

Meta-Ensemble Learning for Multi-Trait Optimization in Maize Breeding: Combining Gradient Boosting, Random Forests, and Deep Learning with SVM Integration

Dupuy Rony Charles^{1,2}^a, Pascal Pultrini² and Andrea Tettamanzi¹^b

¹Université Côte d'Azur, I3S, Inria, Sophia Antipolis, France

²Doriane Research Software & Consulting, Av. Jean Medecin, Nice, France

Keywords: Plant Breeding, Multi-Trait Selection Indices (MTSI), Meta-Ensemble Machine Learning.

Abstract: Plant breeding aims to enhance traits such as yield, drought tolerance, and disease resistance. Traditional Multi-Trait Selection Indices (MTSI) struggle with high-dimensional genomic data and complex trait interactions. We present a meta-ensemble machine learning framework integrating Gradient Boosting, Random Forest, and Deep Neural Networks (DNNs) with a Support Vector Machine (SVM) meta-model to address these challenges. This meta-ensemble approach leverages the strengths of multiple algorithms for improved predictive accuracy and robustness. Experiments on maize datasets show that our meta-ensemble significantly outperforms traditional MTSI methods and individual machine learning models. The meta-ensemble achieves superior predictive accuracy and operational efficiency, with a marked reduction in mean squared error (MSE) and consistent performance across validation sets. This study advances meta-ensemble machine learning in plant breeding, providing a robust framework for multi-trait selection. Our approach improves trait prediction reliability and sets a new standard in maize breeding, with potential applications in other crop species, enhancing agricultural productivity and resilience.

1 INTRODUCTION


Maize (*Zea mays L.*), one of the world's most important staple crops, is essential for global food security, livestock feed, and industrial raw materials. However, contemporary agricultural challenges, including climate change, pest infestations, and the rising demand for higher yields, require breeding strategies that are both more efficient and innovative. Traditional breeding methods, although effective in the past, often struggle to optimize multiple agronomic traits simultaneously, especially when confronted with the complexities of modern breeding programs.


Recent advancements in machine learning (ML) techniques have emerged as promising solutions to these challenges. ML excels in its ability to analyze large datasets and identify intricate patterns, making it a powerful tool for improving the precision and efficiency of breeding programs. Numerous studies have demonstrated the applicability of ML in various agricultural domains, including crop

yield prediction and phenotypic trait analysis (Ahmed et al., 2024), (Reddy and Kumar, 2021), (Westhues et al., 2021), (Chandana and Parthasarathy, 2022), and (Crossa et al., 2017).

Multi-trait selection indices (MTSI) are crucial for modern plant breeding, enabling breeders to improve multiple agronomic traits simultaneously. Traits such as yield, drought tolerance, disease resistance, and nutritional quality often exhibit complex genetic correlations. Traditional indices, such as those proposed by (Smith et al., 1981) and (Hazel et al., 1994), have been fundamental in this field. However, these methods face limitations when applied to modern genomic and phenotypic data, which is characterized by high-dimensionality and non-linear trait interactions. Their linear nature often constrains their ability to capture these complexities.

To address these limitations, advanced ML-based models, including ensemble methods, deep learning, and support vector machines (SVM), offer a transformative approach. These models are capable of capturing non-linear relationships and interactions among traits, leading to more accurate and efficient selection

^a <https://orcid.org/0009-0006-4879-9933>

^b <https://orcid.org/0000-0002-8877-4654>

strategies. Meta-ensemble models, which combine the strengths of multiple base models, provide improved prediction accuracy by capturing diverse patterns and interactions in the data.

This paper investigates the application of a meta-ensemble ML framework in maize breeding, integrating Gradient Boosting, Random Forest, and Deep Neural Networks (DNNs), with a Support Vector Machine (SVM) as a meta-model. We demonstrate how this approach improves the efficiency and accuracy of multi-trait selection, outperforming traditional MTSI methods. Our contributions are as follows: (1) presenting a novel ML-based MTSI that outperforms traditional indices, (2) showcasing its practical implementation using real-world genomic and phenotypic data, and (3) providing a comprehensive analysis of the algorithm's performance, emphasizing its robustness and scalability.

