Application of Artificial Intelligence in Digital Breeding Industry Using Plant Genotype and Phenotype Data

Najeong Chae [1], Sung-Woo Byun [1,*], Jiho Choi [1], Taehoon Lim [2], Ji Hoon Lim [2], Hye In Lee [2], and Hwa Seon Shin [2]

[1] Digital Innovation Support Center, Jeonbuk Regional Branch, Korea Electronics Technology Institute, Jeonju, Korea
[2] Intelligence Integrated Software Research Center, Korea Electronics Technology Institute, Seongnam, Korea
Email: skwjd1356@keti.re.kr (N.C.); swbyun@keti.re.kr (S.-W.B.); choijh1027@keti.re.kr (J.C.); lth9029@keti.re.kr (T.L.); jhlim36573@keti.re.kr (J.H.L.); honeyb7@keti.re.kr (H.I.L.); l544@keti.re.kr (H.S.S.)
*Corresponding author

***Abstract*—With the increasing impact of climate change and population growth, enhancing crop productivity and ensuring food security have become more crucial than ever. To address this, new technologies are being introduced in the agricultural sector, and recent research has been conducted using machine learning and deep learning to analyze the relationships between crop genes and traits, aiming to boost productivity by selecting beneficial genetic information and predicting crop characteristics. In this study, we introduce machine learning-based methods for analyzing the correlation between crop genotype and phenotype, predicting phenotypes, and applying these techniques using data from tomato crops. Our results demonstrate the feasibility of predicting specific crop traits using various machine learning algorithms. Notably, we highlight the effectiveness of a semi-supervised learning approach, where synthetic genotypic data are generated to address the lack of genotype-phenotype-based datasets. Additionally, machine learning models showed similar predictive accuracy for populations cultivated in the same location over multiple years, underscoring the robustness of our approach across consistent environmental conditions.**

***Keywords*—digital breeding, artificial intelligence, genomic selection, phenotype prediction**

I. Introduction

The repercussions of climate change have intensified the need to ensure food security by increasing crop productivity, which, in turn, requires a comprehensive understanding of the genetic makeup and traits of diverse crop varieties. Genome-Wide Association Studies (GWAS) represent a well-established method for identifying associations between genetic variations and specific traits, such as disease resistance and yield characteristics [1]. By focusing on genetic attributes linked to growth, yield, tolerance, and quality, GWAS plays a critical role in the development of superior crop varieties that contribute to enhanced productivity. Phenotype prediction based on genetic information enables researchers to anticipate crop traits by identifying associations between specific genetic markers and observable characteristics [2]. This approach bridges the gap between a crop's genotype and its phenotype by uncovering complex relationships, essential for breeding novel varieties with desired traits [3]. Integrating both genotypic and environmental data facilitates the development of robust predictive models that can significantly enhance crop quality and productivity.

The recent advancements in machine learning and deep learning have expanded the scope of these technologies, with researchers increasingly applying them to accelerate and improve crop breeding processes. Data-driven artificial intelligence techniques offer more efficient and precise methods for predicting crop characteristics, potentially surpassing the traditional GWAS and phenotype prediction approaches [4–7]. This advancement is anticipated to reduce the time required for breeding, allowing for the selection of crop populations optimized for various climates and environments, and ultimately producing varieties better suited to their specific conditions [8]. These technological approaches yield substantial advantages in agriculture, including more efficient cultivar development, increased crop productivity, optimized resource usage, and the preservation of genetic diversity. However, despite the benefits of these approaches, limited research has been conducted. Specifically, there is a lack of studies on using selected Single Nucleotide Polymorphisms (SNPs) to enhance phenotype prediction accuracy via fixed-effects modeling. Additionally, there is limited validation of such models in practical breeding environments [9]. To address these research gaps, this study investigates various factors influencing prediction accuracy, including model selection, SNP characteristics, and the number of SNP markers employed. Using genotype and phenotype data from tomatoes, we further demonstrate the potential applicability of phenotype prediction models within actual breeding scenarios. In this study, we conducted a

Manuscript received October 8, 2024; revised November 2, 2024; accepted December 17, 2024; published March 14, 2025.

GWAS to identify novel alleles associated with five specific traits in tomatoes (fruit weight, width, height, firmness, and brix) using a core collection of 192 tomato varieties, with phenotypic measurements collected over two years [10]. To predict these traits, we employed 19 models for classification tasks, evaluating model performance using 3-fold cross-validation on the training set of 192 varieties. In addition to conducting phenotype prediction using all available SNP markers (51,214), we performed an alternative GWAS utilizing a selected subset of 656 SNP markers linked to the five targeted fruit traits, and re-evaluated phenotype predictions using these markers. To address the challenge of training machine learning models with a relatively small dataset (100 to 200 data points), we generated synthetic genotypic data and augmented phenotypic data through pseudo-labeling. Finally, we validated the model trained on the training population using the actual breeding population to assess its applicability within practical agricultural contexts.

