Predicting Phenotypes From Novel Genomic Markers Using Deep Learning

A thesis submitted to the
College of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Shivani Sehrawat

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building, 110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5C9 Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

Genomic selection (GS) is a powerful method concerned with predicting the phenotypes of individuals from genome-wide markers to select candidates for the next breeding cycle. Previous studies in GS have used single nucleotide polymorphism (SNP) markers to predict phenotypes using conventional statistical or deep learning models. However, these predictive models face challenges due to the high dimensionality of genome-wide SNP marker data and interactions between alleles. Thanks to recent breakthroughs in DNA sequencing and decreased sequencing cost, the study of novel genomic variants such as structural variations (SVs) and transposable elements (TEs) became increasingly prevalent. Here, we present a one-dimensional deep convolutional neural network, `NovGMDeep`, to predict phenotypes using novel genomic markers, such as SVs and TEs. The model is designed to use novel genomic markers to reduce the curse of dimensionality of the SNP genotypic data for GS. The proposed model is trained and tested on the samples of *Arabidopsis thaliana* and *Oryza sativa* using 3-fold cross-validation. The prediction accuracy is evaluated using Pearson's Correlation Coefficient (PCC), Mean Absolute Error (MAE), and Standard Deviation (SD) of MAE on the testing sets. The predicted results showed a higher correlation when the model is trained with SVs and TEs than SNPs. `NovGMDeep` also has higher prediction accuracy when compared with conventional statistical models. We also included an extended study which describes sample size effects when the proposed model is trained on different number of samples for SVs. The results show better PCC values when the model was trained on more than 700 samples. This work sheds light on the unrecognized function of SVs and TEs in genotype-to-phenotype associations, as well as their extensive significance and value in crop development. Moreover, the predictions identified here using SVs and TEs will be useful to investigate the evolution and trait architecture of *A. thaliana* and *O. sativa*.

Acknowledgements

I would like to acknowledge and give my warmest thanks to my supervisor, Dr. Lingling Jin, who made this work possible. Her invaluable advice, continuous support, and guidance carried me through all the stages of my thesis. Her immense knowledge and plentiful experience have encouraged me all the time in my academic research and daily life.

I want to express my thanks to the committee members, Drs. Kusalik, Bett and Liu, for their comments and suggestions.

I would also like to thank my lab mates for a cherished time spent together in the lab, specifically Keyhan Najafian, who helped me brainstorm ideas for the development of the model and writing research papers.

My appreciation also goes out to my family and friends for their encouragement and support all through my studies.

Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
CBN	Cauline Axillary Branch Number
CNN	Convolutional Neural Network
CNV	Copy Number Variants
DBF	Deep Belief Neural Network
DEL	Deletions
DL	Deep Learning
DTH	Days To Heading
DUP	Duplications
FT	Flowering Time
G-BLUP	Genomic Best Linear Unbiased Prediction
GEBV	Genomic-Estimated Breeding Value
GS	Genomic Selection
GWAS	Genome Wide Association Studies
INDEL	Insertion or Deletion
INV	Inversions
LD	Linkage Disequlibrium
LINE	Long Interspersed Nuclear Elements
LTR	Long Terminal Repeat
MAE	Mean Absolute Error
MITE	Miniature Inverted repeat Transposable Element
ML	Machine Learning
MLP	Multi Layer Perceptron
PCC	Pearson's Correlation Coefficient
PH	Plant Height
PL	Panicle Length
ReLU	Rectified Linear Unit
RFLP	Restriction Fragment Length Polymorphisms
RLN	Rosette Leaf Number
RNN	Recurrent Neural Network
RR-BLUP	Ridge Regression Best Linear Unbiased Prediction
SD	Standard Deviation
SINE	Short Interspersed Nuclear Elements

SL	Stem Length
SNP	Single Nucleotide Polymorphism
SN	Spikelet Number
STR	Short Tandem Repeat
SV	Structural Variation
TE	Transposable Element
VNTR	Varying Number of Tandem Repeats

1 Introduction

1.1 Research Problem

The genotype of an organism is described by its entire genetic composition. The phrase can also refer to the set of alleles contained within the genome. In diploid organisms, an individual inherits two alleles at each genetic locus, with one allele inherited from each parent. Each pair of alleles represents the genotype of a specific gene. The genotype is expressed when information encoded in the DNA of genes is utilized to generate RNA and protein molecules. The phenotype, or observable characteristic, is the detectable expression of a genotype. Generally, three variables influence a phenotype: genes are the most effective; inherited epigenetic and acquired environmental factors are the other two [122].

Predicting phenotypes of a crop is an essential step in explaining crop behavior using insights from genome-wide markers such as genome-wide Single Nucleotide Polymorphisms (SNPs). Genomic selection (GS) is a powerful method concerned with predicting the phenotypes from genotypes using the genomic-estimated breeding values (GEBVs) of individuals to select candidates for the next breeding cycle [82]. The GEBV is determined by summing the impacts of genetic markers across the entire genome [52]. The information from the large number of markers dispersed throughout the genome can be utilized to capture variability in that genome, which is adequate to determine the breeding qualities. The idea behind GS is to estimate the phenotypic values of an unseen population using genome-wide markers and phenotypic data from an observed population. GS is beneficial when it is expensive or difficult to measure the phenotype or where the breeding cycle is long (e.g., tree species).

Complex traits are traits that are controlled by multiple genes and their interactions. Each gene can have a minuscule effect on the trait, and it can be challenging to identify which genes are important when dealing with huge genomic data. In addition, environmental factors can also affect complex traits. For example, temperature, water, humidity and nutrition can affect plant growth and development. The need to predict complex traits from large amounts of genomic data has given rise to attempts to develop predictive statistical and machine learning models. Various genomic prediction models and algorithms include least absolute shrinkage and selection operator (LASSO) [116]; Bayesian-based algorithms, such as Bayes A, Bayes B, Bayes C, and Bayes Cπ [47]; and BLUP (best linear unbiased prediction)-based algorithms, such as ridge regression BLUP (rrBLUP) [32] and genomic BLUP (gBLUP) [20]. These methods for predicting phenotypes using genotypes are typically based on regression analysis. The predictor variables are the genotypes, and the response variable is the phenotype. The coefficients of the regression model are estimated using data

from a reference population of individuals with known genotypes and phenotypes. The model can then be used to predict the phenotype of an individual with its genotype.

However, these conventional predictors typically assume that the random effect of genotype follows a normal distribution, and the contribution of each genotype to phenotypes is considered as an independent attribute. Whereas it is unknown in practice how genotype effects behave. Moreover, these conventional statistical models do not consider non-linearity between the variables. For example, SNPs may also interact with other SNPs to cause complex diseases or traits due to epistasis. Also while these models face challenges due to the high dimensionality of genome-wide marker data and interactions between alleles, GS can benefit from Deep Learning (DL).

DL provides novel approaches to process noisy data [70] and handle nonlinearity [60]. It can capture complex relationships within genotypes and between genotypes and phenotypes. DL models show their favorability and superiority over traditional statistical models through their ability to form convoluted and non-linear links. These complex links form relationships among genomic parameters by reducing the rigorous assumptions made by conventional statistical models [35]. These DL models inspired by GS models can consider the non-additive effects in data and can instantly learn from the interconnected links in the dataset. Studies have shown that these methods base their predictions on overall genetic relatedness rather than the effects of particular markers [115]. Moreover, DL models offer feature extraction capabilities for leaf [71] and disease [114] detection in plants and therefore have the potential to capture hidden patterns in large-scale datasets [85]. Also, the most prominent advantage of DL is the high capacity and flexible trainable parameters. Deep learning-based models have shown promising results in a wide variety of tasks such as object detection [130], speech recognition [27], natural language processing [126], etc.

The fact that deep neural networks are primarily black-boxes is one limitation of deep learning. Once a network has been trained, it can be challenging to understand why the model predicts what it does or what input components are important to the output. There are software libraries to resolve the blackbox characteristic of the neural networks. AutoWeka is one example, designed to provide automatic model selection and hyperparameter optimization [111]. Neural-symbolic computation offers another way of implementing white-box nonlinear prediction by means of rule extraction [12].

In addition to SNPs, there are other larger-scale genomic variants. Structural variations (SVs) include variations in the number of chromosomes, the size and shape of chromosomes, and the location of genes. Transposable elements (TEs) include mobile genetic variations that can move between different chromosomes in a genome. Both SVs and TEs are highly abundant and active in plants. Such novel genomic markers could significantly contribute more to plant phenotypic diversity than SNPs. Also, using SVs and TEs will help in reducing the curse of dimensionality of the SNP genotypic data for GS. However, there need to be methods predicting phenotypes using these genomic markers. In this thesis, we developed a deep learning model, `NovGMDeep`, to predict the different phenotypes from novel genomic markers- SVs and TEs, as well as SNPs. This work highlights the role of genomic structural variants and transposable elements

in predicting phenotypes in plants using deep learning. The developed model differs from other statistical models as it spontaneously learns the complex relationships among the genome-wide markers and phenotypic traits from the training data. With advanced deep learning technology, the model avoids the overfitting of the data by using the convolutional, pooling, and dropout layers hence decreasing the complexity of many genomic markers. The predicted results showed a higher correlation when the model is trained with SVs and TEs than with SNPs. `NovGMDeep` also has higher prediction accuracy when compared with conventional statistical models. The genotype-to-phenotype predictions identified here using SVs and TEs will be helpful to investigate *Arabidopsis thaliana* and *Oryza sativa* evolution and trait architecture. This work sheds light on the underappreciated function of SVs and TEs in genotype-to-phenotype associations, as well as their extensive significance and value in crop development.

This work has been presented at ISCB Latin America international conference on Bioinformatics as a full paper. As the first author of the paper, I contributed in looking for dataset, methodology design, implementation, verifying results, and manuscript writing. Our co-author, Keyhan Najafian, contributed in model development and proofreading the manuscript.

- Shivani Sehrawat, Keyhan Najafian, Lingling Jin, 'Predicting Phenotypes from Novel Genomic Markers using Deep Learning', *ISCB Latin America SoIBio BioNetMX*, Queretaro, Mexico, November 3-7 (2022).

1.2 Layout of the Thesis

The thesis is organized as follows: Chapter 1 gives a clear view of the research problem and the encountered challenges. Chapter 2 explains the motivations and objectives behind the work in the thesis. In Chapter 3, we provide the necessary background knowledge to understand our work. Chapter 4 contains the methodology used in the thesis. In Chapter 5 and Chapter 6, we describe the two phenotype prediction case studies and also discuss the results of each experiment. In Chapter 7, we summarize the thesis and discuss potential future works.

2 Motivation and Objectives

In this chapter, we will discuss the importance of genomic prediction as well as the key objectives of the thesis.

2.1 Motivation

Strategies to ensure a consistent food supply for the expanding human population, which is anticipated to reach 9.5 billion by 2050, heavily rely on plant breeding [53]. Plant breeding must achieve the best rates of genetic improvement to optimize its contribution to raising agricultural production if it is to keep up with the anticipated rise in food demand in the upcoming years [98]. Utilizing the possibilities of cutting-edge techniques is a crucial first step in this situation. Genomic selection is a breeding technique that helps breeders find improved breeding solutions. GS is a predictive methodology that involves training models with a reference population that comprises known phenotypic (output) and genotypic (input) data to make predictions for an unseen population. This predictive approach can also reduce the time and cost associated with traditional breeding methods.

The most popular choice of genotypic data, known as genomic markers, is SNPs. However, several difficulties must be resolved or overcome when employing SNPs. For instance, SNPs are often scattered across the genome and have low minor allele frequencies, making it difficult to detect an association between a SNP and a phenotype. Also, SNPs can be in linkage disequilibrium with each other, making it difficult to interpret the results of a SNP association study. Another difficulty is working with polyploid crops, where the relevant SNPs only comprise a small portion of the overall polymorphisms [113]. That is why there is an increasing need to use novel genomic markers that can capture the genome variation better than SNPs.

Non-SNP DNA variation, such as SVs and TEs, accounts for a smaller number of variation events compared to SNPs. However, they involve a much more significant proportion of overall variant bases and could significantly contribute to plant phenotypic diversity. Some examples regarding the impacts of SVs in plants regarding gene regulation towards environmental fitness are: SVs may contribute to biotic [29] and abiotic stress [40], heterosis in maize [67], and increased chlorophyll content in *Brassica napus* [99]. According to a PanSV genome analysis of 100 tomato accessions, SVs altered gene dosage and expression levels, which led to quantitative trait variations in fruit flavor, size, and yield [7].

Genotyping is the method of identifying an individual's genotype by analyzing its DNA sequence through biological experiments and comparing it to another individual or a reference sequence to determine which of

the two alleles is more common in a population. The final product of genotyping is a report that lists the alleles from most to least common. Currently, the most widely used genotyping platforms for SNP calling and validation are KBiosciences' Competitive Allele Specific PCR (KASPar) coupled with the SNP Line platform, OpenArray platform, and Illumina's BeadArray-based Golden Gate [34]. These current genotyping assays and platforms vary from one another in terms of their chemistry, price, the throughput of genotyped samples, and the number of SNPs reported. The length of the SNP context sequence and the total number of SNPs to genotype influence the choice of chemistry and genotyping platform. Most of these chemistries are still expensive. On the other hand, the widespread detection and genotyping of SVs and TEs are now more feasible than ever, thanks to recent developments in molecular and statistical approaches. Numerous computational and experimental techniques have been developed during the past years; most of them concentrate on cost-effective and high-quality structural variations identifications [54], [106], [55].

Deep learning is a potent technology with a lot of potential for discovering relationships between genotype and phenotype. Plant breeding programs will gain from the successful use of deep learning in phenotype prediction since it will enable more accurate phenotype predictions from genomic variations. This allows breeders to develop and breed a more focused selection of lines by better anticipating the lines from their breeding stock with the best phenotypic potential for the trait of interest. If genetic data can be utilized effectively to forecast crop production or crop health, the agricultural industry can better predict and optimize yields. The global advances in agricultural products that have resulted from previous genetic changes could be accelerated by building better prediction models [109]. Deep learning models such as convolutional neural networks (CNNs) have different convolutional layers designed to detect data patterns.

For phenotype prediction challenges, CNN can extract spatial information from genomic variations. In CNNs, the interactions between nearby genomic markers can be taken into account within various ranges of the kernel window throughout the convolving process [77]. The primary motivation for using DL models for GS is that they learn the pattern from the data without any prior assumption; in this way, they include all the variations and their interactions.

2.2 Objectives of the thesis

The objectives of this thesis can be divided into five categories.

1. Provide a literature survey on genomic selection.
2. Highlight the role of novel genomic markers in the prediction of phenotypes in plants which is not just limited to SNPs.
3. Present the proposed DL model, its architecture, and optimization strategies for the prediction of phenotypes.
4. Compare the performance of the newly proposed DL model with other classical GS models.

5. Compare the architecture and performance of other DL methods with the proposed method.

3 Literature Review

In this chapter, we will provide the basic knowledge of the different genomic variances, a summary of deep learning techniques, a survey of genomic selection, theoretical concepts behind the different conventional statistical and deep learning methods for predicting traits.

3.1 Genomic Variances

Genomic variance is the difference in the DNA sequences of two or more individuals or groups [94]. Genomic variances can be categorized into four types, namely genetic drift, selection, genomic mutation, and genetic variation [63]. Genetic drift is a random change in the frequencies of alleles in a population over time [17]. It is more likely to occur in small populations, and can lead to the loss of alleles from a population altogether. Selection is cause when the frequencies of alleles change in a non-random way [13]. Genomic mutation is defined as a change in the DNA sequence that is passed on to the offspring. The most common type of genomic variance is genetic variation. This can include variations in the number of copies of a particular gene which are caused by insertions, deletions, transpositions, translocations, and copy number variations. There are several other factors that can also cause genetic variations. These factors include but are not limited to, environmental factors, radiation, and chemical mutagens.

Any change in a nucleic acid sequence or another genetic characteristic that can be identified and used to classify individuals, populations, or species, or to accurately predict the genes responsible for inherited trait, is referred to as a genetic marker [1]. A genetic marker is a nucleotide or sequence of DNA that can be used to identify a particular gene or chromosome that is associated with a certain physical trait. Markers can be located on any chromosome, including the sex chromosomes, and are used in many different ways. For example, they may be used to track the inheritance of a disease-causing gene in a family or to find out whether an individual is at increased risk of developing a certain disease. Markers can also be used to study the genetics of populations and to map the genes responsible for inherited traits. SNPs, restriction fragment length polymorphisms (RFLPs), varying number of tandem repeats (VNTRs), microsatellites, insertion or deletion (indel), copy number variants (CNVs), SVs, and TEs are a few types of genetic markers.

3.1.1 Single Nucleotide Polymorphisms

A single nucleotide polymorphism, or SNP for short (pronounced “snip”), is a type of DNA sequence variation in which a nucleotide in the genome (sometimes in a particular gene) differs between individuals of the same

biological species or between paired chromosomes in an individual [10]. For example, at a specific base position in the genome, the nucleotide cytosine may appear in most organisms, but in a particular species it may occasionally be replaced by the nucleotide thymine. SNPs and indels are the most prevalent forms of markers in the human genome. SNP is a nucleotide polymorphism that can be defined as a substitution of one nucleotide for another, while an indel is a variation that can be defined as the insertion or deletion of nucleotide(s). A SNP is different from a mutation which causes a permanent alteration of any size in the DNA sequence that makes up a gene, such that the sequence differs from what is found in most individuals. Mutations can be caused by a variety of reasons, including errors that occur during DNA replication, exposure to ionizing radiation, or exposure to certain chemicals [39]. SNPs are also different from copy number variants (CNVs), which happen when an entire gene (or other significant portion of DNA) is duplicated or deleted [38].

SNPs can be found in the coding regions of genes, which can affect the production of proteins. In the human genome, SNPs account for a large percentage of genetic variation and are important for genetic studies and genealogical research. There are around 4 to 5 million SNPs in an human's genome, which implies they typically occur almost once every 1,000 nucleotides [22]. Despite the relative ease with which SNPs are discovered in plants with simple genomes, the identification of SNPs in complex genomes faces significant challenges. The highly repetitive nature of plant genomes is one of the main issues [84]. There are several ways to identify SNPs and indels in next generation sequencing data:

1. Whole genome sequencing: this is the best approach for identifying SNPs and indels since this approach allows for the analysis of the entire genome rather than just a few genes or regions.

2. Targeted sequencing: this approach is used when the focus is on a specific gene or region.

3. SNP arrays: this approach is used when the focus is on specific genes or regions and SNPs are already known.

The impact of SNPs and indels on gene function is not fully understood. However, it is known that SNPs can change the function of a gene, while indels can disrupt the function of a gene. They can change the coding sequence of a gene, resulting in a change in the protein that is produced [89] [76]. They can also change the way that a gene is regulated, which can impact how much protein is produced. For example, a SNP that affects the codon for the methionine can change the way the protein is encoded and can affect the way the protein is expressed. Though a majority of SNPs have no impact on growth or health, some of these genetic variations have turned out to be crucial for understanding an organism's health. SNPs can result in phenotypic variability in plants, such as differences in color and size of the fruit, maturing, plant health, yield components, or susceptibility to various abiotic and biotic stimuli [59].

The type and frequency of these types of variations can be used to determine and measure the relationship between different individuals or groups of individuals, such as different populations. Also, comparing nucleotide sequence between two different genotypes is a common approach for identifying the molecular

basis of a difference in a phenotypic trait. They can be used as markers to track inheritance patterns and help researchers identify genes associated with traits of interest. SNPs may have an impact on the trait by altering the function of the gene when they are found within a gene or in a regulatory region close to a gene.

