction accuracy in a perennial crop: case study of oil palm (elaeis guineensis jacq.). *Theoretical and applied genetics*, 128(3):397–410, 2015.

J. Crossa, G. de los Campos, M. Maccaferri, R. Tuberosa, J. Burgueño, and P. Pérez-Rodríguez. Extending the marker× environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. *Crop Science*, 56(5):2193–2209, 2016a.

J. Crossa, D. Jarquín, J. Franco, P. Pérez-Rodríguez, J. Burgueño, C. Saint-Pierre, P. Vikram, C. Sansaloni, C. Petroli, D. Akdemir, et al. Genomic prediction of gene bank wheat landraces. *G3: Genes, Genomes, Genetics*, 6(7):1819–1834, 2016b.

J. Crossa, P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, G. De Los Campos, J. Burgueño, J. M. González-Camacho, S. Pérez-Elizalde, Y. Beyene, et al. Genomic selection in plant breeding: methods, models, and perspectives. *Trends in plant science*, 22(11):961–975, 2017.

J. Cuevas, J. Crossa, V. Soberanis, S. Pérez-Elizalde, P. Pérez-Rodríguez, G. de los Campos, O. Montesinos-López, and J. Burgueño. Genomic prediction of genotype× environment interaction kernel regression models. *The plant genome*, 9(3):1–20, 2016.

J. Cuevas, J. Crossa, O. A. Montesinos-López, J. Burgueño, P. Pérez-Rodríguez, and G. de los Campos. Bayesian genomic prediction with genotype× environment interaction kernel models. *G3: Genes, Genomes, Genetics*, 7(1):41–53, 2017.

J. Cuevas, O. Montesinos-López, P. Juliana, C. Guzmán, P. Pérez-Rodríguez, J. González-Bucio, J. Burgueño, A. Montesinos-López, and J. Crossa. Deep kernel for genomic and near infrared predictions in multi-environment breeding trials. *G3: Genes, Genomes, Genetics*, pages g3–400493, 2019.

M. F. Danilevicz, M. Gill, R. Anderson, J. Batley, M. Bennamoun, P. E. Bayer, and D. Edwards. Plant genotype to phenotype prediction using machine learning. *Frontiers in Genetics*, 13, 2022.

G. de Los Campos, D. Gianola, G. J. Rosa, K. A. Weigel, and J. Crossa. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel hilbert spaces methods. *Genetics Research*, 92(4):295–308, 2010.

J. C. Dekkers and F. Hospital. Multifactorial genetics: The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics*, 3(1):22, 2002.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

A. Dey. Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies*, 7(3):1174–1179, 2016.

T. G. Dietterich et al. Ensemble learning. *The handbook of brain theory and neural networks*, 2:110–125, 2002.

H. Dong, R. Wang, Y. Yuan, J. Anderson, M. Pumphrey, Z. Zhang, and J. Chen. Evaluation of the potential for genomic selection to improve spring wheat resistance to fusarium head blight in the pacific northwest. *Frontiers in plant science*, 9:911, 2018.

S. Dong, P. Wang, and K. Abbas. A survey on deep learning and its applications.

Computer Science Review, 40:100379, 2021. ISSN 1574-0137. doi: https://doi.org/10.1016/j.cosrev.2021.100379. URL `https://www.sciencedirect.com/science/article/pii/S1574013721000198`.

G. Doquire and M. Verleysen. Mutual information-based feature selection for multi-label classification. *Neurocomputing*, 122:148–155, 2013.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.

S. R. Dubey, S. K. Singh, and B. B. Chaudhuri. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, 2022.

R. Edgar, M. Domrachev, and A. E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1): 207–210, 2002.

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

J. B. Endelman. Ridge regression and other kernels for genomic selection with r package rrblup. *The Plant Genome*, 4(3):250–255, 2011.

S. Esposito, D. Carputo, T. Cardi, and P. Tripodi. Applications and trends of machine learning in genomics and phenomics for next-generation breeding. *Plants*, 9(1):34, 2019.

T. Falkendal, C. Otto, J. Schewe, J. Jägermeyr, M. Konar, M. Kummu, B. Watkins, and M. J. Puma. Grain export restrictions during covid-19 risk food insecurity in many low-and middle-income countries. *Nature Food*, 2(1):11–14, 2021.

FAO. 2022 global report on food crises. *https://www.fao.org/documents/card/en/c/cb9997en/, accessed 2022-11-20*, 2022.

FAO. FAOSTAT. *http://www.fao.org/faostat/en/, accessed 2021-02-04*, 2019. publisher: Food and Agriculture Organization of the United Nations.

FAO. Fao food insecurity map. *https://www.fao.org/fileadmin/templates/SOFI/2022/docs/map-fies-print.pdf, accessed 2022-11-20*, 2021.

FAO. Policy brief—food security. *https://www.fao.org/fileadmin/templates/faoitaly/documents/pdf/pdf_Food_Security_Cocept_Note.pdf, accessed 2022-11-20*, 2006.

M. J. Faville, S. Ganesh, M. Cao, M. Jahufer, T. P. Bilton, H. Easton, D. L. Ryan, J. A. Trethewey, M. P. Rolston, A. G. Griffiths, et al. Predictive ability of genomic selection models in a multi-population perennial ryegrass training set using genotyping-by-sequencing. *Theoretical and Applied Genetics*, 131(3):703–720, 2018.

