Smart Agricultural Technology 10 (2025) 100791

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Improving wheat yield prediction through variable selection using Support Vector Regression, Random Forest, and Extreme Gradient Boosting

Juan Carlos Moreno Sánchez [a], Héctor Gabriel Acosta Mesa [b], Adrián Trueba Espinosa [a,*], Sergio Ruiz Castilla [a], Farid García Lamont [a]

[a] Centro Universitario UAEM Texcoco, Universidad Autónoma del Estado de México, Mexico
[b] Instituto en Investigaciones en Inteligencia Artificial, Universidad Veracruzana, Mexico

ARTICLE INFO

Keywords:
Support vector regression
Random forest
Extreme gradient boosting, grain yield, vegetation indices, climate data

ABSTRACT

Plant breeding centers, in their relentless pursuit of more productive and resilient wheat varieties, have generated vast data repositories that are fundamental to ensuring global food security. This study uses these data to develop a wheat grain yield (GY) prediction model, using machine learning techniques such as Random Forest (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost). The results obtained prove the potential of RF and XGBoost-based models to accurately predict wheat yield. One of the major challenges of this research was to find the most relevant variables for predicting wheat yield. Using clustering, feature selection, and variable combination techniques, particularly agronomic variables such as harvest index (HI) and biomass (BM), provided complementary information to the Normalized Difference Vegetation Index (NDVI). This combination, analyzed through the XGBoost model, resulted in an exceptional performance, with an RMSE of 28.5082 (grams/square meter) and an R^2 of 0.9156, showing the constructive collaboration between these indicators. After a thorough analysis, it was discovered that daily clustering and filtering of climatic variables, especially precipitation rate, were favorable in these types of models.

1. Introduction

Agriculture is facing significant challenges such as food security, environmental degradation, and climate change [1,2]. To mitigate these challenges, initiatives like Climate-Smart Agriculture (CSA) have established three pillars: productivity, defined as the ratio of agricultural outputs to inputs, resilience, or adaptation to climate change and climate change mitigation [3,4].

Considering the productivity pillar of wheat (*Triticum aestivum* L.) as one of the most important cereals worldwide and fundamental for human nutrition [5,6], plant breeding centers have commenced the task of generating new varieties with superior agronomic traits. The creation of new varieties involves a complex research process that requires the collaboration of multiple disciplines, the conduct of large-scale experiments, and the analysis of extensive historical records [7,8].

Experiments focus on improving specific wheat traits, such as increased productivity or enhanced resilience to heat, pests, and/or diseases. To evaluate if the new variety meets these aims, it is cultivated in a controlled environment (E) with specific management practices (M). Historical data for each variety cultivated in breeding centers is obtained by integrating genotypic (G) data with E and M data over several years [9–11]. Specifically, historical data includes phenotypic traits of plants, soil characteristics, climatic conditions, cultural practices, and images of plants set up in plots.

Grain yield (GY) is a crucial indicator for agronomists, as it allows for estimating the total production of a crop based on the analysis of a representative sample. Therefore, plant breeders have utilized historical GY data to develop statistical regression models that predict the yield of new wheat varieties [12,13], as well as crop simulation models, which dynamically interact with variables, serving as powerful tools for predicting yield and physiological processes of plant growth. To execute these experiments, three aspects must be considered: a large amount of data, excellent data quality, and the ability to address high computational costs to obtain results [14,15].

1.1. Background research

Given the complexity and heterogeneity of available data, several

* Corresponding author.
E-mail address: atruebae@uaemex.mx (A.T. Espinosa).

https://doi.org/10.1016/j.atech.2025.100791
Received 5 November 2024; Received in revised form 17 January 2025; Accepted 18 January 2025
Available online 29 January 2025

researchers have employed machine learning (ML) techniques to estimate GY using G x E x M variables, demonstrating that ML can incorporate different types of data and achieving adequate accuracy [16–19]. ML offers generalizability and nonlinear interactions between predictor and target variables, reduces computational requirements, is easy to implement, and works efficiently with large datasets.

Gómez [20] worked with historical data of wheat crops sown in Mexico from 2004 to 2018, using characteristics such as the Normalized Difference Vegetation Index (NDVI) obtained from satellite images, as well as climate data generated by remote sensing (RS). The dataset was filtered using feature selection based on Pearson correlation (0.5, 0.75 and 0.9), resulting in four feature subgroups: three filtered by Pearson correlations and a final group including all unfiltered features to contrast results. The filtered data was used to train and evaluate ML models: Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). The model that presented the best prediction was RF, filtered by a Pearson correlation of 0.50, with a Root Mean Squared Error (RMSE) of 78 and a correlation coefficient (R^2) of 0.84.

Zhou [21] designed a study to evaluate the ability of different ML models to predict wheat yield using a database that included climatic variables (temperature, evaporation, solar radiation, and wind speed) and vegetation indices such as NDVI and solar-induced chlorophyll fluorescence (SIF) obtained from remote sensing. The authors divided China into three zones according to wheat yield and combined the data into three groups: climatic data only, index data only, and a combination of both. With these groups, they trained RF, SVM, and the least absolute shrinkage and selection operator (LASSO) models. The results showed that RF and SVM models, when using combinations of climatic data, NDVI, and SIF, achieved higher accuracy in yield prediction, with R^2 and RMSE values ranging from 0.66 to 0.79 and 63–74, respectively.

Lastly, in the most recent study Ashfaq [22], wheat crops in the Multan region of Punjab, Pakistan, were analyzed from 2017 to 2022. The study used climatic, soil, and NDVI data obtained from remote sensing and employed RF, SVM, and LASSO techniques to predict the annual yield of the entire Multan region. The results proved that RF, using climatic and NDVI data, was the best prediction technique, with the highest R^2 of 0.78–0.88. Additionally, all three techniques individually have the potential to surpass the traditional approach to yield prediction in the winter wheat-growing area of Multan.

Thanks to these researchers, the value of historical data as a predictive tool for wheat yield and the potential of ML techniques to surpass traditional methods have been proved. However, the heterogeneity of historical data repositories poses challenges in terms of preprocessing. While filtering and feature selection techniques or limiting the analysis to confirmed datasets are common approaches, these options, in addition to restricting the scope of the analysis, can significantly increase the mean squared error in predicting wheat yield. To address these challenges, it is necessary to explore new preprocessing techniques and consider the integration of data from multiple sources.

This research has a twofold aim:

- To develop a data analysis platform to use historical repositories from plant breeding centers.
- To construct a high-precision predictive model for wheat yield.

To achieve this, an entity-relationship model was implemented to store and manage tabular data (.csv, .xls, and .xlsx), which has a REST API capable of filtering data using the same model.

Subsequently, the API was used to construct a dataset to which feature selection techniques based on correlation and principal component analysis (PCA), as well as temporal clustering by day, week, fortnight, and month, were applied to reduce dimensionality. The resulting datasets were used to train and evaluate RF, SVR, and XGBoost models, which have shown great potential in crops other than wheat [23,24]. The research found the most influential variables in wheat yield and developed a predictive model capable of supporting decision-making in agriculture.

2. Material and methods

2.1. Computing resources

All tests and training were conducted on a computer system with the following specifications: Intel i7 quad-core processor at 4.0 GHz with 8 threads, an NVIDIA Quadro 600 graphics card, 24 GB of RAM, and a storage capacity of 250 GB SSD and 2 TB HDD. This computing equipment was selected because many plant breeding centers lack surplus computational resources. Using prohibitively expensive equipment would create difficulties for these centers. Therefore, we chose conventional hardware that can be used by any plant breeding center, representing the minimum requirements for developing these types of artificial intelligence models.

2.2. Historical wheat data retrieval

An exhaustive search was conducted in various repositories to find a robust and comprehensive dataset on GY across different wheat varieties. The CIMMYT data portal was selected as the primary source due to the quality and quantity of information available. In particular, the International Durum Yield Nursery (IDYN) project [25] provided a dataset of 52 files containing phenotypic and genotypic data from wheat experiments worldwide. These data, enriched with metadata from the Crop Ontology Project (COP) [26], offer a detailed description of how each crop characteristic was obtained and quantified.