3.1.2 Structural Variations as Novel Genomic Markers

Structural variations (SVs) refer to large-scale genomic changes in the structure of a chromosome that are larger than 50 bp in length. SVs can be classified as deletions, insertions, inversions, duplications, or translocations [81]. It is a type of genetic variation that is frequently caused by genomic rearrangements that occur during meiosis or mitosis such as gene loss, gene duplication, or the creation of new genes. They can also be caused by DNA replication errors, DNA repair errors, exposure to mutagens and rearrangements of chromosomal segments. Structural variants can occur in both coding and non-coding regions of the genome. They can have a range of effects on gene expression and function, depending on their location and size [21].

Structural variants are thought to be a major source of genetic variation and have been implicated in the development of many human diseases [104]. They can also impact an individual's response to medication and their susceptibility to infections. A SV is a section of DNA that differs across individuals in sequence length, copy quantity, orientation, or chromosomal position. Large insertions and deletions, duplications and other copy-number variations (CNVs), short tandem repeat (STR) expansions, translocations of genomic sequences throughout the chromosomes, and inversions of a separated segment of DNA in the reverse position are all examples of SVs [81]. It is crucial to remember that structural changes can simultaneously affect characteristics via changing recombination, gene expression, or protein structure, irrespective of how they affect reproductive isolation per se. This is especially true for inversions since the mutational process that gives birth to inversions can also result in deletions at the breakpoints [118]. The effects of structural variations in plants are many and varied. Some common effects include changes in the plant's appearance, its ability to photosynthesize, its ability to resist pests and diseases, and its growth rate [104].

SVs are more complicated than SNPs and short insertions and deletions (indels). They can impact the coding regions of numerous genes or the functionality of genomic segments in diverse ways. Our knowledge of the evolution of populations, phenotypic polymorphisms, and the functional genome can all be improved by the study of SVs. Nowadays, structural variation is also being acknowledged as a substantial source of genetic diversity, even though SNPs were once considered to account for the most selective variation in a genome. There are various hypotheses to explain how these SVs increase illness risk, and one of them contends that SVs are infrequent mutations that together cause a considerable burden of disease [74], [97]. SVs, as opposed to SNPs, can produce large-scale perturbations of cis-regulatory regions, making them more likely to impact gene expression and phenotypes quantitatively [7].

The widespread detection and genotyping of structural variations is now more feasible than ever due to the recent developments in molecular and statistical approaches. The advances in long-read sequencing technology make it easier to find structural variants, especially complicated structural variants [79]. SMRT-SV [58],

DELLY [101], and LUMPY [69] are examples for detecting SVs using short read sequencing data; while Picky [45], Sniffles [106], and NanoSV [24] are tools using long read data. With these new technologies, software program for structural variant detection is developing quickly. After calling SVs, genotyping can be performed using SVTYPER [18] and SVTOOLS [68]. These genetic markers may be studied in the similar manner and with the same computational models that have been used for SNP data sets once structural variants have been found and structural variant genotypes have been determined.

3.1.3 Transposble Elements as Novel Genomic Markers

Transposable elements (TE) are fragments of DNA that can copy or cut themselves then instert in new chromosome locations. After the emergence of large-scale DNA sequencing, it has become evident that TEs are the single largest component of the genetic material of most eukaryotes, rather than being an unusual component of certain genomes. They make up at least 45% of the human genome and 50–85% of the genomes of certain plants [14]. Barbara McClintock discovered TEs in maize more than seventy years ago as the genetic agents responsible for the altered pigmentation sectors on mutant kernels. In 1983, she was awarded the Nobel Prize for this discovery and characterization of the genetic properties of TEs.

TEs are divided into two groups based on the intermediate substrate propagating insertions (RNA or DNA), and then further divided into families and subfamilies based on structural characteristics. Class I elements (also known as retrotransposons) use a copy-and-paste mechanism to transpose via an RNA intermediate, while Class II elements (also known as DNA transposons) use a 'cut-and-paste' mechanism [36]. Plants have the highest concentration of class I elements. Long Terminal Repeat retrotransposons (LTR-RT), non-LTR retrotransposons, are the two orders of retrotransposons based on structural features. LTR-RT is the most common order, and it can account for up to 80% of the genome size in plants like wheat, barley, and rubber trees [92].

When it comes to humans, the phylogenetic analysis of TEs shows that significant amplification stopped about 35–50 Myr ago [23]. Therefore, it is not surprising that only a small number of human TEs remain active despite accounting for nearly more than 1,200 Mb of the genome. According to one estimate, only a few elements out of 500,000 in the LINE family are functional in the human genome. An analysis of closely related plant genomes (especially grasses) has shown that TE amplification is responsible for the majority of the variation in genome size, both interspecifically and intraspecifically [36]. It's reasonable to expect that active TEs (especially active LTR retrotransposons) could be routinely isolated from plants based on such circumstantial evidence.

TEs in genomes might be employed as a source of genomic markers for modern plant breeding due to their widespread distribution and abundant genetic variation [102]. [30] studied how TEs contribute to phenotypic diversity in tomato. Even though these studies have linked specific TE insertions to a variety of agricultural characteristics, the phenotypic changes caused by TE insertions in crops have yet to be fully understood. In addition, developing high-throughput techniques for detecting TE markers on a large scale has been difficult

but is in progress. `TEmarker` is a robust tool that can aggregate findings from various TE identification tools from whole-genome shotgun sequencing data, identify and eliminate probable false-positive TE insertions, and perform high-throughput genotyping [124]. These findings, taken together, give a paradigm for exploiting TEs to assist molecular breeding efforts while also providing unique evolutionary insights. The authors of [124] concluded through GWAS that TE markers have a similar ability to discover association patterns to SNP markers. Another TE annotation tool is the Extensive de novo TE Annotator (EDTA), which is developed to automate *de-novo* TE annotation with various information such as length, number of exons, and read support [93]. A variety of TE detecting tool were combined with the appropriate filtering and processing scripts to create the EDTA pipeline, which can be used with any novel genome assembly.

There is a growing opportunity and a need for high-quality annotation of TEs as a result of recent innovations in genome sequencing. Also, this high-quality detection and annotation of TEs can give rise to a significant algorithmic and computational challenge [93]. Due to the varying sequence features of TEs of different classes, a plethora of software programs have been developed for almost every type of TE. But there is still a need for methods which can prove to be the gold standard in presenting the comprehensive benchmarking study. Since transposable elements range from up to 85% in plant genomes, their characterization and classification are important for understanding the complexities and processes of genome evolution. Furthermore, the annotation of TEs may boost the accuracy of coding area annotations and promote functional gene studies by relying on the development of different strategies for automatic bioinformatic identification and classification of transposable elements.

3.2 Phenotypes

The phenotype of an organism is the physical appearance or traits of the organism that is largely dictated by the genotype [5]. Each phenotype or trait is the result of any number of genetic variants interacting in complex ways and, given enough analytical power, is predictable to some degree from an organism's genetic information [19]. However, a trait is not always predictable from genotype information alone. There are many other factors that influence how a phenotype is expressed, the simplest example of which is environmental factors such as sunlight or nutrition availability [105]. Additionally, there are other biological factors that influence phenotype expression but are not visible via genome sequencing such as DNA methylation or chromosomal condensation [37]. Finally, there is the concept of epistasis, where the effect of one gene depends on the alleles of another gene, thus affecting the phenotype [72]. For example, some genetic variants of the APOE gene increase the risk of developing Alzheimer's disease [26]. If a person has two variants of the APOE gene that increase risk, then their risk of developing Alzheimer's is increased about three-fold. However, there is also evidence that lifestyle choices (such as diet and exercise) can significantly reduce the risk of Alzheimer's disease. So, even if a person has two risk-increasing variants of the APOE gene, they may be able to reduce their risk of developing Alzheimer's disease to the same level as someone with no

risk-increasing variants by making healthier lifestyle choices.

Phenotype prediction based on genotype information is a complex task, and it is difficult to predict exactly how a phenotype will be expressed. However, as our understanding of the genetics underlying a trait increases, we can develop increasingly accurate models for predicting how different gene variants will affect that trait. Punnett squares are a very simple method for predicting the phenotypes of cross-breeding experiments where all possible combinations of genes are enumerated with pen and paper [121]. For predicting simple traits, simple techniques such as Punnett squares can be used, but for traits that involve the interactions of many different genes, computational methods like GS are required.

3.3 Deep Learning

The ability of traditional machine learning approaches to analyze natural data in its raw form was limited. For many years, building a pattern-recognition system needed meticulous architecture and a great deal of domain knowledge to develop a feature extractor that converted the original data into an appropriate internal description from which the learning subsystem, frequently a classifier, could find or categorize patterns in the input [70]. Deep learning (DL) techniques emerge from Artificial intelligence (AI). DL makes process of learning quicker and simpler, which is very advantageous to data scientists who are entrusted with gathering, analyzing, and interpreting massive amounts of data. Deep learning algorithms are constructed in a hierarchy of evolving complexity and abstraction, as opposed to conventional machine learning algorithms, which are linear [33]. Applications of DL models, however, are still not widely used, as demonstrated by the fact that the majority of these applications are in the domains of computer science and engineering and less so in the fields of medical imaging, robotics, and computational biology [31]. In order to anticipate the structure and function of genomic components including enhancers [108], promoters [120], chromatin interaction predictions [129], levels of gene expression [107], and protein structure prediction with AlphaFold [62], DL applications are making headway in biology.

Multiple layers of "neurons" are combined to create a general DL architecture. The well-known Rosenblatt "perceptron", modeled based on how the brain functions, is the fundamental component of deep learning models [4]. Deep learning is regarded as one of the finest, if not the best, algorithms for tackling perceptual issues like image classification. The main reason for this is that DL models employ numerous hidden layers, which boosts their ability to recognize nonlinear patterns in the data more effectively [86]. This is why the additional layers in the architecture are referred to be "deep" in DL models [46]. When it comes to prediction tasks that use multi-layer neural networks with multiple hidden units, deep learning models need to be seen not as a single technique but rather as a series of learning algorithms that are now quite popular [70]. The DL movement gained momentum in the last ten years as a result of the development of effective algorithms such as backpropagation that can estimate parameters in multi-layer, complex neural networks and the fact that these techniques outperform existing algorithms in a variety of automatic recognition tasks like image

analysis [9]. A feedforward neural network (FNN), also known as multi-layer perceptron is the simplest DL topology shown in Fig. 3.1. Recurrent neural networks (RNN), deep belief neural networks (DBNN), and convolutional neural networks (CNN) are other examples of deep learning topologies.

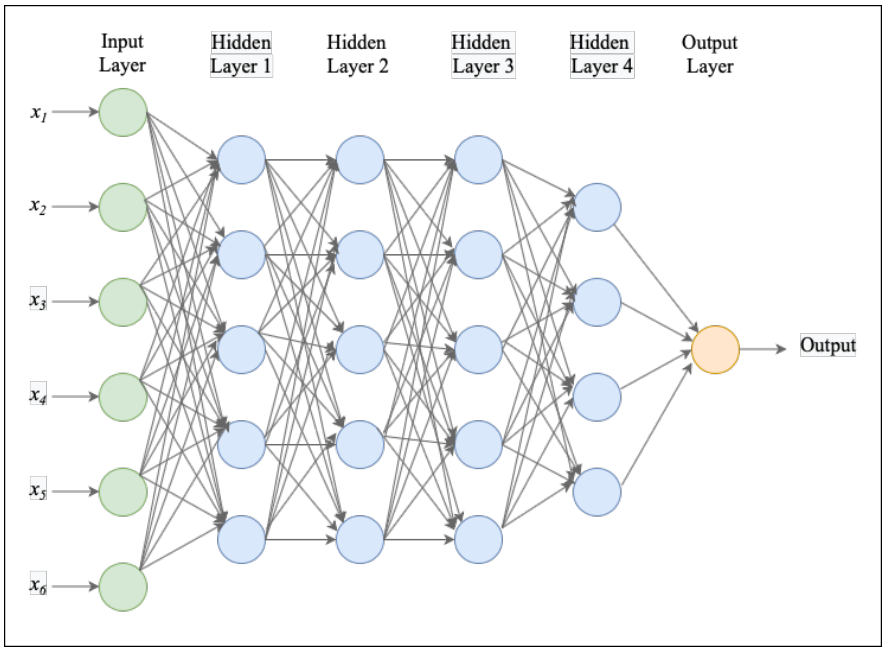


Figure 3.1: A deep feedforward neural network with five layers: one input layer, four hidden layers, and one output layer. This figure is adapted from [96].

Feedforward Neural Network (FNN) FNN are typically fully connected, meaning that each neuron in the network is connected to every other neuron in the network. Information moves in a single path in this network, from the input neurons via the hidden layers to the output layer. Each connecting edge between neurons in one layer and those in the next layer may have varied weights. This kind of artificial deep neural network requires the least amount of training time, often performs well over a wide range of tasks, and is ideally suited for general prediction tasks when it is assumed that the input data does not share any unique relationships. However, these networks are vulnerable to overfitting and less efficient when it comes to train on large datasets. FNNs are limited in their ability to model complex patterns in data. This is because they are limited to a single layer of hidden units, which limits their ability to learn complex patterns. Additionally, feed forward neural networks are also limited by their reliance on a linear activation function, which can limit their ability to learn non-linear patterns.

Recurrent Neural Network (RNN) A RNN is a type of neural network that is used for processing sequences of data. In FNNs, information flows in one direction from input to output. In a RNN, each unit receives input from the previous time step as well as the current input, that means information can flow in multiple directions. This makes RNNs better suited for analyzing temporal data, as they can take

into account information about what has come before in the sequence. The feature that enables this is the recurrent connection, which allows the hidden state of the network to be passed from one time step to the next. These models are well-suited for tasks such as modeling of sequential data, for instance the time series data or natural language processing.

Deep Belief Neural Network (DBNN) A DBNN is a neural network that is composed of multiple layers of stochastic latent variables, with each layer representing a higher level of abstraction than the previous one. A stochastic latent variable is a hidden variable that is randomly generated in order to create a stochastic process. This process can be used to generate data that appears to be generated by a deterministic process. Latent variables are important because they can be used to create models that are more accurate than those that are created using only deterministic variables. A DBNN can learn a representation of data that is much more compact than the original data, while still retaining all of the important information in the data. This is because a DBNN can learn to extract the important features of the data at each level of abstraction, and then use these features to reconstruct the data at the next level. The main limitation of a DBNN is that it can be difficult to train. This is because each layer of the DBNN needs to be trained separately, and each layer is only able to learn a limited amount of information. DBNs are most commonly used for unsupervised learning tasks.

Convolutional Neural Network (CNN) A CNN is a neural network that uses a local connection pattern, where each node is connected to a subset of the nodes in the previous layer. It has a convolutional layer that is responsible for extracting features from the input data. In a convolutional neural network, each layer consisting of many neurons represents complicated relationships among the large datasets. The different neurons in each layer receive information from the lower hierarchical layers, which is triggered by predefined activation functions. Activation functions bring non-linearity to the output of each layer to determine the input of the next layer [86].

For genomics research, a hierarchical learning algorithm like a CNN is attracting attention because it is anticipated that complex features would be produced by combinations of genes as well as regulatory components in the genome [28]. Convolution layers can learn gene connections, chromosome folding patterns, or other biological features. Hierarchical learning enables lower layers to learn simple traits like single SNPs or methylation sites. We will emphasize CNNs for predicting phenotypes using genomic data in this thesis.

CNNs are made up of a series of layers, each of which performs a specific task. CNNs are based on the hierarchical pattern in data and combine smaller and simpler patterns imprinted in their filters to create patterns of increasing complexity. Given how well CNNs capture both temporal and spatial connections of the input, they are effective tools for visual identification tasks. CNNs make use of the data's grid structure and uses images as input. Similar to image data, genomic data naturally has special connections, encoded by genomic locations of markers. CNNs are effective because they are able to learn the relationships between different variables and extract features that are relevant to the task at hand. The fitting process is what

allows the CNN to learn these relationships and extract the features. A typical CNN architecture can be divided into seven parts. The first five layers detailed below are the basic layers needed for the functioning of a CNN. The dropout layer and the activation function, are two additional crucial parameters in addition to these five layers.

1. The input layer: This is where the input data is fed into the network. The input data representation in a CNN is typically a matrix with shape (w, x, y, z), where w, x, y, and z represent input batch size, height, width, channels respectively.

2. The convolutional layer: This is the main computational layer and the foundational component of the CNN. Convolutional layer is the first class of trainable layers. In this layer, a filter is used to apply convolution to the input data in order to create a matrix known as a feature map [64]. This layer make use of a sliding window strategy with a set of learnable filters. Each filter is small in size (usually 3×3 or 5×5), and it is applied to the input to produce a feature map. In other words, the filter is moved across the input's width and height, and at each spatial position, the dot product between the input and filter is computed. The parameters of a convolutional layer include the number, size, stride, and the padding of the filters. The number of filters determines the number of convolutions that will be performed on the input. The size of the filter determines the size of the region of the input that each filter will convolve over. The stride determines the sliding window size that the filters will take when convolving over the input. And the padding of the filters determines how the input matrix is padded before the convolution is performed.

3. The pooling layer: This layer downsamples the feature maps produced by the convolutional layer. Each feature map is treated independently by the pooling layer. This reduces the size of features and helps to reduce overfitting. The pooling layer usually serves as a bridge between the convolutional layer and the fully-connected output layer. A two-dimensional filter is slid over each channel of the feature map during the pooling operation, and the features contained within the area enclosed by the filter are summarised [42]. In max-pooling, the largest element is taken from feature map and is the most used pooling method.

4. The flatten layer: It is a non-trainable layer. A flatten layer only resizes the input matrix into a one-dimensional matrix without doing any numerical change. We must use the flatten layer since fully connected layers can only receive one-dimensional inputs, while the convolutional layers allow two-dimensional inputs.

5. The output layer: This is the final layer of the CNN. In the output layer, the feature maps are fed into a fully connected layer. The output of this layer is the prediction. The fully connected layer is the second trainable layer. It is similar to a convolutional layer, but with the entire input taken into account at once. Each neuron in this layer is linked to every neuron in the previous layer. As each connection's weight is a separate trainable parameter, this leads to a large number of trainable parameters.

6. Dropout layer: The dropout layer is used to improve the effectiveness of network training and prevent overfitting. It has no impact on the size of the subsequent layer. Only during training does the dropout layer run; during inference, this layer executes the identity operation [73]. Neurons are "dropped out" or set to 0 at random in the dropout layer. The dropout rate parameter of the layer indicates the likelihood of keeping or dropping a neuron. By limiting the network from picking up on noise or unrelated features, this layer lessens overfitting.

7. Activation function: This function is used in a CNN to introduce non-linearity into the network. This non-linearity allows the network to learn more complex functions. Activation functions are usually applied to the output of each layer in a CNN. Some common activation functions are Sigmoid, hyperbolic tangent (TanH) and the Rectified Linear Unit (ReLU) function [90]. ReLU is the most commonly used activation function because it is the simplest and fastest to compute. Sigmoid and TanH are more computationally expensive, but they allow for a wider range of values. Each activation function has different properties, we must choose the activation function that is best suited for the task at hand.

A one-dimensional convolutional neural network can be seen in Fig. 3.2.