II. Related Works

It is important to understand the meanings of genotype and phenotype. The genotype of an organism is its complete set of genetic material. It usually refers to genetic characteristics expressed in cells, organisms, individuals, and other entities. Genotypes can be depicted by the actual DNA sequence at a specific location, such as CC, CT, or TT [2]. Phenotype refers to the observable characteristics of an individual such as height, eye color, and blood type. It is determined by both the genome composition (genotype) and environmental factors [3]. Research has been conducted to explore Quantitative Trait Locus (QTL) for fruit traits using tomato varieties with various genetic backgrounds to analyze the genotype-phenotype relationship of crops [11, 12]. Similarly, GWAS research has been conducted on SNP markers of the genotype of chili peppers (Capsicum spp.), resulting in the selection of 25 markers associated with QTL for capsaicinoid biosynthesis genes and capsaicin content [13]. In addition to tomato and chili peppers, various GWAS studies have been conducted on soybeans, including research on improving cyst nematode resistance and chlorophyll content stability [14] understanding the correlation and genetic interactions between agronomic traits for improving crop yield and quality [7]; and exploring genetic diversity associated with seed yield, protein, oil content, and agronomic traits [15]. Additional studies have been conducted regarding phenotype prediction of crops using machine learning and deep learning. In one study, an artificial intelligence-based system was developed to improve the understanding of plant behavior in various climates for plant breeders [10], where statistical and machine learning methods were applied to predict phenotypic traits based on gene markers, environmental data, and images for crop breeding [5]. Research has also been conducted effectively to analyze genotypes using machine learning instead of traditional genomic selection technology methods to identify superior genotypes. A study was conducted on haplotype-based phenotype prediction to improve the prediction accuracy for soybean yield and composition traits by integrating advanced machine learning algorithms and optimized genetic information [16]. Research is also underway to accelerate plant breeding by applying machine learning to phenotype prediction methods [4]; to predict the flavor of sweet pepper (Capsicum annuum) using random forest regression and to identify the major metabolites of flavor differences, providing a foundation for improving specific flavor through breeding [16]; and to develop and optimize a decision method using the Support Vector Machine (SVM) algorithm for determining optimal irrigation during the growth stage of winter wheat [17]. Machine learning methods such as artificial neural networks are also increasingly being studied to improve phenotype prediction performance. Such methods have demonstrated their ability to recognize patterns in data, predict complex functions such as universal approximations, and automatically identify factors such as the superiority or dominance of genome marker information. In particular, applying machine learning to phenotype prediction can effectively estimate the effects of complex interactions without the need for assumptions about the phenotype distribution [9]. AI technologies such as high-throughput genomics and phenomics can contribute to advanced breeding techniques by accelerating the breeding process of new crop varieties. Considering the vastly increased amount of available genomic data, including unique individual characteristics and various environmental variables, it is a difficult task for users to build a highly accurate model directly from this dataset. Machine learning solves this problem to some extent, providing a valuable alternative to existing methods [8].

III. Materials and Methods

A. Genoype and Phenotype of Tomato Datasets

Tomatoes (Solanum lycopersicum L.) are a crop in high demand due to their nutritional value. In 2023, global tomato production reached approximately 190 million tons, making it one of the most widely consumed vegetables worldwide. Additionally, tomatoes exhibit extensive genetic variation in fruit traits, including shape, size, and weight. Due to the tomato crop's economic value, several studies have been conducted on tomato breeding. In this study, two tomato datasets (of Solanum lycopersicum) including genotype and phenotype data were used, both acquired from the Department of Bioresources Engineering at Sejong University, Republic of Korea [10, 12]. The training population consisted of 192 core tomato cultivars selected from contemporary breeding lines, vintage varieties, and wild species. Data included five traits: fruit weight, height, width, firmness, and brix. Genotype-By-Sequencing (GBS) analysis detected 140,072 SNPs in the 192 cultivars. GWAS was performed using a missing data rate of <20% and a small allele frequency of $\geqslant$5%. As a result, 8,550 SNPs were selected [12]. The test population consisted of 162 points

for processing, selected from 337 points collected from 14 countries. The population was obtained from a germplasm collection representing S. lycopersicum and S. pimpinellifolium. Genotyping analysis of these 162 breeding population points was performed using the tomato Axiom 55K SNP array. As a result, 51,214 SNPs from the Axiom tomato array that show polymorphism in the tomato accession were selected.

B. Preprocessing of Datasets

Genotype data typically involves the letters A, C, T, and G, which represent the four bases of DNA (adenine, cytosine, thymine, and guanine). In order to use genotype data as machine learning model input, converting genotype data to integers is required as a preprocessing step. For example, considering a diploid organism and two biallelic loci (such as SNPs) on the same chromosome, a locus with an A or T allele has three possible genotypes: (AA, AT, and TT), and a locus with G or C would have genotypes (GG, GC, and CC). We convert each combination of genotypes by mapping it to an integer; for example. AA = 0, AC = 1, and AG = 2. Phenotype data consists of a continuous series of real-number values. Due to its complexity relative to the total amount of data, phenotype data can be sensitive to outliers, leading to potential limitations in predictive accuracy. Therefore, to simplify these values, we perform preprocessing by categorizing phenotypes based on specific range criteria. We divided the data distribution into three categories for the five phenotypic traits (fruit weight, height, width, firmness, and brix). The fruit weight is measured within a range of 7.819 g to 324.22 g, and is categorized based on the thresholds of 34.013 g and 75.573 g. The fruit height is measured within a range of 20.499 mm to 102.09 mm, with 40.793 mm and 48.393mm serving as the categorization thresholds. The fruit width is measured within a range of 22.209 mm to 89.1 mm, and is categorized based on the thresholds of 37.24 mm and 54.24 mm. The fruit firmness is measured within a range of 0.315 kg to 0.92 kg, and is categorized using 0.56 kg and 0.673 kg as thresholds. The brix is measured within a range of 2.099% to 9.0%, and is categorized with the thresholds of 4.0% and 4.833%.