The IDYN project's CSV and XLS files were used to create a relational database (Fig. 1). The data was categorized into genotypes, locations, phenotypes, and management practices, stored in tables Trail, Genotype, Location, and EnvironmentDefinition, respectively. To accommodate future phenotypic variables, a raw data table (RawCollection) was included, allowing for data transformation and validation before integration into the main tables.

A RESTful API was developed using Python 3.12 and SQLAlchemy to interact with the IDYN project database. The API offers endpoints for creating new data records[1] and querying existing data.[2] The querying endpoints allow for sophisticated filtering based on multiple criteria, such as genotype, planting date, location, and various phenotypic characteristics stored in the database.

An automation process was implemented to consolidate and analyze the agricultural data from the 52 IDYN project files. An agent[3] employed regular expressions to extract key information from CSV and XLS files, which was then submitted to a database using POST requests.

Upon completion of the process, a whole database was generated, encompassing data from 953 wheat genotypes across 62 countries with 192 phenotypical data.

2.3. Dataset construction

Through the API, data from all wheat experiments were retrieved. However, a high percentage of missing data (exceeding 92 %) was detected, limiting global-level analysis. A detailed analysis of characteristics in Crop Ontology revealed that some variables were specific to certain countries or needed analytical technologies not available at all evaluation sites. To overcome this limitation, a Python script was

[1] Phen API Store, code accessible on GitHub: https://github.com/carlosmorenophd/phen_api_store.

[2] Phen API Fetch, code accessible on GitHub: https://github.com/carlosmorenophd/phen_api_fetch.

[3] Phen Field Book, code accessible on GitHub: https://github.com/carlosmorenophd/phen_field_book.

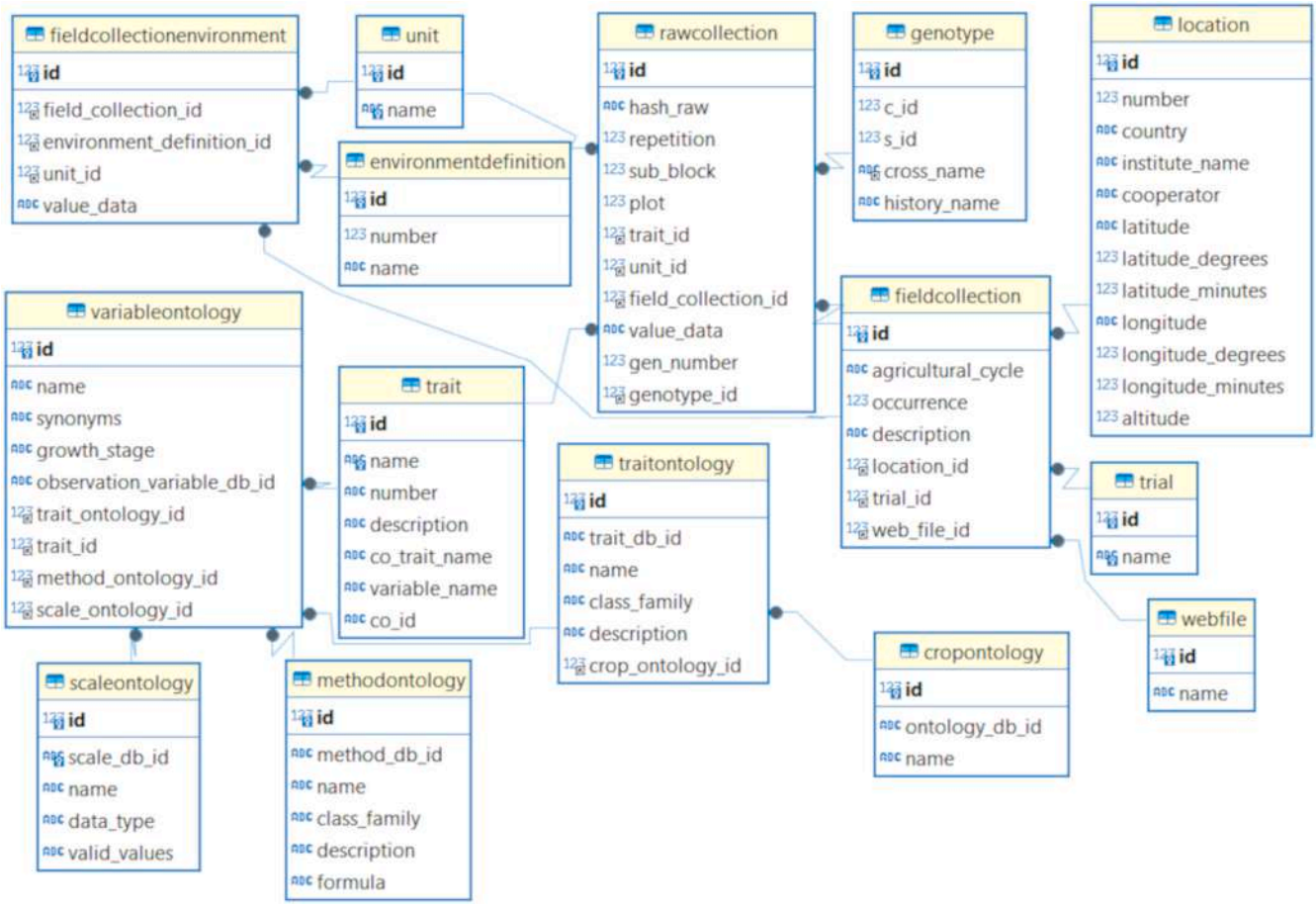

Fig. 1. Entity-relationship diagram for the International Durum Yield Nursery data.
An adaptable entity-relationship model for agricultural data management. Its flexible structure allows for the storage and analysis of a broad spectrum of variables, including phenotypic and climatic data, making it suitable for various agricultural research projects.

implemented to select experiments with the highest number of complete and relevant features.

A script was employed to iteratively remove features and experiments that contained missing data. This rigorous filtering process resulted in a dataset of 350 wheat genotypes evaluated at the Norma E. Borlaug Scientific Station, in Ciudad Obregón, Sonora, Mexico, during the 2015–2016 and 2016–2017 cycles. Experiments were established in an alpha lattice design, replicated three times across six blocks, and cultivated under flood irrigation conditions. To enhance the dataset's contextual richness, climatic data from the nearest weather station, comprising 32 variables, was incorporated.

2.4. ML model definition

According to multiple studies [27–31], in this work, three ML algorithms were employed: SVR, RF, and XGBoost.

2.4.1. Support vector regression

SVR is a machine learning model that extends Support Vector Machines (SVM) to regression tasks. It aims to find a regression hyperplane that minimizes the distance between the hyperplane and the data points. To handle complex relationships, SVR often employs kernel functions to map the data into a higher-dimensional space, where a linear regression model can be applied. [28,32]. The core algorithm of SVR is as follows:

$$\frac{1}{2}\| W\|^2 + C \sum_{i=1}^{N} \left(\xi_i + \xi_i^*\right), \tag{1}$$

With constrains, when $y = wx + b$,

$$y_i - W{\cdot}x_i - b \leq \in + \xi_i$$

$$W{\cdot}\, x_i + b - y_i \leq \in + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0$$

where W represents the weights, C is a regularization parameter, ξ_i and ξ_i^* are slack variables, ϵ represents the margin, and x_i and y_i are the input variables and output (target) values, respectively.

2.4.2. Random forest

RF is an ensemble learning method that can be applied to both regression and classification tasks [33]. Introduced by Breiman [34], RF created a random vector Θ_k for every k^{th} tree, all vectors $\Theta_{1,\ldots\ldots,}\Theta_{k-1}$ are independent but have the same distribution; each tree builds a regression model $h(\boldsymbol{x},\ \Theta_k)$where $\boldsymbol{x}$ is an input vector. The result is a set of regression values, these are evaluated by the mean and this becomes the RF result.

$$Y = \sum_{k=1}^{j} \frac{h(\boldsymbol{x},\ \Theta_k)}{k}, \tag{2}$$

where Y is the result of the regression predict, j is the total of tree, $h(\boldsymbol{x},\ \Theta_k)$ is the regression value of the k^{th} tree.