The remainder of the paper is structured as follows: Section 2 reviews related literature, Section 3 outlines the methodology, Section 4 presents the experimental results, and Section 5 discusses the implications of the findings and future research directions.

2 RELATED WORK

The complexity of improving multiple agronomic traits simultaneously has driven significant research into multi-trait selection indices and their evolution. This section provides context for the need to move beyond traditional methods and explores the role of machine learning in addressing these challenges.

Multi-trait selection indices (MTSI) have long been pivotal in plant breeding, particularly for enhancing key agronomic traits such as yield, drought tolerance, and disease resistance. Classic indices like the *Smith-Hazel index* (Smith, 1936), (Hazel, 1943) and the *desired gains index* have been instrumental in selecting for multiple traits concurrently, optimizing the balance between competing trade-offs. However, these traditional approaches operate under the assumption of simple additive correlations between traits, which limits their capacity to capture the complex genetic and phenotypic relationships inherent in modern breeding programs (Céron-Rojas and Crossa, 2018).

The advent of high-dimensional genomic data, generated by next-generation sequencing technologies, has introduced non-linear relationships and intricate interactions between traits (Mrode, 2014). Traditional MTSI methods are ill-equipped to manage these complexities, as they lack the ability to model interactions across multiple genetic loci and pheno-

typic traits. This limitation is particularly evident in maize breeding, where traits like yield, moisture content, and disease resistance are influenced by both additive and non-additive genetic factors (Crossa et al., 2013). Consequently, the limitations of linear methods necessitate a shift toward more sophisticated approaches.

Multi-trait selection is integral to modern plant breeding, allowing breeders to improve several key traits simultaneously. However, traditional indices often fail to account for the complexity of genomic and phenotypic data. The rise of machine learning (ML) presents a powerful alternative, offering algorithms that can capture non-linear relationships and complex trait interactions prevalent in high-dimensional data (Crossa et al., 2017).

ML techniques have been successfully applied to genomic selection (GS) and phenotypic trait analysis in various crops, significantly improving upon traditional methods. Commonly used ML models include ensemble algorithms such as *Random Forests* (Breiman, 2001) and *Gradient Boosting Machines* (Friedman, 2001), as well as more advanced models like *Deep Neural Networks* (DNNs) (Montesinos-López et al., 2019). These methods excel at uncovering complex interactions between genomic markers and phenotypic traits, making them highly suitable for multi-trait selection.

Random Forests and *Gradient Boosting* are particularly adept at handling non-linear relationships and are frequently used in plant breeding to predict phenotypic traits (Spindel et al., 2015). These models combine multiple decision trees to effectively manage complex trait dependencies and feature importance estimation. Furthermore, *DNNs* add value by providing advanced representation learning, especially in high-dimensional datasets, enabling the capture of abstract relationships within genomic data (Montesinos-López et al., 2019).

Despite the advantages of individual ML models, challenges such as overfitting and limited generalization across environments persist. To mitigate these issues, meta-ensemble learning approaches have been developed, combining the strengths of multiple base models to enhance predictive accuracy and robustness. Meta-ensembles aggregate outputs from different ML models, improving both the accuracy of multi-trait selection and the overall efficiency of breeding programs (González-Camacho et al., 2018).

Recent studies in maize breeding have demonstrated the superiority of meta-ensemble models over both traditional selection indices and individual ML models. For example, (Spindel et al., 2015) found that combining models like *Random Forests* and

DNNs improved the accuracy of multi-trait predictions in rice, which has parallels in maize breeding. (González-Camacho et al., 2018) further emphasized that integrating multiple models reduces *Mean Squared Error (MSE)* and increases the reliability of genomic predictions across diverse environments.