W. D. Fernando, A. O. Oghenekaro, J. R. Tucker, and A. Badea. Building on a foundation: advances in epidemiology, resistance breeding, and forecasting research for reducing the impact of fusarium head blight in wheat and barley. *Canadian Journal of Plant Pathology*, pages 1–32, 2021.

L. F. V. Ferrão, R. G. Ferrao, M. A. G. Ferrão, A. Francisco, and A. A. F. Garcia. A mixed model to multiple harvest-location trials applied to genomic prediction in coffea canephora. *Tree Genetics & Genomes*, 13(5):95, 2017.

T. Gangopadhyay, J. Shook, A. K. Singh, and S. Sarkar. Interpreting the impact of weather on crop yield using attention. 2020.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

D. Gianola, H. Okut, K. A. Weigel, and G. J. Rosa. Predicting complex quantitative traits with bayesian neural networks: a case study with jersey cows and wheat. *BMC genetics*, 12(1):87, 2011.

J. Gillberg, P. Marttinen, H. Mamitsuka, and S. Kaski. Modelling G×E with historical weather information improves genomic prediction in new environments. *Bioinformatics*, 35(20):4045–4052, 2019.

M. Goddard and B. Hayes. Genomic selection. *Journal of Animal breeding and Genetics*, 124(6):323–330, 2007.

J. González-Camacho, G. de Los Campos, P. Pérez, D. Gianola, J. Cairns, G. Mahuku,

R. Babu, and J. Crossa. Genome-enabled prediction of genetic values using radial basis function neural networks. *Theoretical and Applied Genetics*, 125(4):759–771, 2012.

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

J. M. González-Camacho, L. Ornella, P. Pérez-Rodríguez, D. Gianola, S. Dreisigacker, and J. Crossa. Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. *The plant genome*, 11(2):170104, 2018.

Government of Canada. Household food insecurity in canada: Overview. *https://www.canada.ca/en/health-canada/services/food-nutrition/food-nutrition-surveillance/health-nutrition-surveys/canadian-community-health-survey-cchs/household-food-insecurity-canada-overview.html*, *accessed 2022-11-20*, 2020.

F. Gul, I. Ahmed, M. Ashfaq, D. Jan, S. Fahad, X. Li, D. Wang, M. Fahad, M. Fayyaz, and S. A. Shah. 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 Journal of Geosciences*, 13(3):1–14, 2020.

J. Guo, J. Khan, S. Pradhan, D. Shahi, N. Khan, M. Avci, J. Mcbreen, S. Harrison, G. Brown-Guedira, J. P. Murphy, et al. Multi-trait genomic prediction of yield-related traits in us soft wheat under variable water regimes. *Genes*, 11(11):1270, 2020.

C. Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.

M. B. Hassen, J. Bartholomé, G. Valè, T.-V. Cao, and N. Ahmadi. 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: Genes, Genomes, Genetics*, 8(7): 2319–2332, 2018.

B. Hayes, M. Goddard, et al. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

X. He, M. Osman, J. Helm, F. Capettini, and P. K. Singh. Evaluation of canadian barley breeding lines for fusarium head blight resistance. *Canadian Journal of Plant Science*, 95(5):923–929, 2015.

M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

E. L. Heffner, M. E. Sorrells, and J.-L. Jannink. Genomic selection for crop improvement. *Crop Science*, 49(1):1–12, 2009.

C. Henderson. *Applications of linear models in animal breeding.* University of Guelph, Guelph, Ontario, 1984.

N. Heslot, H.-P. Yang, M. E. Sorrells, and J.-L. Jannink. Genomic selection in plant breeding: a comparison of models. *Crop science*, 52(1):146–160, 2012.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997.

A. Hoffstetter, A. Cabrera, M. Huang, and C. Sneller. Optimizing training population data and validation of genomic selection for economic traits in soft winter wheat. *G3: Genes, Genomes, Genetics*, 6(9):2919–2928, 2016.

J. A. Holliday, T. Wang, and S. Aitken. Predicting adaptive phenotypes from multilocus genotypes in sitka spruce (picea sitchensis) using random forest. *G3: Genes, Genomes, Genetics*, 2(9):1085–1093, 2012a. doi: 10.1534/g3.112.002733. URL `http://www.g3journal.org/content/2/9/1085`.

J. A. Holliday, T. Wang, and S. Aitken. Predicting adaptive phenotypes from multilocus genotypes in sitka spruce (picea sitchensis) using random forest. *G3: Genes, Genomes, Genetics*, 2(9):1085–1093, 2012b.

R. Howard, D. Gianola, O. Montesinos-López, P. Juliana, R. Singh, J. Poland, S. Shrestha, P. Pérez-Rodríguez, J. Crossa, and D. Jarquín. Joint use of genome, pedigree and their interaction with environment for predicting the performance of wheat lines in new environments. *G3: Genes, Genomes, Genetics*, pages g3–400508, 2019.

Y. Huang, M. Haas, S. Heinen, B. J. Steffenson, K. P. Smith, and G. J. Muehlbauer. Qtl mapping of fusarium head blight and correlated agromorphological traits in an elite barley cultivar rasmusson. *Frontiers in plant science*, 9:1260, 2018.

M. R. Islam, D. Fujita, S. Watanabe, and S.-H. Zheng. Variation in photosensitivity of flowering in the world soybean mini-core collections (gmwmc). *Plant Production Science*, 22(2):220–226, 2019.