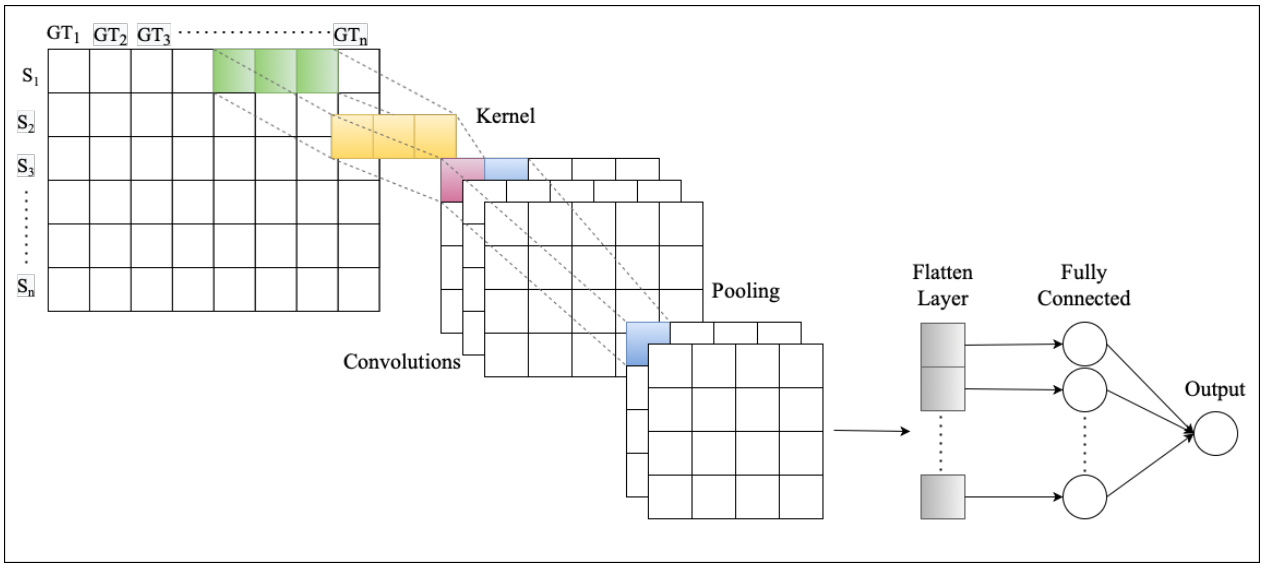


Figure 3.2: A one-dimensional convolutional neural network fully represented for a two-dimensional genotype matrix for n number of samples. Yellow represents the filter of size (3×1) applied over the input features in green color. Pink and blue are used to indicate the outputs of the convolution. Blue represents the output of the pooling layer. Then the layers are flattened and output is given to a fully-connected layer. The result is the predicted output in the end. This figure is adapted from [96]

The training and testing process in a CNN typically works as follows:

1. A dataset is first split into a training set, a validation set and a testing set. Though there is no hard-and-fast rule for splitting the dataset, the amount of samples in the dataset influence the dataset split ratio.

2. The set of data known as the "training set" is used to train the model and help it discover any hidden features or patterns in the data. At each iteration, the model receives the same training data repeatedly, continuing to learn the characteristics of the data. The training set includes a portion of the dataset which have corresponding labels known as ground-truth values. A diverse collection of inputs should be included in the training set so that the model can be trained on all possible scenarios and be able to forecast any unanticipated data sample that might arise in the future. The loss function is used to train the neural network by minimizing the error between the predicted values and the actual values. A loss function maps a set of predicted values to a real number that represents how far off those predictions are from the actual values. The most common loss functions used in CNNs are the mean absolute error (MAE) and the cross-entropy loss [61].

3. The validation set is a set of data that is not part of the training set and is used to verify the performance of the model as it is being trained. The major goal of is to avoid overfitting. At each iteration, the model is evaluated on the validation set while being trained on the training set. The hyper-parameters of the model are adjusted using the data from this validation process. It tells us whether or not the training is on the right track.

4. After the model has been trained, it is tested using a different collection of data called the test set that is never utilized for training or validation. It offers a neutral measure of the final model's performance in terms of precision, accuracy, etc. based on the ground-truth values.

A variety of hyper-parameters affect how training functions. Hyper-parameters that affect training include batch size, epoch count, learning rate, and early stopping functions.

1. Batch size is the number of samples that are processed together in one step in a neural network. The larger the batch-size, the more data a CNN can process at one time, and the training process will be faster.

2. Epoch count is the number of times the entire training dataset is passed through the neural network during training.

3. Learning rate is the speed at which the neural network learns.

4. Early stopping is a technique used to stop the training if the validation error is not decreasing to prevent overfitting in a neural network.

Incorrectly specified hyperparameters might result in delayed training or overfitting. Overfitting happens when a neural network just memorises which inputs correlate to which outputs rather than learning a systematic mapping from major input properties to outputs. Insufficient training data can lead to overfitting, which appears as a neural network that performs well on training data but poorly on validation data.

Deep learning is a game-changing technology with enormous promise for use in predictive data science. By creating particular topologies for the type of data present in plant breeding programs, DL algorithms have a significant ability to increase prediction accuracy. Due to these factors, it is crucial that GS implement this innovative technology to improve its effectiveness and precision.

There are a few reasons why CNNs might be more appropriate for genomic selection than other deep learning topologies discussed earlier. First, CNNs are designed to work with data that has a spatial structure, which is common in genomic data, while RNN is limited to temporal relationships. Second, CNN is less sensitive to overfitting. As FNNs face difficulties in learning complex patterns and RNNs can have very long-term dependencies, which can lead to overfitting if the training data is not very large. Third, CNNs can be parallelized more easily than FNNs and RNNs, which means that they can be trained on larger datasets more quickly. Additionally, CNNs are generally good at learning local patterns, which can be important for detecting patterns in genomic data. Finally, CNN is more efficient at processing high-dimensional data than FNN. This is because CNN can exploit the structure of the data to reduce the number of parameters that need to be learned.

The availability of adequate amounts of high-quality training data is a crucial need for DL models. The fact that deep neural networks are primarily black boxes is one problem with deep learning. After a network has been trained, it can be challenging to examine the network and examine the layers to determine why the model makes the predictions that it does or what input components are important to the output.

3.4 Genomic Selection

Genomic Selection (GS) is a method of predicting the genetic value of individuals for a trait, without having to measure the phenotype of the individuals. It is based on the relationship between phenotypes and genotypes of individuals within a reference population. Breeders can choose breeding lines to propagate during a growing season by assessing and anticipating the phenotypes of that line. When examining the impact of genetic markers sequenced with the most cutting-edge dense marker maps available at the time, Meuwissen *et al.* [83] first came up with the term "Genomic Selection" in 2001. Due to the advantages of dense marker maps, certain genetic markers were possibly found in linkage disequilibrium with the quantitative trait loci and others were close to them. The availability of detailed marker maps and the possibility of genotyping multiple individuals for these markers are both made possible by recent developments in molecular genetic techniques such as Next Generation Sequencing.

In genomic selection methods, the individual effects of each marker are calculated, then the combined marker effects are used to calculate each individual's genotypic value, known as genome estimated breeding value (GEBV) defined in Equation 3.1.

$$GEBV_i = \sum_{j=1}^{m} G_{ij} f_j \tag{3.1}$$

where $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m$ (n is the total number of individuals and m is the total number of markers), G_{ij} is the element in the incidence matrix corresponding to the jth marker of a individual i, and f_j is the estimated effect of marker j.

The first step in the GS process is to create a statistical model using individuals with both genotypic and phenotypic data available (also known as the training set or reference population). This model is then used to estimate GEBVs for individuals with genotypic data. Following a ranking of individuals based on GEBVs, the best candidates are chosen. The effectiveness of GS depends on the choice of a good and large training population, and appropriate prediction model.

GS has become a routine technique in many plant and animal breeding programs. The International Maize and Wheat Improvement Center breeding studies have demonstrated that GS can cut the breeding cycle in half at the very least and create lines with much higher agronomic performance [25]. GS has been used in breeding projects for crops including, chickpea, groundnut [117], soybean [123], barley [15], wheat [48], strawberry [41], millet [75], and maize [80].

3.4.1 Existing methods in GS

Broadly speaking, GS methods can be classified into two categories: conventional statistical modeling and deep learning methods. Existing methods in both categories only consider the SNP variants as genomic markers to predict the phenotypes. In this section, we briefly summarize two models in each category. These four models are compared with our proposed `NovGMDeep` model in Section 5.3 and Section 6.3 to demonstrate the prediction of phenotypes using SV, TE, and SNP markers.

Ridge Regression-Best Linear Unbiased Prediction

One of the most widely used statistical models for GS is Ridge Regression-Best Linear Unbiased Prediction (rrBLUP) [32]. The model is created using the linear regression algorithm, the genotype matrix, and the phenotype vector as defined in Equation 3.2.

$$y = \mu + Xq + e \tag{3.2}$$

where y is the phenotype data vector, μ is the overall mean vector of random effects, X is the design matrix with a vector q of fixed effects, and e is the residual effect.

Under the notion that marker effects follow a normal distribution with a modest but non-zero variance, the ridge regression technique is used to estimate the joint effects of all genotypic markers at the same time. The `mixed.solve` function in the rrBLUP package was used to solve the rrBLUP model.

Genomic Best Linear Unbiased Prediction

Genomic best linear unbiased prediction (gBLUP) [20] is another statistical modeling method for estimating an individual's genetic performance based on its genomic associations. A genomic relationship matrix is used

in gBLUP, which is used to represent genomic markers as defined in Equation 3.3.

$$y = Zb + Wg + e \tag{3.3}$$

where y is the phenotype data vector, Z is a design matrix with a vector b of fixed effects for each individual, W is a design matrix allocating records to genetic values with a vector g of additive genetic effects for each individual, and e is the residual effect.

To make more accurate merit predictions, the matrix specifies covariance between individuals based on observed similarity at the genomic level rather than predicted similarity based on lineage. gBLUP has been used to predict merit in cattle breeding, in disease risk prediction, as well as in the calculation of variance components and genetic heritabilities [51]. The `BGLR` function in the R package was used to solve the GBLUP model.

DeepGS

The DeepGS model has been developed using a deep convolutional neural network with an $8 - 32 - 1$ architecture [78]. The final output represents the predicted phenotypic values for the analyzed individual markers. The DeepGS model has been trained and tested using 10-fold cross-validation on Iranian bread wheat (*Triticum aestivum*) landrace accessions with eight measured phenotypes. The phenotypes include grain length, grain width, grain hardness, thousand-kernel weight, test weight, sodium dodecyl sulphate sedimentation, grain protein, and plant height. A backpropagation algorithm has been used to optimize the parameters of the model. The number of epochs has been kept at 6000, momentum has given a 0.5 value, weight decay is 0.00001 and the learning rate of the model has been set to 0.01. The DeepGS model utilizes hidden variables to collectively represent features in the genome-wide markers while making predictions. Convolutional, dropout, and sampling layers have been employed in the model to reduce the complexity of large volume genomic data.

G2PDeep

The quantitative phenotypic prediction problem was treated by G2PDeep as a regression problem [128]. Data on zygosity and SNPs for Soybean with five phenotypes, including grain yield, height, moisture, oil, and protein were used to train the model. For genotypes and missing data, one-hot binary encoding is employed to express zygosity data. Each genetic marker is displayed as a four-dimensional and five-dimensional vector for zygosity and SNP data, with 1 for the matching genotype and 0 for the other genotypes. The model is made up of a dual-CNN layer and a fully connected neural network. The encoded genotypes are transferred by dual-CNN, which has two concurrent CNN streams, followed by single-CNN to enhance marker representation. The flattened layer with the combined representation of markers is passed to a fully connected neural network with only one neuron in the output layer to estimate the quantitative phenotype. The two CNN layers have kernel sizes of 4 and 20, as well as the same number of 10 filters, in the left stream. In the right stream, a

single CNN layer with a kernel size of 4 and 10 filters is utilized. The add-up layer combines the output from two streams, followed by a single CNN layer with a kernel size of four and ten filters. The fully connected layers with 512 and 1 neurons serve as regression blocks for predicting phenotype.

The two above-discussed statistical methods have been used as predictive models for a plethora of both plant and animal species for the last two decades. However, these conventional models typically assume that the genotype random effects follow a normal distribution, and each genotype's contribution to phenotypes is considered an independent attribute. Whereas, it is unknown in practice how genotype effects behave. Moreover, these models do not consider non-linearity between the variables, as SNPs may also interact with other SNPs to cause complex traits as a result of epistasis. Though both of these BLUP models have given equivalent prediction results in the past, gBLUP is known for its computational efficiency [20]. For the aforementioned DL models, it is still early to compare as there is a limited number of applications for these in the literature.

3.4.2 GS using Novel Genomic Markers

All above-discussed methods only consider the SNP variants as genomic markers to predict the phenotypes. For standard linear models, the number of genome-wide SNP features largely exceeds the sample size, leading to overfitting. For deep learning models, studies have shown that these methods base their predictions on overall genetic relatedness, rather than on the effects of particular markers [115]. To mitigate these limitations, instead of SNPs, we propose to use SVs and TEs to represent genomic variations and explore their effectiveness in genomic predictions. We have provided an overview of SVs and TEs in Section 3.1.2 and Section 3.1.3 and discussed their importance in phenotypic prediction in plants.

This research demonstrates the need of looking further than SNPs for genomic selection. And how significant sources of variation such as SVs and TEs can help the breeders in selection of desired traits more accurately and improve the efficiency of their breeding programs. The proposed model in Chapter 4 makes effective predictions using novel genomic variations. Although our method focuses on representing structural variations and TE markers for predicting phenotypes, SNP variation is also compatible with the model.

Early plant genomic investigations were hampered by technological limitations and a lack of high-quality reference genome assemblies, which prohibited a detailed analysis of both SVs and TEs in plants. Recent improvements in genomic technology, notably whole-genome sequencing and long-read mapping, offer the generation of high-quality plant genome and pangenome assemblies, as well as exposure to a diverse set of SVs for evaluating their possible significance in plant phenotypic diversity [127]. Similarly, the transposition of TEs, which are common in most eukaryote genomes, accounts for a considerable amount of genomic variation. As a result, TE-derived molecular markers are useful resources for unraveling the genomic variations of both plants and animals [102]. The aim of this work is to shed light on the unrecognized function of SVs and TEs in genotype-to-phenotype associations, as well as their extensive significance and value in crop development.

4 NovGMDeep: a convolutional neural network for genomic selection using novel genomic markers

In this chapter, we will discuss general strategies to represent novel genomic markers as the input to the model, the model design and architecture, and performance metrics.

4.1 Genomic marker representation

DNA sequence variation and associated information are stored in variant call format (VCF) files. A VCF file contains information about genetic variants in a tab-delimited text format. Each line in a VCF file represents a single genetic variant. For example, a variant could be a SNP, indel, SV or TE. The file format consists of a header followed by data lines. The header line contains information about the format of the data lines that follow. The data lines contain information about individual genetic variants in ten columns with column headers listed as follows:

CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE

The first five columns are mandatory and must be in order. The remaining columns are optional and can be in any order.

1. CHROM: The name of the chromosome or contig where the variant is located in the reference genome.
2. POS: The position of the variant on the chromosome or contig, with the first base having position 1 in the reference genome.
3. ID: A unique identifier for the variant.
4. REF: The reference allele for the variant.
5. ALT: The alternate allele for the variant. A basic string made up of the letters A, C, G, T, N, $*$, or an angle-braced ID String ("<ID>") are all options to represent an alternate allele. The reserved allele, $*$, is an allele that is not found in the population. The ID string in the ALT column indicates the type of the variant. For example, <DEL> for deletion, <DUP> for duplication, <INV> for inversion, <INS:ME:ALU> for insertion of ALU retrotransposon element, <INS:ME:LINE1> for insertion of LINE1 retrotransposon element, etc.
6. QUAL: A quality score for the variant.

7. FILTER: One or more filters that have been applied to the variant. "PASS" means that the variant called at that position has passed all of the filters that have been applied to the data.

8. INFO: Additional information about the variant.

9. FORMAT: It describes the order of the genotype fields for each individual. It includes information on the individual's genotype, allele depth, quality, and other information.

10. SAMPLE: It contains the genotype data of the variant in the order described in the FORMAT column for a particular sample. There can be multiple columns of different samples for a variant.

Table 4.1: Example of a few types of variants encoded in a VCF file. In the INFO column, SVTYPE indicates the type of structural variant (INV for inversion, DEL for deletion, INS for insertion, DUP for duplication); END marks the end position of the variant; SVLEN is the length of the variant; and CIPOS/CIEND records confidence interval around POS and END positions for imprecise variants. For precisely known variations, the REF and ALT columns contain the full sequences for the alleles. For imprecise variations, the REF field may include a single base, while the ALT fields may contain alleles that are descriptive (<ID>). The last column, NA00001 (sample name), contains the genotype data of the variants in the order (GT:GQ:CN:CQ) described in the FORMAT column (GT is genotype; GQ is genotype quality; CN is the copy number genotype for imprecise events; CQ is the copy number genotype quality for imprecise events).

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001
1	2827694	rs2376870	CGTGGATGCGGGGAC	C	.	PASS	SVTYPE=DEL; END=2827708; HOMLEN=1; HOMSEQ=G; SVLEN=-14	GT:GQ	1/1:14
2	321682	.	T	<INV>	6	PASS	SVTYPE=INV; END=421681; SVLEN=99999; CIPOS=56,20; CIEND=10,62	GT:GQ	0/1:12
2	14477084	.	C	<DEL:ME:ALU>	12	PASS	SVTYPE=DEL; END=14477381; SVLEN=-297; CIPOS=-22,18; CIEND=-12,32	GT:GQ	0/1:12
3	9425916	.	C	<INS:ME:L1>	23	PASS	SVTYPE=INS; END=9431943; SVLEN=6027; CIPOS=-16,22	GT:GQ	1/1:15
3	12665100	.	A	<DUP>	14	PASS	SVTYPE=DUP; END=12686200; SVLEN=21100; CIPOS=-500,500; CIEND=-500,500	GT:GQ:CN:CQ	./.:0:3:16.2

The five rows in Table 4.1 contain information for different variants. The details of these variants are described as follows.

Row 1. Sample NA00001 has a homozygous deletion of length 14 at position 2827694 in chromosome 1. This is an accurate deletion with a known breakpoint, and a micro-homology of one base. Micro-homology is defined as one or more base pairs (bp) of perfectly matching sequence surrounding the breakpoints.

Row 2. An imprecise inversion in sample NA00001 (Figure 4.1). The length of the inverted DNA fragment is 99999 bp.

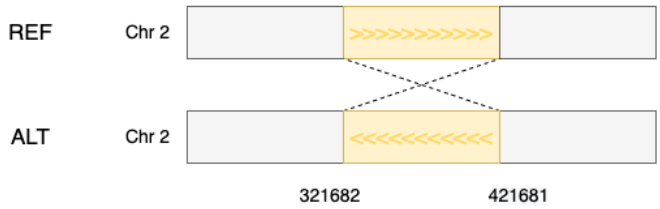


Figure 4.1: An inversion (INV) of 99,999 bp in chromosome 2 of the sample.

Row 3. An imprecise deletion of an ALU retrotransposon component (297 bp) from the reference (Figure 4.2).

Figure 4.2: A deletion (DEL) of 297 bp.

Row 4. An imprecise L1 retrotransposon element insertion of length 6027 bp in relation to the reference (Figure 4.3).

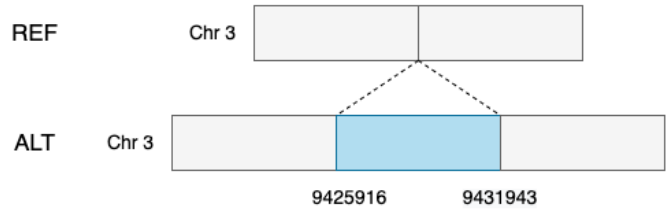


Figure 4.3: An insertion (INS) of 6,027 bp.

Row 5. A roughly 21Kb imprecise duplication in chromosome 3 of sample NA00001 (Figure 4.4). The two unknown haplotypes (./.) indicate the missing genotype for this sample.