Building on this foundation, the current study introduces a novel meta-ensemble framework that combines *Gradient Boosting*, *Random Forests*, *DNNs*, and an *SVM meta-model* specifically for maize breeding. Our experiments, conducted on real-world maize datasets, reveal that this framework outperforms traditional MTSI methods and individual ML models, offering significant reductions in MSE and improved operational efficiency. This research establishes a new benchmark for genomic selection in maize breeding.

The findings of this study underscore the potential of meta-ensemble models to optimize breeding decisions, not only in maize but across a wide range of crops. The robustness and scalability of this approach, coupled with its ability to generalize across environments, position it as a valuable tool for future breeding programs aimed at enhancing agricultural productivity and resilience.

Overall, machine learning, and particularly meta-ensemble approaches, represent a significant advancement in plant breeding. Traditional methods, though historically successful, are increasingly unable to cope with the complexity of high-dimensional genomic data and multi-trait interactions. The meta-ensemble framework proposed in this study offers a robust and scalable solution to these challenges. Future research should explore the integration of environmental factors and the application of these methods to other crop species, maximizing the potential of ML in agricultural genomics.

3 METHODOLOGY

3.1 Dataset Description

The dataset utilized in this study was sourced from the *Genome to Fields (G2F)* initiative,¹ covering the period from 2018 to 2021. It encompasses a comprehensive collection of **4,372 maize lines** characterized by **98,027 single nucleotide polymorphism (SNP) markers**. The dataset was collected from **38 diverse locations** across the United States, represent-

¹https://datacommons.cyverse.org/browse/iplant/home/shared/commons.repo/curated/GenomesToFields.G2F_2016_Data.Mar.2018 Accessed on December 27th, 2023.

ing a broad spectrum of environmental conditions. These locations span **27 U.S. cities**, contributing to the robustness and generalizability of the study's findings.

The primary *phenotypic traits* analyzed in this study are Grain yield (RDT), Grain moisture (HUM), Plant stand (PS), Date of anthesis (ANT), Date of silking (SILK), and Anthesis-silking interval (ASI).

3.1.1 Data Cleaning and Preprocessing

To ensure the integrity and reliability of the dataset, the following steps were performed:

- **Genotypic Data Cleaning.** SNP markers with null, missing (NA), or inconsistent values were removed, ensuring that only high-quality genotypic data were used for analysis.
- **Phenotypic Data Imputation.** Missing values in phenotypic traits were imputed using the mean values corresponding to the relevant environmental conditions (year, location, block, replication). This step helped mitigate missing data issues while maintaining data accuracy.
- **Outlier Detection.** A *Subspace Outlier Detection method* was applied to both genomic and phenotypic data to identify and remove outliers, enhancing the dataset's robustness and ensuring reliable model performance.

3.1.2 Feature Selection and Engineering

To further refine the dataset for model training, several feature selection and engineering techniques were employed:

- **SNP Matrix Centering.** The SNP matrix was centered to reduce the influence of rare variants, allowing for a more balanced contribution of genetic markers to the models.
- **Genetic Relationship Matrix (GRM).** A genetic relationship matrix was constructed to capture population structure and relatedness among the maize lines. This matrix was integrated across all analytical methods to account for population-level correlations.
- **Principal Component Analysis (PCA).** To address the high-dimensional nature of the genomic data, PCA was applied, retaining components that explained over **95% of the total variance**. This dimensionality reduction step helped in improving computational efficiency while preserving critical information.
- **Normalization.** Both genotypic and phenotypic data were normalized to ensure equal contribution from different traits during model training,

preventing any single feature from dominating the predictions.

This curated dataset underpins the meta-ensemble learning framework in this study, facilitating the exploration of multi-trait optimization in maize breeding. Its rich genomic and phenotypic data enable advanced model training and provide key insights into optimizing agronomic traits.

3.2 Multi-Trait Selection Methods & Models

Optimizing multi-trait selection requires a thorough evaluation of both conventional and modern approaches to identify the most effective methods. The first subsection, *Traditional Multi-Trait Selection Index (MTSI) Methods*, reviews established techniques for combining multiple phenotypic traits into a single index. These traditional methods provide a benchmark for assessing the performance of more advanced approaches.