C. Genome-Wide Association Study (GWAS)

GWAS is a set of statistical genomic analysis techniques used to identify genetic variants associated with a particular trait using SNP markers [12]. In this study, QTL mapping was performed using biparental populations for genetic disambiguation of fruit traits, with several major genes identified as a result. However, QTL mapping in structured populations derived from two parents has the disadvantage of low mapping resolution due to limited recombination events. As an effective method for mapping complex traits, GWAS enables identification of close linkages between markers and QTLs through dense genomic coverage in unstructured populations such as germplasm collections and breeding lines. These GWAS panels have a higher recombination rate than parents, which improves mapping resolution. In this study, we performed a GWAS using a large number of SNPs spanning 12 chromosomes. This allowed us to identify alleles across a variety of traits in tomatoes. We examined a total of five fruit traits: fruit weight, width, height, firmness, and brix. We used the TASSEL v5.2.4 program for the analysis and performed LD KNNI imputation to fill in missing values before data analysis. We implemented a compressed Mixed Linear Model (MLM) that integrates information on relationships between individuals by including both fixed and random effects. In this MLM, the population structure matrix (Q) and kinship matrix (K) were used to reduce false positive associations. Subsequently, we analyzed the data independently to identify 656 SNP markers associated with each trait at a $P < 0.0001$ significance level.

D. Phenotype Prediction

In this study, classification tasks were performed for the purpose of phenotype prediction of tomato genotype data. As this genotype and phenotype data is collected through cultivation, it is difficult to acquire a dataset of the necessary size and complexity for study purposes. Considering the risk of overfitting when the model becomes more complex than the data, we used models less complex than deep learning. For the classification task, we used 19 machine learning models including tree-based models: Decision Tree (DT), Random Forest (RF), Extra Tree (ET), Gradient Boosting (GB), Extreme Gradient Boosting (XG), Light Gradient Boosting Machine (LGBM), CatBoost (CAT), and AdaBoost (ADA), linear models: Linear Regression (LR), Linear Discriminant Analysis (LDA), and Ridge (RIDGE), and others: SVM-Linear kernel (SVM-L), SVM-Radial kernel (SVM-R), K-Neighbors (KNN), naive Bayes, Gaussian process, multilayer perceptron, quadratic discriminant analysis, and dummy classifier. Tree-based models rely on the idea of building a series of decision trees that recursively partition the data space. A decision tree splits the data into branches based on feature values, making it easy to interpret and visualize. The goal is to create a tree where each branch node represents a feature decision and each leaf node corresponds to a class label. Linear models assume a linear relationship between the input features and the output. In the case of classification tasks, linear models create a decision boundary that is a straight line or hyperplane, which separates different classes. The models were trained using preprocessed SNP data as input. The output of the models was mapped to the five stated fruit traits. Examples of input data and output data are shown in Table I.

TABLE I. EXAMPLES OF CROP GENOTYPE AND PHENOTYPE DATA

Genotype					Phenotype					
Population ID	SNP ID	Chromosome	Position	Genotype	Population ID	Weight (g)	Height (mm)	Width (mm)	Brix (%)	Firmness (kg)
TC1_001	AX.95808842	1	20288	CC	TC1_001	42.26	33.27	47.36	4.00	0.64
TC1_001	AX.95772159	1	62862	TT	TC1_002	55.22	37.49	47.54	5.30	0.62
TC1_001	AX.95771895	1	65279	TT	TC1_003	12.22	34.41	27.69	5.60	0.57
TC1_001	AX.95770515	1	65409	GG	TC1_004	31.42	32.95	40.33	4.30	0.69
TC1_001	AX.95792562	1	65869	CC	TC1_005	33.70	35.60	39.82	3.90	0.88
TC1_001	AX.95780425	1	66377	CC	TC1_006	58.38	38.80	57.46	4.20	0.76

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

Fig. 1 shows the overall process of the design and experiment in this study. First, we constructed a training population of 192 tomato core collections. This population represents diverse breeding lines, vintage varieties, and wild species. Additionally, we created a testing population with 162 tomato groups grown in the same location across different years. Second, from this core collection, we obtained genotypic data that could distinguish DNA sequence variations and phenotypic data covering weight, width, height, firmness, and brix. Third, we divided a portion of the training population to create a validation population. Then, we trained machine learning models using the remaining training population. Next, we performed GWAS to extract SNP markers that were strongly associated with the traits. Subsequently, we used machine learning models trained on all SNPs and on significant SNP markers from GWAS. These models predicted the phenotypic characteristics of the validation population. Finally, we performed predictions on the test population, using both all SNPs and significant SNPs. In this study, several classification models were used to perform experiments.