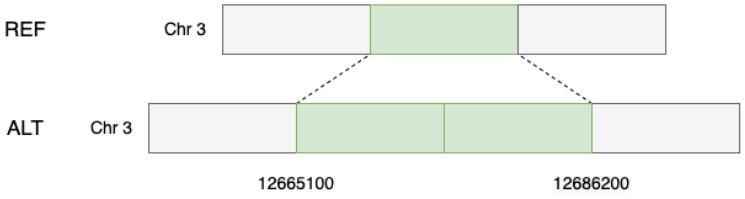


Figure 4.4: A duplication (DUP) of 21,100 bp.

TE variants are mobile genetic elements that can move around within the genome. There are two main types of transposons: those that move around by a "copy-and-paste" mechanism and those that use a "cut-and-paste" mechanism. Retrotransposon variants use RNA as a template to make a copy of themselves and insert the copy into a new location within the genome (Figure 4.5). An example of a L1 retrotransposon element can be seen in Table 4.1. DNA transposons cut themselves out of their current location and insert themselves into a new site (Figure 4.6).

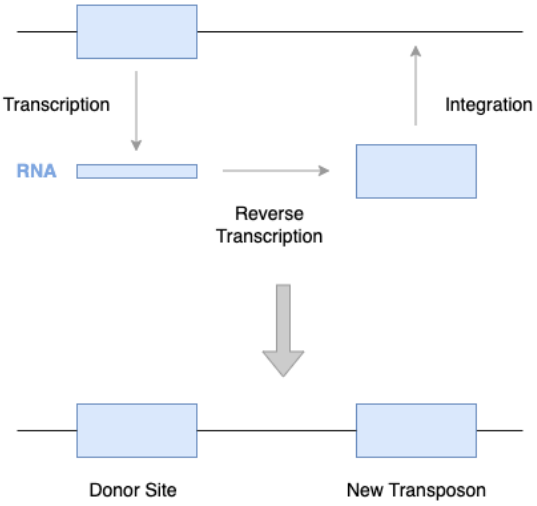


Figure 4.5: Retrotransposons following a "copy-and-paste" mechanism. This figure is adapted from [2].

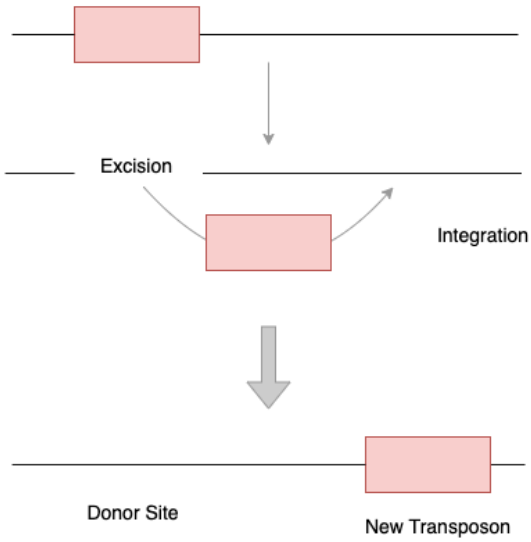


Figure 4.6: DNA transposons following a "cut-and-paste" mechanism. This figure is adapted from [2].

The subsequent combination of alleles that an individual possesses for a specific gene is their genotype [103]. The genotypes of all the samples are stored in the columns after the FORMAT column as seen in Table 4.1. The genotypes in Table 4.1 are in the format '0/0' and '1/1' accompanied by the genotype quality, copy number genotype for imprecise events and copy number genotype quality for imprecise events. Genotypes in plant species are represented as either haploid or diploid or polyploid. A *haploid* genotype comprises a single set of chromosomes for generating the phenotype whereas a *diploid* genotype contains two sets of chromosomes, each of which is required for phenotype generation. Genotypes are usually represented as '0|0' or '0/0', where '|' and '/' indicates the phased and unphased genotypes respectively. To gain a complete picture of genetic variation, phasing entails separating maternally and paternally inherited copies of each chromosome into haplotypes. A genotype with at least one set of explanatory haplotypes is referred to as a phased genotype. Unphased genotypes are those for which no set of explaining haplotypes have been identified [88]. The genotype is considered to be *homozygous* if all the haplotypes for a specific site have the same value. Otherwise, the genotype is considered *heterozygous*. A diploid organism either inherits two copies of the same allele or one copy of two different alleles from its parents. If an individual inherits two identical alleles, their genotype is homozygous at that locus. However, if they possess two different alleles, their genotype is heterozygous for that locus. Alleles of the same gene are either autosomal dominant or recessive. An autosomal dominant allele will always be preferentially expressed over a recessive allele.

There are two ways to represent genotypes in genomic selection models. The first way is to represent each allele as a single integer, with 0 representing the reference allele and 1 representing the alternate allele. For example, the genotype AA would be represented as 0, the genotype Aa would be represented as 1, and the genotype aa would be represented as 2. The second way is to represent the genotype as a one-hot encoding integer. It is a representation of a genotype where each allele is represented by a unique integer. For example, each genotype is represented as a vector of 0s and 1s, with each element of the vector corresponding to a different allele. A detailed explanation of the one-hot encoding representation can be found in Section 5.2 and 6.2.

The number of genomic markers needed to obtain a complete sampling of an organism's genetic variation is usually lower for an organism with a smaller genome than one with a bigger genome. A dataset with fewer genomic markers requires less memory and training time for a neural network, while a dataset with millions of genomic markers is undesirable due to the computer resources they demand. For the case study in Chapter 5, we used the genome of *A. thaliana* to compare the effectivity of SVs and SNPs (which are usually found in millions in most of the genomes) in phenotype prediction. And for the case study in Chapter 6, the genome of *O. sativa* is used to compare the effectivity of TEs and SNPs.

4.2 Phenotypic Data

The availability and association of phenotypic information with each sample is also crucial for a neural network to be trained as it learns to predict labels based on inputs. The genotype serves as the input and the phenotype serves as the label for each individual in a dataset. Labeled samples are utilized for training, validation and testing, allowing a neural network to predict the phenotypic data for unlabelled samples after it has been trained. Each genotype must have corresponding phenotypes as labels to be utilized for training or validation. Most of the current genotype databases lack information on associated phenotypes with matching samples. High costs of genome sequencing technologies and data privacy and security could be some of the challenges.

4.3 Model Architecture

The fundamental idea behind a convolutional neural network architecture is to start with a broad and shallow feature space and gradually make it tighter and deeper as the network progresses. In order to deepen the feature space and aid in the learning of more levels of abstract features, the number of filters is raised. So in the process of model architecture design, as we advance through the layers of our convolutional neural network architecture, the number of channels typically rises or stays the same. Add layers until the model starts overfitting. Once we reach a high accuracy level in the validation set, we can reduce over-fitting by using regularization components, dropout, batch normalization, data augmentation, etc.

Hyper-parameter tuning of CNNs is a challenging task. We must keep checking if we select the appropriate parameters while creating the CNN architecture. The architecture selected for this work was chosen after keeping all these conventions in mind and playing with different hyper-parameters to be able to predict phenotypes with less overfitting.

We also considered RNNs for the model architecture in this thesis. The issue with RNNs is that they only have a single weight matrix that is used by all of the recurrent units, which makes it difficult for them to find a variety of spatial features. While a CNN can use several kernels or filters in a layer to locate a variety of features. The RNN model needed numerous layers and an inordinate amount of time to replicate the same as the proposed CNN architecture because they can only find a small number of sequences at a single layer.

For both the case studies in Chapter 5 and 6 similar CNN network architecture is utilised, however layer shapes were adjusted to accommodate the various input sizes as discussed in Section 5.2 and 6.2. The number of different genomic markers in the input samples is always equal to the number of neurons in the input layer of a CNN. The number of neurons in each consecutive layer is subsequently decreased by convolutional layers that follow the input layer. Although the architecture (the selection and arrangement of layers) is the same for each dataset, the size of those layers must match the input size.

`NovGMDeep`, a deep one-dimensional (1D) convolutional neural network to predict phenotypes, is presented

in this study. It has four 1D convolutional layers, a single 1D max-pooling layer, a flatten layer and one dropout layer followed by a fully connected dense layer (Figure 4.7).

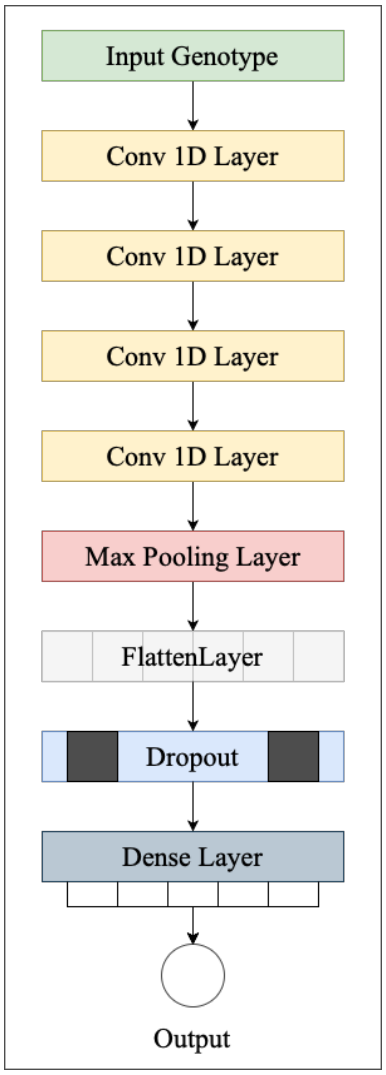


Figure 4.7: `NovGMDeep` architecture. The first convolutional layer has 16 filters of kernel size five. The second layer comes with 32 filters of kernel size five. The third one has 64 filters of kernel size three. And the last convolutional layer has 128 filters of kernel size three. The max-pooling layer has a pool size of two and stride of two. At last, there is one fully connected layer as the regressor with one output neuron.

As the backbone of the model, we use four 1D convolutional layers of the following sizes. The first layer includes 16 filters of size five. For the next layers, we doubled the number of filters in each layer and reduce the filter sizes of the last two layers to three. We double the number of filters in each layer to increase the depth of the network. This allows the network to learn more complex features. We reduce the filter sizes to reduce the amount of resources required to train the network. If the input to the convolution layer is represented as X, filter as F, bias as B, and the output of the convolution layer as Y, the convolution operation between X and F including bias is defined in equation (4.1).

$$Y(N_i) = F * X(N_i) + B \tag{4.1}$$

where N is the batch size and $*$ is the cross-correlation [16] operator.

For all the convolutional layers, ReLU activation function [3] (Equation 4.2), LecunNormal kernel initializer [49], and L1L2 kernel regularizer [66] is used with same padding.

$$ReLU(Y(N_i)) = \max(0, Y(N_i)) \tag{4.2}$$

To reduce the high dimensionality of the prominent features of the last convolutional layers, a one-dimensional max-pooling layer is used after the convolutional layers with a window size of two and a stride of two. It minimizes the amount of the input to the following layer by taking only the maximum values as the most important features, which cuts the size of the input into the following layer in half [6]. Then the extracted feature map from the model's backbone is flattened using a flatten layer and fed into the fully connected layer which is our phenotype predictor. Fully connected layers link every node in the input to every node in the output, but they do not capture spatial information. Due to the high number of input features which results in a large-size feature map, the dropout layer with the dropout ratio of 0.6 is used before the fully connected layer to reduce computational complexity and avoid overfitting. We used the Adam [65] optimizer with a learning rate of 0.0003. The Adam optimization algorithm is applied to iteratively adjust the network weights based on training samples. Moreover, we utilized mean absolute error as the loss function. The `NovGMDeep` is developed with Keras [46], an open-source software library that provides a Python interface for artificial and deep neural networks. The source code of `NovGMDeep` is publicly available on Github (`https://github.com/shivanisehrawat4/NovGMDeep.git`).

Due to a large number of features, four convolutional layers were used to extract the most representative feature maps for the linear predictor. On one side, because of the small number of training samples, increasing the number of layers made the model more complex and led to overfitting. On the other side, decreasing the number of layers as well could not provide the predictor with informative feature maps. Even with four convolutional layers, we had a huge output feature map which resulted in a complex linear predictor. Therefore, we needed to use a dropout layer to decrease the complexity of the predictor and avoid overfitting. Each of these choices were made based on the stability of the model performance and behavior on the validation and test sets during and after model development and training.

The model is trained separately on three different types of data: SVs, TEs and SNPs. Also, we used K-fold cross validation to select the best model. It is a process for model selection that involves partitioning data into K subsets, using K-1 subsets to train the model, and then evaluating the model on the remaining subset. This process is then repeated K times, with each subset serving as the validation set once. The model selection process is then repeated for each value of K.

4.4 Performance Metrics

To evaluate the model, statistical analysis using Pearson's coefficient of correlation was performed among the predicted and observed phenotype values. The results show that SVs and TEs are more informative for phenotype prediction than SNPs. PCC is a model-free method, and it, therefore, shows the nature of the data without depending on any of the existing models. It is commonly used in statistical analysis for the evaluation of the model and measuring the relationship between two variables.

The model predicts quantitative phenotype as a real number which is the model's output. The predicted phenotypes are later compared with the ground-truth values to evaluate the model performance. As an evaluation metric, we use the Pearson's coefficient of correlation (PCC) between the predicted and observed phenotypes. Five assumptions mentioned below must be met before we calculate the PCC between two variables [91]:

1. Level of Measurement
2. Linear Relationship
3. Normality
4. Related Pairs
5. No Outliers

As far as our datasets are concerned, all of the conditions are true and the detailed analysis can be found in Appendix B.

PCC [43], normally denoted by ρ, is a measure of the linear correlation between two variables (X and Y), whose values range from -1 to 1, with 1 indicating total positive correlation, 0 indicating no correlation, and -1 indicating a total negative correlation. It is defined as the covariance of the two variables divided by the product of their standard deviations.

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y}, \tag{4.3}$$

where $cov(X,Y)$ is the covariance of X and Y, σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y.

For a classification task, the ability of a model can be indicated by calculating its accuracy to predict the class label correctly. A regression model's performance or competence can be indicated by calculating the error in those predictions. As we deal with a regression task and try to predict numerical phenotypes in this study, such as plant height, we are more interested in how closely the predictions came to the actual values. This can be measured by error metrics. For assessing and summarizing the effectiveness of our proposed model, mean absolute error (MAE) metric was employed.

The MAE metric is used to calculate the validation loss on the test set. MAE refers to the magnitude of difference between the prediction of observation and the true value of that observation.

$$MAE = \frac{1}{N}\sum_{i=0}^{N}|P_i - O_i|, \tag{4.4}$$

where N represents the total number of individuals present in the training set and P_i and O_i denote the values for the ith predicted and ith observed phenotypes.

The Standard Deviation (SD) is a measure of how spread out numbers are. SD of MAE is calculated to see the deviation among all the calculated MAE values for each sample in the test set.

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \tag{4.5}$$

where x_i is a value in the data set, μ is the mean of the data set, and N is the number of data points in the population.

Thus, the MAE tells us what the average value for the test set is and the SD tells us what the average scatter of MAE values is for the same test set, around the mean.

5 Case Study #1: Phenotype Prediction using SV markers

In this chapter, we will discuss the phenotype prediction experiment for *A. thaliana* using SV markers.

5.1 *A. thaliana* dataset description

The flowering plant, *A. thaliana*, is an ideal plant species for researching genotype-phenotype-environment interactions because their naturally inbred strains allow for repeated phenotyping and have adapted genotypes under a variety of controlled circumstances. *A. thaliana* is quite appealing in scientific research for surveying the molecular and phenotypic effects at the species level [119]. It is found in a wide variety of environments around the world, and representative accessions have been extensively phenotypically and genetically characterized. It matures, replicates, and reacts to stress and diseases in a manner that is similar to that of many agricultural plants. Additionally, because Arabidopsis is simple, cheap, and prolific, it enables large-scale genetic studies that frequently involve several thousand plants. Moreover, the relatively tiny genome of Arabidopsis makes genotyping easier and cheaper compare to some large genomes. The 1001 Genomes "A Catalog of *A. thaliana* Genetic Variation" [100] is a database where whole-genome sequencing has been performed on over 1000 *A. thaliana* accessions which makes the genomic selection and genome-wide association studies on this species possible.

The full VCF variant files containing the structural variants data for the $1,301$ *A. thaliana* samples are publicly available on European Variation Archive (PRJEB38975) [44]. They were used in this study, as well as the SNPs and indels for $1,135$ accessions of *A. thaliana* [8]. The SV dataset includes several types of SVs, such as deletions, duplications, and inversions. The lengths of these variations span from 50 bp to 10 kb.

The phenotypes of matched samples were downloaded from AraPheno, the 1001 genomes project [112]. Flowering time, rosette leaf number, cauline axillary branch number, and stem length were used with their definitions listed as follows.

1. Flowering time: Generally scored as days from seeds in the soil until the first open flower.
2. Rosette leaf number: A shoot system morphology trait which is the number of leaves in the shoot system of *A. thaliana*.
3. Cauline axillary branch number: A shoot axis morphology trait which is the amount or pattern of

branches arising from the shoot axis.

4. Stem length: A stem morphology trait often measured from the soil surface to the highest point on the stem.

All the sources of the datasets used in this experiment can be found in Appendix A.

5.2 SV marker representation

A. thaliana is a diploid species and has phased genotypes as '0|0' and '1|1' which means homozygous for reference allele and homozygous for alternate allele respectively. Although SVs are smaller in numbers than SNPs, which account for almost a million variant sites per genome, they are more informative, as they account for a higher number of nucleotide differences due to their size. Different types of SVs such as inversions, duplications, and deletions may capture the genomic variations better than the SNPs. We propose a strategy as shown in formula 5.1 to represent the SVs. The SV genotypic data for inversions, duplications, and deletions are transformed using one-hot encoding as binary arrays. In one-hot encoding, given a dataset with different features, the encoder finds the unique values per feature and transforms the data to a binary one-hot vector [95]. Different arrays for different types of SVs were chosen so that the details of genomic variations were well represented and understandable by the model. For the SNP and indel data, the same one-hot encoding strategy was used, 0|0 for reference allele, 1|1 for alternate allele, and .|. indicating the missing genotypes. As shown below, INV represents inversions, DUP represents duplications, and DEL represents deletions:

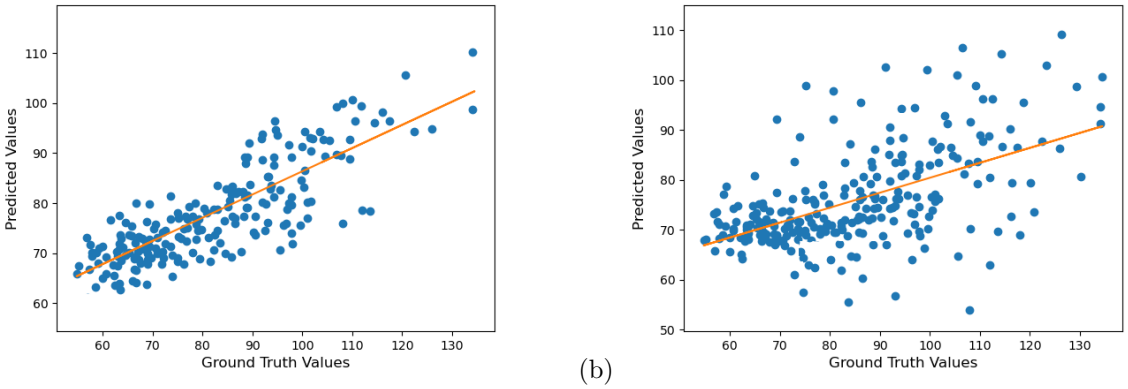
$$
\begin{aligned}
0|0(INV) &\rightarrow [1,0,0,0,0,0] \\
1|1(INV) &\rightarrow [0,1,0,0,0,0] \\
0|0(DUP) &\rightarrow [0,0,1,0,0,0] \\
1|1(DUP) &\rightarrow [0,0,0,1,0,0] \\
0|0(DEL) &\rightarrow [0,0,0,0,1,0] \\
1|1(DEL) &\rightarrow [0,0,0,0,0,1] \\
.|.(MissingValue) &\rightarrow [0,0,0,0,0,0]
\end{aligned}
\tag{5.1}
$$

For the SV dataset, the data is represented in the format of a 3D matrix $A(x \times y \times z)$ where x represents the number of accessions, y represents number the SV/SNP markers, and z represents the size of one-hot-encoded arrays. As one example, in the *A. thaliana* SV dataset, there are 914 accessions, 155, 440 genotypes, and 6 variant types as defined in formula (5.1), therefore, $x = 914$, $y = 155,440$, and $z = 6$ respectively. As another example, for the *A. thaliana* SNP dataset, $x = 923$, $y = 500,000$ (only 500, 000 SNP genotypes were selected randomly from the 10 million genotypes to save computational resources), and $z = 2$ respectively.

