Smart Agricultural Technology 13 (2026) 101936

Contents lists available at ScienceDirect

Smart Agricultural Technology

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

Machine learning approaches for wheat yield prediction integrating biophysical modeling and remote sensing: Effects of sample size, dimensionality, and transferability

Davoud Ashourloo [a,*], Jason Brader [b], Yan Zhao [a], Ruizhu Jiang [a], Miao Zhang [c,*], Scott Chapman [a], Graeme Hammer [a], Andries B Potgieter [a]

[a] The University of Queensland, Queensland Alliance for Agriculture and Food Innovation, Brisbane, QLD, 4072, Australia
[b] Department of Primary Industries, Toowoomba, QLD, 4350, Australia
[c] Chinese Academy of Sciences, Beijing, 9718, China

ARTICLE INFO

Keywords:
Machine learning techniques
Remote sensing
Biophysical variable
Sample size
Dimensionality
Transferability

ABSTRACT

Prediction of wheat yield plays a crucial role in mitigating the impacts of climate extremes and ensuring food security. While vegetation indices (VIs) derived from remote sensing have been widely used for this purpose, integrating biophysical variables, such as crop stress, simulated using a biophysical model has the potential to enhance model robustness. However, the role of integrating biophysical stress information in machine learning based yield prediction, particularly in terms of sample size efficiency, dimensionality reduction, and spatial transferability, remains insufficiently examined. In this study, we address this gap by systematically evaluating the impact of incorporating a Stress Index (SI) using three widely applied machine learning (ML) models Random Forest (RF), Gaussian Process Regression (GPR), and Extreme Gradient Boosting (XGBoost) to predict wheat yield across five Australian states. These models leveraged both Sentinel-2-derived VIs at the peak of greenness and a biophysical factor: the SI.

The integration of the SI led to clear improvements in model performance. For XGBoost, the coefficient of determination (R^2) increased from 0.65 to 0.73, and the RMSE decreased from 0.67 to 0.58. GPR showed an increase in R^2 from 0.66 to 0.70 and a reduction in RMSE from 0.76 to 0.60. RF reached an R^2 of 0.68 and an RMSE of 0.64 after including SI. Additionally, SI contributed to dimensionality reduction, lowering the number of required VIs from 11 to 6 without loss of accuracy. The inclusion of SI also improved spatial transferability. When tested in Western Australia, cross-regional RMSE decreased by 0.09 for XGBoost, 0.16 for GPR, and 0.13 for RF. These results highlight that integrating biophysical crop responses, such as SI, into ML model calibration can enhance the reliability and robustness of yield prediction using remote sensing data.

1. Introduction

Food security has become a critical concern for the growing global population [1–4]. Accurate crop yield predictions are essential for managers and planners to ensure food availability [5–7]. As a staple crop, wheat plays a vital role in supplying daily calories and proteins worldwide [8]. Reliable yield estimates are essential for researchers and policymakers, as they support informed decisions related to import and export strategies, the optimization of planting patterns, the allocation of subsidies during extreme weather events, market regulation, humanitarian response during famines, and supply management at both national and regional levels [5,9].

However, conventional yield prediction methods often fall short due to their dependence on labor-intensive ground measurements, making them impractical for large-scale applications [10,11]. To address these limitations, remote sensing techniques have emerged as a promising alternative, offering timely and scalable solutions for yield estimation [12].

Among the various remote sensing methods, optical satellite imagery has proven particularly effective in agricultural monitoring. In optical wavelengths (450–2500 nm) have been widely used due to the unique characteristics of plants in the visible, near-infrared (NIR), and short-

* Corresponding authors.
E-mail addresses: D.ashourloo@uq.edu.au (D. Ashourloo), zhangmiao@aircas.ac.cn (M. Zhang).

https://doi.org/10.1016/j.atech.2026.101936
Received 28 August 2025; Received in revised form 24 February 2026; Accepted 27 February 2026
Available online 28 February 2026
2772-3755/© 2026 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

wave infrared (SWIR) regions [13]. Sentinel-2 offers 13 spectral bands (10–60 m, 5-day revisit) widely used for crop monitoring [14–16]. VIs (VIs), derived from Sentinel-2, are vital for assessing crop health, chlorophyll content, and canopy structure factors closely linked to yield potential [13,17]. Sentinel-2′s diverse spectral bands support many VIs that are widely used as key inputs for machine-learning–based crop monitoring and yield prediction [18,19]. When integrated with ML algorithms, these VIs significantly enhance the predictive power of the models [20,21]. The combination of Sentinel-2 data and ML techniques offers a powerful approach to enhancing yield prediction reliability and precision [22–24]. These techniques excel in managing high-dimensional data and are widely applied for retrieving biophysical and biochemical parameters, which are essential for predicting crop yields [13,22]. ML techniques such as Extreme Gradient Boosted (XGBoost), Gaussian process regression (GPR), and Random Forest (RF) have demonstrated significant success in quantitative remote sensing due to their ability to handle complex relationships between dependent and independent variables [25–27].

Despite the advancements in using ML techniques for yield prediction, the accuracy and robustness of these models are influenced by various factors, including model selection, type and the size of input data [28,29]. Regarding the input data for the ML models, the use of meteorological data alongside remote sensing data, including climate variables such as rainfall and temperature, as well as biophysical factors that characterize crop stress conditions particularly water and temperature stress quantified by the SI is crucial for achieving accurate yield predictions [1,30,31]. Incorporating biophysical factors such as the SI is particularly valuable, as it captures the effects of environmental stressors, especially those driven by weather conditions like extreme temperatures, water deficits, and solar radiation on crop development. The SI is derived from meteorological data collected at ground-based weather stations and is calculated using key climate variables such as daily air temperature, solar radiation, relative humidity, and wind speed. These variables are integrated to estimate evapotranspiration and water availability during specific phenological stages, enabling the SI to quantify crop water stress dynamically throughout the growing season [32]. According to Genotype × Environment × Management ($G \times E \times M$) concept, crop traits are the result of dynamic interactions between genetic potential, environmental conditions, and management practices. Environmental drivers such as radiation, temperature, soil properties, and water availability influence canopy structure and biochemical processes, while management interventions, including fertilization and irrigation, modify these responses. Spectral reflectance therefore represents a measurable manifestation of $G \times E \times M$ interactions. Reflective bands primarily describe structural canopy properties and background radiometric effects, whereas absorptive bands capture biochemical absorption features associated with pigments, nitrogen, and water content. By explicitly separating and systematically combining these two spectral domains within a physically constrained mathematical framework, the proposed approach enhances sensitivity to functional crop traits [33–35]. Therefore, GEM offers a strong foundation for integrating process-based stress information into data-driven ML models.

Within the GEM framework, SI effectively represents the environmental component, offering a quantifiable measure of how adverse conditions impact crop physiology and yield potential. While VIs primarily provide spatial snapshots of canopy structure and greenness, SI reflects cumulative stress over time, enabling the detection of harmful environmental effects that may not be immediately visible in satellite imagery. Several studies have demonstrated that combining a weather-derived SI or agrometeorological variables with VIs improves crop monitoring and yield estimation by linking environmental constraints with canopy spectral responses [36,37]. These studies show that stress-related variables provide complementary information to VIs, particularly under climate variability where spectral signals alone may not fully capture the effects of crop stress. Chen et al. [38] integrated meteorology-driven stress indices with remotely sensed information to estimate crop yields across the Australian wheatbelt, highlighting the importance of stress indicators for representing spatial agro-environmental heterogeneity [38]. Similarly, Kamir et al. [39] combined satellite vegetation index time series, primarily NDVI, with climate variables such as rainfall and temperature derived from meteorological records to improve wheat yield estimation across Australia. However, climate variables such as rainfall or temperature alone represent individual drivers and do not fully describe crop stress dynamics [39]. In contrast, crop model–derived SI integrate multiple meteorological drivers, including temperature, radiation, humidity, and wind, through crop water balance and evapotranspiration processes, allowing stress to be quantified dynamically across key growth stages [32,34]. As a result, combining SI with VIs provides a more complete and process-based representation of crop growth conditions than using VIs and simple climate variables alone. In this way, SI explicitly accounts for water-limiting environments and their associated stress effects on final yield, leading to improved prediction performance across diverse environmental conditions [40]. Therefore, from an application perspective, it is both timely and important to investigate the effectiveness of integrating the SI with VIs, particularly in relation to three key aspects.

The size of the training dataset significantly limits the reliability of ML models, particularly for non-parametric algorithms that require large sample sizes to achieve stable performance. However, collecting such extensive training data for ML applications in agricultural remote sensing is often costly and resource-intensive [41]. Therefore, it is important to evaluate the reliability and accuracy of input data when using limited training sample sizes. In addition, incorporating independent datasets from different regions or years can help improve model robustness and generalization, allowing reliable predictions to be achieved with fewer training samples and at lower data collection cost [29]. SI is derived from climate station data and provides stress-related information that is independent of spectral remote sensing observations. As a process-based indicator, SI represents environmental constraints on crop growth that are not directly observable from canopy reflectance alone [32]. The inclusion of such independent and physically meaningful information can improve model robustness and data efficiency, which may help achieve reliable predictions with fewer training samples when combined with remote sensing inputs. As a result, we assumed that the inclusion of SI can reduce the required sample size for accurate yield prediction by improving model stability and generalization.

Moreover, spatial transfer learning plays a vital role in ML methods, enabling models to adapt to new regions by leveraging knowledge from diverse environments [42,43]. This is particularly important for yield prediction, where environmental conditions can vary strongly across locations. Input features such as the SI can support this adaptability because they represent climate-driven stress processes that are less dependent on local spectral characteristics. By capturing stress-related variability in a process-based manner, SI can help ML models generalize better and reduce prediction errors when applied to independent regions or datasets.

Another critical consideration in developing robust ML models is reducing the number of input features without compromising prediction accuracy. At the same time, dimension reduction is essential for optimizing model performance by eliminating redundant or less informative features, reducing computational costs, and improving interpretability. More importantly, it helps mitigate overfitting, where models memorize patterns in the training data rather than learning generalizable relationships [44]. By integrating only the most relevant VIs and SI, models can achieve greater robustness, ensuring stable and reliable predictions even when trained on limited data. This balance between efficiency and accuracy is essential for developing scalable and adaptable machine ML–based yield prediction models. Although high-resolution Earth observation data have proven useful for crop yield estimation, integrating functional and physiological crop stress responses into ML models adds further value by improving model

efficiency and robustness [45,46]. This integration is particularly important for reducing model complexity and computational time during model calibration and large-scale application. Hence, we investigate the value of incorporating biophysically derived

SI, grounded in functional crop process understanding, alongside conventional VIs in ML models for predicting wheat yield at the field scale across Australia. We evaluated the contribution of the SI and VIs to model performance through three key objectives:

(I) ML techniques robustness by examining how the inclusion of SI improves performance under reduced training sample sizes, demonstrating its contribution to data-efficient learning and reliable yield prediction in data-limited conditions;
(II) spatial transferability of ML techniques by assessing how SI enhances performance across heterogeneous environments, highlighting its role in improving model generalisation and robustness when applied beyond the training area; and
(III) ML techniques input dimensionality reduction by investigating how SI enables a reduction in input features through efficient capture of stress-related information, thereby simplifying model structure, reducing overfitting, and improving the interpretability of ML models.