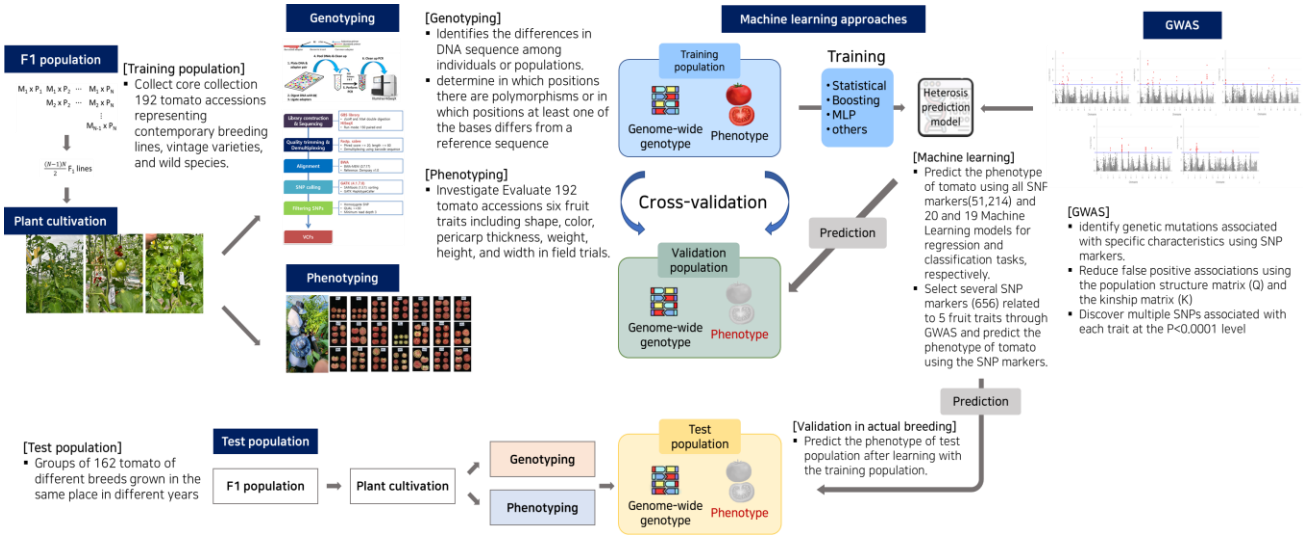


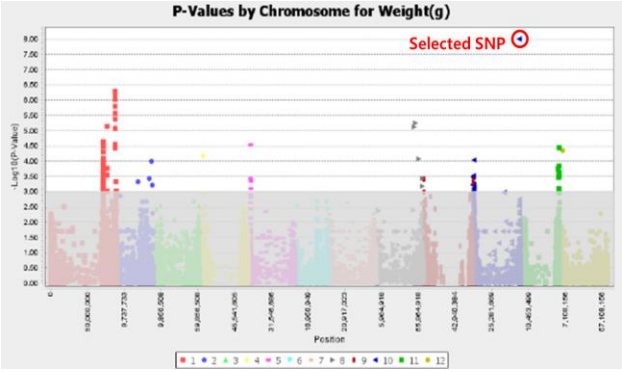
Fig. 1. Experimental process of phenotype prediction.

The hyperparameter settings for each classification model are as follows: The Random Forest uses 100 decision trees, with the number of features considered for splitting each tree set to the square root of the total number of features. Additionally, the minimum number of samples required to split a node is set to 2, and the minimum number of samples required at a leaf node is set to 1. The CatBoost is configured with a border count of 32. The XGBoost uses the gradient boosting tree booster, and the tree method is set to histogram-based algorithm. The Logistic Regression uses a regularization strength C of 1.0, a maximum number of iterations set to 1000, L2 regularization, and the limited memory BFGS solver for optimization. The Ridge uses a regularization strength alpha of 1.0. The Extra Trees uses 100 decision trees, with the number of features considered for splitting each tree set to the square root of the total number of features. The splitting criterion used is Gini impurity. The Linear Discriminant Analysis uses the SVD solver and sets the tolerance to 0.0001 for numerical stability. The AdaBoost uses 50 weak learners and a learning rate of 1.0. The boosting algorithm used is SAMME.R. The Gradient Boosting uses 100 trees, a learning rate of 0.1, and a maximum tree depth of 3. The loss function used is log loss. The K-Nearest Neighbors considers 5 neighbors and applies uniform weights. The LightGBM uses 100 trees, with a learning rate of 0.1. The maximum number of leaves per tree is set to 31, and the boosting type used is gradient boosting decision tree. This experimental setup ensured a comprehensive evaluation of each model's performance in predicting phenotypic traits from genotypic data. By integrating GWAS and machine learning, we aimed to explore the potential of these approaches for improving crop breeding strategies.

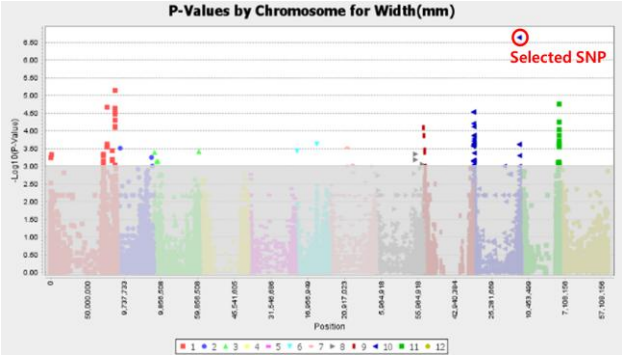
B. GWAS

Using GWAS, we identified 656 significant marker-trait association (MTA) SNPs linked to five fruit traits in the training population (see Fig. 2). The data analysis found that the associations were significant at p-value <0.0001. The SNPs are distributed on Chrs 1 to 12. Analysis using the Manhattan plot shows Chrs 1 to 12 in order, shown from red on the left and ranging to green on the right, with detected SNPs that exceed 3.0 on the *y*-axis. The SNPs discovered in the training population data are associated with the following fruit traits. Fig. 2 (a) represents whole SNPs associated with fruit weight found on 10 Chrs (1, 2, 3, 4, 5, 8, 9, 10, 11, and 12).