5.3 Results and Discussion

We trained the `NovGMDeep` model on both SVs and SNPs variations of the accessions and their associated phenotypes for *A. thaliana*. We splitted all the samples into training and testing subsets with the ratios of 80% and 20%, respectively. Note that we applied the k-fold cross-validation [50] on the training sets in the model development step. Cross-validation process has a single parameter called k that refers to the number of groups that a given data sample is to be split into. In this process, the whole training dataset is randomly split into independent k-folds. Then, k-1 folds are used for the model training and one fold is used for validation. This procedure is repeated k times (iterations) so that k number of performance estimates for each iteration is obtained. In each iteration, the model is evaluated on the validation set while being trained on the training set. In most cases, the choice of k is 5 or 10, but there is no formal rule. However, the value of k relies on the size of the dataset being used. The run-time of cross-validation algorithm and computational cost come along with large values of k, that's why we chose k=3. The best models are saved while training for the three rounds of cross-validation. Then, those best models are used to calculate the PCC value on the test set. Later, the average of the three PCC values obtained from testing was finally used in the results to evaluate the model.

PCC was used to find the correlation between the observed and predicted phenotypes. The correlation between the two quantitative variables assessed for the same samples is depicted by a scatter plot in Figure 5.1.



(a) (b)

Figure 5.1: PCC analysis of *A. thaliana* for predicting flowering time using (a) SV and (b) SNP datasets. (a) The orange line represents the trend of the relationship between predicted and ground-truth values (blue points). The graph shows that predicted values by the model positively follow the ground truth values with a high confidence level (minimum fluctuation). (b) The graph shows that predicted values by the model positively follow the ground-truth values but with a lower confidence level (more fluctuation).

The horizontal axis displays the true values of the phenotypes, while the vertical axis shows the values of the predicted phenotypes. Each value in the data is represented by a point on the graph. The strength of the association between the two variables is a key aspect of a scatterplot. The linear regression line shows the best fit between predicted and true values. The slope conveys information about the strength of

the relationship between true and predicted values. When the slope is 1, the correlation between the two comparable variables is the strongest. Figure 5.1(a) and Figure 5.1(b) show the results between ground truth and predicted phenotypic values for `NovGMDeep` using SVs and SNPs respectively for predicting flowering time of *A. thaliana*. These PCC graphs were created on the testing samples and show the accuracy of the proposed model in predicting phenotypes. The orange line in these graphs represents the trend of the relationship between the observed and predicted phenotypes (blue points) by the NovGMDeep model.

As shown in Figure. 5.1, the predicted values which are close to the fitting line indicate a strong correlation, and values far from the fitting line indicate a weak correlation between the true and predicted values.

The graphs for phenotypes of rosette leaf number (Figure 5.2), cauline axillary branch number (Figure 5.3), and stem length (Figure 5.4), all showed similar results as flowering time for *A. thaliana*.

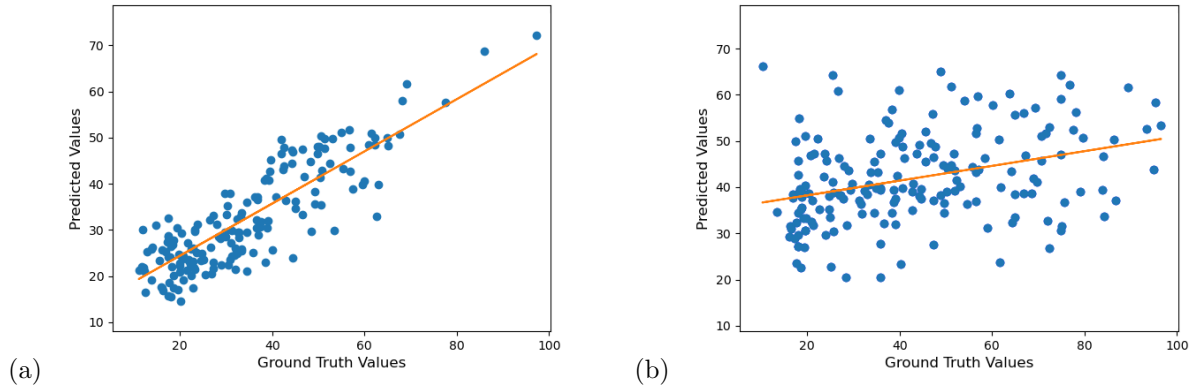
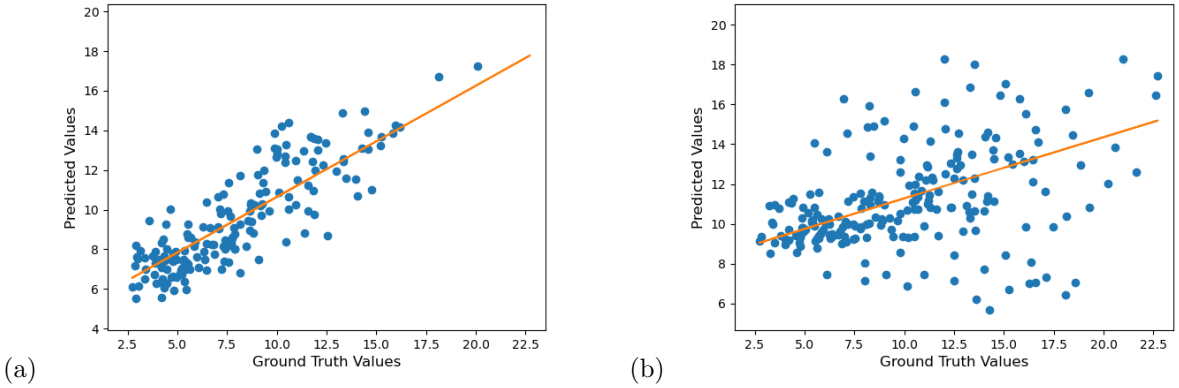
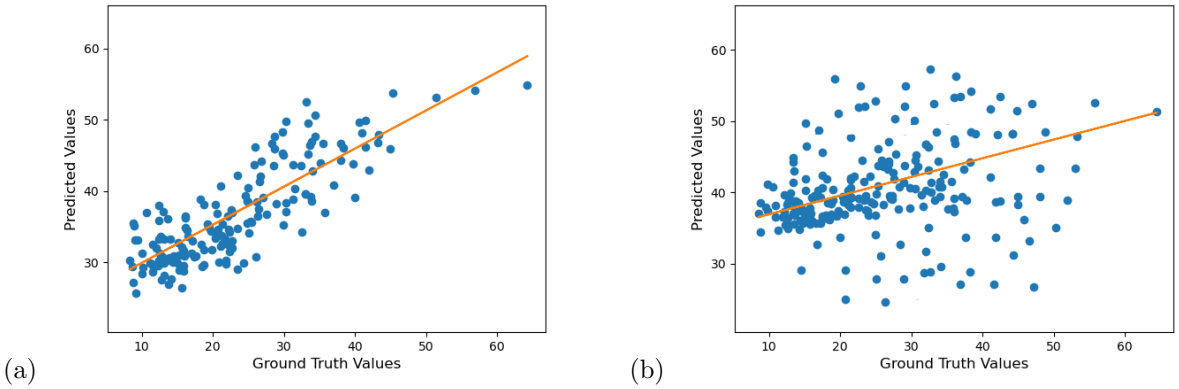


Figure 5.2: PCC analysis of *A. thaliana* for predicting rosette leaf number using (a) SV and (b) SNP datasets. (a) The orange line represents the trend of the relationship between predicted and ground-truth values (blue points). The graph shows that predicted values by the model positively follow the ground-truth values with a high confidence level (minimum fluctuation). (b) The graph shows that predicted values by the model positively follow the ground-truth values but with a lower confidence level (more fluctuation).



(a) (b)

Figure 5.3: PCC analysis of *A. thaliana* for predicting cauline axillary branch number using (a) SV and (b) SNP datasets. (a) The orange line represents the trend of the relationship between predicted and ground-truth values (blue points). The graph shows that predicted values by the model positively follow the ground-truth values with a high confidence level (minimum fluctuation). (b) The graph shows that predicted values by the model positively follow the ground-truth values but with a lower confidence level (more fluctuation).



(a) (b)

Figure 5.4: PCC analysis of *A. thaliana* for predicting stem length using (a) SV and (b) SNP datasets. (a) The orange line represents the trend of the relationship between predicted and ground-truth values (blue points). The graph shows that predicted values by the model positively follow the ground truth values with a high confidence level (minimum fluctuation). (b) The graph shows that predicted values by the model positively follow the ground-truth values but with a lower confidence level (more fluctuation).

Performance Evaluation

The aforementioned two statistical methods in Section 3.4.1, rrBLUP and gBLUP, were evaluated with the same data to compare their overall prediction performance with `NovGMDeep`. The `mixed.solve` and `BGLR` function in the rrBLUP and gBLUP package is used to solve these BLUP models. These packages are ran in R environment. The genotype data is encoded as integer numbers: 0 & 1 indicating the reference & alternate genotype of an inversion respectively; 2 & 3 indicating the reference & alternate genotype of a duplication respectively; 4 & 5 indicating the reference & alternate genotype of a deletion respectively; -1 indicating the missing genotype. The training and testing datasets are splitted with ratios of 80% and 20%, respectively. And 3-fold cross-validation was applied in the same manner as the proposed model described in Section5.3.

As shown in Table 5.1, phenotypes predicted by `NovGMDeep` show the strongest correlations with observed phenotypes. Therefore, `NovGMDeep` outperforms the other statistical models in predicting phenotype with SV data. Moreover, phenotypes predicted by `NovGMDeep` using SV data show stronger correlations with observed phenotypes than the ones predicted using SNP data. This shows the effectiveness of using SV marker data instead of SNPs for phenotype prediction.

Table 5.1: Pearson's coefficients of correlations between the observed and predicted phenotypes on the testing set. SVs and SNPs are used to capture the genomic variations among the accessions of *A. thaliana*. Results show the average of the three PCC values taken from the three-fold cross-validation results of the model. The highlighted values show the highest PCC values obtained using SVs.

Model	Flowering Time		Rosette Leaf Number		Branch Number		Stem Length	
	PCC (SVs)	PCC (SNPs)	PCC (SVs)	PCC (SNPs)	PCC (SVs)	PCC (SNPs)	PCC (SVs)	PCC (SNPs)
rrBLUP	0.664	0.520	0.653	0.547	0.675	0. 513	0.690	0.502
gBLUP	0.657	0.516	0.662	0.512	0.683	0.530	0.681	0.514
NovGMDeep	**0.788**	**0.594**	**0.742**	**0.563**	**0.761**	**0.581**	**0.759**	**0.572**

The MAE loss was watched to select the best model during the training process of 150 epochs with early stopping. The model with the lowest MAE was selected to evaluate the model on the test set as shown in Table 5.2.

The SD of all the MAE values on the test set was also calculated to see the deviation among those values. The MAE measures the average absolute error over the dataset while the SD measures how far the absolute error on each training point is from the MAE. A low SD means that errors across the dataset tend to have similar values close to the mean. A high SD tells that the errors are spread over a bigger range. This can provide insight into the model: a model with a low MAE indicates a good "average" performance over the dataset, and if that model also has a low SD then it tells that the performance is not only good on average but also uniform on the dataset. The low values of MAE and SD of MAE for the SV markers show higher prediction accuracy of `NovGMDeep` model over the SNPs.

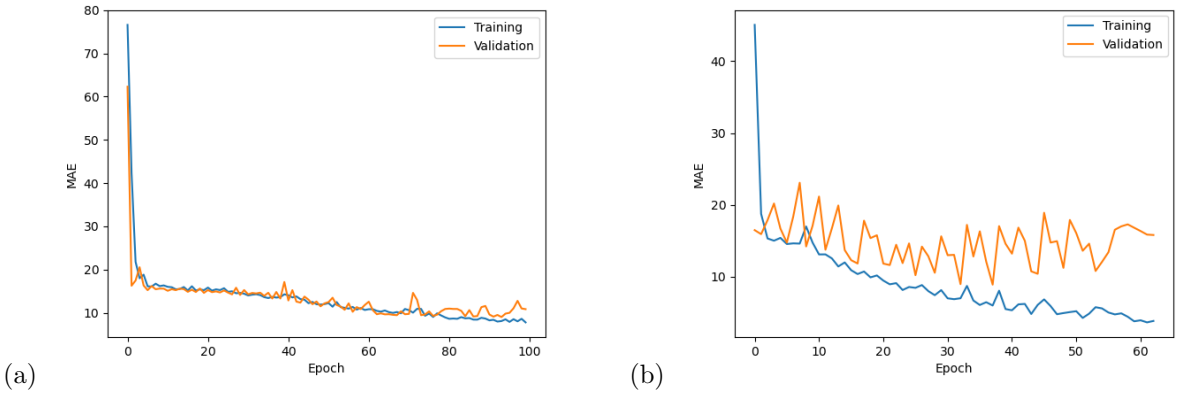
Figure 5.5 demonstrates a typical training process for all the datasets. The trend shows that the model

Table 5.2: Mean Absolute Error (MAE) and Standard Deviation of MAE calculated by `NovGMDeep` model on the test sets for predicting different phenotypes of *A. thaliana* using SV and SNP datasets.

Phenotypes	SVs		SNPs	
	MAE	**SD**	**MAE**	**SD**
Flowering Time	7.64	6.51	13.71	11.56
Rosette Leaf Number	8.23	7.02	14.89	12.47
Branch Number	7.89	6.52	14.84	12.62
Stem Length	7.92	6.71	14.31	13.09

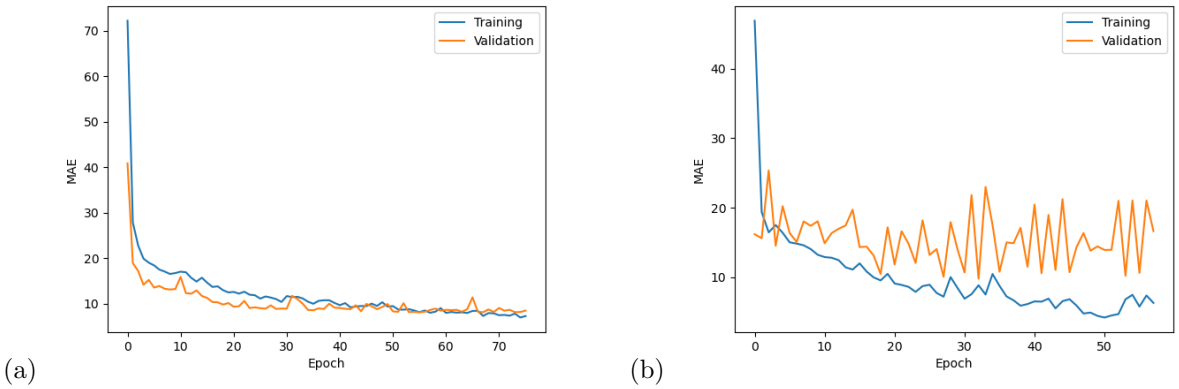
keeps learning the discriminative features of the data for the first few epochs as both the training and validation losses are decreasing to a significant amount at the start. For the rest of the epochs, the flattened part of the graph illustrates that the model is capable to handle the co-variate shift between the training and validation set because now the training and validation losses are decreasing steadily. The closeness of the training and validation curves demonstrates that the model is neither overfitted nor under-fitted in Figure 5.5(a). In Figure 5.5(b) the training curve is lower than the validation curve, which shows that the model is overfitted. An under-fitted model will have a high training and high validation loss while an overfitted model will have an extremely low training loss but a high validation loss. A large number of features (SNPs) expands the hypothesis space, making the data more sparse and this might also lead to overfitting problems.

The loss graphs for the rest of the phenotypes in Figure 5.6, 5.7, and 5.8 also showed similar trends as flowering time for *A. thaliana*.



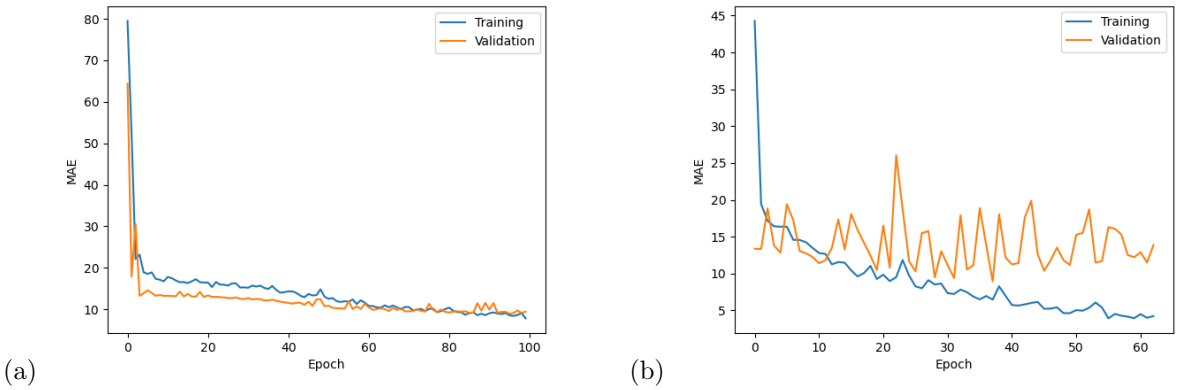
(a) (b)

Figure 5.5: Training and validation losses of `NovGMDeep` model using (a) SV and (b) SNP dataset for predicting flowering time of *A. thaliana*. The first few epochs illustrate that the model is learning quickly as both the training and validation losses are decreasing faster. Later it shows the potential of the model in handling the co-variate shift between sets as the training and validation losses are decreasing steadily now.



(a) (b)

Figure 5.6: Training and validation losses of `NovGMDeep` model using (a) SV and (b) SNP dataset for predicting rosette leaf number of *A. thaliana*. The first few epochs illustrate that the model is learning quickly as both the training and validation losses are decreasing faster. Later it shows the potential of the model in handling the co-variate shift between sets as the training and validation losses are decreasing steadily now.



(a) (b)

Figure 5.7: Training and validation losses of `NovGMDeep` model using (a) SV and (b) SNP dataset for predicting cauline axillary branch number of *A. thaliana*. The first few epochs illustrate that the model is learning quickly as both the training and validation losses are decreasing faster. Later it shows the potential of the model in handling the co-variate shift between sets as the training and validation losses are decreasing steadily now.