2.4.3. eXtreme gradient boosting

XGB, a powerful machine learning technique introduced by Chen and Guestrin [35], is widely used for regression tasks. It leverages

ensemble learning to combine multiple decision trees, iteratively improving the model's predictive accuracy. By assigning weights to individual trees based on their performance, XGB effectively minimizes prediction errors [2]. The prediction process was defined by:

$$\widehat{y} = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i), \quad f_k \in F, \tag{3}$$

where $\boldsymbol{x}_i$ represents the vector of the i^{th} samples f_k is the k^{th} regression tree and F be the set of all decision trees. Each tree is defined recursively as:

$$f_k(\boldsymbol{x}) = w_{q(x:\theta_k),} \tag{4}$$

where w represents the lead weights and $q(x:\theta_k)$ is a function that maps the input features $x = (x_1, x_2, x_3, x_4)$ to a leaf in the k^{th} tree. θ_k represents the parameters of the k^{th} tree. The final prediction is the sum of all the tree predictions:

$$Y = \sum_{k=1}^{K} w_{q(x;\theta_k)\cdot,} \tag{5}$$

where Y is the final prediction, K is the general set of trees and $w_{q(x;\theta_k)\cdot}$ is the k^{th} tree result.

The objective function of the XGB model typically includes a loss term to minimize the difference between the predicted and actual values, as well as regularization terms to control the complexity of the trees:

$$\mathscr{L}(\Theta) = \sum_{i=1}^{N} L(y_i, \widehat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) + \lambda \|\Theta_k\|_2, \tag{6}$$

where:

N: Number of training samples,
L: Loss function,
y_i: Actual wheat yield for sample i,
$\widehat{y}_i$: Predict wheat yield for sample i,
K: Number of trees,
$\Omega(f_k)$: Regularization term for tree k,
λ: Regularization parameter,
$\| \Theta_k\|_2$: $L2$ norm of parameters for tree k,

The model parameters Θ_k are learned by optimizing the objective function $\mathscr{L}(\Theta)$ using techniques such as gradient boosting.

2.5. Running and configuring ML systems

These default parameters were selected for the ML models due to the computational capabilities of the equipment, which allowed the use of the algorithms default's values. The available computational ability allowed us not to impose more restrictions on the parameter definition.

- **SVR**: Four kernels (sigmoid, polynomial, linear, and RBF) were experimented with, fixing epsilon at 0.1 and the regularization parameter at 1.0.
- **RF**: A forest with 1000 trees were constructed, allowing each tree to grow until it reaches pure leaves.
- **XGBoost**: A boosted gradient tree was used with a learning rate of 0.3, a shrinkage rate of zero, and a maximum depth of 6.

To evaluate the models, the following metrics were employed: RMSE, R^2, and mean absolute percentage error (MAPE). The RMSE measures the difference between the actual and predicted values, highlighting outliers. The value of R^2 indicates how much of the variance between the variables can be explained by the linear fit. The MAPE value measures the average error in percentage terms compared to the actual values. These error measurements are frequently used for agricultural systems and crop models [36–38]. To help the interpretation of the results and evaluate the performance of ML models, was used equation number one, which stands for the accuracy function.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}, \tag{7}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right|, \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{9}$$

$$Accuracy = 1 - MAPE, \tag{10}$$

where $\widehat{y}_i$ represents the predicted on i^{th} sample, $\bar{y}$ represents the mean of actual predicted on y_i

A two-stage experiment was conducted to determine the optimal predictors of GY. In the first stage, a comprehensive analysis of both phenotypic and climatic variables was performed. Various grouping and selection techniques were applied to identify the most influential variables associated with GY. Subsequently, a focused analysis was conducted on phenotypic variables, employing feature reduction techniques to pinpoint the most informative features. For each stage, a random 80/20 split of the data was used for training and testing, allowing for a rigorous evaluation of the different models.

2.6. Construction of data subsets

2.6.1. Phenotypic data

The analysis initially focused on the 68 phenotypic characteristics of the crop, paying particular attention to the NDVI and canopy temperature (CT). Both variables were obtained using two complementary techniques: RS and unmanned aerial vehicles (UAV). Multiple measurements were taken at various stages of the crop cycle, both before and after flowering (VG and GF, respectively). To visualize the relationships between these variables and GY, a correlation matrix was constructed and represented in a heatmap (Fig. 2). The results showed a high correlation (greater than 0.90) between these characteristics.

Following the recommendations of Kheir [39], a selection of variables was made to reduce the dimensionality of the phenotypic data and mitigate multicollinearity problems. Variables with a correlation coefficient greater than 0.95 were removed, resulting in a final set of 12 independent variables (Table 1) that captures most of the variability in the data.

A PCA was performed to complement the correlation-based analysis. Principal components that together explained at least 95 % of the total variance were selected, following the recommendations of Adilova and Aravind [40,41]. After preprocessing, two phenotypic datasets were generated: one filtered by correlation and another one filtered by PCA.

2.6.2. Climatic data

Climatic data was collected at a 5-minute interval for 66 days, resulting in a large dataset. Following the guidelines of Zhou [21], this data was aggregated into daily, weekly, bi-weekly, and monthly intervals (Table 2). The time-based groups were merged with phenotypic data. Subsequently, both PCA and correlation filtering were applied to each combined dataset. The resulting eight datasets (four from PCA and four from correlation) were served as input for ML models to assess their ability to forecast GY.

2.7. Conducting experiments

Experiments with SVR, RF, and XGBoost models were divided into three stages. In the first (control) stage, phenotypic datasets were

