International Journal of Design & Nature and Ecodynamics
Vol. 19, No. 6, December, 2024, pp. 1909-1917
Journal homepage: http://iieta.org/journals/ijdne

# Impact of Feature Selection on Wheat Yield Prediction Using Machine Learning

Yashraj Patil[1], Rani Fathima[2], Sridhevi Sundarajan[1], P. Sridevi Ponmalar[3], Harikrishnan Ramachandran[1*]

[1] Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune 412115, Maharashtra, India
[2] Electrical Engineering Program, EDICT Department, Bahrain Polytechnic, Isa Town 33349, Bahrain
[3] School of Computing, SRM Institute of Science and Technology, Chennai 603203, Tamil Nadu, India

Corresponding Author Email: dr.rhareish@gmail.com



https://doi.org/10.18280/ijdne.190607

**Received:** 18 September 2024
**Revised:** 15 October 2024
**Accepted:** 25 October 2024
**Available online:** 27 December 2024

***Keywords:***
*crop yield prediction, feature selection in machine learning, Gradient Boosting regression, Linear Regression, Random Forest Regression, wheat dataset*

**ABSTRACT**

The purpose of this study was to determine how feature selection and model complexity affect the predictive performance of several regression models for crop yield prediction. Two experiments were conducted on a wheat production dataset: one with a limited collection of features and one with an expanded set. The study used Gradient Boosting, Random Forest, and Linear Regression models. Performance measurements included Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared ($R^2$). In Experiment 1, Linear Regression outperformed complex models, with an MSE of 4,316, RMSE of 65.7, and $R^2$ of 0.98. In Experiment 2, Linear Regression showed considerable improvement, lowering the MSE to 1,757 and improving $R^2$ to 0.99. These findings challenge the widespread preference for complex models in crop production prediction, highlighting the importance of feature selection. The findings have far-reaching implications for agricultural planning, indicating that simpler models can produce more reliable predictions, improving resource allocation, food security, and economic stability especially in regions with limited technological infrastructure.

## 1. INTRODUCTION

Crop yield prediction plays a crucial role in agricultural decision-making, guiding farmers, policymakers, and stakeholders in optimizing crop production and resource management. Accurate yield predictions are essential for addressing challenges such as food security, climate change adaptation, and sustainable agricultural practices. Predictive modeling techniques offer powerful tools for forecasting crop yields based on various factors, including environmental conditions, soil properties, and agronomic practices.

In the realm of agricultural production, understanding key metrics such as area, production, and yield is fundamental to assessing and managing crop systems effectively [1]. Area refers to the total land under cultivation for a particular crop within a defined geographic region. It is measured in hectares, acres, or other appropriate units of land measurement. The area under cultivation can vary widely depending on factors such as agricultural practices, land availability, and government policies. Production represents the total quantity of a crop harvested from a specific area within a given time frame, typically measured in terms of weight (e.g., kilograms) or volume (e.g., litres). It reflects the output of agricultural activities and is influenced by factors such as crop variety, planting density, pest management, and weather conditions. Yield is a crucial indicator that measures the efficiency of crop production by quantifying the output per unit of land area. It is determined by dividing the total production by the corresponding area under cultivation. Yield can be expressed in various units depending on the crop and region, such as kilograms per hectare or bushels per acre. Yield serves as a critical performance metric for farmers, researchers, and policymakers, offering insights into productivity levels, resource utilization, and overall agricultural sustainability.

Current crop yield prediction studies reveal a number of research gaps that require more examination. One noticeable shortcoming is the little investigation into the impact of feature selection on model performance across various crop datasets. For example, Shah et al. [2] used meteorological data for multivariate regression models, but they did not investigate how different feature selections affected performance. Similarly, Shyamala and Rajeshwar [3] investigated increased Gradient Boosting algorithms but did not evaluate how simpler models perform with constrained feature sets. Charoen-Ung and Mittrapiyanuruk [4] utilized Random Forest and Gradient Boosting tree algorithms to predict sugarcane yield grade, achieving accuracies surpassing non-machine-learning baselines, with acknowledged limitations on generalizability and performance analysis. Puligudla et al. [5] applied data mining techniques, specifically Gradient Boosting and regression, to predict crop yields and enhance harvesting methods, focusing on automation and alleviating farmer burden. Yasaswy et al. [6] demonstrated the superiority of an innovative Gradient Boosting algorithm over Random Forest in predicting crop yield rates, with limitations on sample size and reliance on historical data. Huber et al. [7]

demonstrated that XGBoost outperformed deep learning approaches in predicting agricultural yield through remote sensing, showcasing superior performance compared to other deep learning algorithms. Our study aimed to close this gap by running tests with different feature sets and directly comparing the prediction performance of simpler and more complicated models like Linear Regression, Random Forest, and Gradient Boosting Regression.