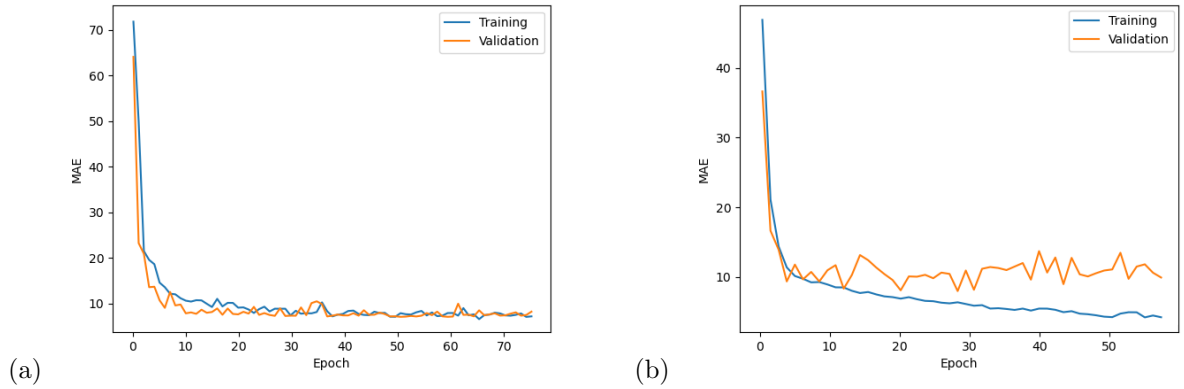


Figure 5.8: Training and validation losses of `NovGMDeep` model using (a) SV and (b) SNP dataset for predicting stem length of *A. thaliana*. The first few epochs illustrate that the model is learning quickly as both the training and validation losses are decreasing faster. Later it shows the potential of the model in handling the co-variate shift between sets as the training and validation losses are decreasing steadily now.

The `NovGMDeep` model was also compared with the aforementioned DL models. However, the comparison was possible just on the SNP data as these models are specially designed for that. G2PDeep, is a web-based framework where users can upload their datasets and predict phenotypes by creating a deep learning model according to the data. Because of the extremely high-dimensional SNP dataset, the web-based framework did not work, therefore, we ran the code uploaded into their repository that is publicly available on GitHub. DeepGS package was also ran on the command line. PCC values were calculated for all the phenotypes of *A. thaliana* in Table 5.3.

Table 5.3: Pearson's coefficients of correlations between the observed and predicted phenotypes on the testing set. SNPs are used to capture the genomic variations among the accessions of *A. thaliana*. Results show the average of the three PCC values taken from the three-fold cross validation results of the model.

Model	Flowering Time	Rosette Leaf Number	Branch Number	Stem Length
DeepGS	0.542	0.559	0.544	0.563
G2PDeep	0.561	0.557	0.571	0.569
NovGMDeep	**0.594**	**0.563**	**0.581**	**0.572**

The results in the Table 5.3 showed that `NovGMDeep` performed better than the other existing DL models. Also, it can be seen that G2PDeep has a better performance than DeepGS.

5.4 Study of sample size effects

This section describes a comprehensive study showing sample size effects when the proposed model is trained on the different numbers of samples for SVs. The increased number of training samples could improve the model to get even better predictions. Adding more samples adds diversity to the model. It decreases the generalization error because the model becomes more general under training on more samples.

For this study, we selected the first 100 samples from the whole dataset and then kept increasing the sample size by 100 for training the model on nine different datasets. Table 5.4 shows the results for the different number of samples for the SVs for predicting the flowering time of *A. thaliana*. The results improved when the number of samples increased. Also, the PCC values for 700, 800, and total samples (914) are high compared to the other sample sizes. The PCC values were also quite close to each other for the datasets with 700, 800, and 914 samples. The results showed better PCC values when the model is trained on more than 700 samples. Also, the MAE decreased when the sample size increased, which shows improvement in the predicted results of the proposed model.

Table 5.4: Different number of samples for the SV markers to train `NovGMDeep` to predict flowering time of *A. thaliana*.

Number of Samples	PCC	MAE
100	-0.370	15.79
200	-0.219	15.04
300	0.186	13.81
400	0.385	10.24
500	0.433	9.95
600	0.662	8.70
700	0.740	7.93
800	0.767	7.82
914	0.788	6.51

5.5 Perspectives

The low PCC values for the traditional statistical models are potentially because simple regressions cannot capture the complex relationships between genotypes and phenotypes. Though we tuned the hyperparameters for SVs, TEs, and SNPs so that the current model works with all these three types of data, it might give different results when working with another dataset/organism as different DL model architectures impact the results significantly. Also, the more significant number of input features might lead to overfitting

because the model has extra potential to become complex. DL model is often very hungry for training data [57] if it has to learn a large number of features. Training deep learning models frequently demands a large amount of data. The performance of the model will continually improve with more quantity of data. Lack of data poses a challenge because the deep learning model may not detect patterns from the data, which could result in poor performance on unseen data. Similarly, our proposed model's performance increased with the increment in the number of samples.

6 Case Study #2: Phenotype Prediction using TE markers

In this chapter, we will discuss the phenotype prediction experiment for *O. sativa* using TE markers.

6.1 *O. sativa* dataset description

Oryza sativa, commonly known as Asian rice, is a monocotyledonous flowering plant in the *Poaceae* family. It is one of the world's most significant agricultural plants, providing the primary source of nutrition for half of the world's population. *Oryza sativa subsp. japonica* is one of three primary rice subspecies: *indica*, *javanica*, and *japonica*. The grains of *Oryza sativa* are short and high in amylopectin, causing them to stick together when cooked. It is known for being simple to genetically modify and is a model organism for the botany of grains. It has a genome that is 430 Mbp in size and is distributed across 12 chromosomes. It is the most important crop for human food production as more energy is provided by rice to humans globally than by any other crop, including wheat and maize. Worldwide, 3.1% of the total agricultural land is used to grow it. Around the world, rice cultivation, production, and yield are all constantly rising [87].

The TE markers computed from the 176 *O. sativa* accessions were used in this study [124]. The SNPs for the same number of accessions were also obtained in this study in comparison to TE markers.

Four phenotypes are used in the study: panicle length, spikelet number, days to heading, and plant height measured for the same accessions [125].

1. Panicle length: The length measured from a terminal component of the rice tiller which is an inflorescence called a panicle.

2. Spikelet number: The number of panicle rice spikelets, which develop into grains.

3. Days to heading: Characterized together by the vegetative growth phase. It is the period from germination to panicle initiation and the reproductive phase of rice development, meaning the time from panicle initiation to heading.

4. Plant height: Defined as the shortest distance between the upper boundary (the highest point) of the main photosynthetic tissues (excluding inflorescences) and the ground level.

All the sources of the datasets used in this experiment can be found in Appendix A.

6.2 TE marker representation

Orzya sativa is a diploid species and has unphased genotypes. The TE and SNP markers in the raw dataset were coded as integers 1, 0, and -1 which means the reference genotype (1/1), the first (most common) alternative genotype (0/1), and the second most frequent alternative genotype (0/0) respectively. TE and SNP variants are encoded using one-hot-encoded binary arrays as shown in formula (6.1).

$$\begin{aligned} 1/1 &\rightarrow [1,0,0] \\ 0/1 &\rightarrow [0,1,0] \\ 0/0 &\rightarrow [0,0,1] \\ ./. &\rightarrow [0,0,0] \end{aligned} \tag{6.1}$$

For the TE markers, the data is represented in the format of a 3D matrix $B(u \times v \times w)$ where u refers to the accessions, v indicates the TE/SNP markers, and w shows the depth of one-hot-encoded vector. So, for the *O. sativa* TE dataset $u = 176$, $v = 6,074$, and $w = 3$ respectively. And for the *O. sativa* SNP dataset, $u = 176$, $v = 493,882$, and $w = 3$ respectively.

6.3 Results and Discussion

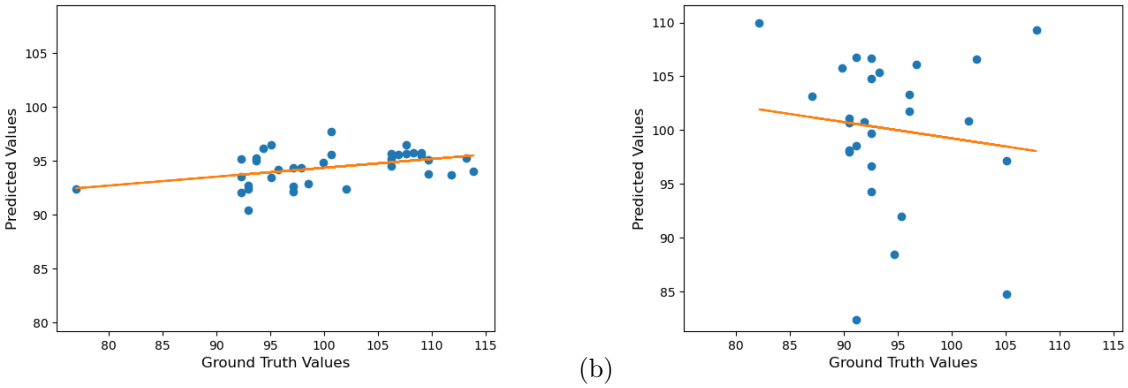
Here we trained the model on both TE and SNP data of *O. sativa*. Similar strategies as Chapter 5 are used here for splitting all the samples into training and testing subsets, applying cross-validation, saving best models, and evaluating the model.

The correlation between the two quantitative variables assessed for the same samples is depicted by scatter plots in Figure 6.1. Figure 6.1(a) and 6.1(b) show the results between observed and predicted phenotypic values for `NovGMDeep` using TEs and SNPs respectively for predicting heading time of *O. sativa*. These PCC graphs were created on the testing samples and show the accuracy of the proposed model in predicting phenotypes. The orange line in these graphs represents the trend of the relationship between the observed and predicted phenotypes (blue points) by the NovGMDeep model.

As shown in Figure 6.1, the predicted values which are close to the fitting line indicate a strong correlation, and values far from the fitting line indicate a weak correlation between the true and predicted values.

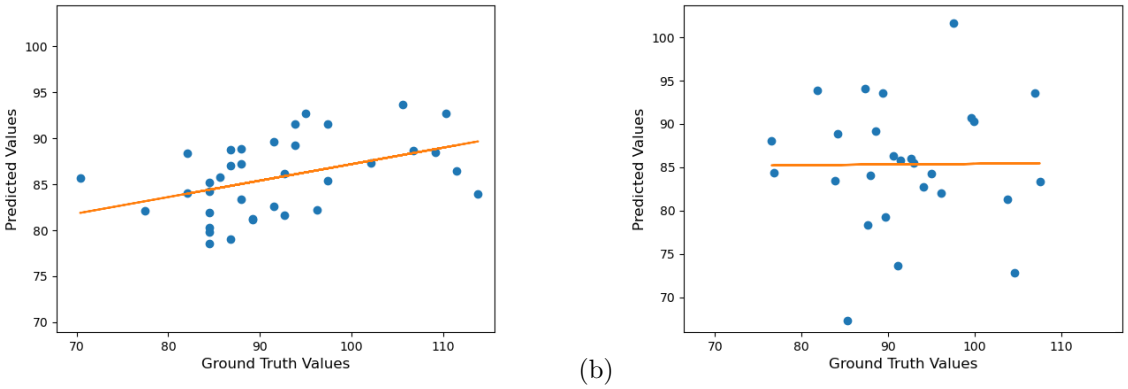
The graphs for phenotypes of plant height (Figure 6.2), spikelet number (Figure 6.3), and panicle length (Figure 6.4), all showed similar results as days to heading for *O. sativa*.

These graph show that the phenotypes predicted by `NovGMDeep` using TEs data have stronger correlations with observed phenotypes than the ones predicted using SNPs data. This shows the effectiveness of using TE marker data instead of SNPs for phenotype prediction.



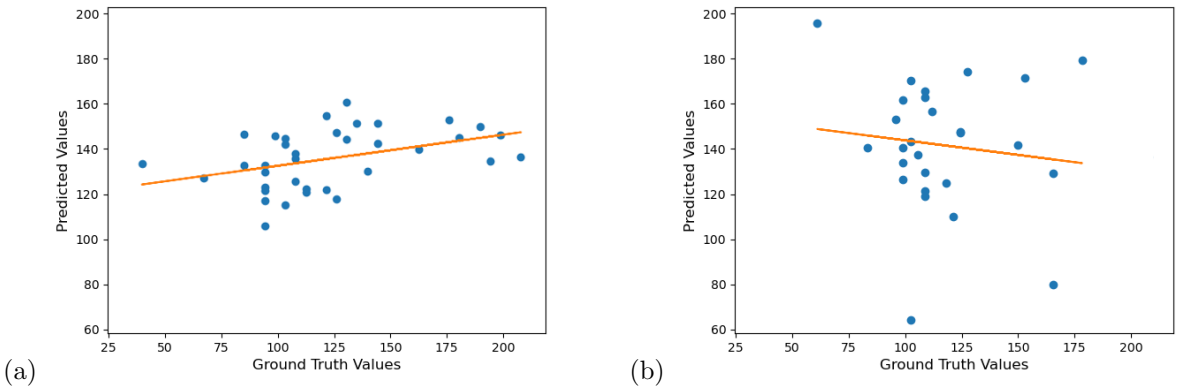
(a) (b)

Figure 6.1: PCC analysis of *O. sativa* for predicting days to heading using (a) TE and (b) SNP datasets. (a) The graph shows that predicted values by the model are positively following the ground truth values but have a low confidence level. (b) The graph shows that predicted values by the model are negatively following the ground truth values and have no confidence level at all.



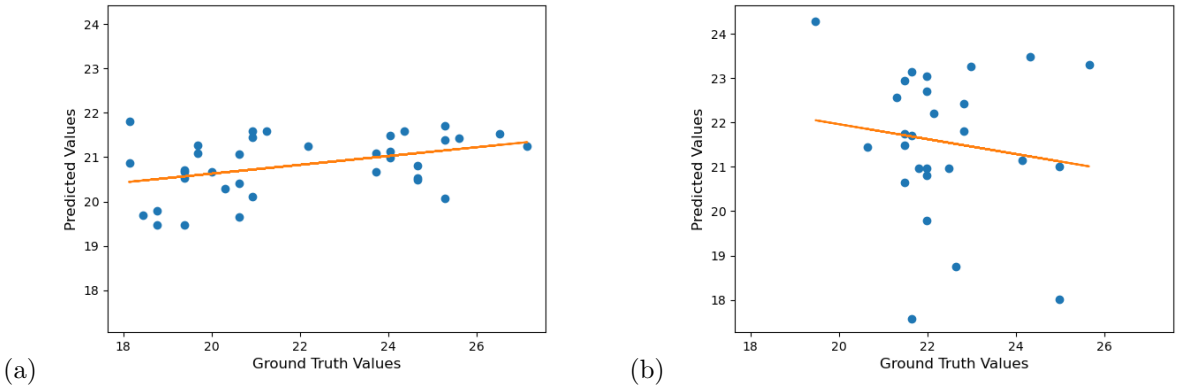
(a) (b)

Figure 6.2: PCC analysis of *O. sativa* for predicting plant height using (a) TE and (b) SNP datasets. (a) The graph shows that predicted values by the model are positively following the ground truth values but have a low confidence level. (b) The graph shows that predicted values by the model are negatively following the ground truth values and have no confidence level at all.



(a) (b)

Figure 6.3: PCC analysis of *O. sativa* for predicting spikelet number using (a) TE and (b) SNP datasets. (a) The graph shows that predicted values by the model are positively following the ground truth values but have a low confidence level. (b) The graph shows that predicted values by the model are negatively following the ground truth values and have no confidence level at all.



(a) (b)

Figure 6.4: PCC analysis of *O. sativa* for predicting panicle length using (a) TE and (b) SNP datasets. (a) The graph shows that predicted values by the model are positively following the ground truth values but have a low confidence level. (b) The graph shows that predicted values by the model are negatively following the ground truth values and have no confidence level at all.

Performance Evaluation

The aforementioned two statistical methods in Section 3.4.1, rrBLUP and gBLUP, were evaluated with the same data to compare their overall prediction performance with `NovGMDeep`. The `mixed.solve` and `BGLR` function in the rrBLUP and gBLUP package is used to solve these BLUP models. These packages are ran in R environment. The genotype data for TEs and SNPs is encoded as integer numbers: 1, 0, -1, -2 which means the reference genotype, the first (most common) alternative genotype, the second most frequent alternative genotype, and missing genotype respectively. The training and testing datasets are splitted with ratios of 80% and 20%, respectively. And 3-fold cross-validation was applied in the same manner as the proposed model described in Section5.3. TE data showed the lowest PCC values compared to the two statistical models because of the mere 176 samples as there are not enough training samples for the DL model (Table 6.1).

Table 6.1: Pearson's coefficients of correlations between the observed and predicted phenotypes on the testing set. TEs and SNPs are used to capture the genomic variations among the accessions of *O. sativa*. Results show the average of the three PCC values taken from the three-fold cross validation results of the model.

Model	Days to Heading		Plant Height		Spikelet Number		Panicle Length	
	PCC (TEs)	PCC (SNPs)	PCC (TEs)	PCC (SNPs)	PCC (TEs)	PCC (SNPs)	PCC (TEs)	PCC (SNPs)
rrBLUP	**0.439**	**0.281**	0.391	**0.275**	**0.387**	**0.294**	**0.405**	**0.209**
gBLUP	0.417	0.212	0.402	0.266	0.371	0.282	0.387	0.215
NovGMDeep	0.436	-0.014	**0.406**	0.002	0.372	-0.169	0.397	-0.148

The MAE loss was watched to select the best model during the training process of 150 epochs with early stopping. The model with the lowest MAE was selected to evaluate the model on the test set as shown in Table 6.2. The SD of all the MAE values on the test set was also calculated to see the deviation among those values. The low values of MAE and SD of MAE for the TE markers show higher prediction accuracy of `NovGMDeep` model over the SNPs.

Figure 6.5 demonstrates a typical training process for all the datasets. The closeness of the training and validation curves demonstrates that the model is neither overfitted nor under fitted in Figure 6.5(a). In Figure 6.5(b) the training curve is lower than the validation curve, which shows that the model is overfitted. For *O. sativa* also, a large number of features (SNPs) expands the hypothesis space, making the data more sparse and this might also lead to overfitting problems.

The loss graphs for rest of the phenotypes in Figure 6.6, 6.7, and 6.8 also showed similar trends as days to heading for *O. sativa*.

The `NovGMDeep` model was also compared with the aforementioned DL models. However, the comparison was possible just on the SNP data as these models are specially designed for that. G2PDeep, is a web-based framework where users can upload their datasets and predict phenotypes by creating a deep learning model according to the data. Because of the extremely high-dimensional SNP dataset, the web-based framework

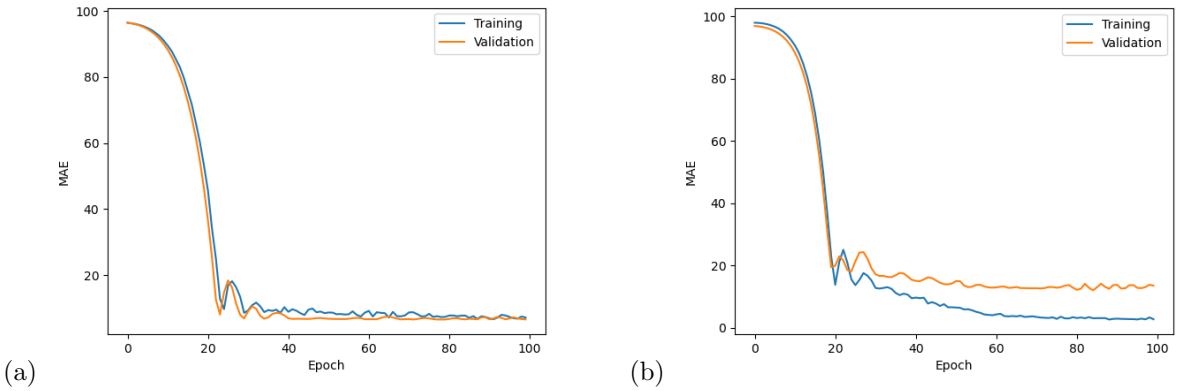


Figure 6.5: Training and validation losses of `NovGMDeep` model using (a) TE and (b) SNP dataset for predicting days to heading of *O. sativa*. The graph shows that the model is learning slowly at the start as it took 20 epochs for it to be at a steady state. Later it shows the potential of the model in handling the co-variate shift between sets as the training and validation losses are decreasing steadily now.