M. S. Izydorczyk and M. Edney. Chapter 9 - barley: Grain-quality characteristics and management of quality requirements. In C. Wrigley, I. Batey, and D. Miskelly, editors, *Cereal Grains (Second Edition)*, Woodhead Publishing Series in Food Science, Technology and Nutrition, pages 195–234. Woodhead Publishing, second edition edition, 2017. ISBN 978-0-08-100719-8. doi: https://doi.org/10.1016/B978-0-08-100719-8.00009-7. URL `https://www.sciencedirect.com/science/article/pii/B9780081007198000097`.

J.-L. Jannink, A. J. Lorenz, and H. Iwata. Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics*, 9(2):166–177, 2010.

K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, volume 51, pages 243–250. ACM New York, NY, USA, 2017.

Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

Y. Jiang, Y. Zhao, B. Rodemann, J. Plieske, S. Kollers, V. Korzun, E. Ebmeyer, O. Argillier, M. Hinze, J. Ling, et al. Potential and limits to unravel the genetic architecture and predict the variation of fusarium head blight resistance in european winter wheat (triticum aestivum l.). *Heredity*, 114(3):318–326, 2015.

JMP Genomics. *Version 9*. SAS Institute Inc., Cary, NC, USA, 1989-2021.

A. Jović, K. Brkić, and N. Bogunović. A review of feature selection methods with

applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1200–1205. IEEE, 2015.

S. Jubair and M. Domaratzki. Crop genomic selection with deep learning and environmental data: A survey. *Frontiers in Artificial Intelligence*, 5, 2023. ISSN 2624-8212. doi: 10.3389/frai.2022.1040295. URL https://www.frontiersin.org/articles/10.3389/frai.2022.1040295.

S. Jubair and M. Domaratzki. Ensemble supervised learning for genomic selection. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1993–2000. IEEE, 2019.

S. Jubair, J. R. Tucker, N. Henderson, C. W. Hiebert, A. Badea, M. Domaratzki, and W. D. Fernando. Gptransformer: A transformer-based deep learning method for predicting fusarium related traits in barley. *Frontiers in plant science*, 12, 2021a.

S. Jubair, J. R. Tucker, N. Henderson, C. W. Hiebert, A. Badea, M. Domaratzki, and W. G. D. Fernando. Gptransformer: A transformer-based deep learning method for predicting fusarium related traits in barley. *Frontiers in Plant Science*, 12, 2021b. ISSN 1664-462X. doi: 10.3389/fpls.2021.761402. URL https://www.frontiersin.org/articles/10.3389/fpls.2021.761402.

S. Jubair, O. Tremblay-Savard, and M. Domaratzki. Gxenet: Novel fully connected neural network based approaches to incorporate gxe for predicting wheat yield. *Artificial Intelligence in Agriculture*, 8:60–76, 2023. ISSN 2589-7217. doi: https:

//doi.org/10.1016/j.aiia.2023.05.001. URL https://www.sciencedirect.com/science/article/pii/S2589721723000168.

P. Juliana, R. P. Singh, P. K. Singh, J. Crossa, J. Huerta-Espino, C. Lan, S. Bhavani, J. E. Rutkoski, J. A. Poland, G. C. Bergstrom, et al. Genomic and pedigree-based prediction for leaf, stem, and stripe rust resistance in wheat. *Theoretical and Applied Genetics*, 130(7):1415–1430, 2017.

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

S. Khaki and L. Wang. Crop yield prediction using deep neural networks. *Frontiers in plant science*, 10:621, 2019.

S. Khaki, L. Wang, and S. V. Archontoulis. A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750, 2020.

P. Kim. Convolutional neural network. In *MATLAB deep learning*, pages 121–147. Springer, 2017.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman. 1d convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, 151:107398, 2021.

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

J. Kukačka, V. Golkov, and D. Cremers. Regularization for deep learning: A taxonomy. *arXiv preprint arXiv:1710.10686*, 2017.

Q. B. Kwong, A. L. Ong, C. K. Teh, F. T. Chew, M. Tammi, S. Mayes, H. Kulaveerasingam, S. H. Yeoh, J. A. Harikrishna, and D. R. Appleton. Genomic selection in commercial perennial crops: applicability and improvement in oil palm (elaeis guineensis jacq.). *Scientific Reports*, 7(1):1–9, 2017.

N. Q. K. Le, Q.-T. Ho, V.-N. Nguyen, and J.-S. Chang. Bert-promoter: An improved sequence-based predictor of dna promoter using bert pre-trained model and shap feature selection. *Computational Biology and Chemistry*, 99:107732, 2022.

K. Le Nguyen, A. Grondin, B. Courtois, and P. Gantet. Next-generation sequencing accelerates crop gene discovery. *Trends in plant science*, 24(3):263–274, 2019.

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

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

P. Lenz, J. Beaulieu, S. D. Mansfield, S. Clément, M. Desponts, and J. Bousquet. 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(1):1–17, 2017.

B. Li, N. Zhang, Y.-G. Wang, A. W. George, A. Reverter, and Y. Li. Genomic prediction of breeding values using a subset of snps identified by three machine learning methods. *Frontiers in Genetics*, 9, 2018. ISSN 1664-8021. doi: 10.3389/fgene.2018.00237. URL https://www.frontiersin.org/articles/10.3389/fgene.2018.00237.