This framework demonstrates how integrating remote sensing and biophysical indices enhances yield prediction by improving data efficiency, model simplicity, and spatial generalization within ML techniques.

2. Material and methods

Fig. 1 illustrates the overall workflow and methodology used in this study for wheat yield prediction. A time series of Sentinel-2 imagery was processed to determine the date of peak greenness by identifying the maximum NDVI value during the growing season. VIs (VIs) were then computed based on the corresponding date for each field. In parallel, ground-truth yield data collected over five years at the field scale, along with meteorological data from nearby weather stations, were incorporated into the analysis. SI was calculated using the Oz-Wheat model for each climate station across Australia [32]. As Fig. 1 shows, the computation of the SI involves four steps: (1) estimating daily crop water requirement (CWR) inputs of these steps are Temperature, Solar Radiation, Wind Speed, Relative Humidity and Crop Coefficient (Kc), (2) updating plant available water (PAW) based on rainfall and soil properties, (3) estimating actual evapotranspiration (Ea) depending on water availability (4) calculating the daily SI. In this study, the SI during the flowering period corresponding to the peak of greenness was used as an input variable in the model.

The computed VIs and SI served as input features for three ML models: Extreme Gradient Boosting (XGBoost), Gaussian Process Regression (GPR), and Random Forest (RF). These models were evaluated through a multi-step analytical framework, including correlation analysis, model performance under varying sample sizes, spatial transfer learning assessments, dimensionality reduction, and bias–variance decomposition. Finally, the added value of incorporating SI was assessed by comparing model accuracy and robustness with and without SI under different scenarios. This workflow provides a systematic and comprehensive approach to wheat yield prediction by integrating remote sensing data, meteorological stress indicators, and advanced ML techniques.

2.1. Studying area, field data and computing stress index

The study area encompasses major wheat-producing regions across five Australian states, covering vast agricultural zones that span diverse climatic conditions. Variability in rainfall and temperature contributes to substantial differences in wheat yield and grain quality across these regions [47]. Australia is one of the world's largest wheat exporters, with the wheat industry playing a vital role in the national economy. Wheat cultivation in Australia predominantly relies on rainfed systems, making it particularly sensitive to fluctuations in rainfall patterns especially in semi-arid areas [48]. The growing wheat season typically begins in May, with planting occurring through July. Harvesting usually takes place between late October and early January, depending on the region [49,50].

In this study, SI was computed using an established optimization-based calibration approach, which estimates the balance between water supply and demand throughout the wheat growth cycle [51,52]. It is defined as:

$$SI = \frac{E_a}{E_t} \quad (1)$$

where E_a represents actual evapotranspiration, and E_t denotes potential

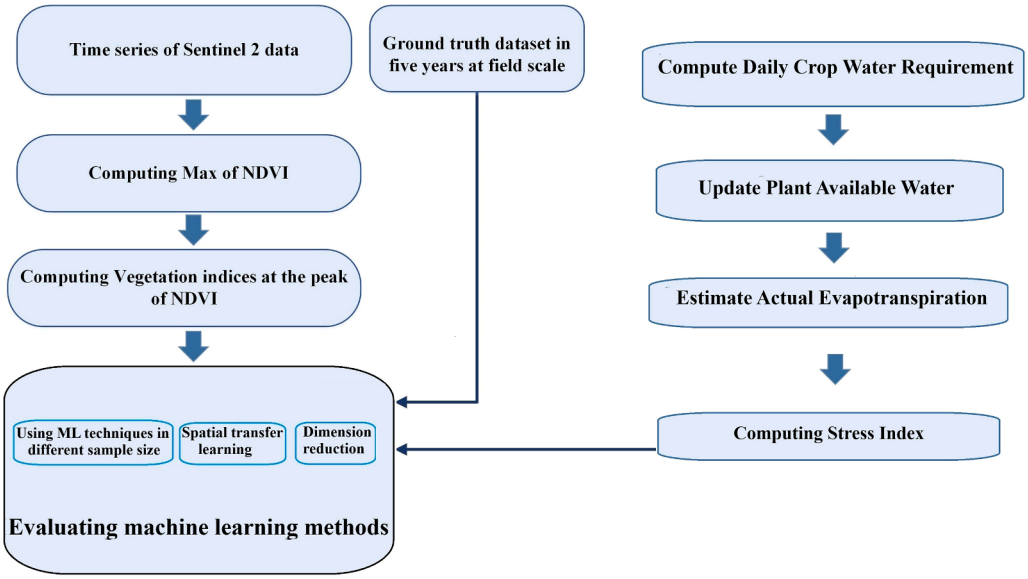

Fig. 1. Workflow of current study to estimate wheat yield at peak of greenness using ML techniques. SI was simulated using the Oz-Wheat model for each climate station across Australia [32].

evapotranspiration. This ratio serves as an indicator of crop water stress and corresponds well with the physiological responses of wheat under water deficit conditions. The Oz-Wheat model simulates this index on a daily scale, averaging it over an optimized critical growth window around the flowering stage. SI values range from 0 to 1, where a value of 1 ($E_a = E_t$) signifies no water stress, while values below 1 indicate varying degrees of water limitation affecting the wheat crop [51,52].

To support wheat yield prediction, field-level harvester data were collected over five years from 375 wheat fields across Australia. This dataset includes 65 fields in 2016, 45 in 2017, 85 in 2021, 143 in 2022, and 46 in 2023. Yield data were collected within known field boundaries located in New South Wales (NSW), Victoria (VIC), Western Australia (WA), Queensland (QLD), and South Australia (SA). All combine harvesters were equipped with GPS, and the average dry yield reported in tons per hectare (t/ha) was calculated at field scale. Fig. 2 illustrates the distribution of field data across five Australian states used for wheat yield prediction, along with a zoomed-in view of the WA region.

2.2. Data

In this study, cloud-free Sentinel-2 Level-2A Bottom-of-Atmosphere (BOA) surface reflectance products from Sentinel-2A and Sentinel-2B, captured around the peak of greenness, were used. These images were downloaded from the Copernicus Open Access Hub. We utilized spectral bands with a resolution of 10 or 20 m, and all data were resampled to 20 m using the nearest neighbor method. In the following, the average reflectance value of each spectral band was computed for each field, and these values in various bands were used to calculate VIs.

2.3. Machine learning techniques

2.3.1. Random forest regression

Random Forest improves bagging by reducing the overall variance of the model. It does this by lowering the similarity between individual trees. To achieve this, each tree is built using a random selection of features at every split. The prediction of model $\widehat{f}_{rf}^{B}$ is eventually as a mean of the all trees grown [53].

$$\widehat{f}_{rf}^{B} = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b) \tag{2}$$

where B, $T(x;\Theta_b)$, x and Θ_b are the total number of trees, b-th tree, the target point that requires making prediction and random vector generated for b-th tree independent of previous vectors, respectively. To build each tree, a random sample of features is chosen at every split. The data not used in building the tree are called out-of-bag (OOB) observations. The average error from these OOB observations gives an estimate of how well the model performs on unseen data [54].

Bayesian optimization is used to find the best settings for the Random Forest model, like the smallest leaf size and how many features to consider at each split [55]. The model, built with 500 trees, chooses these settings by reducing the quantile regression error.

2.3.2. Gaussian process regression

Gaussian Process Regression (GPR) is a kernel-based method used for nonlinear regression. It's a non-parametric model that starts with a prior Gaussian process based on the training data and updates it to form a posterior Gaussian process [56,57]. Given the training inputs $(\{\boldsymbol{x}_i^{(b)}\}_{i=1}^{N})$ and target variables $(\{\boldsymbol{y}_i\}_{i=1}^{N})$, GPR captures the relationships between data points by modeling their correlations, which helps it better represent the patterns in the data [58].

$$\hat{\boldsymbol{y}} = f(\boldsymbol{X}) = \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}_i.\boldsymbol{x}_t) + \omega_0 \tag{3}$$

where, ω_0 is a bias term, $\alpha_i \in \mathbb{R}$ is the weight of the training inputs, and function K evaluates the similarities between the training inputs $\boldsymbol{x}_i^{(b)}$ and

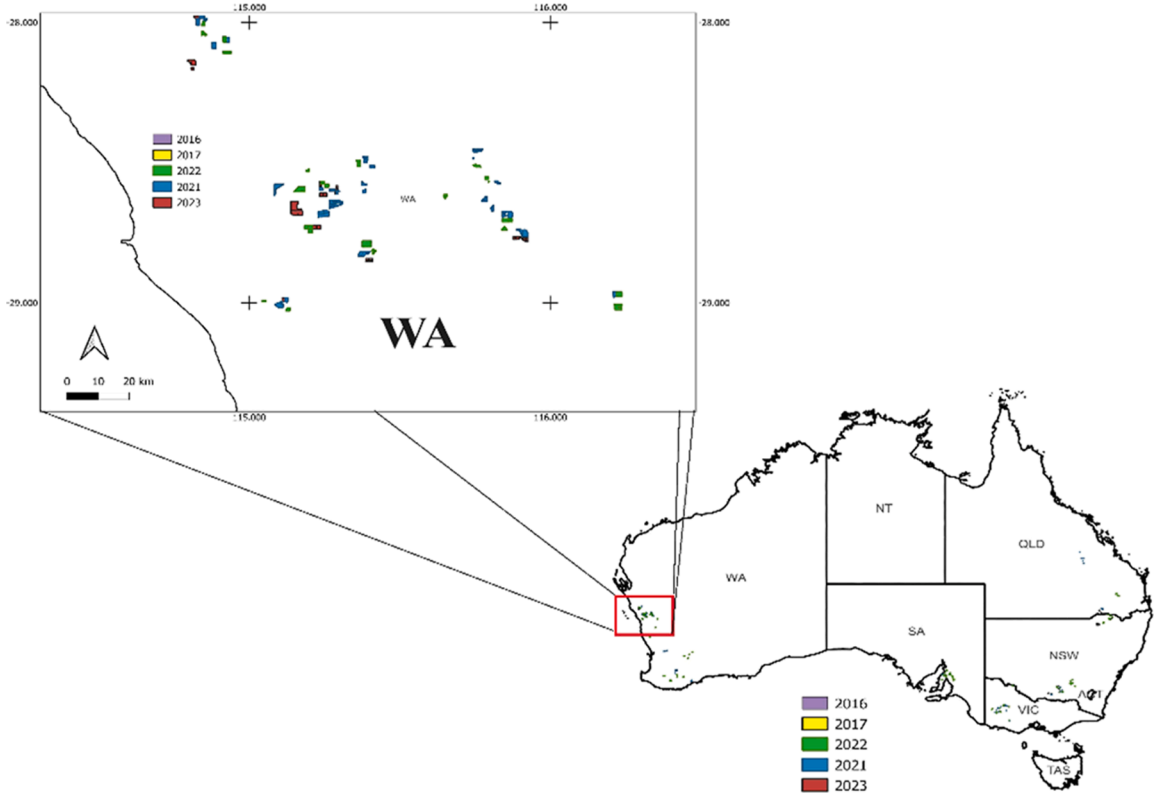

Fig. 2. Distribution of field data in Australia to predict wheat yield for five years including 2016,2027,2021,2022 and 2023.

the test set $\boldsymbol{x}_t$.