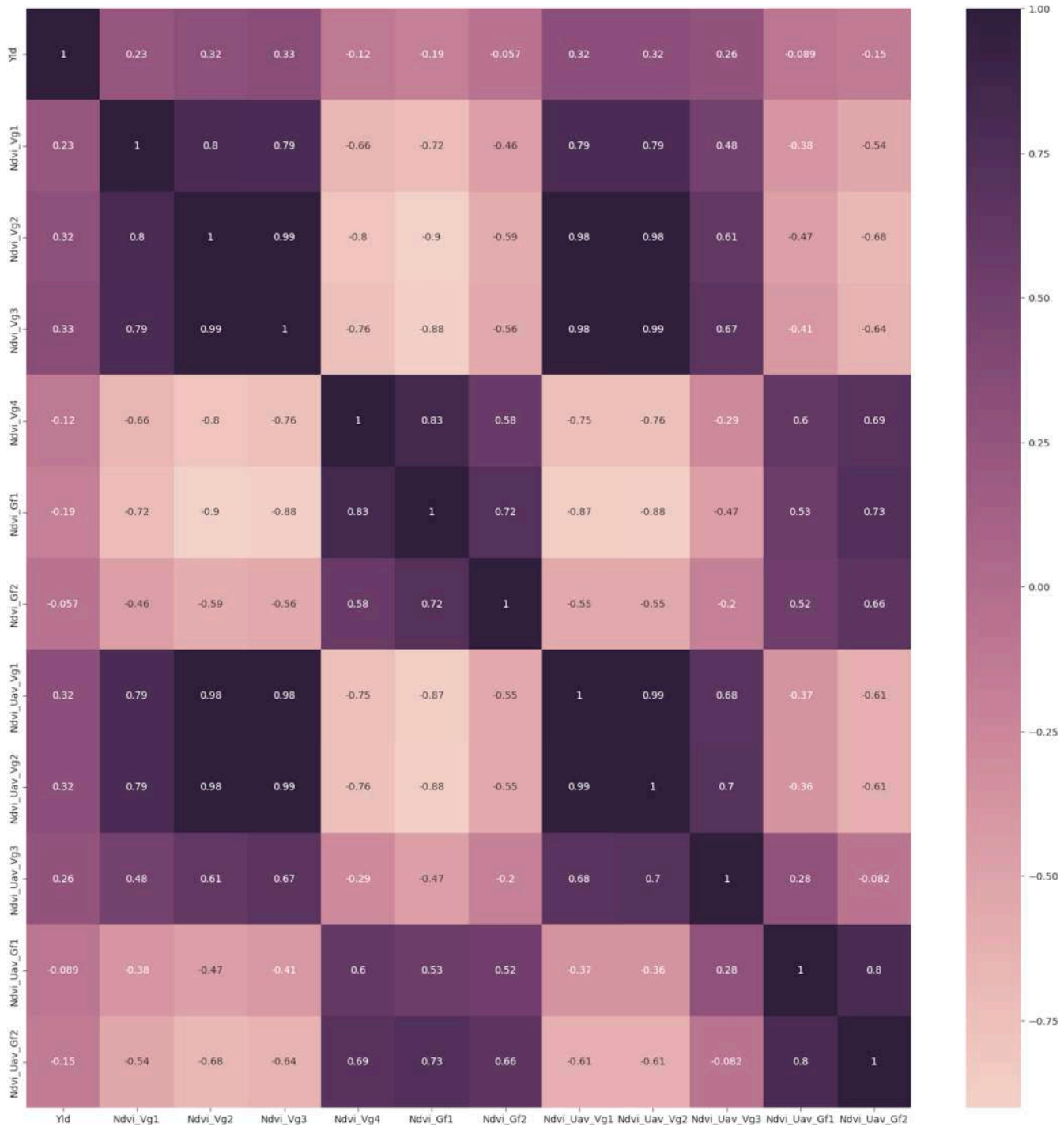

Fig. 2. Heatmap of NDVI and GY characteristics.
The heatmap highlights the strength of the correlation between NDVI and GY. The most intense colors signify a very strong positive or negative association between these variables.

evaluated using correlation filtering and PCA, setting up a baseline for comparing model performance in next stages. The lowest MAPE obtained in this stage was selected as the reference value.

In the second stage (integration), the eight filtered datasets, created in the previous step, were subjected to further analysis. By comparing the performance of models trained on these different datasets, the impact of various variable combinations on model accuracy was evaluated.

Finally, in the third stage (evaluation), an analysis was conducted to assess the individual importance of each phenotypic variable on model performance. Given the complexity of the model and the large amount of data, a two-phase approach was adopted:

- **Evaluation without NDVI:** Phenotypic data was evaluated excluding the NDVI variable to find the most relevant phenotypic variables in this subset of data.
- **Evaluation with NDVI:** Subsequently, the impact of the NDVI variable was assessed by including it in the model, along with the phenotypic variables selected in the earlier stage.

Table 1
12 selected features with low pairwise correlation.

Variable	Description
TGW	Thousand-grain weight
HI	Harvest index
BM	Biomass
NDVI_VG1	NDVI by SR before flowering first measurement
NDVI_VG2	NDVI by SR before flowering second measurement
NDVI_VG4	NDVI by SR before flowering fourth measurement
NDVI_GF1	NDVI by SR after flowering first measurement
NDVI_GF2	NDVI by SR after flowering second measurement
NDVI_GF3	NDVI by SR after flowering third measurement
NDVI_UAV_3	NDVI by UAV after flowering third measurement
NDVI_UAV_4	NDVI by UAV after flowering fourth measurement
NDVI_UAV_5	NDVI by UAV after flowering fifth measurement

Table 2
Climate features selected through correlation analysis.

Time-based clustering of climate features	Number of climate records	Number of filtered climate records
Daily average	2068	219
Weekly average	322	37
Bi-weekly average	167	23
Monthly average	105	17

This approach allowed for the identification of both the most influential phenotypic variables in the absence of NDVI and those that, in combination with NDVI, improve GY prediction.

3. Results

The control stage produced the best results when using an RF model with feature reduction by correlation and PCA (Table 3). This model achieved a MAPE of 0.0381, an RMSE of 35.78 (g/m^2), and an R^2 of 0.8670. Additionally, it was found that the linear kernel is the most suitable for the SVM model, and therefore it was used in the next experiments.

Having achieved a MAPE of 0.0381 in the control stage, a benchmark was set up for evaluating the integration of phenotypic and climatic variables. The results of the integration stage are presented in (Tables 4 and 5), highlighting the outcomes of feature reduction using PCA and correlation analysis, respectively.

When compared to the benchmark MAPE, both daily and monthly climatic variables, when integrated with an RF model and correlation-based feature reduction, achieved performance comparable to the benchmark. However, a more in-depth analysis of the RMSE and R^2 revealed that the model using daily climatic variables showed a slightly superior fit, with an RMSE of 35.7214 and an R^2 of 0.8675. Models employing PCA for feature reduction did not reach the level of performance achieved by correlation-based models.

Upon application of a correlation filter to daily climatic data, precipitation rate, rainfall, and cooling and heating degree-days appeared as the most influential variables in the analysis (Fig. 3). The higher frequency of these variables in the results shows a greater significance in relation to other assessed variables, including atmospheric pressure, relative humidity, and heat index.

During the first phase of the evaluation, a comprehensive analysis was undertaken, excluding NDVI data. Experiments were selected based on a MAPE threshold of 0.0381. Detailed findings of these selected combinations are tabulated in (Table 6).

The variables HI, BM, and CT_UAV_1, identified in the initial phase, served as the foundation for the subsequent stage of analysis. A comprehensive exploration of all possible combinations of these variables with NDVI indices was conducted to determine the optimal

Table 4
PCA for feature reduction: Phenotypic-Climatic.

Clustering of climatic variables	ML	Accuracy	Metrics MAPE	RMSE	R^2
Daily average	SVR	0.9191	0.0809	61.6549	0.6053
	RF	0.9273	0.0727	55.4664	0.6806
	XGBoost	0.9343	0.0657	52.6943	0.7117
Weekly average	SVR	0.9191	0.0809	61.6516	0.6054
	RF	0.9273	0.0727	54.4538	0.6921
	XGBoost	0.9264	0.0736	57.623	0.6553
Bi-weekly average	SVR	0.9192	0.0808	61.6379	0.6056
	RF	**0.9291**	**0.0709**	**53.0556**	**0.7078**
	XGBoost	0.9283	0.0717	55.8784	0.6758
Monthly average	SVR	0.9177	0.0823	63.183	0.5855
	RF	0.9245	0.0755	56.6681	0.6666
	XGBoost	0.9314	0.0686	53.9475	0.6978

Table 5
Correlation for feature reduction: Phenotypic-Climatic.

Clustering of climatic variables	ML	Metrics Accuracy	MAPE	RMSE	R^2
Daily average	SVR	0.9191	0.0809	61.6549	0.605
	RF	**0.9619**	**0.0381**	**35.7214**	**0.868**
	XGBoost	0.9574	0.0426	40.9659	0.826
Weekly average	SVR	0.9191	0.0809	61.6516	0.605
	RF	0.9618	0.0382	35.7218	0.868
	XGBoost	0.9574	0.0426	40.9659	0.826
Bi-weekly average	SVR	0.9192	0.0808	61.6379	0.606
	RF	0.9618	0.0382	35.833	0.867
	XGBoost	0.9574	0.0426	40.9659	0.826
Monthly average	SVR	0.9177	0.0823	63.183	0.586
	RF	**0.9619**	**0.0381**	**35.8191**	**0.867**
	XGBoost	0.9574	0.0426	40.9659	0.826

Table 3
Model metrics in the control stage.