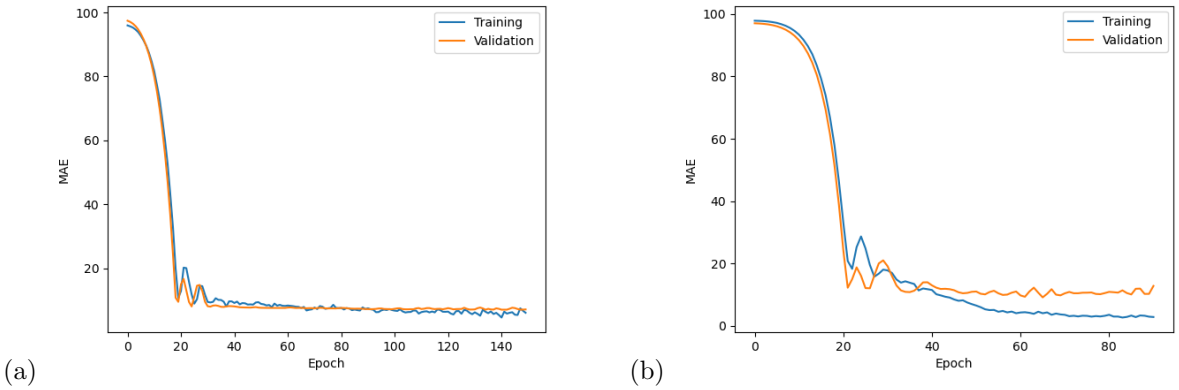
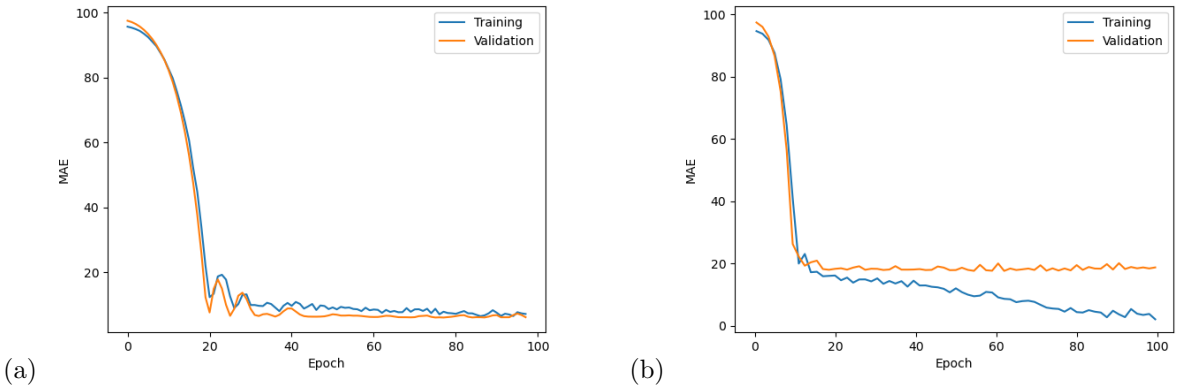


Figure 6.6: Training and validation losses of `NovGMDeep` model using (a) TE and (b) SNP dataset for predicting plant height of *O. sativa*. The graph shows that the model is learning slowly at the start as it took 20 epochs for it to be at a steady state. Later it shows the potential of the model in handling the co-variate shift between sets as the training and validation losses are decreasing steadily now.

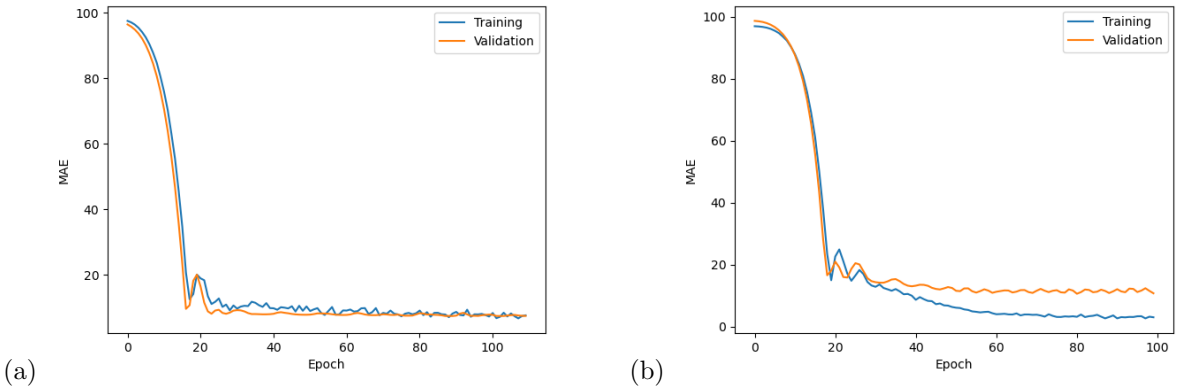
Table 6.2: Mean Absolute Error (MAE) and Standard Deviation of MAE calculated by `NovGMDeep` model on the test sets for predicting different phenotypes of *O. sativa* using TE and SNP datasets.

Phenotypes	**TEs**		**SNPs**	
	MAE	**SD**	**MAE**	**SD**
Days to Heading	8.65	6.41	16.84	18.51
Plant Height	8.01	6.57	15.92	19.74
Spikelet Number	9.26	7.22	19.46	20.25
Panicle Length	9.51	6.38	16.78	19.02



(a) (b)

Figure 6.7: Training and validation losses of `NovGMDeep` model using (a) TE and (b) SNP dataset for predicting spikelet number of *O. sativa*. The graph shows that the model is learning slowly at the start as it took 20 epochs for it to be at a steady state. Later it shows the potential of the model in handling the co-variate shift between sets as the training and validation losses are decreasing steadily now.



(a) (b)

Figure 6.8: Training and validation losses of `NovGMDeep` model using (a) TE and (b) SNP dataset for predicting panicle length of *O. sativa*. The graph shows that the model is learning slowly at the start as it took 20 epochs for it to be at a steady state. Later it shows the potential of the model in handling the co-variate shift between sets as the training and validation losses are decreasing steadily now.

did not work, therefore, we ran the code uploaded into their repository that is publicly available on GitHub. DeepGS package was also ran on the command line. PCC values were calculated for all the phenotypes of *O. sativa* in Table 6.3.

Table 6.3: Pearson's coefficients of correlations between the observed and predicted phenotypes on the testing set. SNPs are used to capture the genomic variations among the accessions of *O. sativa*. Results show the average of the three PCC values taken from the three-fold cross-validation results of the model.

Model	**Days to Heading**	**Plant Height**	**Spikelet Number**	**Panicle Length**
DeepGS	-0.004	-0.015	-0.171	-0.165
G2PDeep	0.023	-0.011	-0.036	-0.182
NovGMDeep	**-0.014**	**0.002**	**-0.169**	**-0.148**

In Table 6.3 the results for DL models suggest that statistical models are better for datasets with fewer samples as seen in Table 6.1

7 Conclusions and Future Directions

In this chapter we conclude our findings and discuss about future work.

7.1 Synopsis

Genomic selection has revolutionized the applications of breeding programs in plants and livestock [56]. Novel methods for predicting complex traits of extensive genotypic data have increasingly become essential. DL, as an advanced technique of machine learning algorithms, has the potential to find hidden patterns in massive datasets. It has the potential to serve the purpose of the principal need for breeders to improve their practices. This work aimed to develop a DL model and evaluate its accuracy in predicting the specific traits from genome-wide SV and TE markers.

In this thesis, the applications of deep learning and its relationship with GS have been explored. Although the literature suggests DL algorithms as a reliable method for predicting complex phenotypes, there are a few limitations to using deep learning techniques. In large-scale datasets, the most important consideration for high prediction accuracy is the model's architecture which requires excellent knowledge of deep learning techniques. A well-constructed network in any model is the critical source for the model's high performance. Another crucial consideration is the choice of applying convolutional, fully connected, dropout, and sampling layers with distinct sets of hyperparameters that handle specific characteristics of the data [11]. Interpreting the biological relevance of the data is also a helpful consideration to minimize the limitations of deep learning techniques in bioinformatics. The small number of DL for GS applications demonstrates the enormous potential for these models to enhance early candidate genotype selection and knowledge of the intricate biological mechanisms underlying the link between genotypes and phenotypes. This potential is partially explained by how such models are constructed, which allows them to recognize more intricate data patterns.

There are some challenges that must be overcome to employ deep learning techniques for phenotype prediction. First, finding a dataset with enough samples and related phenotypic labels to train a neural network to be efficient and broadly applicable. Several datasets with significant numbers of sequenced genomes are available, but most of these databases need more reliable phenotypic data for the linked samples. Second, the cost of sequencing sufficient depth of coverage for SV and TE calls is also high.

In this study, we developed `NovGMDeep` as a method to predict phenotypes based on different types of SVs (inversions, duplications, and deletions) and TEs. This work was motivated by the importance of SVs and TEs in genome evolution and that they could better capture variants among genomes than SNPs [110]. Also,

the authors of [124] concluded through GWAS that TE markers have a similar ability to discover association patterns to SNP markers. Therefore, the model is not only for predicting phenotypes from SVs and TEs but also shows that they can be more beneficial in genomic selection than SNPs. With the proposed model, it is still challenging to link phenotypes to genetic structural variation and environmental variables, which will be a future direction to explore.

Compared with traditional statistical models, rrBLUP and gBLUP have low PCC values compared to the `NovGMDeep` model for the *A. thaliana* data. This is potentially because simple regressions cannot capture the complex relationships between genotypes and phenotypes. For the *O. sativa* data, the PCC value is low for TEs and even negative for SNPs because of the low sample size to train the DL model. The results for the TE markers are better than the SNPs, which makes the former more useful. The performance comparison between `NovGMDeep` and other DL models, such as: DeepGS [78] and G2PDeep [128], was only possible with SNP data. `NovGMDeep` outperformed the other two DL models for *A. thaliana* whereas the low sample size problem persisted for *O. sativa*. Though we tuned the hyper-parameters for SVs, TEs and SNPs datasets in a way that the current model works with all these three types of data, it might give different results when working with other dataset/organism. As different model architectures impact the results significantly.

The search for datasets with more than 176 samples and associated phenotypes for TEs was unsuccessful at the time of writing this thesis. It was challenging to develop a generalizable prediction model for TEs due to the issues with overfitting and short validation sets caused by using datasets with small sample sizes. Also, the study of sample size effects for SVs explains the low performance of the proposed model for TE markers as the PCC value is very low for less number of samples to train `NovGMDeep` for the SV data as well.

Due to a large number of SNP markers across the genome used as features, overfitting increases in the model. This is potentially because the number of SNPs is significantly larger than the number of samples, causing the "small-n-large-p problem".

Overall, the `NovGMDeep` model was well implemented in predicting phenotypes using SVs genotypic markers and could be added to the toolkits of crop breeders. Also, the results showed the importance of selecting algorithms and hyperparameters for genotype-to-phenotype predictions. In conclusion, this study paves the way for the novel use of SVs and TEs in the field of GS.

7.2 Future Work

The study also shows that the nature of data, network architecture, and the type of trait which is being predicted play an influential role in model training and prediction. It is essential to focus on individuals with high phenotypic trait values so that they can serve as selected assets for different breeding programs for other vital crops. In future work, we will explore finding top-ranked individuals with high phenotypic values by incorporating all types of genomic data, such as SVs, TEs, and SNPs together. Moreover, environmental data, such as soil, PH-value, and temperature can be taken into consideration by incorporating an environment

vector while training the model in phenotype prediction.

Other potential future work includes reducing the size of SNP features by turning those into haplotypes. This can be done by collapsing group of SNPs to reduce their number. A haplotype refers to the inheritance of a cluster of SNPs. Also two SNPs show linkage disequilibrium (LD) more often when they are not together. LD is the non-random association of alleles at different loci within haplotypes. It measures the degree to which alleles at two loci are associated based on the allele frequencies at those two loci. It would be an interesting direction to explore as SVs capture rearrangements but SNPs within LD can not. Also, another direction is to assess if SV markers are better than TE markers or vice-versa and if there are ways to combine them together.

References

[1] https://www.britannica.com/science/genetic-marker/. Accessed: 2022-8-24.

[2] https://www.judithrecht.com/blog/jumping-pieces-in-our-dna-called-transposons. Accessed: 2022-11-22.

[3] Abien Fred Agarap. Deep learning using rectified linear units (relu). *CoRR*, 2018.

[4] Charu C Aggarwal et al. Neural networks and deep learning. *Springer*, 10:978–3, 2018.

[5] Timothy J Aitman, Charles Boone, Gary A Churchill, Michael O Hengartner, Trudy FC Mackay, and Derek L Stemple. The future of model organisms in human disease research. *Nature Reviews Genetics*, 12(8):575–582, 2011.

[6] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. IEEE, 2017.

[7] Michael Alonge, Xingang Wang, Matthias Benoit, Sebastian Soyk, Lara Pereira, Lei Zhang, Hamsini Suresh, Srividya Ramakrishnan, Florian Maumus, Danielle Ciren, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, 182(1):145–161, 2020.

[8] Carlos Alonso-Blanco, Jorge Andrade, Claude Becker, Felix Bemm, Joy Bergelson, Karsten M Borgwardt, Jun Cao, Eunyoung Chae, Todd M Dezwaan, Wei Ding, et al. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, 166(2):481–491, 2016.

[9] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):1–74, 2021.

[10] Seied Abdolhamid Angaji. Single nucleotide polymorphisms and association studies: A few critical points. *Biomacromolecular Journal*, 3(1):1–4, 2017.

[11] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular Systems Biology*, 12(7):878, 2016.

[12] Bart Baesens, David Martens, Rudy Setiono, and Jacek M Zurada. Special section on white box nonlinear prediction models. *IEEE Transactions on Neural Networks*, 22(12):2406–2408, 2011.

[13] Nicholas H Barton. Genetic hitchhiking. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355(1403):1553–1562, 2000.

[14] Jeffrey L Bennetzen. Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology*, 42(1):251–269, 2000.

[15] Madhav Bhatta, Lucia Gutierrez, Lorena Cammarota, Fernanda Cardozo, Silvia Germán, Blanca Gómez-Guerrero, María Fernanda Pardo, Valeria Lanaro, Mercedes Sayas, and Ariel J Castro. Multitrait genomic prediction model increased the predictive ability for agronomic and malting quality traits in barley (hordeum vulgare l.). *G3: Genes, Genomes, Genetics*, 10(3):1113–1124, 2020.

[16] R Bracewell. Pentagram notation for cross correlation. the fourier transform and its applications. *New York: McGraw-Hill*, 46:243, 1965.

[17] Brian Charlesworth. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205, 2009.

[18] Colby Chiang, Ryan M Layer, Gregory G Faust, Michael R Lindberg, David B Rose, Erik P Garrison, Gabor T Marth, Aaron R Quinlan, and Ira M Hall. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods*, 12(10):966–968, 2015.

[19] Mete Civelek and Aldons J Lusis. Systems genetics approaches to understand complex traits. *Nature Reviews Genetics*, 15(1):34–48, 2014.

[20] Samuel A Clark and Julius van der Werf. Genomic best linear unbiased prediction (gblup) for the estimation of genomic breeding values. In *Genome-wide association studies and genomic prediction*, pages 321–330. Springer, 2013.

[21] Donald F Conrad, Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts, T Daniel Andrews, Chris Barnes, Peter Campbell, et al. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, 2010.

[22] 1000 Genomes Project Consortium et al. A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061, 2010.

[23] Richard Cordaux and Mark A Batzer. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10):691–703, 2009.

[24] Mircea Cretu Stancu, Markus J Van Roosmalen, Ivo Renkens, Marleen M Nieboer, Sjors Middelkamp, Joep De Ligt, Giulia Pregno, Daniela Giachino, Giorgia Mandrile, Jose Espejo Valle-Inclan, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature Communications*, 8(1):1–13, 2017.

[25] José Crossa, Diego Jarquín, Jorge Franco, Paulino Pérez-Rodríguez, Juan Burgueño, Carolina Saint-Pierre, Prashant Vikram, Carolina Sansaloni, Cesar Petroli, Deniz Akdemir, et al. Genomic prediction of gene bank wheat landraces. *G3: Genes, Genomes, Genetics*, 6(7):1819–1834, 2016.

[26] Elise Cuyvers and Kristel Sleegers. Genetic variations underlying alzheimer's disease: evidence from genome-wide association studies and beyond. *The Lancet Neurology*, 15(8):857–868, 2016.

[27] Li Deng, Geoffrey Hinton, and Brian Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8599–8603. IEEE, 2013.

[28] Raquel Dias and Ali Torkamani. Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine*, 11(1):1–12, 2019.

[29] Aria Dolatabadian, Dhwani Apurva Patel, David Edwards, and Jacqueline Batley. Copy number variation and disease resistance in plants. *Theoretical and Applied Genetics*, 130(12):2479–2490, 2017.

[30] Marisol Domínguez, Elise Dugas, Médine Benchouaia, Basile Leduque, José M Jiménez-Gómez, Vincent Colot, and Leandro Quadrana. The impact of transposable elements on tomato diversity. *Nature Communications*, 11(1):1–11, 2020.

[31] Frank Emmert-Streib, Zhen Yang, Han Feng, Shailesh Tripathi, and Matthias Dehmer. An introductory review of deep learning for prediction models with big data. *Frontiers in Artificial Intelligence*, 3:4, 2020.

[32] Jeffrey B Endelman. Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*, 4(3), 2011.

[33] Zubair Md Fadlullah, Fengxiao Tang, Bomin Mao, Nei Kato, Osamu Akashi, Takeru Inoue, and Kimihiro Mizutani. State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems. *IEEE Communications Surveys & Tutorials*, 19(4):2432–2455, 2017.

[34] J-B Fan, A Oliphant, R Shen, BG Kermani, F Garcia, KL Gunderson, M Hansen, F Steemers, SL Butler, P Deloukas, et al. Highly parallel SNP genotyping. In *Cold Spring Harbor symposia on quantitative biology*, volume 68, pages 69–78. Cold Spring Harbor Laboratory Press, 2003.

[35] Livio Favaro, Elodie F Briefer, and Alan G McElligott. Artificial neural network approach for revealing individuality, group membership and age information in goat kid contact calls. *Acta Acustica United with Acustica*, 100(4):782–789, 2014.

[36] Cédric Feschotte, Ning Jiang, and Susan R Wessler. Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics*, 3(5):329–341, 2002.

[37] EJ Finnegan, RK Genger, WJ Peacock, and ES Dennis. DNA methylation in plants. *Annual Review of Plant Biology*, 49:223, 1998.

[38] Jennifer L Freeman, George H Perry, Lars Feuk, Richard Redon, Steven A McCarroll, David M Altshuler, Hiroyuki Aburatani, Keith W Jones, Chris Tyler-Smith, Matthew E Hurles, et al. Copy number variation: new insights in genome diversity. *Genome Research*, 16(8):949–961, 2006.

[39] Errol C Friedberg. DNA damage and repair. *Nature*, 421(6921):436–440, 2003.

[40] Iulian Gabur, Harmeet Singh Chawla, Rod J Snowdon, and Isobel AP Parkin. Connecting genome structural variation with complex traits in crop plants. *Theoretical and Applied Genetics*, 132(3):733–750, 2019.

[41] Salvador A Gezan, Luis F Osorio, Sujeet Verma, and Vance M Whitaker. An experimental validation of genomic selection in octoploid strawberry. *Horticulture Research*, 4, 2017.

[42] Hossein Gholamalinezhad and Hossein Khosravi. Pooling methods in deep neural networks, a review. *arXiv preprint arXiv:2009.07485*, 2020.