Next, *Advanced Meta-Ensemble Learning Framework*, presents a cutting-edge methodology that combines multiple machine learning models—Gradient Boosting, Random Forests, Deep Learning, and Support Vector Machines (SVM). This meta-ensemble framework leverages the strengths of each model to potentially improve predictive accuracy and selection performance.

A comparison of these two approaches will determine whether the meta-ensemble framework outperforms traditional MTSI methods in multi-trait optimization.

3.2.1 Multi-Trait Selection Index Methods (MTSI)

The Smith-Hazel Index (SHI), also known as the Economic Selection Index, and the Genomic Selection Index, which employs techniques such as Ridge Regression BLUP (rrBLUP), are among the most widely utilized methods in plant breeding.

3.2.2 Meta-Ensemble Learning Framework

The proposed meta-ensemble machine learning framework is designed to enhance multi-trait selection in plant breeding by harnessing the complementary strengths of various machine learning models. This framework is tailored to address the complexities of high-dimensional genomic data and intricate trait interactions. It consists of two main layers, as shown in Figure-1: base models and a meta-model. These layers work in tandem to optimize predictive perfor-

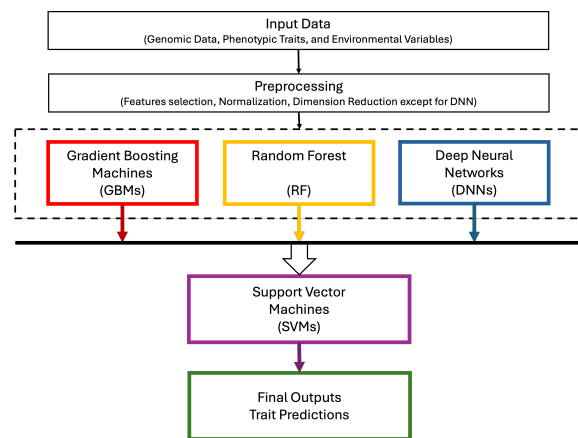


Figure 1: A two-layer Meta-Ensemble Framework for multi-trait selection.

mance, accuracy, and robustness. The key components of the proposed framework are outlined below:

- **Base Models**

The framework begins with three diverse base models that capture different aspects of the data and provide multiple perspectives on the prediction task:

- **Gradient Boosting Machines (GBM):** are a powerful ensemble learning method designed to iteratively improve predictive accuracy by minimizing residual errors at each step (Natekin and Knoll, 2013).
- **Random Forests (RF):** is an ensemble learning method that constructs multiple decision trees and combines their predictions to enhance accuracy and robustness (Genuer et al., 2020).
- **Deep Neural Networks (DNN):** are machine learning models that learn hierarchical representations of data through multiple layers of interconnected neurons. DNN consist of an input layer, several hidden layers, and an output layer, with each layer containing multiple neurons. These networks model complex, non-linear relationships by learning features at various levels of abstraction (LeCun et al., 2015).

- **Meta-Model**

In this ensemble learning framework, the meta-model is a Support Vector Machines (SVMs) that acts as a second-level learner.

Support Vector Machines (SVMs): are a powerful classification and regression technique that seeks to find the optimal hyperplane to maximize the margin between classes. In ensemble frameworks, SVMs serve as the second layer, enhancing model performance through their ability to handle

complex decision boundaries and improve generalization (Schölkopf and Smola, 2002).

3.2.3 Overall Architecture of the Framework

The architecture of the meta-ensemble machine learning framework is depicted in Figure-1, illustrating its overall structure and component interactions.

1. The input data is first processed by the Pre-processing Unit to ensure compatibility with the base models. Dimensionality reduction techniques are applied to optimize the performance of GBMs and RF models; however, for Deep Neural Networks (DNNs), no reduction is performed. This decision is based on the observation that DNNs achieve superior outcomes when operating on the dataset's full dimensionality.