According to the literature, changes in the initial parameters do not affect the final results of GPR [59]. In this study, the length-scale—an important hyperparameter of the squared-exponential covariance function—is considered crucial for model performance ℓ, the signal variance σ_f and the noise variance σ_n are set to 1, 1 and 0.1, respectively. We utilized freely available GPR software for our yield prediction at http://www.gaussianprocess.org/gpml/ [60].

2.3.3. XGBoosted regression trees

Boosting is a powerful statistical learning method based on the principle that combining multiple weak learners can produce a strong predictive model [53]. When integrated with regression trees, it forms Boosted Regression Trees (BRT), which enhance predictive performance by sequentially refining weak learners [61]. BRTs can model complex nonlinear relationships, highlight feature importance, handle highly correlated predictors, and operate without the need for outlier removal or data transformation [62].

Extreme Gradient Boosting (XGBoost) is an advanced and scalable implementation of BRTs that boosts performance through regularization, helping prevent overfitting and improving generalization [63]. XGBoost builds the model by adding trees iteratively to minimize a loss function, thus enhancing predictive accuracy over time [63]. The general form of the XGBoost model is:

$$\widehat{f}^{A}_{BRT} = \sum_{m=1}^{M} T(x;\Theta_m) \tag{4}$$

Where $T(x;\Theta_m)$ is the m-th tree, and Θ_m represents the set of parameters for this tree. The model is updated at each step by minimizing the following cost function:

$$\widehat{\Theta}_m = arg\ \min_{\Theta_m} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x;\Theta_m)) \tag{5}$$

where, $f_{m-1}(x)$ is the current model, $\Theta_m = \left\{R_{jm}, \gamma_{jm}\right\}_1^{J_m}$ as a set of the region set R_{jm} and constants γ_{jm} in m-th tree, J is a meta-parameter, R_{jm} refers to j-th disjoint non-overlapping region in m-th tree, y_i is the response associated with x_i, and L is the loss function. XGBoost controls model complexity using hyperparameters such as maximum tree depth, learning rate (shrinkage), and the number of boosting rounds. It also handles missing data effectively, maintaining strong performance even with incomplete inputs. In this study, XGBoost was implemented using Bayesian optimization to fine-tune key hyperparameters including learning rate, tree depth, and regularization terms to improve performance and reduce overfitting [63].

2.4. Computing VIs based on Sentinel 2 spectral bands and feature selection

The VIs listed in Table 1 were selected to represent a broad range of biophysical crop properties relevant to wheat yield formation. These indices capture complementary information related to canopy greenness and biomass (e.g., NDVI, SR, DVI), chlorophyll content and photosynthetic activity (e.g., CI, CIgreen, GNDVI, NDRE, RENDVI, EVI), and crop water and moisture status (e.g., NDWI, NDMI, NBR). Indices based on visible, red-edge, near-infrared (NIR), and shortwave infrared (SWIR) bands were included to ensure sensitivity to both structural and physiological crop responses. This diverse set of indices has been widely used in previous crop monitoring and yield prediction studies and provides a comprehensive feature space for subsequent feature selection and ML analysis.

Table 1
Spectral VIs were used to compute wheat yield.

Vegetation Index	Formula	Explanation	Reference
SR	(B8 / B4)	Simple Ratio (SR)	[64]
NDVI	(B8 - B4) / (B8 + B4)	Normalized Difference Vegetation Index	(Rouse [65])
EVI	2.5 (B8 - B4) / (B8 + 6 * B4 - 7.5 B2 + 1)	Enhanced Vegetation Index	[66]
NDGI	(B8 - B5) / (B8 + B5)	Normalized Difference Green Index	[67]
NDWI	(B8 - B11) / (B8 + B11)	Normalized Difference Water Index	[68]
CI	(B7 / B5) – 1	Chlorophyll Index	[69]
GNDVI	(B8 - B3) / (B8 + B3)	Green Normalized Difference Vegetation Index	[69]
CIgreen	(B8 / B3) – 1	Green Chlorophyll Index	[69]
NDRE	(B8 - B5) / (B8 + B5)	Normalized Difference Red Edge	[70]
PRI	(B3 - B4) / (B3 + B4)	Photochemical Reflectance Index	[71]
NBR	(B8 - B12) / (B8 + B12)	Normalized Burn Ratio	[72]
GRVI	B8 / B3	Green Ratio Vegetation Index	[73]
RENDVI	(B6 - B5) / (B6 + B5)	Red-Edge Normalized Difference Vegetation Index	[74]
DVI	B8 - B4	Difference Vegetation Index	[73]
NDMI	(B8 - B11) / (B8 + B11)	Normalized Difference Moisture Index	[75]
VARI	(B3 - B4) / (B3 + B4 - B2)	Visible Atmospherically Resistant Index.	[76]

2.5. Evaluating the effectiveness of SI on yield prediction

In order to evaluate the impact of the SI on ML methods for yield prediction, the following steps were implemented:

First, the correlation between SI, VIs (VIs), and yield was assessed by computing a correlation matrix to evaluate the strength of relationships among different indices. The coefficient of determination (R^2) was used to quantify these relationships, providing insights into how SI complements VIs in capturing yield variability.

Next, the effect of SI across different sample sizes was evaluated by training ML models with varying dataset sizes, ranging from small (30 samples) to larger datasets (260 samples). The Random Forest (RF), Gaussian Process Regression (GPR), and XGBoost models were trained using only VIs and then with a combination of VIs and SI. Model performance was assessed based on R^2 (coefficient of determination) and RMSE (Root Mean Square Error) to measure the predictive power and error reduction achieved by incorporating SI. This approach ensures a systematic evaluation of SI's role in improving yield prediction across different data availability conditions.

To evaluate the spatial transferability of the model with SI, ML models were trained using data from multiple regions and tested in an independent region, Western Australia (WA). This spatial transfer learning approach tested whether SI improves model generalization when applied to different geographic locations. The evaluation was carried out by comparing observed and predicted yields using scatter plots, with and without SI, to assess its effect on prediction accuracy. In addition, to evaluate the impact of the SI on dimensionality reduction, a Sequential Forward Feature Selection (SFS) method was applied using three ML models: Random Forest (RF), Gaussian Process Regression (GPR), and XGBoost. In this approach, models were first trained using only VIs (VIs), where features were added one by one based on their contribution to reducing the Root Mean Square Error (RMSE). This process was repeated with SI introduced as the initial feature, followed by the stepwise addition of VIs. The RMSE was recorded at each step to determine the optimal number of features required for accurate yield prediction. This method enabled a clear comparison of model performance and complexity, revealing how the inclusion of SI accelerated the

reduction in RMSE and decreased the number of required features.

This methodological approach ensures a comprehensive evaluation of SI's role in enhancing yield prediction under varying conditions.

2.6. Validation

Yield predictions in this study were first generated at the pixel level (20 m resolution) using VIs (VIs) and the biophysical SI, with values expressed in tons per hectare (T/ha). These pixel-level estimates were then aggregated by averaging all predicted values within each field boundary to produce a single yield estimate per field. Similarly, the observed yield data derived from harvester measurements were averaged across each field and reported in T/ha. This approach ensured a consistent comparison at the field scale between model predictions and observed yields.

The coefficient of determination (R^2) (Eq. (8)) and the root mean square error (RMSE) (Eq. (6)) are employed to evaluate the performance of the regression methods to estimate water content. Additionally, the mean absolute error (MAE), was employed to describe the average model-performance error (Eq. (7)).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i-1}^{N}(y_i - o_i)^2} \tag{6}$$

$$MAE = \frac{1}{N}\sum_{i-1}^{N}|y_i - o_i| \tag{7}$$

Where, y_i and $o_i$are the predicted and the observed samples, respectively.

$$R^2 = \frac{\left(\sum_{i=1}^{N}(o_i - \overline{o}_i)(y_i - \overline{y}_i)\right)^2}{\sum_{i=1}^{N}(o_i - \overline{o}_i)^2\sum_{i=1}^{N}(y_i - \overline{y}_i)^2} \tag{8}$$

Here, $\overline{y}_i$ and $\overline{o}_i$are the average of the predicted and the observed samples, respectively.

3. Results

3.1. Correlation matrix of indices with Yield

Fig. 3 shows the R^2 correlation matrix of VIs (VIs), SI, and Yield at field scale. The highest R^2 values with Yield are observed for NDMI (0.57), followed by NDWI (0.56), DIVI (0.57), and NDLI (0.57). The lowest R^2 is for NDRE (0.36). The internal R^2 correlations among the VIs (excluding self-correlations) range from 0.05 (VARI with SI) to 0.98 (NDMI with NDLI), indicating generally strong internal relationships among most VIs. Regarding the correlation with SI, VIs displays a relatively low R^2 range from 0.11 (VARI with SI) to 0.24 (CI with SI). The SI itself demonstrates a moderate R^2 with Yield (0.42). Overall, while VIs exhibit low R^2 correlation with SI, both VIs and SI show moderate R^2 correlations with Yield. This suggests that although SI captures stress-related characteristics distinct from VIs, both types of indices contribute to explaining yield variability.

3.2. Evaluating ML techniques in various sample sizes

The results show that as the sample size increases (Table 2), the R^2 values for all three methods (GPR, XGBoost, and RF) consistently improve when using only VIs, indicating that the models are better able to explain the variance in the yield data with larger datasets. For smaller sample sizes (e.g., 30), the R^2 values are lower across all methods, but as the sample size reaches 260, the R^2 improves significantly for each method, demonstrating that the predictive power of the models increases with more data. Specifically, the R^2 for GPR reaches 0.69 with 260 samples, while XGBoost and RF reach 0.63 and 0.63, respectively.