Feature reduction	Model	Model settings	Accuracy	Metrics MAPE	RMSE	R^2
Correlation	SVR	Lineal	0.9583	0.0417	40.6686	0.8283
		Poli	0.9064	0.0936	68.5616	0.512
		Sigmoid	0.8918	0.1082	78.7287	0.3565
		RBF	0.8768	0.1232	90.7393	0.1452
	RF		**0.9619**	**0.0381**	**35.7831**	**0.8671**
	XGBoost		0.9574	0.0426	40.9659	0.8258
PCA	SVR	Lineal	0.9177	0.0823	63.1897	0.5854
		Poli	0.8884	0.1116	83.6499	0.2735
		Sigmoid	0.8882	0.1118	80.3868	0.3291
		RBF	0.8881	0.1119	82.7321	0.2894
	RF		0.9238	0.0762	58.5873	0.6436
	XGBoost		0.9302	0.0698	53.7795	0.6997

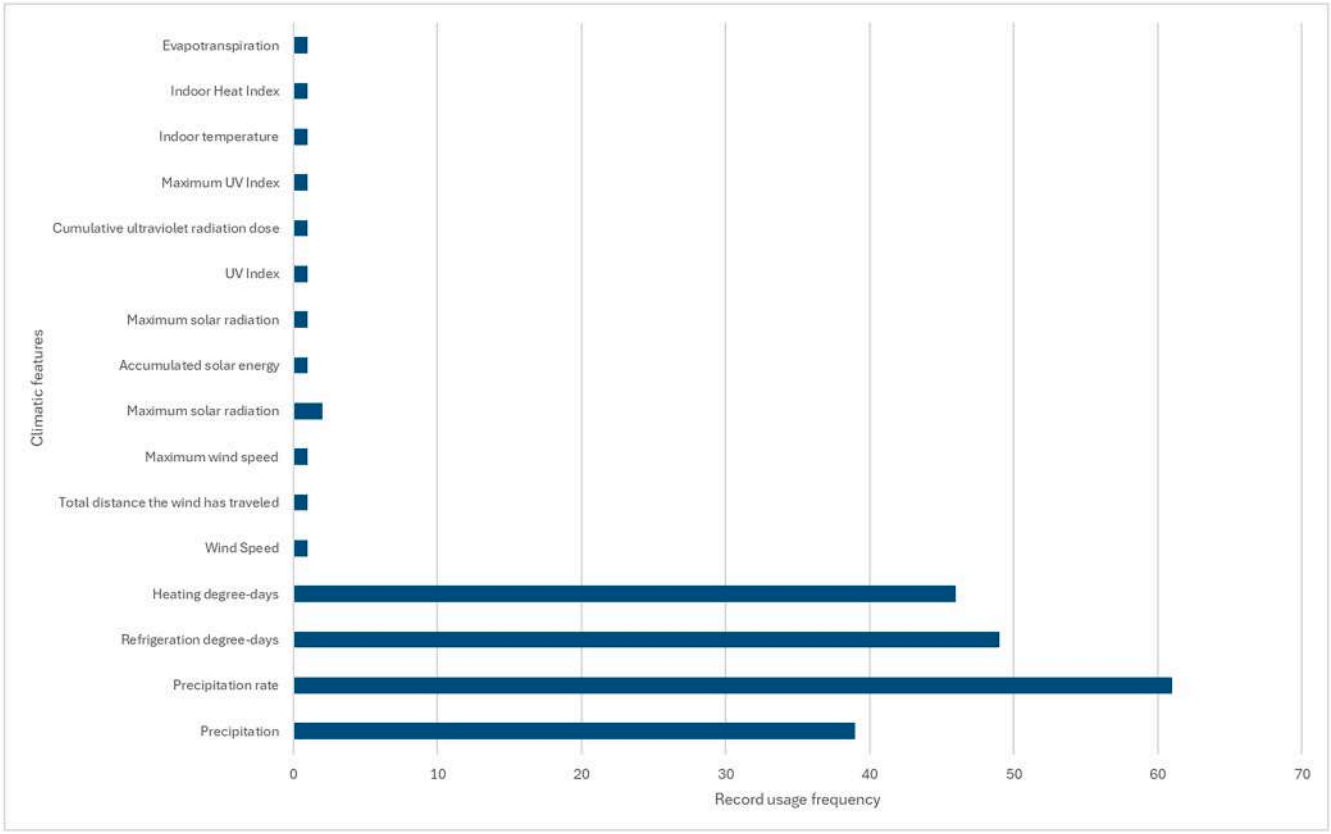

Fig. 3. Occurrence of daily climate variables after correlation filtering.
A bar graph visualizing the frequency with which each climatic variable was selected in the feature selection process based on correlation.

Table 6
Exploration of diverse phenotypic variable combinations, excluding NDVI.

Phenotypic features		Metrics			
	ML	Accuracy	MAPE	RMSE	R^2
HI, BM, CT_UAV_1	**RF**	**0.9648**	**0.0352**	**34.1861**	**0.8787**
HI, BM, CT_UAV_1	XGB	0.9640	0.0360	34.5892	0.8758
TGW, HI, BM, CT_UAV_1	RF	0.9632	0.0368	35.0134	0.8727
HI, BM, CT_UAV_2	RF	0.9627	0.0373	37.7283	0.8522
HI, BM, CT_UAV_1, CT_UAV_2	RF	0.9626	0.0374	37.9395	0.8506
HI, BM, CT_UAV_2	XGB	0.9626	0.0374	37.7163	0.8523
HI, BM, CT_UAV_1, CT_UAV_3	RF	0.9625	0.0375	36.4081	0.8624
TGW, HI, BM, CT_UAV_1, CT_UAV_3	RF	0.9621	0.0379	34.9847	0.8729

Table 7
Exploration of diverse agronomic variable combinations, including NDVI.

Indices NDVI		Metrics			
	ML	Accuracy	MAPE	RMSE	R^2
NDVI_GF1, NDVI_UAV_1, NDVI_UAV_2, NDVI_UAV_3, NDVI_UAV_4, NDVI_UAV_5	**XGB**	**0.9677**	**0.0323**	28.5082	**0.9156**
NDVI_GF1, NDVI_UAV_2, NDVI_UAV_3, NDVI_UAV_4	RF	0.9669	0.0331	30.8055	0.9015
NDVI_GF1, NDVI_UAV_2, NDVI_UAV_3	RF	0.9667	0.0333	30.8463	0.9012
NDVI_VG1, NDVI_GF1, NDVI_UAV_2, NDVI_UAV_3	RF	0.9667	0.0333	31.1541	0.8992
NDVI_VG1, NDVI_GF1, NDVI_UAV_2, NDVI_UAV_3, NDVI_UAV_4	RF	0.9666	0.0334	31.0162	0.9001
NDVI_VG, NDVI_GF1, NDVI_GF3, NDVI_GF4, NDVI_GF, NDVI_UAV_1, NDVI_UAV_2, NDVI_UAV_3, NDVI_UAV_4	XGB	0.9665	0.0335	31.1075	0.8995
NDVI_GF1, NDVI_UAV_3	RF	0.9665	0.0335	31.8240	0.8949
NDVI_GF1, NDVI_UAV_3, NDVI_UAV_4	RF	0.9665	0.0335	31.1607	0.8992
NDVI_GF1, NDVI_GF4, NDVI_UAV_4	RF	0.9664	0.0336	32.6567	0.8893
NDVI_GF1, NDVI_GF4, NDVI_UAV_2, NDVI_UAV_3, NDVI_UAV_4	RF	0.9664	0.0336	32.3657	0.8912

configuration that minimized MAPE and maximized GY prediction accuracy. The resulting top 10 configurations are detailed in (Table 7).

Training times for the models were surprisingly fast, with most completing the process in less than a second. Even when working with large datasets (221 variables and 560 genotypes), the models proved high computational efficiency. The RF model, due to its structure of multiple decision trees, needed slightly longer training times, reaching a maximum of 16 ss (Fig. 4).