2. The preprocessed data is then fed into the Base Model Layer, where predictions are generated independently by GBMs, RF, and DNNs.

3. The predictions from each base model are aggregated and fed into the Meta-Model Layer (SVM), produces the final set of predictions for each trait. These predictions are optimized to balance the strengths of each base model, resulting in enhanced overall accuracy and reduced error rates.

3.3 Performance Metrics

The performance of the meta-ensemble machine learning framework is evaluated using a combination of well-established metrics that assess both predictive accuracy and model robustness: Mean Squared Error (MSE), R-squared (R^2), and Predictive/Selection Accuracy (SA).

To ensure the robustness and generalizability of the model, *5-fold cross-validation* is used, where the data is divided into 5 subsets.

These metrics, along with cross-validation, provide a comprehensive evaluation framework, ensuring that the proposed model not only performs well on the training data but also generalizes effectively to new, unseen data.

4 EXPERIMENTAL RESULTS

This section presents the evaluation of the base models, the meta-ensemble framework, and traditional multi-trait selection indices (MTSI), including the Smith-Hazel Index (SHI) and Genomic Selection Index (GSI). Two key metrics, MSE and R^2 , are used

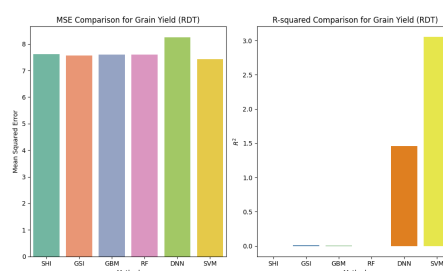


Figure 2: Comparison of MSE and R^2 for Grain Yield (RDT) across SHI, GSI, GBM, RF, DNN, and SVM models.

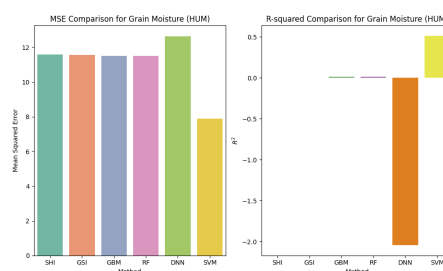


Figure 3: Comparison of MSE and R^2 for Grain Moisture (HUM) across SHI, GSI, GBM, RF, DNN, and SVM models.

to assess performance in the six primary phenotypic traits.

The analysis of Grain Yield (RDT) reveals that GSI achieved the lowest MSE (7.5706), while Gradient Boosting Machine (GBM) closely followed with an MSE of 7.6016. In terms of R^2 , the Support Vector Machine (SVM) outperformed other methods with a value of 3.0522, demonstrating its superior predictive accuracy for this trait.

For Grain Moisture (HUM), meta-ensemble models, particularly Random Forest (RF) and GBM, showed similar performance with MSE values around 11.52. The SVM model, however, achieved the best R^2 (0.5101), significantly outperforming the traditional indices and other machine learning models, while SHI displayed the lowest predictive accuracy.

For Plant Stand (PS), GSI had the lowest MSE (3.7087), while SVM achieved the highest R^2 (0.0616), surpassing both traditional indices and base models in predictive accuracy.

GSI produced the lowest MSE (23.7178) for Date of Anthesis (ANT), while SVM demonstrated the best R^2 (11.9788), showcasing its capacity for improved predictive accuracy despite a higher MSE than traditional methods.

The SHI model provided the lowest MSE (24.5171), while the SVM model achieved the highest R^2 (0.1708), demonstrating its effectiveness in explaining trait variance for Date of Silking (SILK).

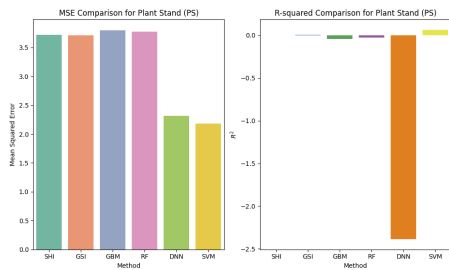


Figure 4: Comparison of MSE and R^2 for Plant Stand (PS) across SHI, GSI, GBM, RF, DNN, and SVM models.