The introduction of SI has a clear positive effect, especially for smaller sample sizes. The R^2 values with SI are generally higher than those without SI across all sample sizes, reflecting that incorporating SI

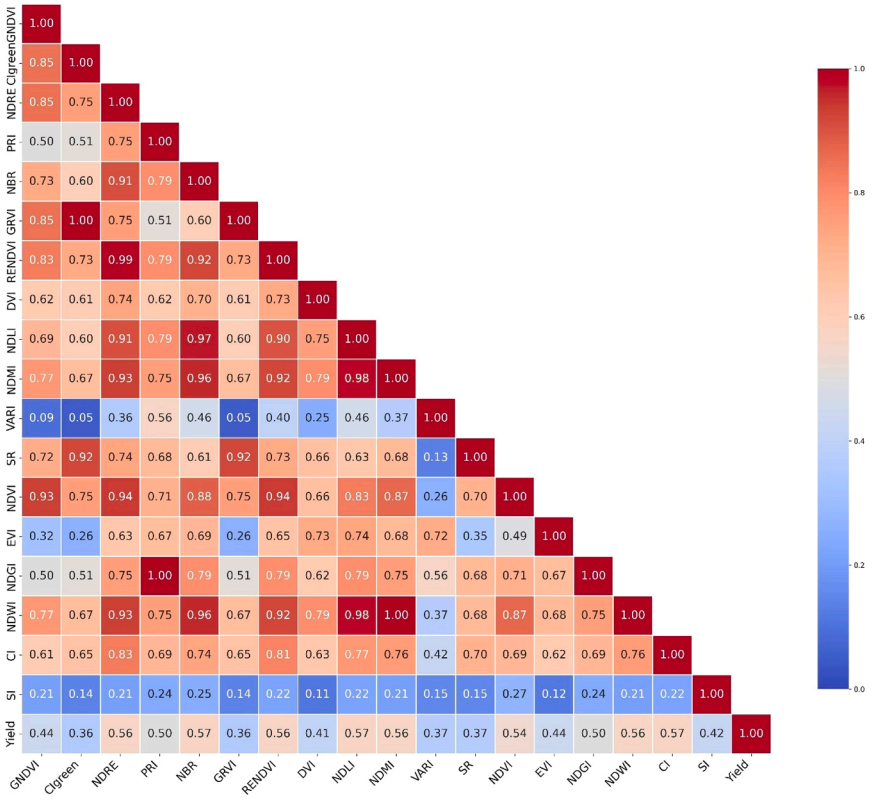

Fig. 3. The correlation matrix of VIs, stress index and yield.

Table 2
Computed RMSE and R2 for GPR, XGBoost and RF in various sample sizes; and increase in R^2 (parentage) when Stress Index (SI) is added.

ML Method	Input	Statistical Parameter	Sample size												
			30	50	70	90	110	120	140	160	180	200	220	240	260
GPR	VIs	R2	0.14	0.54	0.50	0.56	0.53	0.59	0.60	0.64	0.65	0.71	0.70	0.69	0.69
		RMSE	1.17	0.94	0.92	0.86	0.93	0.88	0.86	0.84	0.85	0.78	0.80	0.80	0.80
	VIs+ SI	R2	0.39	0.72	0.61	0.64	0.65	0.70	0.73	0.76	0.77	0.80	0.80	0.78	0.80
		RMSE	0.99	0.74	0.81	0.78	0.80	0.76	0.71	0.69	0.69	0.66	0.66	0.68	0.65
	Increase	*R^2 (%)*	**179**	**33**	**22**	**14**	**23**	**19**	**22**	**19**	**18**	**13**	**14**	**13**	**16**
XGBoost	VIs	R2	0.13	0.50	0.42	0.49	0.48	0.50	0.51	0.54	0.56	0.60	0.61	0.62	0.63
		RMSE	1.18	0.98	0.99	0.92	0.98	0.97	0.96	0.94	0.95	0.92	0.91	0.88	0.88
	VIs+ SI	R2	0.49	0.70	0.56	0.63	0.60	0.62	0.65	0.69	0.70	0.72	0.73	0.73	0.74
		RMSE	0.90	0.76	0.86	0.78	0.86	0.85	0.81	0.78	0.78	0.76	0.75	0.75	0.73
	Increase	*R^2 (%)*	**277**	**40**	**33**	**29**	**25**	**24**	**27**	**28**	**25**	**20**	**20**	**18**	**17**
RF	VIs	R2	0.20	0.45	0.43	0.52	0.50	0.52	0.52	0.55	0.57	0.61	0.61	0.62	0.63
		RMSE	1.13	0.93	0.98	0.90	0.97	0.95	0.94	0.93	0.94	0.91	0.90	0.88	0.88
	VIs+ SI	R2	0.44	0.65	0.55	0.63	0.61	0.62	0.66	0.70	0.71	0.74	0.74	0.74	0.75
		RMSE	0.95	0.82	0.87	0.79	0.85	0.84	0.80	0.77	0.77	0.74	0.74	0.74	0.72
	Increase	*R^2 (%)*	**120**	**44**	**28**	**21**	**22**	**19**	**27**	**27**	**25**	**21**	**21**	**19**	**19**

improves the model's ability to predict yield more accurately. For example, with only 30 samples, the R^2 for GPR increases from 0.14 (VIs only) to 0.39 (with SI), showing that SI adds substantial value when data is limited. Similar trends are observed for XGBoost and RF, where the inclusion of SI boosts the R^2 even with lower sample sizes, providing more reliable predictions earlier in the training process. The positive influence of the SI is clearly demonstrated in Table 2 by computing the percentage increase in R^2 when SI is added. The inclusion of SI consistently improves the R^2 values across all sample sizes for each ML method, confirming its overall effectiveness in enhancing yield prediction. While the impact is most significant at smaller sample sizes where data scarcity limits the models' ability to learn the benefits of incorporating SI remain evident even as the sample size increases. For instance, GPR shows a remarkable 179 % increase in R^2 at 30 samples, which gradually tapers to a still meaningful 16 % at 260 samples. Similarly, XGBoost and RF exhibit 277 % and 120 % improvements at 30 samples, respectively, with gains of 17 % and 19 % still observed at the largest sample size. These consistent improvements across the entire sample size range highlight SI's strong complementary value, particularly when training data is limited, while also reinforcing its continued contribution to prediction accuracy in more data-rich scenarios.

Regarding RMSE, as Table 2 and Fig. 4 shows there is a steady decrease as the sample size increases, indicating that the prediction error reduces with larger datasets. Without SI, the RMSE values are generally higher, but with the introduction of SI, the RMSE values drop, signaling more accurate predictions. For example, for GPR, the RMSE decreases from 1.17 (30 samples, VIs only) to 0.99 with SI, reflecting a reduction in the prediction error. Similarly, XGBoost and RF show improved RMSE values when SI is included, highlighting the positive impact of SI on reducing errors.

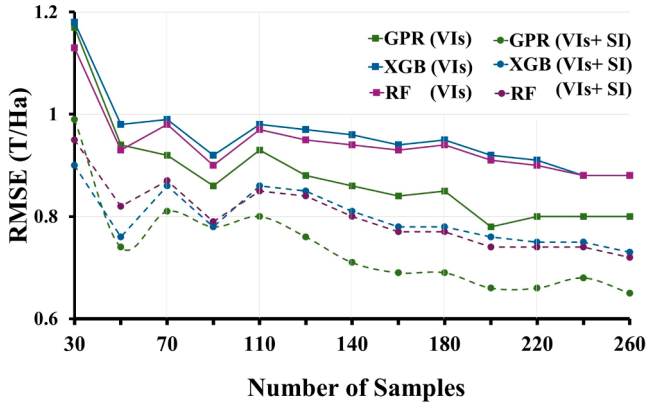

Fig. 4. The effect of sample size on RMSE of indices using GPR, XGBoost and RF.

When comparing the three methods, GPR generally provides higher R^2 values and lower RMSE values with SI compared to XGBoost and RF, making it more robust in yield prediction. XGBoost and RF also benefit from SI, but the performance improvement is slightly less pronounced compared to GPR. Overall, the results suggest that adding SI improves model accuracy and stability across methods, especially when dealing with smaller datasets. After establishing the influence of sample size on model performance, it is important to assess the models' overall robustness and reliability when utilizing the entire dataset. Therefore, to provide a comprehensive evaluation beyond sample size effects. To further assess the effect of the SI on model performance, an 80/20 cross-validation was applied and repeated over 100 iterations for each ML model: Random Forest (RF), Gaussian Process Regression (GPR), and XGBoost. This procedure enabled a comprehensive and consistent evaluation of predictive accuracy across repeated samples. The results are summarised in Fig. 5, which presents boxplots of three key performance metrics RMSE, R^2, MAE for each model, first using only VIs (VIs) (Fig. 5a, 5c, and 5e) and then with the inclusion of SI alongside the VIs (Fig. 5b, 5d, and 5f).

The inclusion of SI resulted in a consistent improvement in all three-performance metrics across the models. For RF, the RMSE range narrowed from 0.60 to 0.80 in Fig. 5a to 0.65–0.75 in Fig. 5b, the R^2 values increased from 0.60 to 0.80 to 0.65–0.85, and the MAE decreased from 0.50 to 0.70 to 0.55–0.65. In the GPR model, the RMSE range was reduced from 0.65 to 0.85 in Fig. 5c to 0.70–0.80 in Fig. 5d, R^2 increased from 0.65 to 0.80 to 0.70–0.85, and MAE improved from 0.52 to 0.70 to 0.55–0.65. Similarly, for XGBoost, the RMSE decreased from 0.68 to 0.85 in Fig. 5e to 0.70–0.80 in Fig. 5f, R^2 increased from 0.65 to 0.80 to 0.70–0.85, and MAE was reduced from 0.53 to 0.70 to 0.56–0.65.

While the overall variance in the boxplots remains relatively similar after adding SI, the central tendency clearly shifts, indicating that the models performed more accurately on average when SI was included during the 100 repeated runs. These improvements across all three models and performance metrics reinforce the important role of the SI in enhancing predictive performance. Incorporating SI alongside traditional VIs improves both the effectiveness and reliability of yield prediction models, making them more robust across repeated training and testing cycles.

3.3. *Spatial transfer learning of machine learning techniques*

Transfer learning is essential for yield prediction, enabling models trained in one region to be adapted for another with limited data. It improves accuracy by leveraging existing knowledge and reduces the need for extensive retraining. In 2023, 32 samples belonged to Western Australia (WA). Therefore, other areas were used to train ML models.