H. Li, S. Tian, Y. Li, Q. Fang, R. Tan, Y. Pan, C. Huang, Y. Xu, and X. Gao. Modern deep learning in bioinformatics. *Journal of Molecular Cell Biology*, 12 (11):823–827, 06 2020. ISSN 1759-4685. doi: 10.1093/jmcb/mjaa030. URL https://doi.org/10.1093/jmcb/mjaa030.

Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 2021.

M. Liang, T. Chang, B. An, X. Duan, L. Du, X. Wang, J. Miao, L. Xu, X. Gao, L. Zhang, et al. A stacking ensemble learning framework for genomic prediction. *Frontiers in genetics*, 12:600040, 2021.

T. Lin, R. Zhong, Y. Wang, J. Xu, H. Jiang, J. Xu, Y. Ying, L. Rodriguez, K. Ting, and H. Li. Deepcropnet: a deep spatial-temporal learning framework for county-level corn yield estimation. *Environmental Research Letters*, 15(3):034016, 2020.

V. Linehan, S. Thorpe, N. Andrews, and F. Beaini. Food demand to 2050: Op-

portunities for australian agriculture–algebraic description of agrifood model. In *ABARES Research Report, May 2012, Canberra*. 2012.

A. Linkmeyer, M. Götz, L. Hu, S. Asam, M. Rychlik, H. Hausladen, M. Hess, and R. Hückelhoven. Assessment and introduction of quantitative resistance to fusarium head blight in elite spring barley. *Phytopathology*, 103(12):1252–1259, 2013.

Y. Liu and D. Wang. Application of deep learning in genomic selection. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2280–2280. IEEE, 2017.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019a.

Y. Liu, D. Wang, F. He, J. Wang, T. Joshi, and D. Xu. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Frontiers in genetics*, 10:1091, 2019b.

M. Lopez-Cruz, J. Crossa, D. Bonnett, S. Dreisigacker, J. Poland, J.-L. Jannink, R. P. Singh, E. Autrique, and G. de los Campos. Increased prediction accuracy in wheat breeding trials using a marker× environment interaction genomic selection model. *G3: Genes, Genomes, Genetics*, 5(4):569–582, 2015.

D. Ly, S. Huet, A. Gauffreteau, R. Rincent, G. Touzy, A. Mini, J.-L. Jannink, F. Cormier, E. Paux, S. Lafarge, et al. Whole-genome prediction of reaction norms

to environmental stress in bread wheat (triticum aestivum l.) by genomic random regression. *Field Crops Research*, 216:32–41, 2018.

W. Ma, Z. Qiu, J. Song, J. Li, Q. Cheng, J. Zhai, and C. Ma. A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta*, 248 (5):1307–1318, Nov 2018. ISSN 1432-2048. doi: 10.1007/s00425-018-2976-9. URL `https://doi.org/10.1007/s00425-018-2976-9`.

H. Måløy, S. Windju, S. Bergersen, M. Alsheikh, and K. L. Downing. Multimodal performers for genomic selection and crop yield prediction. *Smart Agricultural Technology*, 1:100017, 2021.

C. Martin, T. Schöneberg, S. Vogelgsang, R. Morisoli, M. Bertossa, B. Mauch-Mani, and F. Mascher. Resistance against fusarium graminearum and the relationship to β-glucan content in barley grains. *European Journal of Plant Pathology*, 152(3): 621–634, 2018.

M. Mascher, H. Gundlach, A. Himmelbach, S. Beier, S. O. Twardziok, T. Wicker, V. Radchuk, C. Dockter, P. E. Hedley, J. Russell, et al. A chromosome conformation capture ordered sequence of the barley genome. *Nature*, 544(7651):427–433, 2017.

J. Massman, B. Cooper, R. Horsley, S. Neate, R. Dill-Macky, S. Chao, Y. Dong, P. Schwarz, G. Muehlbauer, and K. Smith. Genome-wide association mapping of fusarium head blight resistance in contemporary barley breeding germplasm. *Molecular breeding*, 27(4):439–454, 2011.

R. F. McCormick, S. K. Truong, J. Rotundo, A. P. Gaspar, D. Kyle, F. Van Eeuwijk,

and C. D. Messina. Intercontinental prediction of soybean phenology via hybrid ensemble of knowledge-based and data-driven models. *in silico Plants*, 3(1):diab004, 2021.

R. Mcdowell. Genomic selection with deep neural networks. Master's thesis, 2016.

R. McDowell. *Genomic selection with deep neural networks.* PhD thesis, Iowa state university, 2016.

T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, 157(4):1819–1829, 04 2001. ISSN 1943-2631. doi: 10.1093/genetics/157.4.1819. URL `https://doi.org/10.1093/genetics/157.4.1819`.

E. J. Millet, W. Kruijer, A. Coupel-Ledru, S. A. Prado, L. Cabrera-Bosquet, S. Lacube, A. Charcosset, C. Welcker, F. van Eeuwijk, and F. Tardieu. Genomic prediction of maize yield across european environmental conditions. *Nature genetics*, 51(6):952–956, 2019.

P. J. Milton. *Breeding field crops.* 1979.

V. Mirdita, S. He, Y. Zhao, V. Korzun, R. Bothe, E. Ebmeyer, J. C. Reif, and Y. Jiang. Potential and limits of whole genome prediction of resistance to fusarium head blight and septoria tritici blotch in a vast central european elite winter wheat population. *Theoretical and Applied Genetics*, 128(12):2471–2481, 2015.

N. R. Monteiro, J. L. Oliveira, and J. P. Arrais. Dtitr: End-to-end drug–target

binding affinity prediction with transformers. *Computers in Biology and Medicine*, 147:105772, 2022.