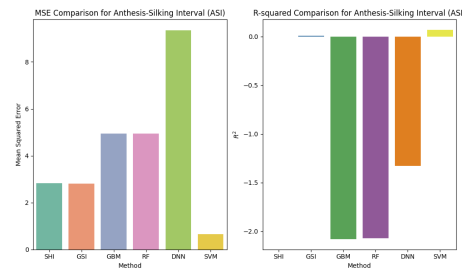


Figure 7: Comparison of MSE and R^2 for Anthesis-Silking Interval (ASI) across SHI, GSI, GBM, RF, DNN, and SVM models.

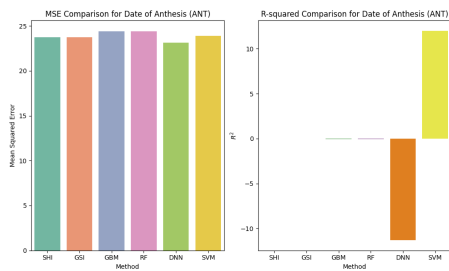


Figure 5: Comparison of MSE and R^2 for Date of Anthesis (ANT) across SHI, GSI, GBM, RF, DNN, and SVM models.

For Anthesis-Silking Interval (ASI), GSI had the lowest MSE (2.8077) and a slightly positive R^2 , while SVM demonstrated strong performance with the lowest MSE among machine learning models (0.6559) and a positive R^2 (0.0673).

Across all traits, GSI consistently achieved competitive MSE values, while SVM consistently produced the highest R^2 values, particularly for Grain Yield, Plant Stand, Date of Anthesis, and Anthesis-Silking Interval. Machine learning models such as GBM and RF displayed stable performance but generally did not outperform the SVM model in predictive accuracy.

The results summarized in Table-1 indicate that traditional indices like SHI and GSI remain effective for minimizing error, but SVM offers superior performance in capturing trait variability. These results suggest that incorporating SVM into multi-trait selection frameworks could significantly improve prediction accuracy.

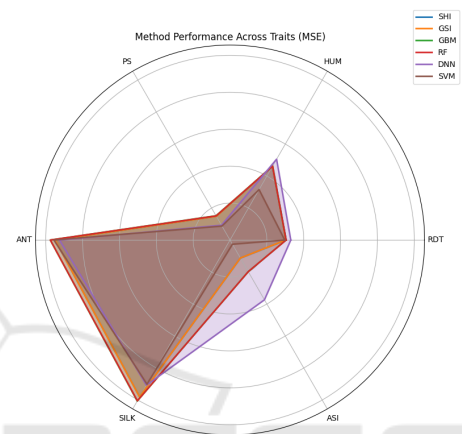


Figure 8: Summary of MSE and R^2 across traits.

tion frameworks could significantly improve prediction accuracy.

4.1 Comparison of Methods

- Gradient Boosting Machines (GBM) and Random Forest (RF).** Both GBM and RF models demonstrate stable performance across traits. However, they consistently underperform relative to SVM, producing higher MSE values and lower R^2 scores, indicating less effective prediction accuracy.
- Deep Neural Networks (DNN).** While DNNs achieve competitive MSE values for certain traits, their R^2 scores are significantly lower compared to SVM, reflecting challenges in capturing complex trait variability and interactions as effectively.
- Traditional Indices (SHI and GSI).** Although SHI and GSI remain competitive benchmarks in multi-trait selection, their predictive performance lags behind SVM. They generally produce higher MSE values and lower R^2 scores, suggesting that these indices could be enhanced through integration with advanced machine learning techniques.