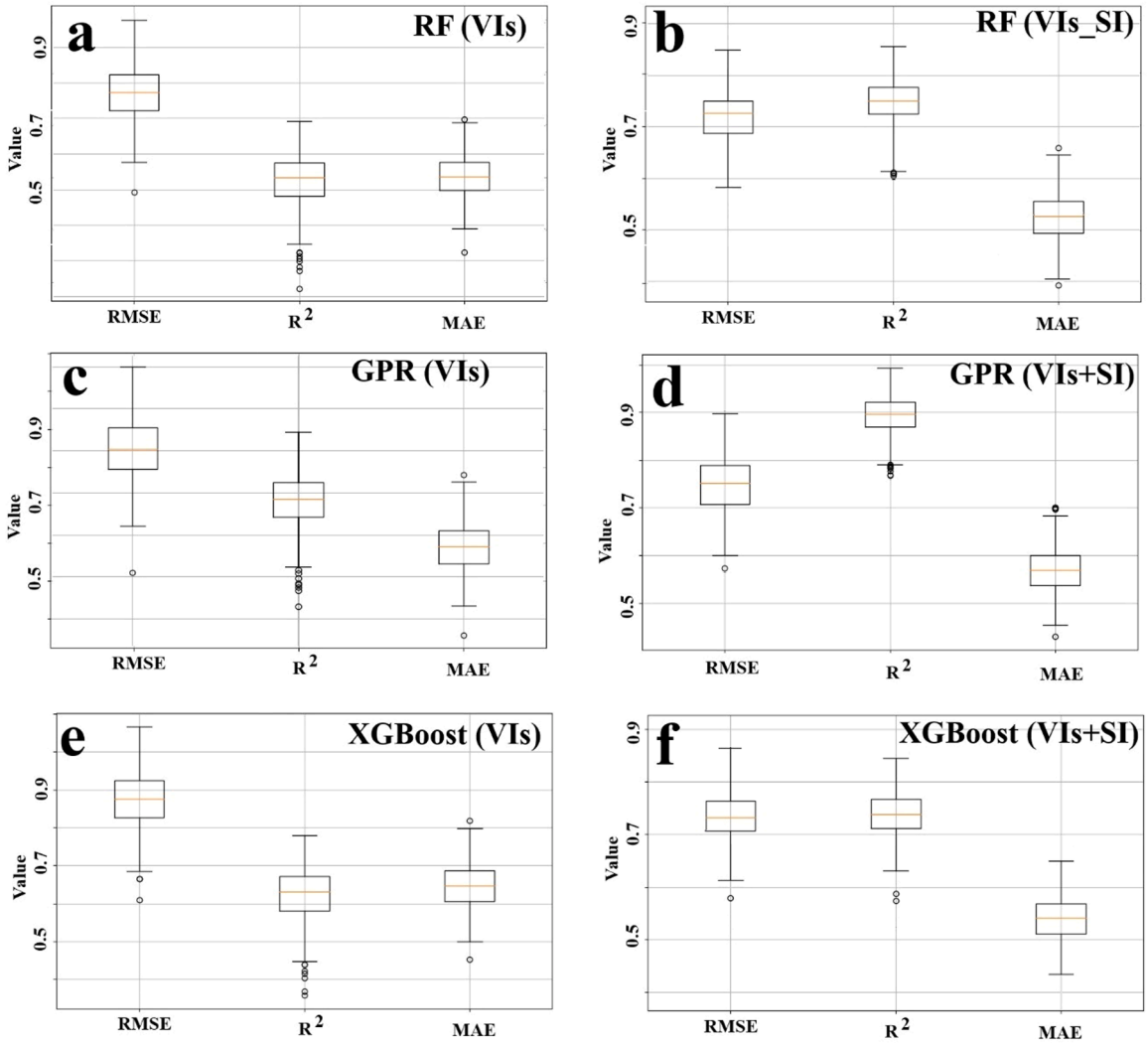

Fig. 5. RMSE, R2 and MAE in 100 repetitions for RF, GPR and XGBoost, first using only VIs, and then using both SI and VIs.

The models were then tested in WA, once with SI and once without it. Fig. 6 compares the relationship between the observed and estimated yield, first using only VIs (VIs) and then using both SI and VIs.

The performance of the XGBoost model is shown in Figs. 6a and 6b Using only VIs (Fig. 6a), the model achieves an R^2 of 0.65, explaining 65 % of the variance in yield, with an RMSE of 0.67. When SI is added (Fig. 6b), the R^2 increases to 0.73, and the RMSE drops to 0.58, highlighting significant improvements in prediction accuracy and consistency. The reduced scatter in the plot further confirms the positive impact of SI. For the Gaussian Process Regression (GPR) model, Figs. 6c and 6d show the effect of incorporating SI. Without SI (Fig. 6c), the model achieves an R^2 of 0.66 and an RMSE of 0.76, suggesting moderate predictive power with notable prediction error. However, when SI is added (Fig. 6d), the R^2 improves to 0.70, and the RMSE decreases to 0.60, indicating that SI contributes to a more accurate and reliable prediction. The tighter clustering of data points along the regression line further supports this improvement. Figs. 6e and 6f illustrate the performance of the Random Forest (RF) model in predicting yield. Using only VIs (Fig. 6e), the model demonstrates moderate predictive accuracy with an R^2 of 0.54, meaning 54 % of the variance in yield is explained by VIs. The RMSE of 0.77 reflects the average prediction error, indicating substantial deviations between predicted and actual yield values. However, when SI is incorporated (Fig. 6f), model performance improves noticeably. The R^2 increases to 0.68, indicating that 68 % of the variance is now explained, while the RMSE decreases to 0.64, reflecting reduced prediction error. The scatter points cluster more closely around the regression line, indicating improved accuracy when SI is included.

Overall, Fig. 6 clearly demonstrates that the inclusion of SI significantly enhances model performance across all techniques. Incorporating SI consistently leads to higher R^2 values, indicating better model fit, and lower RMSE values, reflecting reduced prediction errors. These improvements underscore the critical role of SI in reducing prediction uncertainty and increasing model robustness for yield prediction in diverse environments. Generally, the highest accuracy using only VIs (VIs) was achieved by the XGBoost model, with an R^2 of 0.65 and an RMSE of 0.67. After incorporating the SI, the highest accuracy was again observed in the XGBoost model, where the R^2 increased to 0.73 and the RMSE decreased to 0.58. Based on these results, R^2 and RMSE improved by approximately 12 % and 13 %, respectively

3.4. *Dimension reduction of machine learning techniques using SI*

The provided graphs in Fig. 7 highlight the significant role of the SI in improving dimension reduction for yield prediction. When using only VIs (VIs), the models demand minimum 11 or 12 VIs to reach the Root Mean Square Error (RMSE) of around 0.83, 0.81 and 0.8 for RF (Fig. 7a), GPR (Fig. 7b) and XGBoost (Fig. 7c) respectively. Also, there is gradual decrease in RMSE for all three models suggests that more features are necessary to achieve accurate predictions, which increases model complexity and the risk of overfitting.

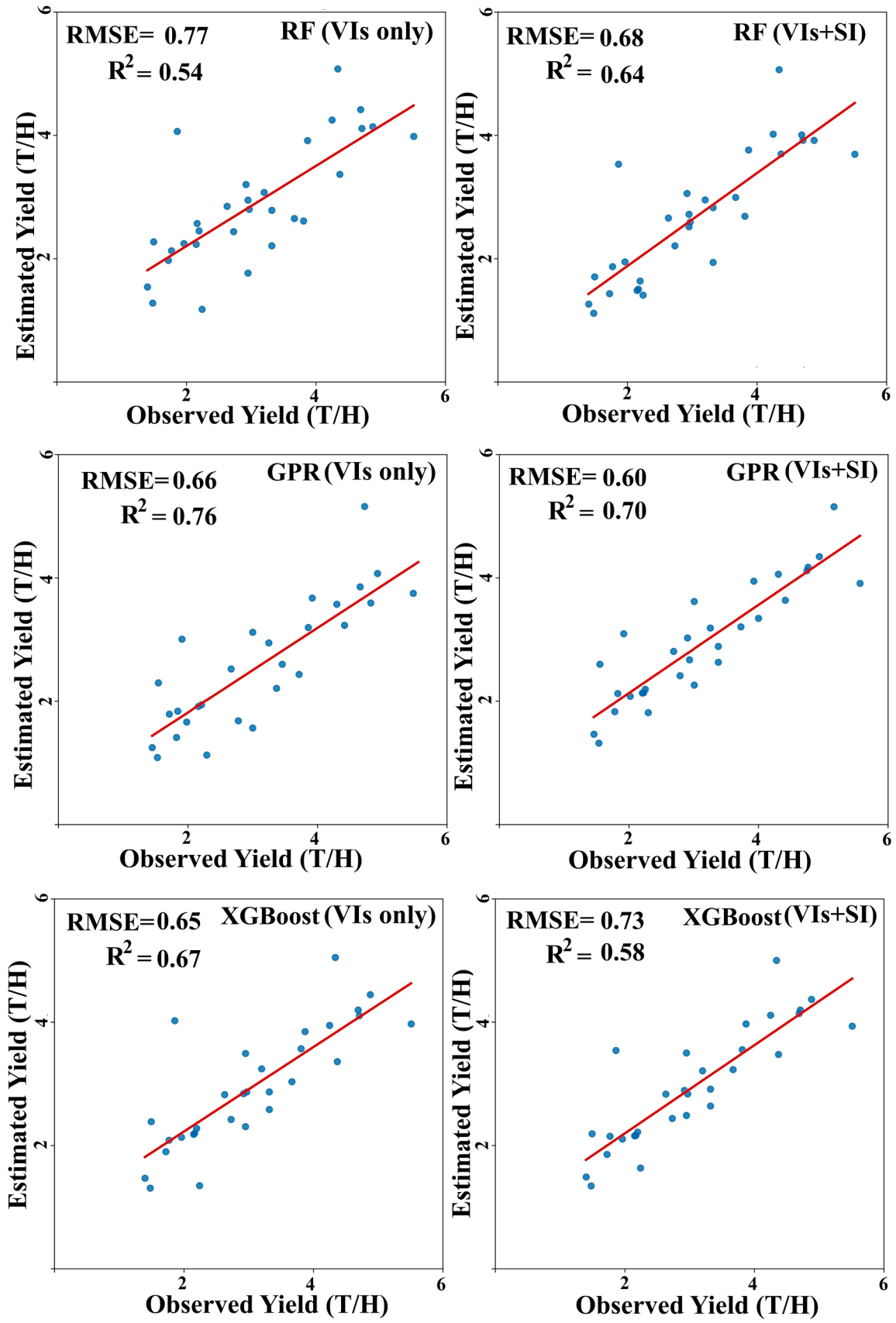

Fig. 6. The relationship between the observed and estimated yield based on indices was analyzed using Random Forest, GPR, and XGBoost, first using only VIs, and then using both SI and VIs.