A. Montesinos-López, O. A. Montesinos-López, D. Gianola, J. Crossa, and C. M. Hernández-Suárez. Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3: Genes, Genomes, Genetics*, 8(12):3813–3828, 2018a. doi: 10.1534/g3.118.200740.

A. Montesinos-López, O. A. Montesinos-López, D. Gianola, J. Crossa, and C. M. Hernández-Suárez. Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3: Genes, Genomes, Genetics*, 8(12):3813–3828, 2018b.

O. A. Montesinos-López, J. Martín-Vallejo, J. Crossa, D. Gianola, C. M. Hernández-Suárez, A. Montesinos-López, P. Juliana, and R. Singh. A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding. *G3: Genes, Genomes, Genetics*, 9(2):601–618, 2019a. doi: 10.1534/g3.118.200998.

O. A. Montesinos-López, J. Martín-Vallejo, J. Crossa, D. Gianola, C. M. Hernández-Suárez, A. Montesinos-López, P. Juliana, and R. Singh. New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes. *G3: Genes, Genomes, Genetics*, 2019b. doi: 10.1534/g3.119.300585.

O. A. Montesinos-López, A. Montesinos-López, R. Tuberosa, M. Maccaferri, G. Sciara, K. Ammar, and J. Crossa. Multi-trait, multi-environment genomic

prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods. *Frontiers in Plant Science*, 10:1311, 2019c.

O. A. Montesinos-López, A. Montesinos-López, P. Pérez-Rodríguez, J. A. Barrón-López, J. W. Martini, S. B. Fajardo-Flores, L. S. Gaytan-Lugo, P. C. Santana-Mancilla, and J. Crossa. A review of deep learning applications for genomic selection. *BMC genomics*, 22(1):1–23, 2021.

O. A. Montesinos-Lopez, J. C. Montesinos-Lopez, E. Salazar, J. A. Barron, A. Montesinos-Lopez, R. Buenrostro-Mariscal, and J. Crossa. Application of a poisson deep neural network model for the prediction of count data in genome-based prediction. *The Plant Genome*, 14(3):e20118, 2021.

O. A. Montesinos-López, A. Montesinos-López, Kismiantini, A. Roman-Gallardo, K. Gardner, M. Lillemo, R. Fritsche-Neto, and J. Crossa. Partial Least Squares Enhances Genomic Prediction of New Environments. *Frontiers in Genetics*, 13, 2022. ISSN 1664-8021. URL `https://www.frontiersin.org/articles/10.3389/fgene.2022.920689`.

D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

C. E. Moore, K. Meacham-Hensold, P. Lemonnier, R. A. Slattery, C. Benjamin, C. J. Bernacchi, T. Lawson, and A. P. Cavanagh. The effect of increasing temperature on crop photosynthesis: from enzymes to ecosystems. *Journal of Experimental Botany*, 72(8):2822–2844, 2021.

R. Moradi, R. Berangi, and B. Minaei. A survey of regularization strategies for deep models. *Artificial Intelligence Review*, 53(6):3947–3986, 2020.

J. Moreno-Amores, S. Michel, T. Miedaner, C. F. H. Longin, and H. Buerstmayr. Genomic predictions for fusarium head blight resistance in a diverse durum wheat panel: An effective incorporation of plant height and heading date as covariates. *Euphytica*, 216(2):1–19, 2020.

M. A. Nawaz and G. Chung. Genetic improvement of cereals and grain legumes. *Genes*, 11(11), 2020. ISSN 2073-4425. doi: 10.3390/genes11111255. URL `https://www.mdpi.com/2073-4425/11/11/1255`.

H. Oakey, B. Cullis, R. Thompson, J. Comadran, C. Halpin, and R. Waugh. Genomic selection in multi-environment crop trials. *G3: Genes, Genomes, Genetics*, 6(5): 1313–1326, 2016. doi: 10.1534/g3.116.027524. URL `http://www.g3journal.org/content/6/5/1313`.

J. O. Ogutu, H.-P. Piepho, and T. Schulz-Streeck. A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proceedings*, 5 (3):S11, May 2011a. ISSN 1753-6561. doi: 10.1186/1753-6561-5-S3-S11. URL `https://doi.org/10.1186/1753-6561-5-S3-S11`.

J. O. Ogutu, H.-P. Piepho, and T. Schulz-Streeck. A comparison of random forests, boosting and support vector machines for genomic selection. In *BMC proceedings*, volume 5, page S11, 2011b.

O. A. Otekunrin, O. A. Otekunrin, S. Momoh, and I. A. Ayinde. How far has africa

gone in achieving the zero hunger target? evidence from nigeria. *Global Food Security*, 22:1–12, 2019.

O. A. Otekunrin, O. A. Otekunrin, B. Sawicka, and I. A. Ayinde. Three decades of fighting against hunger in africa: Progress, challenges and opportunities. *World Nutrition*, 11(3):86–111, 2020.

A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

P. Pérez and G. de Los Campos. Genome-wide regression and prediction with the bglr statistical package. *Genetics*, 198(2):483–495, 2014.

S. Pérez-Elizalde, J. Cuevas, P. Pérez-Rodríguez, and J. Crossa. Selection of the bandwidth parameter in a Bayesian kernel regression model for genomic-enabled prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, 20(4): 512–532, 2015.