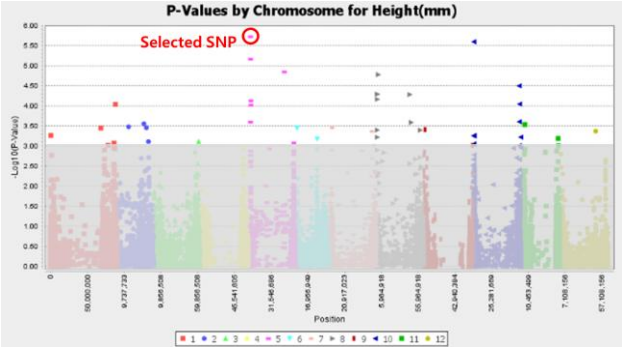
The SNP most closely associated with fruit weight was AX.95792429 locus on Chr 10, with a p-value of 1.66×10^{-10}. Fig. 2 (b) represents whole SNPs associated with fruit width found on 8 Chrs (1, 2, 3, 8, 9, 10, 11, and 12). The SNP most closely associated with fruit width was AX.95792429 locus on Chr 10, with a p-value of 1.31×10^{-8}.



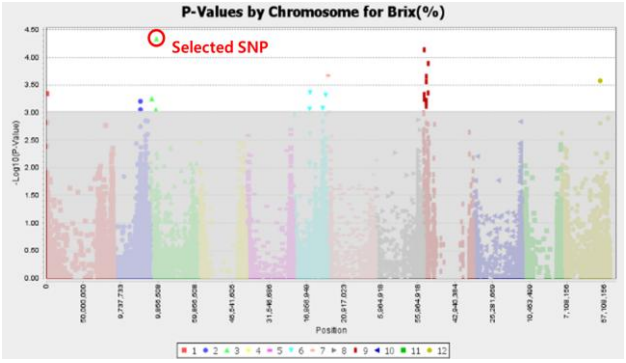
(a)



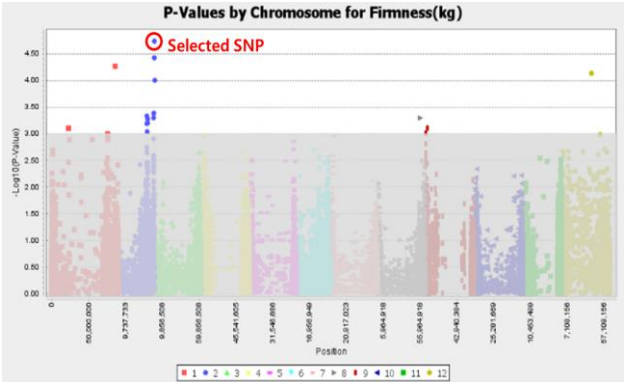
(b)



(c)



(d)



(e)

Fig. 2. The Manhattan plot of the training population of GWAS results according to traits: (a) weight, (b) width, (c) height, (d) brix, and (e) firmness.

Fig. 2 (c) represents whole SNPs associated with fruit height found on 8 Chrs (1, 2, 5, 8, 9, 10, 11, and 12). The SNP most closely associated with fruit height was AX.95815708 locus on Chrs 5, with a *p*-value of 1.99×10^{-7}. Fig. 2 (d) represents whole SNPs associated with brix found on 7 Chrs (1, 3, 5, 6, 9, 10, and 12). The SNP most closely associated with brix was AX.95809690 locus on Chr 9, with a *p*-value of 9.75×10^{-8}. Fig. 2 (e) represents whole SNPs associated with fruit firmness found on 9 Chrs (1, 2, 3, 4, 5, 8, 10, 11, and 12). The SNP most closely associated with fruit firmness was AX.95796914 locus on Chr 2, with a *p*-value of 7.76×10^{-6}. SNP AX.95792429 was found to be most closely associated with both fruit weight and fruit width traits. This suggests that the traits of fruit weight and fruit width may share some genetic factors. No SNPs associated with these traits were found on Chr 7. This suggests that there is no significant association between Chr 7 and the traits of tomato fruit.

C. Phenotype Prediction

We validated 19 classification models, using all 51,214 SNP markers, to predict each trait in the tomato dataset. We used 3-fold cross-validation, where the tomato dataset was randomly divided into three groups. We used F1-Score as the main performance evaluation metric while accuracy and AUC were used as auxiliary performance evaluation metrics. As shown in Table II, LDA had the highest score of 0.7067 for the fruit weight trait. CAT and ET had scores of 0.6290 and 0.5850 for

fruit width and height, respectively. RF had scores of 0.5497 and 0.5505 for fruit firmness and brix. We also conducted experiments using the test population to validate the model trained on the training population. As mentioned in Section III.A, the test population is distinct from the training population, collected in the same location in different years. This approach validates the trained model using the actual breeding population. As shown in Table III, CAT had a score of 0.6493 for fruit weight, GB had a score of 0.6881 for fruit width, RIDGE had a score of 0.5775 for fruit height, RF had a score of 0.5889 for fruit firmness, and LDA had a score of 0.5766 for the brix trait. These results demonstrate that different machine learning models excel in predicting different phenotypic traits, highlighting the importance of selecting appropriate models based on the target trait. The findings underscore the potential of combining genomic data and machine learning to enhance phenotypic prediction accuracy in crop breeding programs.