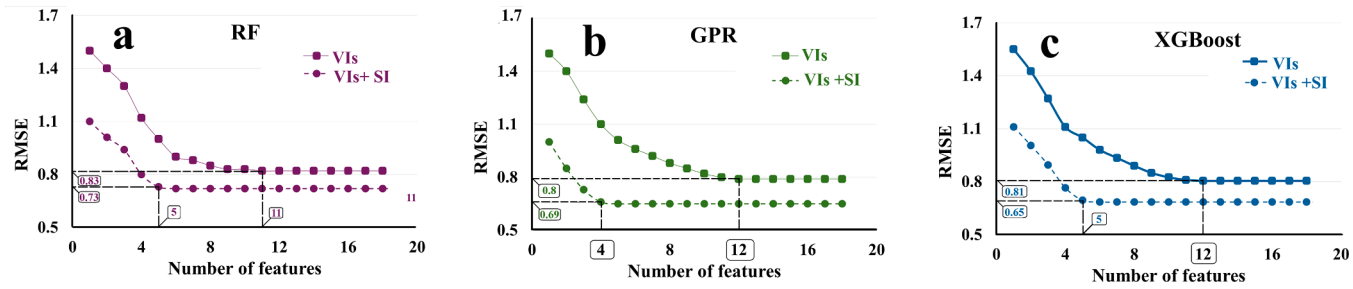

Fig. 7. The relationship between the RMSE and number of features for Random Forest (4.a), GPR (7.b), and XGBoost (7.c), first using only VIs (VIs), and then using both Stress Index and VIs.

In contrast, when SI was introduced as the initial feature, followed by the gradual addition of VIs, the RMSE reduced much faster. The models required only 4 or 5 VIs to reach the minimum RMSE, around 0.73, 0.69 and 0.65 for RF, GPR, XGBoost respectively, not only achieving better prediction accuracy but also reducing the feature count. This approach simplifies the model, enhances interpretability, and demonstrates the efficiency of SI in capturing essential stress-related information that contributes significantly to yield prediction.

Overall, the graph illustrates that including SI accelerates dimension reduction and improves model performance. It enables the model to reach optimal accuracy with fewer features while achieving a lower RMSE than using VIs alone. This finding emphasizes the importance of incorporating SI for efficient and effective yield prediction. In this step, we employed the Sequential Forward Selection (SFS) method for predicting yield. The first five selected indices across all three methods were NDGI, CI, NDVI, VARI, and NDRE when the models were run with SI. Without SI, the models reached the minimum RMSE and the highest R^2 using these indices along with EVI, SR, GNDVI, PRI, and MGR. The difference in selected VIs between models with and without SI is related to the role of SI in representing crop stress information. When SI is not used, the models depend only on VIs to describe crop growth and stress conditions. Because stress information is not directly included, more VIs are needed to capture these effects, and the feature selection process chooses a larger number of indices such as EVI, SR, GNDVI, PRI, and MGR. When SI is included as an input, it already represents important stress effects caused by environmental conditions. This reduces the need for many VIs that provide similar stress-related information. As a result, the feature selection process selects fewer VIs, mainly NDGI, CI, NDVI, VARI, and NDRE, which describe canopy structure and chlorophyll status. This explains why different and fewer VIs are selected when SI is used and why the models achieve better accuracy with lower model complexity.

4. Discussion

4.1. Overview of the study approach

In this study, we employed three ML methods Random Forest (RF), Gaussian Process Regression (GPR), and Extreme Gradient Boosting (XGBoost) to predict wheat yield across five Australian states at field scale. The models were developed using Sentinel-2 derived VIs (VIs) at the peak of greenness, and the biophysical variable, a crop SI was simulated the Oz-Wheat regional crop model during peak of greenness. Model performance was systematically evaluated across multiple dimensions, including correlation analysis, sample size sensitivity, spatial transfer learning and dimensionality reduction.

4.2. Effectiveness of SI on small sample sizes

The results presented in Table 2 and Fig. 4 demonstrate that SI plays a crucial role in improving model reliability when training data is limited. For instance, with only 30 samples, GPR's R^2 increased from 0.14 (VIs only) to 0.39 with SI included. XGBoost and RF models similarly exhibited considerable improvements in R^2 and reductions in RMSE at small sample sizes. The steady decrease in RMSE as the sample size increased (Fig. 4) highlights that while larger datasets naturally improve model performance, SI substantially strengthens prediction accuracy even with minimal training samples, enabling faster and more efficient modelling under data-scarce conditions. Based on Vabalas et al. [29], achieving high model accuracy with limited data in ML techniques is valuable, especially in high-dimensional settings where data collection is costly [29]. These improvements indicate that SI stabilizes model behavior, resulting in more reliable and consistent yield predictions across repeated training and testing cycles. Working model with limited size of data.

4.3. Impact of SI on model performance

The inclusion of SI significantly improved model accuracy across all ML techniques. As shown in Table 2, for XGBoost, the coefficient of determination (R^2) increased from 0.65 to 0.73, and RMSE decreased from 0.67 to 0.58 after SI integration. Similar improvements were observed for GPR (R^2 increased from 0.66 to 0.70, RMSE decreased from 0.76 to 0.60) and RF (R^2 increased from 0.54 to 0.68, RMSE decreased from 0.77 to 0.64). Scatter plots in Fig. 5 further confirm the improvement in prediction accuracy with the addition of SI across all models, emphasizing SI's effectiveness in enhancing overall model performance. To assess whether incorporating SI improves model performance, we use a paired *t*-test to compare the RMSE values of models with and without SI. The *t*-test evaluates whether the mean RMSE of the two models differs significantly. The null hypothesis (H_0) assumes that there is no difference in RMSE, while the alternative hypothesis (H_1) states that the RMSE of the model with SI is significantly lower. The assumptions of the *t*-test include normality of differences, independence of observations, and equal variance when using a standard *t*-test. Given that both models are trained on the same dataset, a paired *t*-test is appropriate for this comparison. The results of the *t*-test for each method confirm a statistically significant improvement when SI is added. For Random Forest (RF), the t-statistic is 12, with a p-value of 0.00, indicating a strong reduction in RMSE. XGBoost shows an even greater improvement, with a t-statistic of 17 and a p-value of 0.00, suggesting that SI significantly enhances the model's predictive ability. Gaussian Process Regression (GPR) also benefits from SI, with a t-statistic of 14 and a p-value of 0.00, reinforcing the conclusion that adding SI consistently improves performance across models. Among the three methods, XGBoost exhibits the most significant improvement based on the highest t-statistic of 17, suggesting that SI has the greatest impact on reducing RMSE in this model. This means that, compared to RF and GPR, XGBoost benefits the most from incorporating SI, leading to the lowest error and more accurate yield predictions. These findings strongly support the inclusion of SI in predictive modeling, demonstrating its role in enhancing model precision and reliability. The observed improvement in model performance with the inclusion of the SI can be attributed to its ability to capture cumulative environmental stressors such as drought, heat, and

radiation that directly affect crop physiological processes but are not fully represented by VIs alone [12,32]. Unlike VIs, which provide snapshot views of canopy greenness or structure at specific time points, SI integrates stress conditions over critical phenological stages, offering a more comprehensive representation of yield-limiting factors [77].

4.4. Contribution of SI to spatial transfer learning

Spatial transfer learning assessments illustrated in Fig. 6 highlight the critical role of SI in enhancing model generalizability across regions. When models trained on multiple regions were applied to an independent region (Western Australia), the inclusion of SI consistently improved outcomes. For example, in XGBoost, R^2 increased from 0.65 to 0.73 and RMSE dropped from 0.67 to 0.58. Similar improvements were observed for GPR and RF models. These results confirm that SI captures stress-related environmental variability that VIs alone may not fully account for, leading to more robust yield predictions under different agro-climatic conditions. Recent studies have shown that incorporating such stress indicators or climate-based variables enhances model robustness and predictive performance across varying climatic zones [77,78]. Additionally, transfer learning techniques that include environmental variability can help address data scarcity issues in under-sampled regions [79] and support generalizable crop monitoring frameworks [80].

4.5. Effect of SI on dimensionality reduction and feature efficiency

As demonstrated in Fig. 7, the SI significantly enhances yield prediction by improving dimensionality reduction and model stability. With SI included, the number of features required to reach optimal Root Mean Square Error (RMSE) decreased from around 11 to just 5 or 6 indices, substantially minimizing model complexity, reducing computational costs, and lowering the risk of overfitting while maintaining or improving predictive accuracy. This enables models to generalize better across regions, even when trained with limited sample sizes. Additionally, SI played a crucial role in reducing variance, bias, and uncertainty, as reflected in more stable RMSE, consistently higher R^2, and lower Mean Absolute Error (MAE) across 100 iterations. Fig. 8 further illustrates SI's independent contribution to model inputs: the correlation between SI and wheat yield reached $R^2 = 0.42$ ($P < 0.05$), highlighting SI's strength as a standalone predictor (Fig. 8a). Meanwhile, SI showed moderate but distinct correlations with VIs (VIs) at the peak of greenness, ranging from 0.52 with SR to 0.61 with NDMI, respectively, indicating that SI captures unique environmental stress information not fully represented by spectral indices alone. The Principal Component Analysis (PCA) biplot (Fig. 8b) confirmed that SI and VIs are distributed across separate principal components (PC1 and PC2), which together explain 79 % of the variance, reinforcing the idea that SI adds a new, independent dimension to the feature space. Overall, by introducing biophysical stress information, SI enriches ML models, enabling more efficient, faster, and more spatially transferable wheat yield predictions.

Adding the SI to yield prediction models brings in an important environmental factor from the GEM framework [81].Unlike VIs (VIs), which mainly show how green or healthy the crop looks at a certain time, SI shows the total effect of stress like heat, drought, and sunlight on the crop during key growth stages [82]. SI is made using weather data from ground stations and helps measure how bad conditions affect the plant's growth and yield in ways that satellite images can't always show [83,84]. This means SI gives extra, useful information that VIs might miss. Including SI makes models stronger, simpler, and better at working in different places. By adding this real stress information, SI helps ML models predict crop yield more accurately, even when growing conditions change.

4.6. Utility and broader importance of this research

Beyond predictive improvements, this study provides a new methodological framework for remote sensing-based agricultural monitoring. As shown in the correlation matrix (Fig. 3), VIs displayed moderate to strong correlations with wheat yield, while SI exhibited a somewhat lower correlation individually. However, the low internal correlation between SI and VIs (R^2 range 0.11–0.24) confirms that SI offers largely independent information. When SI was added, as demonstrated in Fig. 6, the number of features required to reach optimal RMSE dropped from 11 to 12 VIs to just 4–5 indices enhancing model simplicity and reducing computational cost. Importantly, the spatial transfer learning results (Fig. 5) highlighted that SI substantially improved model adaptability across regions. These findings are consistent with recent research demonstrating that combining spectral and biophysical variables improves model interpretability, reduces input dimensionality, and enhances spatial transferability in yield forecasting systems [77,85,86]. This study, therefore, introduces a scalable, efficient, and transferable approach to crop yield prediction, advancing the use of remote sensing and ML in precision agriculture and large-scale food security forecasting.