M. Pérez-Enciso and L. M. Zingaretti. A guide for using deep learning for complex trait genomic prediction. *Genes*, 10(7):553, 2019.

P. Pérez-Rodríguez, D. Gianola, J. M. González-Camacho, J. Crossa, Y. Manès, and

S. Dreisigacker. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3: Genes, Genomes, Genetics*, 2(12): 1595–1605, 2012.

J. J. Pestka. Deoxynivalenol: mechanisms of action, human exposure, and toxicological relevance. *Archives of toxicology*, 84(9):663–679, 2010.

J. Poland, J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sánchez-Villeda, M. Sorrells, et al. Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome*, 5(3):103–113, 2012.

J. Poland, S. Dreisigacker, S. Shrestha, S. Wu, R. Singh, S. Mondal, P. Juliana, J. Crossa, B. R. Basnet, L. Crespo, and et al. Genotypic data from cimmyt bread wheat breeding lines used in the feed the future innovation lab for applied wheat genomics, Oct 2021. URL `https://hdl.handle.net/11529/10695`.

T. Pook, J. Freudenthal, A. Korte, and H. Simianer. Using local convolutional neural networks for genomic prediction. *Frontiers in genetics*, 11:561497, 2020.

A. R. Putra, J. D. Yen, and A. Fournier-Level. Forecasting trait responses in novel environments to aid seed provenancing under climate change. *Molecular Ecology Resources*, 2022.

M. Pérez-Enciso and L. M. Zingaretti. A guide on deep learning for complex trait genomic prediction. *Genes*, 10(7), 2019. ISSN 2073-4425. doi: 10.3390/genes10070553. URL `https://www.mdpi.com/2073-4425/10/7/553`.

X. Qing and Y. Niu. Hourly day-ahead solar irradiance prediction using weather forecasts by lstm. *Energy*, 148:461–468, 2018.

J. R. Quinlan et al. Bagging, boosting, and c4. 5. In *AAAI/IAAI, Vol. 1*, pages 725–730, 1996.

H. Rachmatia, W. Kusuma, and L. Hasibuan. Prediction of maize phenotype based on whole-genome single nucleotide polymorphisms using deep belief networks. In *Journal of Physics: Conference Series*, volume 835, page 012003. IOP Publishing, 2017a.

H. Rachmatia, W. A. Kusuma, and L. S. Hasibuan. Prediction of maize phenotype based on whole-genome single nucleotide polymorphisms using deep belief networks. *Journal of Physics: Conference Series*, 835:012003, May 2017b. doi: 10.1088/1742-6596/835/1/012003.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

H. Rahman, R. A. Bennett, and B. Kebede. Molecular mapping of qtl alleles of brassica oleracea affecting days to flowering and photosensitivity in spring brassica napus. *PLoS One*, 13(1):e0189723, 2018.

J. Ranstam and J. Cook. Lasso regression. *Journal of British Surgery*, 105(10): 1348–1348, 2018.

P. Refaeilzadeh, L. Tang, and H. Liu. Cross-validation. *Encyclopedia of database systems*, 5:532–538, 2009.

S. Rezayi, Z. Liu, Z. Wu, C. Dhakal, B. Ge, C. Zhen, T. Liu, and S. Li. Agribert: Knowledge-infused agricultural language models for matching food and nutrition. In L. D. Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5150–5156. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/715. URL `https://doi.org/10.24963/ijcai.2022/715`. AI for Good.

R. Rincent, E. Kuhn, H. Monod, F.-X. Oury, M. Rousset, V. Allard, and J. Le Gouis. Optimization of multi-environment trials for genomic selection based on crop models. *Theoretical and Applied Genetics*, 130(8):1735–1752, 2017.

M. Roorkiwal, D. Jarquin, M. K. Singh, P. M. Gaur, C. Bharadwaj, A. Rathore, R. Howard, S. Srinivasan, A. Jain, V. Garg, et al. Genomic-enabled prediction models using multi-environment trials to estimate the effect of genotype× environment interaction on prediction accuracy in chickpea. *Scientific reports*, 8(1):11701, 2018.

B. C. Ross. Mutual information between discrete and continuous data sets. *PloS one*, 9(2):e87357, 2014.

J. Rudd, R. Horsley, A. McKendry, and E. Elias. Host plant resistance genes for fusarium head blight: sources, mechanisms, and utility in conventional breeding systems. *Crop Science*, 41(3):620–627, 2001.

S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

D. Runcie and H. Cheng. Pitfalls and remedies for cross validation with multi-trait genomic prediction methods. *G3: Genes, Genomes, Genetics*, 9(11):3727–3741, 2019.

J. Rutkoski, J. Benson, Y. Jia, G. Brown-Guedira, J.-L. Jannink, and M. Sorrells. Evaluation of genomic prediction methods for fusarium head blight resistance in wheat. *The Plant Genome*, 5(2), 2012.

Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

A. H. Sallam and K. P. Smith. Genomic selection performs similarly to phenotypic selection in barley. *Crop Science*, 56(6):2871–2881, 2016.

A. G. Salman, Y. Heryadi, E. Abdurahman, and W. Suparta. Single layer & multi-layer long short-term memory (lstm) model with intermediate variables for weather forecasting. *Procedia Computer Science*, 135:89–98, 2018.

M. Salmerón and L. C. Purcell. Simplifying the prediction of phenology with the dssat-cropgro-soybean model based on relative maturity group and determinacy. *Agricultural Systems*, 148:178–187, 2016.