4. Discussion

As a first step in the preliminary control phase, the random forest model presented the best results, being the best result using the feature reduction technique of filtering by heat maps, obtaining an accuracy of 0.96 and a RMSE of 35.78 with an R2 of 0.86. With this result, we started evaluating the climatic variables using the two filtering techniques, both

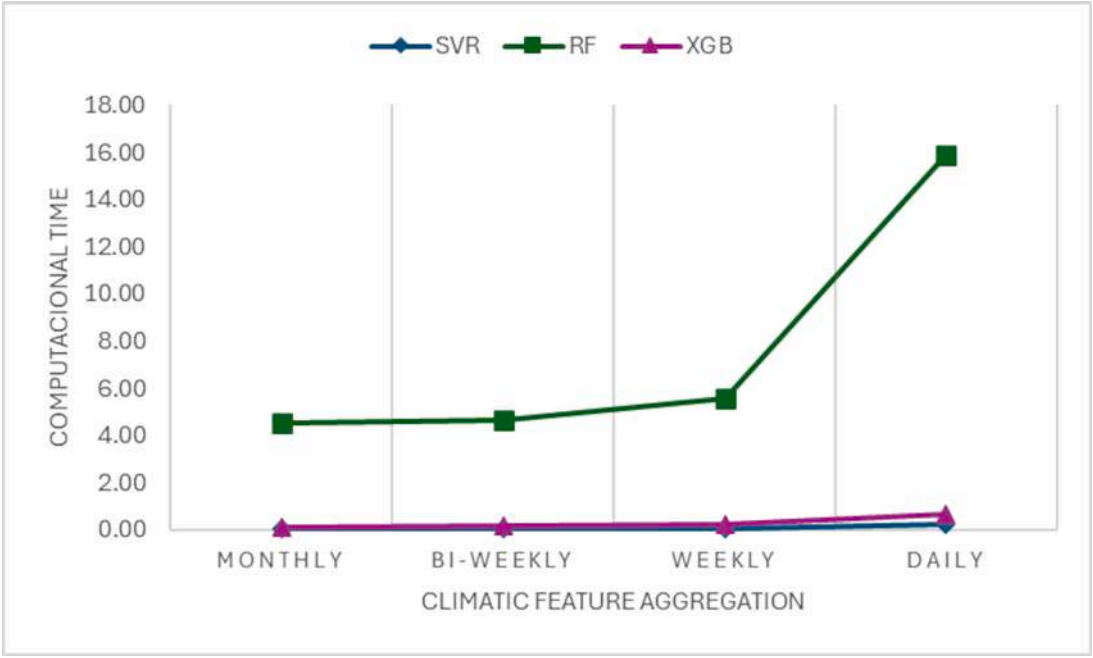

Fig. 4. Computational time of models with phenotypic and climatic variables.
Graph showing the runtime on second of RF, SVR and XGB with different groupings of climatic variables.

PCA and heat maps. The result showed that the grouping of daily and monthly climatic data using feature selection by heat maps obtained the same accuracy. This agrees with different works [39,42] that suggest that climatic variables have a contribution, although this was not very high but was substantial to maintain the same accuracy that had been obtained before. PCA feature reduction appears suboptimal in this case. To generalize these findings, recommend testing other datasets and alternative feature selection methods, such as selection by correlation of Pearson [43].

Among the climatic variables, precipitation was frequently selected by the feature reduction technique. However, its impact was less pronounced than expected because all agronomic experiments in the dataset were conducted under irrigation. This suggests that the influence of precipitation on wheat yield may have been masked by the controlled irrigation conditions. Therefore, recommend future studies utilize datasets that include both irrigated and non-irrigated (seasonal crops) conditions to better evaluate the impact of precipitation on wheat yield.

Datasets incorporating phenotypic variables, whether augmented with daily or monthly aggregated climatic variables, have proved satisfactory performance in our analysis. Common to these datasets is the inclusion of variables such as NDVI, BM, and HI. Notably, NDVI has emerged as a pivotal variable in numerous studies [44–46] due to its capacity to characterize the relationship between photosynthesis and active biomass. Derived from multiple captures throughout the crop growth cycle, NDVI has been instrumental in forecasting GY [47–49]. Although >17 NDVI captures were available, correlation-based feature selection enabled the model to be streamlined to just 9 variables, resulting in robust GY predictions.

During the first evaluation phase, a comprehensive analysis of non-NDVI variables revealed HI, BM, and CT_UAV_1 as the most significant predictors of yield. In the next phase, an exhaustive exploration of all possible combinations of these three variables with various NDVI samples was conducted, generating over 86,000 models. Optimal performance was reached by combining the first pre-flowering NDVI sample and the fifth drone-acquired NDVI sample, utilizing XGBoost and RF algorithms. These models showed a minimum accuracy of 0.96 and an RMSE of 28.50, underscoring the substantial contribution of NDVI variables to wheat yield prediction.

The SVR method showed suboptimal performance, likely due to its susceptibility to outliers. SVM models are inherently sensitive to extreme values or measurement errors, which are prevalent in historical datasets derived from diverse sources and collected over extended periods. Inconsistencies in data collection protocols, variations in measurement instruments, and equipment degradation can introduce noise into the data, compromising algorithm efficacy. Researchers using artificial intelligence algorithms should prioritize the development of techniques to mitigate the adverse effects of such irregularities. Potential strategies include outlier detection and removal, or the adoption of algorithms, such as RF, which are more resilient to outliers.

RF proved superior performance across most datasets. This exceptional performance can be attributed to RF's inherent ability to mitigate overfitting by employing an ensemble of 1000 decision trees. The averaging of predictions from multiple trees effectively reduces the impact of individual classification errors, enhancing model accuracy and robustness. These findings are consistent with previous research [22,34,43,50], which has established RF as a highly effective method for estimating GY.

The findings of this study say that the selection of an ML algorithm is contingent upon the specific characteristics of the dataset. While XGBoost can achieve high predictive accuracy, its performance is susceptible to the curse of dimensionality. Conversely, RF proved superior robustness and efficiency when dealing with high-dimensional data. It is recommended to employ XGBoost for datasets with a limited number of predictors when a comprehensive exploration of the feature space is desired. However, for high-dimensional datasets, RF is the right choice, as it enables the rapid identification of the best hyperparameters and shows strong generalization capabilities.

The main goal of this research was to pinpoint the most significant factors influencing GY. The results of this study can be applied to develop more accurate and robust predictive models. Future research could benefit from using hyperparameter optimization and evolutionary algorithms to fine-tune these models.

5. Conclusion

Plant breeding centers support historical data repositories that offer

a valuable resource for agricultural research. While traditional analysis of these data has relied on complex statistical methods such as multiple regression and simulation modeling, the emergence of artificial intelligence has eased more efficient and exact data extraction.

Our findings prove that the variables HI, BM, CT, and NDVI were pivotal in predicting wheat yield, resulting in a minimum RMSE of 28.5082 and an R^2 of 0.9156. These results highlight the critical importance of meticulous variable selection in the development of predictive models, as it significantly enhances the accuracy and generalizability of the outcomes.

It is imperative to recognize that the selection of a proper artificial intelligence technique is contingent upon the specific context and the nature of the available data. Given the unique strengths and weaknesses of different algorithms, experimentation with various techniques is essential to find the best solution for a given problem. In this study, RF and XGBoost have appeared as robust tools for predicting wheat yield, particularly when integrating crop and climatic variables.

CRediT authorship contribution statement

Juan Carlos Moreno Sánchez: Writing – original draft, Software, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Héctor Gabriel Acosta Mesa:** Writing – review & editing, Investigation, Formal analysis, Data curation. **Adrián Trueba Espinosa:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis. **Sergio Ruiz Castilla:** Visualization, Supervision. **Farid García Lamont:** Visualization, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics Statement

No this manuscript does not include human or animal research.

Acknowledgements

This research was supported by the Universidad Autónoma del Estado de México (**UAEMex**), Consejo Nacional de Humanidades Ciencia y Tecnología **(CONAHCYT)**, Universidad Veracruzana (**UV**) and Centro de Investigación de Mejoramiento del Maíz y Trigo **(CIMMYT)**.

Data availability

The authors do not have permission to share data.

References

[1] Z. Cui, H. Zhang, X. Chen, C. Zhang, W. Ma, C. Huang, W. Zhang, G. Mi, Y. Miao, X. Li, Q. Gao, J. Yang, Z. Wang, Y. Ye, S. Guo, J. Lu, J. Huang, S. Lv, Y. Sun, Y. Liu, X. Peng, J. Ren, S. Li, X. Deng, X. Shi, Q. Zhang, Z. Yang, L. Tang, C. Wei, L. Jia, J. Zhang, M. He, Y. Tong, Q. Tang, X. Zhong, Z. Liu, N. Cao, C. Kou, H. Ying, Y. Yin, X. Jiao, Q. Zhang, M. Fan, R. Jiang, F. Zhang, Z. Dou, Nature, Pursuing sustainable productivity with millions of smallholder farmers, 555, Springer Science and Business Media LLC, England, 2018, pp. 363–366, https://doi.org/10.1038/nature25785.