5. Conclusion

This study evaluated the predictive performance of three ML models Random Forest (RF), Gaussian Process Regression (GPR), and Extreme Gradient Boosting (XGBoost) for estimating wheat yield across 380 Australian fields using VIs (VIs), with and without a functionally derived biophysical SI. Model performance was assessed across varying sample sizes, input dimensionalities, and spatial contexts.

Incorporating the SI led to consistent and significant improvements in model accuracy across all algorithms. Notably, R^2 increased from 0.65 to 0.73 for XGBoost, from 0.66 to 0.70 for GPR, and from 0.54 to 0.68 for RF. These improvements underscore the value of embedding biophysical crop stress responses in data-driven frameworks, yielding nearly a threefold increase in explanatory power for RF in particular.

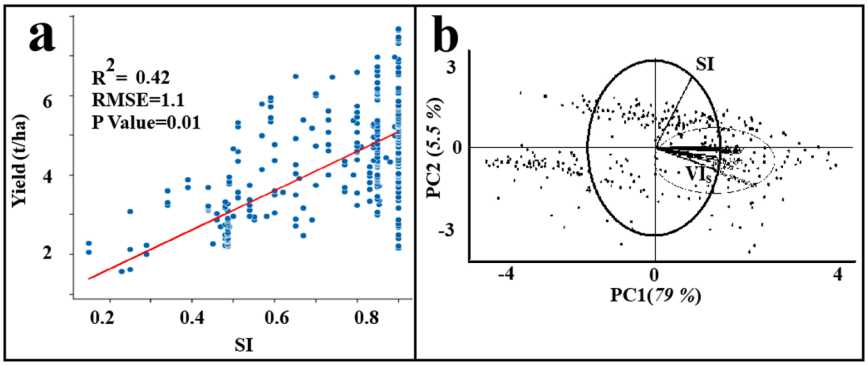

Fig. 8. Correlation of stress index (SI) and Yield (a), Bi plot of SI and VIs (VIs) using PCA analysis (b).

Furthermore, SI integration reduced input feature dimensionality by one-third and improved spatial transferability by approximately 12 %.

These gains were especially pronounced under data-limited conditions, where SI enhanced model robustness, predictive power, and generalizability. Across all tested scenarios, models incorporating SI consistently outperformed those using VIs alone, demonstrating the importance of physiologically meaningful inputs in improving learning efficiency.

Ultimately, this study highlights the limitations of black-box ML approaches in agricultural modelling and demonstrates the critical role of incorporating targeted, functional biophysical indicators such as stress indices to strengthen model interpretability, accuracy, and transferability across complex agro-ecological systems.

CRediT authorship contribution statement

Davoud Ashourloo: Writing – original draft, Visualization, Software, Resources, Investigation, Formal analysis, Data curation, Conceptualization. **Jason Brader:** Visualization, Validation, Software, Formal analysis, Data curation. **Yan Zhao:** Writing – original draft, Validation, Methodology, Formal analysis, Data curation. **Ruizhu Jiang:** Validation, Resources, Investigation, Data curation. **Miao Zhang:** Supervision, Resources, Investigation, Formal analysis. **Scott Chapman:** Visualization, Validation, Supervision. **Graeme Hammer:** Writing – original draft, Validation, Supervision, Investigation, Conceptualization. **Andries B Potgieter:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation.

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 titled *"Quantifying the Efficacy of Machine Learning Methods Driven by Biophysical Modeling and Remote Sensing to Predict Wheat Yield at Field Scale across Australia."*

This study “Quantifying the Efficacy of Machine Learning Methods Driven by Biophysical Modelling and Remote Sensing to Predict Wheat Yield at Field Scale across Australia” did not involve any experiments on humans or animals. Ethical approval was therefore not required.

Data availability

The data that has been used is confidential.

References

[1] A. Araghi, A. Daccache, Remote sensing and TerraClimate datasets for wheat yield prediction using machine learning, Smart Agric. Technol. 11 (2025) 100909.

[2] P.R. Ehrlich, A.H. Ehrlich, G.C. Daily, Food security, population and environment, Popul. Dev. Rev. (1993) 1–32.

[3] M.J. Kazemi, P. Samouei, A new bi-level mathematical model for government-farmer interaction regarding food security and environmental damages of pesticides and fertilizers: case study of rice supply chain in Iran, Comput. Electron. Agric. 219 (2024) 108771.

[4] U. Mc Carthy, I. Uysal, R. Badia-Melis, S. Mercier, C. O'Donnell, A. Ktenioudaki, Global food security–Issues, challenges and technological solutions, Trends. Food Sci. Technol. 77 (2018) 11–20.

[5] W. Guo, Y. Huang, Y. Huang, Y. Li, X. Song, J. Shen, X. Qi, B. Zhang, Z. Zhu, S. Peng, Develop agricultural planting structure prediction model based on machine learning: the aging of the population has prompted a shift in the planting structure toward food crops, Comput. Electron. Agric. 221 (2024) 108941.

[6] H. Jiang, L. Jiang, L. He, B.G. Murengami, X. Jing, P.A. Misiewicz, F.A. Cheein, L. Fu, Yield prediction of root crops in field using remote sensing: a comprehensive review, Comput. Electron. Agric. 227 (2024) 109600.

[7] L. Karthikeyan, I. Chawla, A.K. Mishra, A review of remote sensing applications in agriculture for food security: crop growth and yield, irrigation, and crop losses, J Hydrol 586 (2020) 124905.

[8] P.R. Shewry, S.J. Hey, The contribution of wheat to human diet and health, Food Energy Secur. 4 (3) (2015) 178–202.

[9] D.B. Lobell, K.G. Cassman, C.B. Field, Crop yield gaps: Their importance, magnitudes, and causes, Annu. Rev. Environ. Resour. 34 (1) (2009) 179–204.

[10] P.C. Doraiswamy, T.R. Sinclair, S. Hollinger, B. Akhmedov, A. Stern, J. Prueger, Application of MODIS derived parameters for regional crop yield assessment, Remote Sens. Environ. 97 (2) (2005) 192–202.

[11] X. Zhang, M.A. Friedl, C.B. Schaaf, A.H. Strahler, J.C. Hodges, F. Gao, B.C. Reed, A. Huete, Monitoring vegetation phenology using MODIS, Remote Sens. Environ. 84 (3) (2003) 471–475.

[12] D.B. Lobell, D. Thau, C. Seifert, E. Engle, B. Little, A scalable satellite-based crop yield mapper, Remote Sens. Environ. 164 (2015) 324–333.

[13] M. Manafifard, J. Huang, A comprehensive review on wheat yield prediction based on remote sensing, Multimed. Tools. Appl. (2024) 1–74.

[14] P. Hu, B. Zheng, Q. Chen, S. Grunefeld, M.R. Choudhury, J. Fernandez, A. Potgieter, S.C. Chapman, Estimating aboveground biomass dynamics of wheat at small spatial scale by integrating crop growth and radiative transfer models with satellite remote sensing data, Remote Sens. Environ. 311 (2024) 114277.

[15] D. Phiri, M. Simwanda, S. Salekin, V.R. Nyirenda, Y. Murayama, M. Ranagalage, Sentinel-2 data for land cover/use mapping: a review, Remote Sens. (Basel) 12 (14) (2020) 2291.

[16] J. Segarra, M.L. Buchaillot, J.L. Araus, S.C. Kefauver, Remote sensing for precision agriculture: sentinel-2 improved features and applications, Agronomy 10 (5) (2020) 641.

[17] S.S. Panda, D.P. Ames, S. Panigrahi, Application of vegetation indices for agricultural crop yield prediction using neural network techniques, Remote Sens. (Basel) 2 (3) (2010) 673–696.

[18] D. Ashourloo, M. Manafifard, M. Behifar, M. Kohandel, Wheat yield prediction based on Sentinel-2, regression, and machine learning models in Hamedan, Iran, Scientia Iranica 29 (6) (2022) 3230–3243.

[19] A. Kayad, M. Sozzi, S. Gatto, F. Marinello, F. Pirotti, Monitoring within-field variability of corn yield using Sentinel-2 and machine learning techniques, Remote Sens. (Basel) 11 (23) (2019) 2873.

[20] A.M. Kheir, A. Govind, V. Nangia, M.A. El-Maghraby, A. Elnashar, M. Ahmed, H. Aboelsoud, R. Gamal, T. Feike, Hybridization of process-based models, remote sensing, and machine learning for enhanced spatial predictions of wheat yield and quality, Comput. Electron. Agric. 234 (2025) 110317.

[21] T. Van Klompenburg, A. Kassahun, C. Catal, Crop yield prediction using machine learning: a systematic literature review, Comput. Electron. Agric. 177 (2020) 105709.

[22] P. Mondal, X. Liu, T.E. Fatoyinbo, D. Lagomasino, Evaluating combinations of sentinel-2 data and machine-learning algorithms for mangrove mapping in West Africa, Remote Sens. (Basel) 11 (24) (2019) 2928.

[23] A. Moumni, M. Oujaoura, J. Ezzahar, A. Lahrouni, A new synergistic approach for crop discrimination in a semi-arid region using Sentinel-2 time series and the multiple combination of machine learning classifiers, J. Phys.: Conf. Ser. (2021).

[24] 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.

[25] D. Ashourloo, H. Aghighi, A.A. Matkan, M.R. Mobasheri, A.M. Rad, An investigation into machine learning regression techniques for the leaf rust disease detection using hyperspectral measurement, IEEe J. Sel. Top. Appl. Earth. Obs. Remote Sens. 9 (9) (2016) 4344–4351.

[26] 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, Int. J. Appl. Earth Obs. Geoinf. 118 (2023) 103269.

[27] N. Prasad, N. Patel, A. Danodia, Crop yield prediction in cotton for regional level using random forest approach, Spat. Inf. Res. 29 (2021) 195–206.

[28] S.J. Raudys, A.K. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, IEEe Trans. Pattern. Anal. Mach. Intell. 13 (3) (1991) 252–264.

[29] A. Vabalas, E. Gowen, E. Poliakoff, A.J. Casson, Machine learning algorithm validation with a limited sample size, PLoS. One 14 (11) (2019) e0224365.

[30] 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.

[31] Y. Cai, K. Guan, D. Lobell, A.B. Potgieter, S. Wang, J. Peng, T. Xu, S. Asseng, Y. Zhang, L. You, Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches, Agric. For. Meteorol. 274 (2019) 144–159.

[32] Potgieter, A., & Hammer, G.L. (2006). Oz-wheat: a regional-scale crop yield simulation model for Australian wheat.