K. Sandhu, S. S. Patil, M. Pumphrey, and A. Carter. Multitrait machine-and deep-learning models for genomic selection using spectral information in a wheat breeding program. *The Plant Genome*, 14(3):e20119, 2021a.

K. S. Sandhu, M. Aoun, C. F. Morris, and A. H. Carter. Genomic selection for end-use quality and processing traits in soft white winter wheat breeding program with machine and deep learning models. *Biology*, 10(7):689, 2021b.

K. S. Sandhu, D. N. Lozada, Z. Zhang, M. O. Pumphrey, and A. H. Carter. Deep learning for predicting complex traits in spring wheat breeding program. *Frontiers in Plant Science*, 11:613325, 2021c.

K. S. Sandhu, S. S. Patil, M. Aoun, and A. H. Carter. Multi-trait multi-environment genomic prediction for end-use quality traits in winter wheat. *Frontiers in genetics*, page 41, 2022.

S. Sawitri, N. Tani, M. Na'iem, W. Widiyatno, S. Indrioko, K. Uchiyama, R. Suwa, K. K. S. Ng, S. L. Lee, and Y. Tsumura. 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 (2):239, 2020.

C. C. Schön, H. F. Utz, S. Groh, B. Truberg, S. Openshaw, and A. E. Melchinger. Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. *Genetics*, 167(1):485–498, 2004.

T. Searchinger, R. Waite, C. Hanson, J. Ranganathan, P. Dumas, E. Matthews, and C. Klirs. Creating a sustainable food future: A menu of solutions to feed nearly 10 billion people by 2050. final report. *https://agritrop.cirad.fr/593176/1/WRR_Food_Full_Report_0.pdf, accessed 2022-12-12*, 2019.

A. Sharma, A. Jain, P. Gupta, and V. Chowdary. Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, 9:4843–4873, 2020.

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

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

S. Shrestha, F. Asch, J. Dusserre, A. Ramanantsoanirina, and H. Brueck. Climate effects on yield components as affected by genotypic responses to variable environmental conditions in upland rice systems at different altitudes. *Field Crops Research*, 134:216–228, 2012.

A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

K. P. Smith, A. Budde, R. Dill-Macky, D. Rasmusson, E. Schiefelbein, B. Steffenson, J. Wiersma, J. Wiersma, and B. Zhang. Registration of 'quest'spring malting barley

with improved resistance to fusarium head blight. *Journal of Plant Registrations*, 7(2):125–129, 2013.

H. Song, X. Wang, Y. Guo, and X. Ding. G × EBLUP: A novel method for exploring genotype by environment interactions and genomic prediction. *Frontiers in Genetics*, 13, 2022. ISSN 1664-8021. URL `https://www.frontiersin.org/articles/10.3389/fgene.2022.972557`.

G. Sonkar, R. Mall, T. Banerjee, N. Singh, T. L. Kumar, and R. Chand. Vulnerability of indian wheat against rising temperature and aerosols. *Environmental Pollution*, 254:112946, 2019.

J. E. Spindel and S. R. McCouch. When more is better: how data sharing would accelerate genomic selection of crop plants. *New Phytologist*, 212(4):814–826, 2016.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

B. Steiner, S. Michel, M. Maccaferri, M. Lemmens, R. Tuberosa, and H. Buerstmayr. Exploring and exploiting the genetic variation of fusarium head blight resistance for genomic-assisted breeding in the elite durum wheat gene pool. *Theoretical and Applied Genetics*, 132(4):969–988, 2019.

R.-Y. Sun. Optimization for deep learning: An overview. *Journal of the Operations Research Society of China*, 8(2):249–294, 2020.

T. Szandała. Review and comparison of commonly used activation functions for deep neural networks. In *Bio-inspired Neurocomputing*, pages 203–224. Springer, 2021.

C. R. Tacarindua, T. Shiraiwa, K. Homma, E. Kumagai, and R. Sameshima. The effects of increased temperature on crop growth and yield of soybean grown in a temperature gradient chamber. *Field Crops Research*, 154:74–81, 2013.

W. Tadesse, Y. Manes, R. Singh, T. Payne, and H. Braun. Adaptation and performance of cimmyt spring wheat genotypes targeted to high rainfall areas of the world. *Crop science*, 50(6):2240–2248, 2010.

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

J. Tang, S. Alelyani, and H. Liu. Feature selection for classification: A review. *Data classification: Algorithms and applications*, page 37, 2014.

X. Tang, G. Liu, J. Zhou, Q. Ren, Q. You, L. Tian, X. Xin, Z. Zhong, B. Liu, X. Zheng, et al. A large-scale whole-genome sequencing analysis reveals highly specific genome editing by both cas9 and cpf1 (cas12a) nucleases in rice. *Genome Biology*, 19(1):1–13, 2018.

V. Tarasuk, A.-A. F. St-Germain, and T. Li. Moment of reckoning for household food insecurity monitoring in canada. *Health Promotion and Chronic Disease Prevention in Canada: Research, Policy and Practice*, 42(10):445, 2022.

S. Tittlemier, J. Brunkhorst, B. Cramer, M. DeRosa, V. Lattanzio, R. Malone,

C. Maragos, M. Stranska, and M. Sumarah. Developments in mycotoxin analysis: an update for 2019-2020. *World Mycotoxin Journal*, 14(1):3–26, 2021.

H. Tong and Z. Nikoloski. Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *Journal of plant physiology*, 257:153354, 2021.