D. Phenotype Prediction with Semi-Supervised Learning

Semi-supervised learning is a methodology that aims to improve performance by applying supervised learning to small amounts of labeled data and unsupervised learning to large amounts of unlabeled data. This approach posits that by modeling the essential characteristics of the data, we can enhance generalization performance with minimal guidance from a small amount of labeled data. In general, training machine learning models with only 100 to 200 data points is insufficient. Pseudo-labeling is a semi-supervised learning technique where a model trained on labeled data is used to generate labels for unlabeled data. We used these pseudo-labels to further train the model, enhancing its performance with limited labeled data. In this section, we arbitrarily generated genotype data and labeled the corresponding phenotypic data using models trained in Section IV.C through pseudo-labeling.

Given that tomato genotypic data consists of 12 chromosomes, we randomly shuffled the 12 chromosomes from the 192 available genotype datasets to generate 1,000 random genotype datasets. Additionally, we used PG-cGAN to create another 1,000 genotype datasets, providing a complementary set for further analysis. PG-cGAN is a method that synthesizes and augments genomic data for any specific population or group by transforming existing real genomic data without setting any prior or model parameters [18]. The generated data was labeled for phenotype using pseudo-labeling with the CatBoost model trained on the training population described in Section IV.C. The cross-validation method, which separates the training and test population sets, was used to divide the data into a 70:30 ratio (70% training and 30% testing). Namely, when semi-supervised learning was applied, 70% of the original data was combined with 1,000 generated data points (a total of 1,134) to train the model, which was then tested on the remaining 30% of the original data.

According to Table IV, applying semi-supervised learning across the five traits leads to increased accuracy, demonstrating that this approach improves phenotype prediction performance.

TABLE II. TOP-5 CLASSIFICATION RESULTS FOR ALL SNPS PER TRAIT: 3-FOLD CROSS VALIDATION USING TRAINING POPULATION

Trait	Model	Performance		
		F1-Score	Accuracy	AUC
Weight	LDA	0.7067	0.7059	0.8393
	CAT	0.6963	0.6928	0.8551
	RF	0.6918	0.6928	0.8426
	XG	0.6903	0.6928	0.8322
	RIDGE	0.6657	0.6667	0.0000
Width	CAT	0.6290	0.6340	0.8064
	LR	0.6136	0.6209	0.7878
	RF	0.6116	0.6209	0.7947
	RIDGE	0.6060	0.6144	0.0000
	XG	0.6009	0.6144	0.8049
Height	ET	0.5850	0.5882	0.7356
	LDA	0.5757	0.5752	0.7282
	LR	0.5670	0.5752	0.7434
	RF	0.5574	0.5621	0.7242
	ADA	0.5571	0.5556	0.7009
Firmsness	RF	0.5497	0.5621	0.7384
	CAT	0.5303	0.5425	0.7409
	ET	0.5266	0.5359	0.7156
	XG	0.5245	0.5294	0.6869
	GB	0.5225	0.5229	0.6998
Brix	RF	0.5505	0.5686	0.7455
	KNN	0.5471	0.5686	0.7476
	ET	0.5408	0.5556	0.7433
	LDA	0.5318	0.5490	0.7122
	LGBM	0.5309	0.5556	0.7347

TABLE III. TOP-5 CLASSIFICATION RESULTS FOR ALL SNPS PER TRAIT: 3-FOLD CROSS VALIDATION USING TRAINING POPULATION AND TEST POPULATION

Trait	Model	Performance		
		F1-Score	Accuracy	AUC
Weight	CAT	0.6493	0.6771	0.8338
	LDA	0.6253	0.6458	0.8116
	LGBM	0.6152	0.6354	0.8020
	LR	0.6142	0.6354	0.8096
	XG	0.6072	0.6354	0.8123
Width	GB	0.6881	0.6927	0.8377
	CAT	0.6317	0.6458	0.8313
	LGBM	0.6270	0.6354	0.8257
	XG	0.6219	0.6302	0.8099
	LDA	0.6127	0.6198	0.7866
Height	RIDGE	0.5775	0.5938	0.0000
	LR	0.5654	0.5781	0.7623
	ET	0.5518	0.5729	0.7302
	LDA	0.5451	0.5677	0.7596
	KNN	0.5230	0.5312	0.6977
Firmness	RF	0.5889	0.6094	0.7533
	CAT	0.5624	0.5833	0.7392
	XG	0.5489	0.5729	0.7272
	LGBM	0.5341	0.5573	0.7300
	ET	0.5331	0.5573	0.7286
Brix	LDA	0.5766	0.5885	0.7352
	KNN	0.5389	0.5625	0.7187
	ET	0.5272	0.5573	0.7630
	RIDGE	0.5233	0.5521	0.0000
	CAT	0.5156	0.5573	0.7598

TABLE IV. PERFORMANCE COMPARISON OF PHENOTYPE PREDICTION ACROSS DIFFERENT METHODS FOR TOMATO TRAITS

Metric	Trait	Performance		
		Original	Add generated data (PG-cGAN)	Proposed method
F1-Score	Weight	0.6560	0.7073	**0.8282**
	Width	0.7116	0.7605	**0.7777**
	Height	0.5396	0.5735	**0.7056**
	Firmness	0.5636	0.5885	**0.5997**
	Brix	0.6089	0.6744	**0.6947**

E. Phenotype Prediction with Fixed Effect

In Section IV.B, we selected a total of 656 SNPs significantly associated with five fruit traits in the training population data. To validate the significant SNP markers as fixed effects related to these five traits, we conducted an additional experiment to predict the traits for each genotype data using only these 656 SNP markers. The experimental setup and performance metrics are equal to those mentioned in Section IV.C. The results indicate that the performance of the phenotype prediction was improved when using the significant 656 SNP markers as compared to when using all SNP markers. As shown in Table V, the CAT model achieved the highest F1-Score for predicting traits such as weight (0.7323) and width (0.7121) during 3-fold cross-validation using the training population, demonstrating superior performance compared to other models. For example, comparing with Table III, the F1-Score for phenotype prediction of fruit width increased significantly from 0.6493 to 0.7323, and the performance of phenotype prediction for most other traits also improved. As shown in Table VI, we also validated the model trained by the training population using the test population, considering the selected 656 SNP markers.