[2] S.A. Gyamerah, C. Asare, D. Mintah, B. Appiah, F.A. Kayode, Exploring the optimal climate conditions for a maximum maize production in Ghana: Implications for food security, Smart Agricult. Technol. 6 (2023) 100370. https://doi.org/10.1016/j.atech.2023.100370.

[3] L.A. Germer, C.E. van Middelaar, S.J. Oosting, P.J. Gerber, When and where are livestock climate-smart? A spatial-temporal framework for comparing the climate change and food security synergies and tradeoffs of Sub-Saharan African livestock systems, Agricul. Syst. 210 (2023) 103717. https://doi.org/10.1016/j.agsy.2023.103717.

[4] M.M. Guja, S.B. Bedeke, Smallholders' climate change adaptation strategies: exploring effectiveness and opportunities to be capitalized, Environ. Dev. Sust. (2024). https://doi.org/10.1007/s10668-024-04750-y.

[5] N.N. Zakharova, N.G. Zakharov, Wheat grain production in the world and its dynamics, E3S Web. Conf. 480 (2024) 03001. https://doi.org/10.1051/e3sconf/202448003001.

[6] Z. Li, Z. Chen, Q. Cheng, F. Duan, R. Sui, X. Huang, H. Xu, UAV-Based Hyperspectral and Ensemble Machine Learning for Predicting Yield in Winter Wheat, Agronomy 12 (2022). https://doi.org/10.3390/agronomy12010202.

[7] H.E. Villaseñor-Mir, R. Hortelano-Santa Rosa, E. Martínez-Cruz, J. Huerta-Espino, E. Espitia-Rangel, E. Solís-Moya, J.I. Alvarado-Padilla, J. Calle-Bellido, P.H. S. Angel, GRATA S2022: Nueva variedad de trigo suave galletero para áreas de riego en México, Revista Fitotecnia Mexicana, Sociedad Mexicana de Fitogen. A.C. 47 (2024) 93. https://doi.org/10.35196/rfm.2024.1.93.

[8] M. Kaur, H. Ram, Grain yield, heat and water use efficiency of wheat cultivars in relation to different sowing environments, Bangladesh J. Bot. 52 (2023) 701–707. https://doi.org/10.3329/bjb.v52i3.68878.

[9] V.S. Manivasagam, O. Rozenstein, Practices for upscaling crop simulation models from field scale to large regions, Comp. Electron. Agricul. 175 (2020) 105554. https://doi.org/10.1016/j.compag.2020.105554.

[10] S. Cheng, C. Feng, L.U. Wingen, H. Cheng, A.B. Riche, M. Jiang, M. Leverington-Waite, Z. Huang, S. Collier, S. Orford, X. Wang, R. Awal, G. Barker, T. O'Hara, C. Lister, A. Siluveru, J. Quiroz-Chávez, R.H. Ramírez-González, R. Bryant, S. Berry, U. Bansal, H.S. Bariana, M.J. Bennett, B. Bicego, L. Bilham, J.K.M. Brown, A. Burridge, C. Burt, M. Buurman, M. Castle, L. Chartrain, B. Chen, W. Denbel, A. F. Elkot, P. Fenwick, D. Feuerhelm, J. Foulkes, O. Gaju, A. Gauley, K. Gaurav, A. N. Hafeez, R. Han, R. Horler, J. Hou, M.S. Iqbal, M. Kerton, A. Kondic-Spica, A. Kowalski, J. Lage, X. Li, H. Liu, S. Liu, A. Lovegrove, L. Ma, C. Mumford, S. Parmar, C. Philp, D. Playford, A.M. Przewieslik-Allen, Z. Sarfraz, D. Schafer, P. R. Shewry, Y. Shi, G.A. Slafer, B. Song, B. Song, D. Steele, B. Steuernagel, P. Tailby, S. Tyrrell, A. Waheed, M.N. Wamalwa, X. Wang, Y. Wei, M. Winfield, S. Wu, Y. Wu, B.B.H. Wulff, W. Xian, Y. Xu, Y. Xu, Q. Yuan, X. Zhang, K.J. Edwards, L. Dixon, P. Nicholson, N. Chayut, M.J. Hawkesford, C. Uauy, D. Sanders, S. Huang, S. Griffiths, Harnessing landrace diversity empowers wheat breeding, Nature 632 (2024) 823–831. https://doi.org/10.1038/s41586-024-07682-9.

[11] E.A. Ogutu, S.L. Madahana, S. Bhavani, G. Macharia, Genotype × environment interaction: trade-offs between the agronomic performance and stability of durum (Triticum turgidum) wheat to stem-rust resistance in Kenya, Front. Plant Sci. 15 (2024). https://doi.org/10.3389/fpls.2024.1427483.

[12] N. Darra, E. Anastasiou, O. Kriezi, E. Lazarou, D. Kalivas, S. Fountas, Can Yield Prediction Be Fully Digitilized? A Systematic Review, Agronomy 13 (2023). https://doi.org/10.3390/agronomy13092441.

[13] M Kumar, G Prakash, SK Rana, Statistical Modeling for Analysis of Growth and Trend Pattern of Wheat Production in Selected States of India, Asian J. Res. Crop Sci. 9 (1) (2024) 66–75.

[14] B. Muller, P. Martre, Plant and crop simulation models: powerful tools to link physiology, genetics, and phenomics, J. Experiment. Bot. 70 (2019) 2339–2344. https://doi.org/10.1093/jxb/erz175.

[15] J. Huang, J.L. Gómez-Dans, H. Huang, H. Ma, Q. Wu, P.E. Lewis, S. Liang, Z. Chen, J.-H. Xue, Y. Wu, F. Zhao, J. Wang, X. Xie, Assimilation of remote sensing into crop growth models: Current status and perspective, Agric. For. Meteorol. 276–277 (2019) 107609. https://doi.org/10.1016/j.agrformet.2019.06.008.

[16] D. Paudel, H. Boogaard, A. de Wit, S. Janssen, S. Osinga, C. Pylianidis, I. N. Athanasiadis, Machine learning for large-scale crop yield forecasting, Agric. Syst. 187 (2021) 103016. https://doi.org/10.1016/j.agsy.2020.103016.

[17] L. Schmidt, M. Odening, J. Schlanstein, M. Ritter, Exploring the weather-yield nexus with artificial neural networks, Agric. Syst. 196 (2022) 103345. https://doi.org/10.1016/j.agsy.2021.103345.

[18] C. Trentin, Y. Ampatzidis, C. Lacerda, L. Shiratsuchi, Tree crop yield estimation and prediction using remote sensing and machine learning: A systematic review, Smart Agric. Technol. 9 (2024) 100556. https://doi.org/10.1016/j.atech.2024.100556.

[19] L.S. Cedric, W.Y.H. Adoni, R. Aworka, J.T. Zoueu, F.K. Mutombo, M. Krichen, C.L. M. Kimpolo, Crops yield prediction based on machine learning models: Case of West African countries, Smart Agric. Technol. 2 (2022) 100049. https://doi.org/10.1016/j.atech.2022.100049.

[20] D. Gómez, P. Salvador, J. Sanz, J.L. Casanova, Modelling wheat yield with antecedent information, satellite and climate data using machine learning methods in Mexico, Agric. For. Meteorol. 300 (2021) 108317. https://doi.org/10.1016/j.agrformet.2020.108317.

[21] W. Zhou, Y. Liu, S.T. Ata-Ul-Karim, Q. Ge, X. Li, J. Xiao, Integrating climate and satellite remote sensing data for predicting county-level wheat yield in China using machine learning methods, Int. J. Appl. Earth Obs. Geoinf. 111 (2022) 102861. https://doi.org/10.1016/j.jag.2022.102861.

[22] M. Ashfaq, I. Khan, A. Alzahrani, M.U. Tariq, H. Khan, A. Ghani, Accurate Wheat Yield Prediction Using Machine Learning and Climate-NDVI Data Fusion, IEEE Access 12 (2024) 40947–40961. https://doi.org/10.1109/ACCESS.2024.3376735.