K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas. Multi-label classification of music into emotions. In *ISMIR*, volume 8, pages 325–330, 2008.

J. R. Tucker, A. Badea, R. Blagden, K. Pleskach, S. A. Tittlemier, and W. Fernando. Deoxynivalenol-3-glucoside content is highly associated with deoxynivalenol levels in two-row barley genotypes of importance to canadian barley breeding programs. *Toxins*, 11(6):319, 2019.

J. Ubbens, I. Parkin, C. Eynck, I. Stavness, and A. G. Sharpe. Deep neural networks for genomic prediction do not estimate marker effects. *The Plant Genome*, 14(3): e20147, 2021.

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

United Nations. Department of Economic and Social Affairs — Food security and nutrition and sustainable agriculture. *https://sdgs.un.org/topics/food-security-and-nutrition-and-sustainable-agriculture, accessed 2021-02-04*, 2019.

A. D. J. van Dijk, G. Kootstra, W. Kruijer, and D. de Ridder. Machine learning in plant science and plant breeding. *Iscience*, 24(1):101890, 2021.

H. Van Meijl, P. Havlik, H. Lotze-Campen, E. Stehfest, P. Witzke, I. P. Domínguez, B. L. Bodirsky, M. van Dijk, J. Doelman, T. Fellmann, et al. Comparing impacts of climate change and mitigation on global agriculture by 2050. *Environmental Research Letters*, 13(6):064021, 2018.

R. K. Varshney, M. Roorkiwal, and M. E. Sorrells. Genomic selection for crop improvement: An introduction. In *Genomic Selection for Crop Improvement*, pages 1–6. Springer, 2017.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

S. Wang, J. Wei, R. Li, H. Qu, J. M. Chater, R. Ma, Y. Li, W. Xie, and Z. Jia. Identification of optimal prediction models using multi-omic data for selecting hybrid rice. *Heredity*, 123(3):395–406, 2019.

X. Wang, Y. Xu, Z. Hu, and C. Xu. Genomic selection methods for crop improvement: Current status and prospects. *The Crop Journal*, 6(4):330–340, 2018.

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

J. D. Washburn, E. Cimen, G. Ramstein, T. Reeves, P. O'Briant, G. McLean, M. Cooper, G. Hammer, and E. S. Buckler. Predicting phenotypes from genetic, environment, management, and historical data using cnns. *Theoretical and Applied Genetics*, 134(12):3997–4011, 2021.

C. C. Westhues, H. Simianer, and T. M. Beissinger. learnmet: an r package to apply machine learning methods for genomic prediction using multi-environment trial data. *G3*, 12(11):jkac226, 2022.

WFP. A global food crisis — conflict, covid, the climate crisis and rising costs have combined in 2022 to create jeopardy for up to 828 million hungry people across the world. *https://www.wfp.org/global-hunger-crisis, accessed 2022-11-20*, 2022.

W. WFP, UNICEF, et al. The state of food security and nutrition in the world 2022. 2022.

C. Wong and R. Bernardo. Genomewide selection in oil palm: Increasing selection gain per unit time and cost with small populations. *Theoretical and Applied Genetics*, 116(6):815–824, 2008.

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

Y. Xu, X. Zhang, H. Li, H. Zheng, J. Zhang, M. S. Olsen, R. K. Varshney, B. M.

Prasanna, and Q. Qian. Smart breeding driven by big data, artificial intelligence and integrated genomic-enviromic prediction. *Molecular Plant*, 2022.

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

Y. Yu, X. Si, C. Hu, and J. Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019b.

J. C. Zadoks, T. T. Chang, C. F. Konzak, et al. A decimal code for the growth stages of cereals. *Weed research*, 14(6):415–421, 1974.

M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

W. A. Zegeye, Y. Zhang, L. Cao, and S. Cheng. Whole genome resequencing from bulked populations as a rapid qtl and gene identification method in rice. *International Journal of Molecular Sciences*, 19(12):4000, 2018.

H. Zhang, X. Wang, Q. Pan, P. Li, Y. Liu, X. Lu, W. Zhong, M. Li, L. Han, J. Li, et al. Qtg-seq accelerates qtl fine mapping through qtl partitioning and whole-genome sequencing of bulked segregant samples. *Molecular plant*, 12(3):426–437, 2019.

N. Zhang, S. Ding, J. Zhang, and Y. Xue. An overview on restricted boltzmann machines. *Neurocomputing*, 275:1186–1199, 2018.

R. Zhong, Y. Zhu, X. Wang, H. Li, B. Wang, F. You, L. F. Rodríguez, J. Huang, K. Ting, Y. Ying, et al. Detect and attribute the extreme maize yield losses based on spatio-temporal deep learning. *Fundamental Research*, 2022.

H. Zhu, L. Gilchrist, P. Hayes, A. Kleinhofs, D. Kudrna, Z. Liu, L. Prom, B. Steffenson, T. Toojinda, and H. Vivar. Does function follow form? principal qtls for fusarium head blight (fhb) resistance are coincident with qtls for inflorescence traits and plant height in a doubled-haploid population of barley. *Theoretical and Applied Genetics*, 99(7-8):1221–1232, 1999.

L. M. Zingaretti, S. A. Gezan, L. F. V. Ferrão, L. F. Osorio, A. Monfort, P. R. Muñoz, V. M. Whitaker, and M. Pérez-Enciso. Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Frontiers in plant science*, 11: 25, 2020.