TABLE V. TOP-5 CLASSIFICATION RESULTS FOR SIGNIFICANT SNPS PER TRAIT: 3-FOLD CROSS VALIDATION USING TRAINING POPULATION

Trait	Model	Performance		
		F1-Score	Accuracy	AUC
Weight	CAT	0.7323	0.7320	0.8733
	XG	0.7208	0.7190	0.8614
	RF	0.7056	0.7059	0.8562
	ADA	0.6921	0.6863	0.8164
	ET	0.6892	0.6928	0.8624
Width	CAT	0.7121	0.7124	0.8408
	RF	0.6972	0.6993	0.8187
	LGBM	0.6749	0.6732	0.8335
	ET	0.6590	0.6601	0.7985
	GB	0.6560	0.6601	0.8228
Height	CAT	0.5961	0.5948	0.7534
	GB	0.5732	0.5752	0.7532
	LGBM	0.5657	0.5621	0.7509
	ET	0.5473	0.5490	0.7059
	RF	0.5424	0.5425	0.7076
Firmness	LR	0.5188	0.5229	0.6703
	CAT	0.5027	0.5098	0.6969
	GB	0.4941	0.4967	0.7131
	XG	0.4835	0.4902	0.6770
	LGBM	0.4806	0.4902	0.6384
Brix	SVM-R	0.5706	0.5817	0.7693
	RF	0.5408	0.5621	0.7408
	RIDGE	0.5324	0.5229	0.0000
	KNN	0.5287	0.5359	0.7090
	LDA	0.5253	0.5359	0.6989

TABLE VI. TOP-5 CLASSIFICATION RESULTS FOR SIGNIFICANT SNPS PER TRAIT: 3-FOLD CROSS VALIDATION USING TEST POPULATION

Trait	Model	Performance		
		F1-Score	Accuracy	AUC
Weight	CAT	0.6606	0.6823	0.8584
	LGBM	0.6560	0.6771	0.8527
	XG	0.6483	0.6667	0.8411
	SVM-L	0.6397	0.6510	0.0000
	LR	0.6293	0.6354	0.8230
Width	LGBM	0.6814	0.6927	0.8334
	CAT	0.6715	0.6771	0.8476
	GB	0.6505	0.6615	0.8224
	ADA	0.6470	0.6562	0.7954
	RF	0.6463	0.6562	0.8331
Height	ET	0.5475	0.5521	0.7134
	XG	0.5303	0.5312	0.7503
	ADA	0.5231	0.5365	0.7107
	RF	0.5201	0.5260	0.7303
	DT	0.5475	0.5052	0.6297
Firmness	GB	0.5532	0.5677	0.7249
	LR	0.5297	0.5365	0.6769
	RIDGE	0.5162	0.5208	0.0000
	SVM-R	0.5151	0.5260	0.6618
	LGBM	0.5150	0.5260	0.6815
Brix	SVM-R	0.6049	0.6146	0.7879
	DT	0.5681	0.5833	0.6867
	ET	0.5608	0.5990	0.7761
	LR	0.5338	0.5521	0.6908
	RF	0.5308	0.5729	0.7695

V. CONCLUSION

With advancements in machine learning and deep learning technologies, researchers have increasingly applied these methods across various industries. In plant breeding, digital breeding aims to leverage machine learning technology to analyze the relationship between plant genetics and traits, thereby enhancing the accuracy of phenotype prediction. This study examined several parameters influencing prediction accuracy, including model type, trait characteristics, and the number of SNP markers, using genotype and phenotype data from tomato crops. Additionally, we demonstrated the applicability of our phenotype prediction model within a practical agricultural breeding environment. Using GWAS, we identified 656 SNP markers linked to five fruit traits, predicted phenotypes based on these SNP markers, and validated the machine learning model trained on the training population with an independent breeding population. Experimental results indicated that trait-specific phenotype prediction in tomatoes is feasible using various machine learning algorithms. The findings further revealed that focusing solely on genetic data related to specific traits improved prediction accuracy compared to using all available genotype data. Moreover, machine learning models displayed similar predictive accuracy between training and testing data across different populations cultivated in the same location over multiple years.

This study also explored the effectiveness of various machine learning algorithms for predicting different

tomato traits and focused on improving classification accuracy by utilizing SNPs specifically associated with those traits.

Future studies should investigate methods to further enhance classification performance, potentially by incorporating virtual data generation and advanced deep learning techniques, such as convolutional neural networks and transformers.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Najeong Chae designed the study, supervised the project, and wrote the manuscript; Sung-Woo Byun coordinated the project and prepared the manuscript; Jiho Choi developed the machine learning model and performed data analysis; Taehoon Lim collected and analyzed data, and drafted the manuscript; Ji Hoon Lim selected SNP markers and interpreted the results; Hye In Lee validated the model, conducted independent testing, and contributed to the discussion; and Hwa Seon Shin developed the software and provided technical support; all authors had approved the final version.