[23] W. Zhou, Z. Yan, L. Zhang, A comparative study of 11 non-linear regression models highlighting autoencoder, DBN, and SVR, enhanced by SHAP importance analysis in soybean branching prediction, Sci. Rep. 14 (2024) 5905. https://doi.org/10.1038/s41598-024-55243-x.

[24] Y. Li, H. Zeng, M. Zhang, B. Wu, Y. Zhao, X. Yao, T. Cheng, X. Qin, F. Wu, A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering, IInt. J. Appl. Earth Obs. Geoinf. 118 (2023) 103269. https://doi.org/10.1016/j.jag.2023.103269.

[25] G.W.P. GWP, I. Collaborators, K. Ammar, and T. Payne, 52nd International Durum Screening Nursery [Dataset], (2021) [Online]. Available: https://hdl.handle.net/11529/10548599.
[26] CO, Crop Ontology, 2014 [Online]. Available, https://cropontology.org/.
[27] G. Lischeid, H. Webber, M. Sommer, C. Nendel, F. Ewert, Machine learning in crop yield modelling: A powerful tool, but no surrogate for science, Agri. For. Meteorol. 312 (2022) 108698. https://doi.org/10.1016/j.agrformet.2021.108698.
[28] E. Cheng, B. Zhang, D. Peng, L. Zhong, L. Yu, Y. Liu, C. Xiao, C. Li, X. Li, Y. Chen, H. Ye, H. Wang, R. Yu, J. Hu, S. Yang, Wheat yield estimation using remote sensing data based on machine learning approaches, Front. Plant Sci. 13 (2022) 1090970. https://doi.org/10.3389/fpls.2022.1090970.
[29] A. Vannoppen, A. Gobin, L. Kotova, S. Top, L. De Cruz, A. Vīksna, S. Aniskevich, L. Bobylev, L. Buntemeyer, S. Caluwaerts, R. De Troch, N. Gnatiuk, R. Hamdi, A. Reca Remedio, A. Sakalli, H. Van De Vyver, B. Van Schaeybroeck, and P. Termonia, Wheat Yield Estimation from NDVI and Regional Climate Models in Latvia. *Remote Sensing*, 12 (2020) https://doi.org/10.3390/rs12142206.
[30] M.S. Chiu, J. Wang, Evaluation of Machine Learning Regression Techniques for Estimating Winter Wheat Biomass Using Biophysical, Biochemical, and UAV Multispectral Data, Drones 8 (2024). https://doi.org/10.3390/drones8070287. [31].
[31] S. Yang, L. Li, S. Fei, M. Yang, Z. Tao, Y. Meng, Y. Xiao, Wheat Yield Prediction Using Machine Learning Method Based on UAV Remote Sensing Data, Drones 8 (2024). https://doi.org/10.3390/drones8070284.
[32] A.S. Kharal, S.A. Mahar, M.I. Mushtaque, A. Magsi, J.A. Mahar, A Model for Wheat Yield Prediction to Reduce the Effect of Climate Change Using Support Vector Regression, VFAST Trans. Softw. Eng. 12 (2024) 192–212. https://doi.org/10.21015/vtse.v12i2.1855.
[33] J. Cao, H. Wang, J. Li, Q. Tian, D. Niyogi, Improving the Forecasting of Winter Wheat Yields in Northern China with Machine Learning–Dynamical Hybrid Subseasonal-to-Seasonal Ensemble Prediction, Remote Sens. 14 (2022). https://doi.org/10.3390/rs14071707.
[34] L. Breiman, Random Forests. *Machine Learning*, 45 (2001) 5–32. https://doi.org/10.1023/A:1010933404324.
[35] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System. https://doi.org/10.1145/2939672.2939785.
[36] H. Burdett, C. Wellen, Statistical and machine learning methods for crop yield prediction in the context of precision agriculture, Precision Agric. 23 (2022) 1553–1574. https://doi.org/10.1007/s11119-022-09897-0.
[37] N. Chergui, Durum wheat yield forecasting using machine learning, Artif. Intell. Agric. 6 (2022) 156–166. https://doi.org/10.1016/j.aiia.2022.09.003.
[38] S. Gawdiya, D. Kumar, B. Ahmed, R.K. Sharma, P. Das, M. Choudhary, M.A. Mattar, Field scale wheat yield prediction using ensemble machine learning techniques, Smart Agric. Technol. 9 (2024) 100543. https://doi.org/10.1016/j.atech.2024.100543.
[39] A. M. S. Kheir, A. Govind, V. Nangia, M. Devkota, A. Elnashar, M. E. D. Omar, and T. Feike, Developing automated machine learning approach for fast and robust crop yield prediction using a fusion of remote sensing, soil, and weather dataset. *Environmental Research Communications*, IOP Publishing, 6 (2024) 041005. https://doi.org/10.1088/2515-7620/ad2d02.
[40] K.S. Aravind, A. Vashisth, P. Krishanan, B.Das, Wheat yield prediction based on weather parameters using multiple linear, neural network and penalised regression models, J. Agrometeorol. 24 (2022) 18–25. https://doi.org/10.54386/jam.v24i1.1002.
[41] S.S. Adilova, D.E. Qulmamatova, S.K. Baboev, T.A. Bozorov, A.I. Morgunov, Multivariate Cluster and Principle Component Analyses of Selected Yield Traits in Uzbek Bread Wheat Cultivars, Am. J. Plant Sci. 11 (06) (2020) 10. https://doi.org/10.4236/ajps.2020.116066.
[42] B. Czernecki, J. Nowosad, K. Jabłońska, Machine learning modeling of plant phenology based on coupling satellite and gridded meteorological dataset, Int. J. Biometeorol. 62 (2018) 1297–1309. https://doi.org/10.1007/s00484-018-1534-2.
[43] G. Badagliacca, G. Messina, S. Praticò, E. Lo Presti, G. Preiti, M. Monti, G. Modica, Multispectral Vegetation Indices and Machine Learning Approaches for Durum Wheat (Triticum durum Desf.) Yield Prediction across Different Varieties, Agric. Eng. 5 (2023) 2032–2048. https://doi.org/10.3390/agriengineering5040125.
[44] Y.Ö. Durgun, A. Gobin, G. Duveiller, B. Tychon, A study on trade-offs between spatial resolution and temporal sampling density for wheat yield estimation using both thermal and calendar time, Int. J. Appl. Earth Observ. Geoinf. 86 (2020) 101988. https://doi.org/10.1016/j.jag.2019.101988.
[45] J. Shen, F.H. Evans, The Potential of Landsat NDVI Sequences to Explain Wheat Yield Variation in Fields in Western Australia, Remote Sens. 13 (2021), https://doi.org/10.3390/rs13112202.
[46] M. Aranguren, A. Castellón, A. Aizpurua, Wheat Yield Estimation with NDVI Values Using a Proximal Sensing Tool, Remote Sens. 12 (2020), https://doi.org/10.3390/rs12172749.
[47] C. Trentin, C. Bredemeier, A.L. Vian, M.A. Drum, F.L. Santos, Biomass production and wheat grain yield and its relationship with NDVI as a function of nitrogen availability, Revista Brasileira de Ciências Agrárias - Brazilian J. Agric. Sci. 16 (2021) 1–7, https://doi.org/10.5039/agraria.v16i4a34.
[48] M.M. Kostić, N. Ljubičić, V. Aćin, M. Mirosavljević, M. Budjen, M. Rajković, N. Dedović, An active-optical reflectance sensor in-field testing for the prediction of winter wheat harvest metrics, J. Agric. Eng. PAGEPress Public. (2024), https://doi.org/10.4081/jae.2024.1559.
[49] E. Roma, P. Catania, M. Vallone, S. Orlando, Unmanned aerial vehicle and proximal sensing of vegetation indices in olive tree (Olea europaea), J. Agric. Eng. PAGEPress Public. 54 (2023), https://doi.org/10.4081/jae.2023.1536.
[50] S. Fei, M.A. Hassan, Y. Xiao, X. Su, Z. Chen, Q. Cheng, F. Duan, R. Chen, Y. Ma, UAV-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat, Precision Agric. 24 (2023) 187–212, https://doi.org/10.1007/s11119-022-09938-8.