[43] Stephanie Glen. Correlation coefficient: Simple definition, formula, easy steps. *StatisticsHowTo.com. Available online: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/(accessed on 3 August 2020)*, 2021.

[44] Mehmet Göktay, Andrea Fulgione, and Angela M Hancock. A new catalog of structural variants in 1,301 *A. thaliana* lines from africa, eurasia, and north america reveals a signature of balancing selection at defense response genes. *Molecular Biology and Evolution*, 38(4):1498–1511, 2021.

[45] Liang Gong, Chee-Hong Wong, Wei-Chung Cheng, Harianto Tjong, Francesca Menghi, Chew Yee Ngan, Edison T Liu, and Chia-Lin Wei. Picky comprehensively detects high-resolution structural variants in nanopore long reads. *Nature Methods*, 15(6):455–460, 2018.

[46] Antonio Gulli and Sujit Pal. *Deep learning with Keras*. Packt Publishing Ltd, 2017.

[47] David Habier, Rohan L Fernando, Kadir Kizilkaya, and Dorian J Garrick. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12(1):1–12, 2011.

[48] Jemanesh K Haile, Amidou N'Diaye, Fran Clarke, John Clarke, Ron Knox, Jessica Rutkoski, Filippo M Bassi, and Curtis J Pozniak. Genomic selection for grain yield and quality traits in durum wheat. *Molecular breeding*, 38(6):1–18, 2018.

[49] Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture. *Advances in Neural Information Processing Systems*, 31, 2018.

[50] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[51] Ben J Hayes, Phillip J Bowman, Amanda C Chamberlain, Klara Verbyla, and Mike E Goddard. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution*, 41(1):1–9, 2009.

[52] Ben J Hayes, Phillip J Bowman, Amanda J Chamberlain, and Michael E Goddard. Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, 92(2):433–443, 2009.

[53] Tianhua He and Chengdao Li. Harness the power of genomic selection and the potential of germplasm in crop breeding for global food security in the era with rapid climate change. *The Crop Journal*, 8(5):688–700, 2020.

[54] David Heller and Martin Vingron. SVIM: structural variant identification using mapped long reads. *Bioinformatics*, 35(17):2907–2915, 2019.

[55] Glenn Hickey, David Heller, Jean Monlong, Jonas A Sibbesen, Jouni Sirén, Jordan Eizenga, Eric T Dawson, Erik Garrison, Adam M Novak, and Benedict Paten. Genotyping structural variants in pangenome graphs using the VG toolkit. *Genome Biology*, 21(1):1–17, 2020.

[56] William G Hill. Applications of population genetics to animal breeding, from wright, fisher and lush to genomic prediction. *Genetics*, 196(1):1–16, 2014.

[57] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[58] John Huddleston, Mark JP Chaisson, Karyn Meltz Steinberg, Wes Warren, Kendra Hoekzema, David Gordon, Tina A Graves-Lindsay, Katherine M Munson, Zev N Kronenberg, Laura Vives, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*, 27(5):677–685, 2017.

[59] Md Amdadul Huq, Shahina Akter, Ill Sup Nou, Hoy Taek Kim, Yu Jin Jung, and Kwon Kyoo Kang. Identification of functional SNPs in genes and their effects on plant phenotypes. *Journal of Plant Biotechnology*, 43(1):1–11, 2016.

[60] Md Mohaiminul Islam, Pingzhao Hu, and Yang Wang. Deep learning models for predicting phenotypic traits and diseases from omics data. In *Artificial Intelligence-Emerging Trends and Applications*. IntechOpen, 2018.

[61] Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. *arXiv preprint arXiv:1702.05659*, 2017.

[62] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[63] Marty Kardos, Ellie E Armstrong, Sarah W Fitzpatrick, Samantha Hauser, Philip W Hedrick, Joshua M Miller, David A Tallmon, and W Chris Funk. The crucial role of genome-wide genetic variation in conservation. *Proceedings of the National Academy of Sciences*, 118(48):e2104642118, 2021.

[64] Sehee Kim. *Deep learning with R, François Chollet, Joseph J. Allaire, Shelter Island, NY: Manning*. Wiley Periodicals, Inc., 2020.

[65] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[66] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR 2011*, pages 233–240. IEEE, 2011.

[67] Jinsheng Lai, Ruiqiang Li, Xun Xu, Weiwei Jin, Mingliang Xu, Hainan Zhao, Zhongkai Xiang, Weibin Song, Kai Ying, Mei Zhang, et al. Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature Genetics*, 42(11):1027–1030, 2010.

[68] David E Larson, Haley J Abel, Colby Chiang, Abhijit Badve, Indraniel Das, James M Eldred, Ryan M Layer, and Ira M Hall. SVtools: population-scale analysis of structural variation. *Bioinformatics*, 35(22):4782–4787, 2019.

[69] Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, 15(6):1–19, 2014.

[70] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[71] Sue Han Lee, Chee Seng Chan, Simon Joseph Mayo, and Paolo Remagnino. How deep learning extracts and learns leaf features for plant classification. *Pattern Recognition*, 71:1–13, 2017.

[72] Ben Lehner. Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*, 27(8):323–331, 2011.

[73] Yangkun Li, Weizhi Ma, Chong Chen, Min Zhang, Yiqun Liu, Shaoping Ma, and Yuekui Yang. A survey on dropout methods and experimental verification in recommendation. *arXiv preprint arXiv:2204.02027*, 2022.

[74] Yun Rose Li, Joseph T Glessner, Bradley P Coe, Jin Li, Maede Mohebnasab, Xiao Chang, John Connolly, Charlly Kao, Zhi Wei, Jonathan Bradfield, et al. Rare copy number variants in over 100,000 european ancestry subjects reveal multiple disease associations. *Nature Communications*, 11(1):1–9, 2020.

[75] Zhikai Liang, Shashi K Gupta, Cheng-Ting Yeh, Yang Zhang, Daniel W Ngu, Ramesh Kumar, Hemant T Patil, Kanulal D Mungra, Dev Vart Yadav, Abhishek Rathore, et al. Phenotypic data from inbred parents can improve genomic prediction in pearl millet hybrids. *G3: Genes, Genomes, Genetics*, 8(7):2513–2522, 2018.

[76] Maoxuan Lin, Sarah Whitmire, Jing Chen, Alvin Farrel, Xinghua Shi, and Jun-tao Guo. Effects of short indels on protein structure and function in human genomes. *Scientific Reports*, 7(1):1–9, 2017.

[77] Yang Liu, Duolin Wang, Fei He, Juexin Wang, Trupti Joshi, and Dong Xu. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Frontiers in Genetics*, 10:1091, 2019.

[78] Wenlong Ma, Zhixu Qiu, Jie Song, Jiajia Li, Qian Cheng, Jingjing Zhai, and Chuang Ma. A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta*, 248(5):1307–1318, 2018.

[79] Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J Sedlazeck. Structural variant calling: the long and the short of it. *Genome biology*, 20(1):1–14, 2019.

[80] Adriano T Mastrodomenico, Martin O Bohn, Alexander E Lipka, and Frederick E Below. Genomic selection using maize ex-plant variety protection germplasm for the prediction of nitrogen-use traits. *Crop Science*, 59(1):212–220, 2019.

[81] Claire Mérot, Rebekah A Oomen, Anna Tigano, and Maren Wellenreuther. A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution*, 35(7):561–572, 2020.

[82] Theo HE Meuwissen. Accuracy of breeding values of unrelated individuals predicted by dense snp genotyping. *Genetics Selection Evolution*, 41(1):1–9, 2009.

[83] Theo HE Meuwissen, Ben J Hayes, and ME1461589 Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

[84] Blake C Meyers, Scott V Tingey, and Michele Morgante. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Research*, 11(10):1660–1676, 2001.

[85] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in Bioinformatics*, 18(5):851–869, 2017.

[86] Osval Antonio Montesinos-López, Abelardo Montesinos-López, Paulino Pérez-Rodríguez, José Alberto Barrón-López, Johannes WR Martini, Silvia Berenice Fajardo-Flores, Laura S Gaytan-Lugo, Pedro C Santana-Mancilla, and José Crossa. A review of deep learning applications for genomic selection. *BMC Genomics*, 22(1):1–23, 2021.

[87] NM Nayar. *Origins and phylogeny of rices*. Elsevier, 2014.

[88] Jost Neigenfind, Gabor Gyetvai, Rico Basekow, Svenja Diehl, Ute Achenbach, Christiane Gebhardt, Joachim Selbig, and Birgit Kersten. Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. *BMC Genomics*, 9(1):1–26, 2008.

[89] Pauline C Ng and Steven Henikoff. Predicting the effects of amino acid substitutions on protein function. *Annual Review of Genomics and Human Genetics*, 7(1):61–80, 2006.

[90] Chigozie Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*, 2018.

[91] Anthony J Onwuegbuzie and Larry G Daniel. Uses and misuses of the correlation coefficient. 1999.

[92] Simon Orozco-Arias, Gustavo Isaza, and Romain Guyot. Retrotransposons in plant genomes: structure, identification, and classification through bioinformatics and machine learning. *International Journal of Molecular Sciences*, 20(15):3837, 2019.

[93] Shujun Ou, Weija Su, Yi Liao, Kapeel Chougule, Jireh RA Agda, Adam J Hellinga, Carlos Santiago Blanco Lugo, Tyler A Elliott, Doreen Ware, Thomas Peterson, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, 20(1):1–18, 2019.

[94] Ju-Hyun Park, Mitchell H Gail, Clarice R Weinberg, Raymond J Carroll, Charles C Chung, Zhaoming Wang, Stephen J Chanock, Joseph F Fraumeni Jr, and Nilanjan Chatterjee. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences*, 108(44):18026–18031, 2011.

[95] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[96] Miguel Pérez-Enciso and Laura M Zingaretti. A guide on deep learning for complex trait genomic prediction. *Genes*, 10(7):553, 2019.

[97] Annapurna Poduri, Gilad D Evrony, Xuyu Cai, and Christopher A Walsh. Somatic mutation, genomic variation, and neurological disease. *Science*, 341(6141):1237758, 2013.

[98] Matin Qaim. Role of new plant breeding technologies for food security and sustainable agricultural development. *Applied Economic Perspectives and Policy*, 42(2):129–150, 2020.

[99] Lunwen Qian, Kai Voss-Fels, Yixin Cui, Habib U Jan, Birgit Samans, Christian Obermeier, Wei Qian, and Rod J Snowdon. Deletion of a stay-green gene associates with adaptive selection in brassica napus. *Molecular Plant*, 9(12):1559–1569, 2016.

[100] Leandro Quadrana, Amanda Bortolini Silveira, George F Mayhew, Chantal LeBlanc, Robert A Martienssen, Jeffrey A Jeddeloh, and Vincent Colot. The *Arabidopsis thaliana* mobilome and its impact at the species level. *eLife*, 5:e15716, 2016.

[101] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 2012.

[102] Neha Samir Roy, Ji-Yeong Choi, Sung-Il Lee, and Nam-Soo Kim. Marker utility of transposable elements for plant genetics, breeding, and ecology: a review. *Genes & Genomics*, 37(2):141–151, 2015.

[103] Peter D Sasieni. From genotypes to genes: doubling the sample size. *Biometrics*, pages 1253–1261, 1997.

[104] Rachit K Saxena, David Edwards, and Rajeev K Varshney. Structural variations in plant genomes. *Briefings in Functional Genomics*, 13(4):296–307, 2014.

[105] Carl D Schlichting. The evolution of phenotypic plasticity in plants. *Annual Review of Ecology and Systematics*, pages 667–693, 1986.

[106] Fritz J Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt Von Haeseler, and Michael C Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15(6):461–468, 2018.

[107] Ritambhara Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, 32(17):i639–i648, 2016.

[108] Shashank Singh, Yang Yang, Barnabás Póczos, and Jian Ma. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quantitative Biology*, 7(2):122–137, 2019.

[109] JE Spindel, H Begum, D Akdemir, B Collard, E Redoña, JL Jannink, and S McCouch. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity*, 116(4):395–408, 2016.

[110] Paweł Stankiewicz and James R Lupski. Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61:437–455, 2010.

[111] Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–855, 2013.

[112] Matteo Togninalli, Ümit Seren, Jan A Freudenthal, J Grey Monroe, Dazhe Meng, Magnus Nordborg, Detlef Weigel, Karsten Borgwardt, Arthur Korte, and Dominik G Grimm. AraPheno and the AraGWAS catalog 2020: a major database update including RNA-Seq and knockout mutation data for *Arabidopsis thaliana*. *Nucleic Acids Research*, 48(D1):D1063–D1068, 2020.

[113] Martin Trick, Yan Long, Jinling Meng, and Ian Bancroft. Single nucleotide polymorphism (SNP) discovery in the polyploid brassica napus using solexa transcriptome sequencing. *Plant biotechnology Journal*, 7(4):334–346, 2009.

[114] Muammer Türkoğlu and Davut Hanbay. Plant disease and pest detection using deep learning-based features. *Turkish Journal of Electrical Engineering and Computer Sciences*, 27(3):1636–1651, 2019.

[115] Jordan Ubbens, Isobel Parkin, Christina Eynck, Ian Stavness, and Andrew G Sharpe. Deep neural networks for genomic prediction do not estimate marker effects. *The Plant Genome*, 14(3):e20147, 2021.

[116] M Graziano Usai, Mike E Goddard, and Ben J Hayes. LASSO with cross-validation for genomic selection. *Genetics Research*, 91(6):427–436, 2009.

[117] Rajeev K Varshney, Mahendar Thudi, Manish K Pandey, Francois Tardieu, Chris Ojiewo, Vincent Vadez, Anthony M Whitbread, Kadambot HM Siddique, Henry T Nguyen, Peter S Carberry, et al. Accelerating genetic gains in legumes for the development of prosperous smallholder agriculture: integrating genomics, phenotyping, systems modelling and agronomy. *Journal of Experimental Botany*, 69(13):3293–3312, 2018.

[118] Romain Villoutreix, Clarissa F de Carvalho, Víctor Soria-Carrasco, Dorothea Lindtke, Marisol De-la Mora, Moritz Muschick, Jeffrey L Feder, Thomas L Parchman, Zach Gompert, and Patrik Nosil. Large-scale mutation in the evolution of a gene complex for cryptic coloration. *Science*, 369(6502):460–466, 2020.

[119] Detlef Weigel and Richard Mott. The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biology*, 10(5):1–5, 2009.

[120] Sean Whalen, Rebecca M Truty, and Katherine S Pollard. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics*, 48(5):488–496, 2016.

[121] William C Wimsatt. The analytic geometry of genetics: part i: the structure, function, and early evolution of punnett squares. *Archive for History of Exact Sciences*, 66(4):359–396, 2012.

[122] Albert HC Wong, Irving I Gottesman, and Arturas Petronis. Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Human Molecular Genetics*, 14(suppl_1):R11–R18, 2005.

[123] Alencar Xavier, William M Muir, and Katy Martin Rainey. Assessing predictive properties of genome-wide selection in soybeans. *G3: Genes, Genomes, Genetics*, 6(8):2611–2616, 2016.

[124] Haidong Yan, David C Haak, Song Li, Linkai Huang, and Aureliano Bombarely. Exploring transposable element-based markers to identify allelic variations underlying agronomic traits in rice. *Plant Communications*, 3(3):100270, 2022.

[125] Kenji Yano, Eiji Yamamoto, Koichiro Aya, Hideyuki Takeuchi, Pei-ching Lo, Li Hu, Masanori Yamasaki, Shinya Yoshida, Hidemi Kitano, Ko Hirano, et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature Genetics*, 48(8):927–934, 2016.

[126] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.

[127] Yuxuan Yuan, Philipp E Bayer, Jacqueline Batley, and David Edwards. Current status of structural variation studies in plants. *Plant Biotechnology Journal*, 19(11):2153–2163, 2021.

[128] Shuai Zeng, Ziting Mao, Yijie Ren, Duolin Wang, Dong Xu, and Trupti Joshi. G2PDeep: a web-based deep-learning framework for quantitative phenotype prediction and discovery of genomic markers. *Nucleic Acids Research*, 49(W1):W228–W236, 2021.

[129] Wanwen Zeng, Mengmeng Wu, and Rui Jiang. Prediction of enhancer-promoter interactions via natural language processing. *BMC Genomics*, 19(2):13–22, 2018.

[130] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

Appendix A

Data Sources

The phenotypes for *A. thaliana* were downloaded from `https://arapheno.1001genomes.org/study/12/` and `https://arapheno.1001genomes.org/study/38/`.

The TE markers for *O. sativa* were downloaded from `https://data.cyverse.org/dav-anon/iplant/home/yanhaidong1991/opt_genos_fltmissing_fltmaf_fltukn_TE.txt` and the SNPs were downloaded from `https://data.cyverse.org/dav-anon/iplant/home/yanhaidong1991/176GATK.indelsnpsFiltered0.05M0.25.txt`.

The G2PDeep web-based framework can be found at `https://g2pdeep.org`.

The standalone G2PDeep and DeepGS packages can be found at `https://github.com/kateyliu/DL_gwas` and `https://github.com/cma2015/DeepGS` respectively.

The source code of `NovGMDeep` is publicly available on Github (`https://github.com/shivanisehrawat4/NovGMDeep.git`).

Appendix B

Detailed Analysis of the assumptions of PCC

Five assumptions must be met before we calculate the PCC between two variables [91]. As far as our datasets are concerned, all of the conditions are true. First, while measuring the two variables, we should consider the interval or ratio level. All the phenotypes used in the paper can describe by using intervals for example, plant height, branch number, stem length etc. on the real number line. These all are physical measures which fall into general continuous category.

Second, there must be a linear relationship between both variables. All the phenotypes when plotted using scatter plot, fell roughly along a straight line, which can also be seen in Figure 5.1 and Figure 6.1. The main thing to consider here is that the variables should not exhibit some other type of relationship, like quadratic.

Third, both variables should have a roughly normal distribution. We have plotted the values of flowering time for *A. thaliana* and days to heading for *O. sativa*. Figure B.1 and Figure B.2 show a bell-shaped curve for both of these phenotypes that indicates that the variables follow a normal distribution.

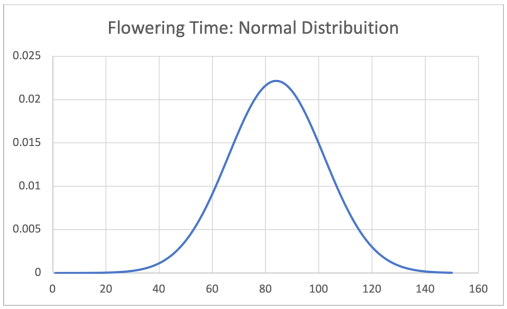


Figure B.1: The bell-shaped curve show that the values of flowering time are normally distributed.

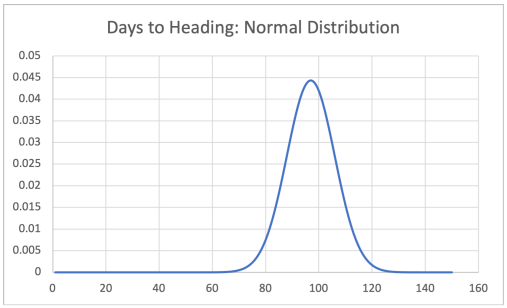


Figure B.2: The bell-shaped curve show that the values of days to heading are normally distributed.

Fourth, each observation in the dataset needs to have a pair of related values. It simply means that to calculate the correlation between the variables, each observation in the dataset has one measurement for the observed value and one measurement for the predicted value. This one was easy to check as `NovGMDeep` predicted phenotypes for all the corresponding ground truth values.

Fifth, the dataset should not contain any extreme outliers. An extreme value in the dataset substantially changes the PCC between the two variables. We did not find any extreme outlier in the dataset of predicted and observed phenotypes.