FUNDING

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00227464, Development of new varieties breeding technology with AI for strengthening food sovereignty).

REFERENCES

[1] Tomato production by country 2023. (2023). Technical report, National Human Genome Research Institute. [Online]. Available: https://www.worldostats.com/post/tomato-production-by-country-2023

[2] Genotype. (2024). Technical report, National Human Genome Research Institute [Online]. Available: https://www.genome.gov/genetics-glossary/genotype.

[3] Phenotype. (2024). Technical report, National Human Genome Research Institute. [Online]. Available: https://www.genome.gov/genetics-glossary/Phenotype

[4] P. Bayer, J. Petereit, M. Danilevicz, R. Anderson, J. Batley, and D. Edwards, "The application of pangenomics and machine learning in genomic selection in plants," *The Plant Genome*, vol. 14, no. 3, e20112, 2021.

[5] M. F. Danilevicz, M. Gill, R. Anderson, J. Batley, M. Bennamoun, P. E. Bayer, and D. Edwards, "Plant geno type to phenotype prediction using machine learning," *Frontiers in Genetics*, vol. 13, 822173, 2022.

[6] P. M. Eggink, C. Maliepaard, Y. Tikunov, J. P. W. Haanstra, L. M. M. Pohu-Flament, S. C. de Wit-Maljaars, F. Willeboordse-Vos, S. Bos, C. B. de Waard, P. J. de G. van Leeuwen, G. Freymark, A. G. Bovy, and R. G. F. Visser, "Prediction of sweet pepper (capsicum annuum) flavor over different harvests," *Euphytica*, vol. 187, pp. 117–131, 2012.

[7] C. Fang, Y. Ma, S. Wu, Z. Liu, Z. Wang, R. Yang, G. Hu, Z. Z. H. Yu, M. Zhang, Y. Pan, G. Zhou, H. Ren, W. Du, H. Yan, Y. Wang, D. Han, Y. Shen, S. Liu, T. Liu, J. Zhang, H. Qin, J. Yuan, X. Yuan, F. Kong, B. Liu, J. Li, Z. Zhang, G. Wang, B. Zhu, and Z. Tian, "Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean," *Genome Biology*, vol. 18, 161, 2017.

[8] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, "A guide to machine learning for biologists," *Nature Reviews Molecular Cell Biology*, vol. 23, no. 1, pp. 40–55, 2022.

[9] D. Jeon, Y. Kang, S. Lee, S. Choi, Y. Sung, T. H. Lee, and C. Kim, "Digitalizing breeding in plants: A new trend of next-generation breeding based on genomic prediction," *Frontiers in Plant Science*, vol. 14, 1092584, 2023.

[10] M. H. U. Khan, S. Wang, J. Wang, S. Ahmar, S. Saeed, S. U. Khan, X. Xu, H. Chen, J. A. Bhat, and X. Feng, "Applications of artificial intelligence in climate-resilient smart-crop breeding," *International Journal of Molecular Sciences*, vol. 23, no. 19, 11156, 2022.

[11] M. Kim, T. T. P. Nguyen, J.-H. Ahn, G.-J. Kim, and S.-C. Sim, "Genome-wide association study identifies qtl for eight fruit traits in cultivated tomato (solanum lycopersicum l.)," *Horticulture Research*, vol. 8, 203, 2021.

[12] N. T. Phan, L. T. Trinh, M. Y. Rho, T. S. Park, O. R. Kim, J. Zhao, H.-M. Kim, and S.-C. Sim, "Identification of loci associated with fruit traits using genome-wide single nucleotide polymorphisms in a core collection of tomato (solanum lycopersicum l.)," *Scientia Horticulturae*, vol. 243, pp. 567–574, 2019.

[13] G. W. Kim, J. P. Hong, H. Y. Lee, J. K. Kwon, D. A. Kim, and B. C. Kang, "Genomic selection with fixed effect markers improves the prediction accuracy for cap saicinoid contents in capsicum annuum," *Horticulture Research*, vol. 9, 2022.

[14] W. S. Ravelombola, J. Qin, A. Shi, L. Nice, Y. Bao, A. Lorenz, J. H. Orf, N. D. Young, and S. Chen, "Genome-wide association study and genomic selection for soybean chlorophyll content associated with soybean cyst nematode tolerance," *BMC Genomics*, vol. 20, 904, 2019.

[15] C. Priyanatha, D. Torkamaneh, and I. Rajcan, "Genome wide association study of soybean germplasm derived from Canadian × Chinese crosses to mine for novel alleles to improve seed yield and seed quality traits," *Frontiers in Plant Science*, vol. 13, 866300, 2022.

[16] M. Yoosefzadeh-Najafabadi, I. Rajcan, and M. Eskandari, "Optimizing genomic selection in soybean: An important improvement in agricultural genomics," *Heliyon*, vol. 8, no. 11, e11873, 2022

[17] H. Shen, K. Jiang, W. Sun, Y. Xu, and X. Ma, "Irrigation decision method for winter wheat growth period in a supplementary irrigation area based on a support vector machine algorithm," *Computers and Electronics in Agriculture*, vol. 182, 106032, 2021.

[18] J. Chen, M. E. Mowlaei, and X. Shi, "Population-scale genomics data augmentation based on conditional generative adversarial networks," in *Proc. the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB'20)*, 2020