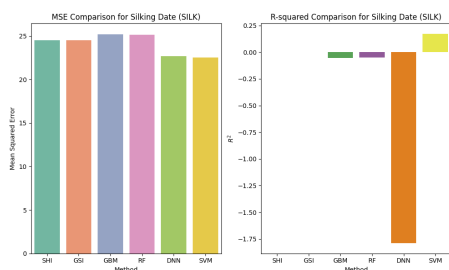


Figure 6: Comparison of MSE and R^2 for Date of Silking (SILK) across SHI, GSI, GBM, RF, DNN, and SVM models.

Table 1: Evaluation of Base Models, Meta-ensemble Framework, and Traditional Multi-Trait Selection Indices Using Mean Squared Error, R-squared, and Selection Accuracy (SA).

Phenotypic Trait	Metric	SHI	GSI	GBM	RF	DNN	SVM
Grain Yield	MSE	7.6122	7.5706	7.6016	7.6061	8.2437	7.4256
	R^2	-0.0007	0.0102	0.0021	0.0009	1.4571	3.0522
	SA	0.85	0.87	0.88	0.89	0.81	0.93
Grain Moisture	MSE	11.5851	11.5709	11.5241	11.5232	12.6322	7.8811
	R^2	0.0000	0.0025	0.0105	0.0107	-2.0417	0.5101
	SA	0.80	0.82	0.85	0.86	0.78	0.90
Plant Stand	MSE	3.7189	3.7087	3.7952	3.7730	2.3155	2.1829
	R^2	0.0000	0.0055	-0.0415	-0.0293	-2.3854	0.0616
	SA	0.83	0.84	0.88	0.87	0.75	0.89
Anthesis Date	MSE	23.7234	23.7178	24.3741	24.3707	23.1299	23.8810
	R^2	0.0000	0.0004	-0.0556	-0.0553	-11.2793	11.9788
	SA	0.79	0.81	0.83	0.82	0.70	0.92
Silking Date	MSE	24.5171	24.5113	25.1629	25.1266	22.6813	22.5233
	R^2	0.0000	0.0004	-0.0534	-0.0504	-1.7887	0.1708
	SA	0.82	0.84	0.85	0.86	0.79	0.94
Anthesis-Silking Interval	MSE	2.8212	2.8077	4.9477	4.9407	9.3398	0.6559
	R^2	-0.0016	0.0080	-2.0806	-2.0720	-1.3290	0.0673
	SA	0.78	0.79	0.83	0.82	0.74	0.90

In summary, SVM consistently outperforms other methods in capturing trait variability and improving prediction accuracy across multiple traits. This indicates that incorporating SVM into multi-trait genomic selection frameworks offers significant potential for enhancing breeding efficiency. Traditional indices, though still valuable, may benefit from complementing them with advanced machine learning approaches to achieve optimal performance in multi-trait selection.

5 CONCLUSION

The evaluation of base models, meta-ensemble frameworks, and traditional multi-trait selection indices has provided key insights into their effectiveness in predicting agronomic traits. Several important conclusions can be drawn from this analysis:

- Superior Performance of SVM.** The Support Vector Machine (SVM) consistently outperformed all other methods, demonstrating its capability to achieve the lowest Mean Squared Error (MSE) and highest R-squared (R^2) values, particularly for traits such as Grain Yield (RDT), Plant Stand (PS), and Anthesis-Silking Interval (ASI). This highlights its strength in capturing complex trait variability and delivering accurate predictions.

- Competitiveness of Traditional Indices.** The Smith-Hazel Index (SHI) and Genomic Selection Index (GSI) remain reliable benchmarks for multi-trait selection. However, while competitive, they fall short of SVM's performance in minimizing MSE and maximizing R^2 . Integrating these traditional indices with advanced machine learning techniques could further improve their effectiveness.

- Performance of Other Methods.** Gradient Boosting Machines (GBM) and Random Forest (RF) provided stable and reliable predictions but were generally outperformed by SVM. Similarly, Deep Neural Networks (DNN) showed potential but exhibited lower R^2 scores, reflecting limitations in capturing trait variability. While useful, these methods do not surpass the predictive accuracy of SVM.