[33] G.L. Hammer, E. van Oosterom, G. McLean, S.C. Chapman, I. Broad, P. Harland, R. C. Muchow, Adapting APSIM to model the physiology and genetics of complex adaptive traits in field crops, J. Exp. Bot. 61 (8) (2010) 2185–2202 [Record #550 is using a reference type undefined in this output style.].

[34] B.A. Keating, P.S. Carberry, G.L. Hammer, M.E. Probert, M.J. Robertson, D. Holzworth, N.I. Huth, J.N. Hargreaves, H. Meinke, Z. Hochman, An overview of APSIM, a model designed for farming systems simulation, Eur. J. Agron. 18 (3–4) (2003) 267–288.

[35] M. Reynolds, M. Kropff, J. Crossa, J. Koo, G. Kruseman, A. Molero Milan, J. Rutkoski, U. Schulthess, K. Sonder, H. Tonnang, Role of modelling in international crop research: overview and some case studies, Agronomy 8 (12) (2018) 291.

[36] Y. Cai, K. Guan, J. Peng, S. Wang, C. Seifert, B. Wardlow, Z. Li, A high-performance and in-season classification system of field-level crop types using time-series

Landsat data and a machine learning approach, Remote Sens. Environ. 210 (2018) 35–47.
[37] M. Shahhosseini, G. Hu, S.V. Archontoulis, Forecasting corn yield with machine learning ensembles, Front. Plant Sci. 11 (2020) 1120.
[38] Y. Chen, R.J. Donohue, T.R. McVicar, F. Waldner, G. Mata, N. Ota, A. Houshmandfar, K. Dayal, R.A. Lawes, Nationwide crop yield estimation based on photosynthesis and meteorological stress indices, Agric. For. Meteorol. 284 (2020) 107872.
[39] E. Kamir, F. Waldner, Z. Hochman, Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods, ISPRS J. Photogramm. Remote Sens. 160 (2020) 124–135.
[40] Y. Zhao, A.B. Potgieter, M. Zhang, B. Wu, G.L. Hammer, Predicting wheat yield at the field scale by combining high-resolution Sentinel-2 satellite imagery and crop modelling, Remote Sens. (Basel) 12 (6) (2020) 1024.
[41] D.R. Stockwell, A.T. Peterson, Effects of sample size on accuracy of species distribution models, Ecol. Modell. 148 (1) (2002) 1–13.
[42] A. Orynbaikyzy, U. Gessner, C. Conrad, Spatial transferability of random forest models for crop type classification using Sentinel-1 and Sentinel-2, Remote Sens. (Basel) 14 (6) (2022) 1493.
[43] K. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, J. Big. Data 3 (2016) 1–40.
[44] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (Mar) (2003) 1157–1182.
[45] M.G. Kallenberg, B. Maestrini, R. van Bree, P. Ravensbergen, C. Pylianidis, F. van Evert, I.N. Athanasiadis, Integrating processed-based models and machine learning for crop yield prediction, arXiv preprint, arXiv:2307.13466, 2023.
[46] M. Shahhosseini, G. Hu, I. Huber, S.V. Archontoulis, Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt, Sci. Rep. 11 (1) (2021) 1606.
[47] Z. Hochman, H. Horan, Causes of wheat yield gaps and opportunities to advance the water-limited yield frontier in Australia, Field. Crops. Res. 228 (2018) 20–30.
[48] Y. Oliver, M. Robertson, Quantifying the spatial pattern of the yield gap within a farm in a low rainfall Mediterranean climate, Field. Crops. Res. 150 (2013) 29–41.
[49] J.R. Hunt, Winter wheat cultivars in Australian farming systems: a review, Crop Pasture Sci. 68 (6) (2017) 501–515.
[50] H. Sprigg, R. Belford, S. Milroy, S.J. Bennett, D. Bowran, Adaptations for growing wheat in the drying climate of Western Australia, Crop Pasture Sci. 65 (7) (2014) 627–644.
[51] A.B. Potgieter, G.L. Hammer, A. Doherty, P. De Voil, A simple regional-scale model for forecasting sorghum yield across North-Eastern Australia, Agric. For. Meteorol. 132 (1–2) (2005) 143–153.
[52] A.B. Potgieter, D.B. Lobell, G.L. Hammer, D.R. Jordan, P. Davis, J. Brider, Yield trends under varying environmental conditions for sorghum and wheat across Australia, Agric. For. Meteorol. 228 (2016) 276–285.
[53] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani, J. Friedman, Random forests, The elements of statistical learning: Data mining, inference, and prediction (2009) 587–604.
[54] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Publisher: Taylor & Francis, (1984). https://books.google.com.au/books?id=JwQx-WOmSyQC.
[55] J. Snoek, H. Larochelle, R.P. Adams, Practical bayesian optimization of machine learning algorithms, Adv. Neural Inf. Process. Syst. (2012) 25.
[56] M. Lázaro-Gredilla, M.K. Titsias, J. Verrelst, G. Camps-Valls, Retrieval of biophysical parameters with heteroscedastic Gaussian processes, IEEe Geosci. Remote Sens. Lett. 11 (4) (2013) 838–842.
[57] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, University Press Group Limited, 2006. https://books.google.com.au/books?id=vWtwQgAACAAJ.
[58] C.K. Williams, Prediction with Gaussian processes: from linear regression to linear prediction and beyond. Learning in Graphical Models, Springer, 1998, pp. 599–621.
[59] Y. Bu, J. Pan, Stellar atmospheric parameter estimation using gaussian process regression, Mon. Not. R. Astron. Soc. 447 (1) (2015) 256–265.
[60] J. Rasmussen. Non destructive testing methods for integrity determination of concrete structures, Luleå tekniska universitet, 2003.
[61] J. Elith, J.R. Leathwick, T. Hastie, A working guide to boosted regression trees, J. Anim. Ecol. 77 (4) (2008) 802–813.
[62] J.H. Friedman, J.J. Meulman, Multiple additive regression trees with application in epidemiology, Stat. Med. 22 (9) (2003) 1365–1381.
[63] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016.
[64] Pearson, R.L., & Miller, L.D. (1972). Remote mapping of standing crop biomass for estimation of the productivity of the shortgrass prairie, Pawnee National Grasslands, Colorado.
[65] Rouse Jr, J. W., Haas, R.H., Deering, D., Schell, J., & Harlan, J.C. (1974). *Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation*.
[66] A. Huete, K. Didan, T. Miura, E.P. Rodriguez, X. Gao, L.G. Ferreira, Overview of the radiometric and biophysical performance of the MODIS vegetation indices, Remote Sens. Environ. 83 (1–2) (2002) 195–213.
[67] M. Karaman, Capability assessment of Sentinel-2 imagery for graphite deposits exploration, Geochemistry (2024) 126117.
[68] B.-C. Gao, NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space, Remote Sens. Environ. 58 (3) (1996) 257–266.
[69] A.A. Gitelson, A. Viña, V. Ciganda, D.C. Rundquist, T.J. Arkebauer, Remote estimation of canopy chlorophyll content in crops, Geophys. Res. Lett. (8) (2005) 32.
[70] F. Jiang, H. Sun, C. Li, K. Ma, S. Chen, J. Long, L. Ren, Retrieving the forest aboveground biomass by combining the red edge bands of Sentinel-2 and GF-6, *Acta Ecol.* Sin 41 (2021) 8222–8236.
[71] J. Gamon, J. Penuelas, C. Field, A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency, Remote Sens. Environ. 41 (1) (1992) 35–44.
[72] J.E. Keeley, Fire intensity, fire severity and burn severity: a brief review and suggested usage, Int. J. Wildland. Fire 18 (1) (2009) 116–126.
[73] C.J. Tucker, Red and photographic infrared linear combinations for monitoring vegetation, Remote Sens. Environ. 8 (2) (1979) 127–150.
[74] A.A. Gitelson, Y.J. Kaufman, M.N. Merzlyak, Use of a green channel in remote sensing of global vegetation from EOS-MODIS, Remote Sens. Environ. 58 (3) (1996) 289–298.
[75] Y. Gu, E. Hunt, B. Wardlow, J.B. Basara, J.F. Brown, J.P. Verdin, Evaluation of MODIS NDVI and NDWI for vegetation drought monitoring using Oklahoma Mesonet soil moisture data, Geophys. Res. Lett. (22) (2008) 35.
[76] D. Stow, M. Niphadkar, J. Kaiser, MODIS-derived visible atmospherically resistant index for monitoring chaparral moisture content, Int. J. Remote Sens. 26 (17) (2005) 3867–3873.
[77] B. Peng, K. Guan, J. Tang, E.A. Ainsworth, S. Asseng, C.J. Bernacchi, M. Cooper, E. H. Delucia, J.W. Elliott, F. Ewert, Towards a multiscale crop modelling framework for climate change adaptation assessment, Nat. Plants. 6 (4) (2020) 338–348.
[78] A. Joshi, B. Pradhan, S. Chakraborty, R. Varatharajoo, S. Gite, A. Alamri, Deep-transfer-learning strategies for crop yield prediction using climate records and satellite image time-series data, Remote Sens. (Basel) 16 (24) (2024) 4804.
[79] A.X. Wang, C. Tran, N. Desai, D. Lobell, S. Ermon, Deep transfer learning for crop yield prediction with remote sensing data, in: Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, 2018.
[80] F. Huber, A. Inderka, V. Steinhage, Leveraging remote sensing data for yield prediction with deep transfer learning, Sensors 24 (3) (2024) 770.
[81] C. Messina, G. Hammer, Z. Dong, D. Podlich, M. Cooper, Modelling crop improvement in a G× E× M framework via gene-trait-phenotype relationships. Crop Physiology: Applications for Genetic Improvement and Agronomy, Elsevier, Netherlands, 2009, pp. 235–265.
[82] D.B. Lobell, The use of satellite data for crop yield gap analysis, Field. Crops. Res. 143 (2013) 56–64.
[83] S. Asseng, F. Ewert, C. Rosenzweig, J.W. Jones, J.L. Hatfield, A.C. Ruane, K. J. Boote, P.J. Thorburn, R.P. Rötter, D. Cammarano, Uncertainty in simulating wheat yields under climate change, Nat. Clim. Chang. 3 (9) (2013) 827–832.
[84] H. Campos, M. Cooper, J. Habben, G. Edmeades, J. Schussler, Improving drought tolerance in maize: a view from industry, Field. Crops. Res. 90 (1) (2004) 19–34.
[85] K. Kuwata, R. Shibasaki, Estimating crop yields with deep learning and remotely sensed data, in: 2015 IEEE international geoscience and remote sensing symposium (IGARSS), 2015.
[86] J. You, X. Li, M. Low, D. Lobell, S. Ermon, Deep gaussian process for crop yield prediction based on remote sensing data, in: Proceedings of the AAAI conference on artificial intelligence, 2017.