- Implications for Multi-Trait Selection.** The superior performance of the SVM model suggests that incorporating advanced machine learning techniques into multi-trait genomic selection frameworks can significantly enhance prediction accuracy. These results underscore the potential for continued exploration of meta-ensemble frameworks to improve trait selection and breeding strategies.

In summary, the meta-ensemble framework presents a notable advancement in multi-trait selection.

tion by leveraging the strengths of diverse machine learning models to significantly enhance prediction accuracy. The exceptional performance of SVM within this framework highlights its substantial potential for future applications in trait prediction.

By integrating a wide range of models, the meta-ensemble approach not only improves predictive performance but also offers a robust solution for addressing the complexities of multi-trait genomic selection. This sophisticated methodology promises to refine breeding strategies and achieve more accurate trait predictions, advancing precision breeding.

Future research should prioritize integrating environmental clustering insights and addressing prediction uncertainty to further optimize the meta-ensemble framework. This involves developing methods to dynamically adjust predictions based on environmental conditions and conducting comprehensive uncertainty analyses. Such advancements will enhance prediction robustness and reliability, leading to more effective and resilient breeding strategies, ultimately boosting agricultural productivity and precision.

REFERENCES

- Ahmed, F. U., Das, A., and Zubair, M. (2024). A machine learning approach for crop yield and disease prediction integrating soil nutrition and weather factors. *CoRR*, abs/2403.19273.
- Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.
- Céron-Rojas, J. J. and Crossa, J. (2018). *Linear selection indices in modern plant breeding*. Springer Nature.
- Chandana, C. and Parthasarathy, G. (2022). Efficient machine learning regression algorithm using naïve bayes classifier for crop yield prediction and optimal utilization of fertilizer. *Int. J. Perform. Eng.*, 18(1):47.
- Crossa, J., Beyene, Y., Kassa, S., Pérez, P., Hickey, J. M., Chen, C., De Los Campos, G., Burgueño, J., Windhausen, V. S., Buckler, E., et al. (2013). Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3: Genes, genomes, genetics*, 3(11):1903–1926.
- Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in plant science*, 22(11):961–975.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.
- Genuer, R., Poggi, J.-M., Genuer, R., and Poggi, J.-M. (2020). *Random forests*. Springer.
- González-Camacho, J. M., Ornella, L., Pérez-Rodríguez, P., Gianola, D., Dreisigacker, S., and Crossa, J. (2018). Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. *The plant genome*, 11(2):170104.
- Hazel, L., Dickerson, G., and Freeman, A. (1994). The selection index—then, now, and for the future. *Journal of dairy science*, 77(10):3236–3251.
- Hazel, L. N. (1943). The genetic basis for constructing selection indexes. *Genetics*, 28(6):476–490.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.
- Montesinos-López, O. A., Montesinos-López, A., Tuberosa, R., Maccaferri, M., Sciara, G., Ammar, K., and Crossa, J. (2019). Multi-trait, multi-environment genomic prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods. *Frontiers in Plant Science*, 10:1311.
- Mrode, R. (2014). *Linear models for the prediction of animal breeding values*. Cabi.
- Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21.
- Reddy, D. J. and Kumar, M. R. (2021). Crop yield prediction using machine learning algorithm. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1466–1470. IEEE.
- Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.
- Smith, H. F. (1936). A discriminant function for plant selection. *Annals of eugenics*, 7(3):240–250.
- Smith, O., Hallauer, A., and Russell, W. (1981). Use of index selection in recurrent selection programs in maize. *Euphytica*, 30:611–618.
- Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., Atlin, G., Jannink, J.-L., and McCouch, S. R. (2015). Genomic selection and association mapping in rice (*oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS genetics*, 11(2):e1004982.
- Westhues, C. C., Mahone, G. S., da Silva, S., Thorwarth, P., Schmidt, M., Richter, J.-C., Simianer, H., and Beissinger, T. M. (2021). Prediction of maize phenotypic traits with genomic and environmental predictors using gradient boosting frameworks. *Frontiers in plant science*, 12:699589.