Feature selection is a crucial step in machine learning that involves choosing the most relevant and informative features from a given dataset [8]. This step helps in improving the performance of the model by reducing dimensionality, decreasing overfitting, and enhancing its interpretability. By selecting the ideal subset of features, the model could focus more on the crucial factors that have a significant impact on prediction outcomes. By doing so, the model not only becomes more efficient but also saves considerable runtime. Additionally, feature selection plays a vital role in saving computational resources and reducing the complexity and cost of data analysis. By utilizing methods such as filters, wrappers, and embedding, researchers can effectively select the most influential features. This process, also known as feature extraction [9], can be achieved through techniques like Principal Component Analysis or Genetic Algorithm. By using these feature selection techniques, the dimensionality of the feature space can be reduced while preserving the most relevant information. This allows the learning algorithm to work more efficiently and accurately, leading to better model performance and prediction outcomes. By leveraging human ingenuity and prior knowledge, feature engineering [10] compensates for the weakness of machine learning algorithms. It expands the scope and ease of applicability of machine learning, making it less dependent on manual feature engineering. This ultimately contributes to faster development of novel applications and progress towards artificial intelligence.

There are many different techniques for feature selection. These techniques aim to identify the most relevant and informative features from a given dataset. Some commonly used feature selection techniques like The Univariate Selection method that selects features based on their independent associations with the target variable [11]. The Recursive Feature Elimination technique iteratively removes less impactful features to enhance model performance [12]. Principal Component Analysis (PCA) is a dimensionality reduction technique that identifies the most important features by transforming the original features into a new set of uncorrelated variables called principal components [9]. The Lasso Regression method uses regularization to shrink the coefficients of less important features, effectively selecting the most relevant features [13]. Genetic Algorithm is inspired by the process of natural selection and evolution. It involves creating a population of potential feature subsets and iteratively applying selection, crossover, and mutation operations to evolve towards the optimal subset of features [14].

When it comes to the crop dataset for yield prediction, feature selection is typically done through the techniques that aim to identify the subset of features that are most relevant and informative for predicting crop yield. Correlation analysis method involves calculating the coefficient of correlation between every feature and the target variable (e.g., crop yield) and selecting features with the highest correlation [15]. In the domain knowledge-based selection, domain experts can provide valuable insights into which features are likely to be important for predicting crop yield [16]. Another approach is through statistical techniques, such as the use of information gain or chi-square tests [17], to determine the dependency between each feature and the target variable.

The current studies on crop yield prediction reveals several research gaps that warrant further investigation. One notable gap is the absence of experiments focused on influence of feature selection [18] on model performance across the different crop datasets. Feature selection is like picking the most important pieces of information from a big pile of data to help a computer make better predictions [19]. Additionally, there is limited exploration of how varying the complexity of models impacts performance, including experiments with simplified or complex models. There is a lack of evaluation on the generalizability of the developed models to different regions or crops, limiting their applicability beyond the specific datasets used in the studies.

This study focuses on estimating wheat production, a staple crop that is critical to Indian food security. Wheat accounts for the bulk of Indian caloric intake, so precise yield estimates are critical for satisfying food demand, especially in the face of climate change. The dataset used in this study contains a variety of wheat production-related parameters, including meteorological data, which allow us to examine the impact of different feature subsets on prediction accuracy. This dataset is particularly well suited to filling the highlighted gaps since it includes a wide variety of variables that influence yield, providing a solid platform for evaluating the impact of feature selection. The study investigates the performance of supervised machine learning models - Linear Regression [20], Random Forest Regression [21] and Gradient Boosting Regression [22], across two experiments with varying feature sets. The experiments aim to elucidate the impact of feature selection and model complexity on predictive accuracy and provide insights for agricultural researchers and practitioners.

In addition to the technical aspects of machine learning algorithms, Garcia-Miralles [23] conducted a study on peri-urban agriculture. This research emphasized the importance of essential green infrastructure in promoting sustainable transitions within cities. This focus will improve decision-making in precision farming for crop recommendations and enable the exploration of additional potential features. The research by Caka [24] highlights the vital role of urban agriculture in promoting sustainable urban development. By incorporating agricultural practices into city environments, urban agriculture enhances food security through the provision of fresh, locally sourced produce and reduces the carbon footprint associated with food transportation. Ultimately, urban agriculture contributes to improved biodiversity, supports local ecosystems, and enhances the quality of life for city residents, addressing many challenges faced by modern urban centers.

## 2. METHODOLOGY

Two experiments were conducted to compare the performance of predictive modeling techniques for wheat yield prediction. Figure 1 illustrates the flow of yield forecasting model evaluation commences with data preparation and progresses through variable selection and experimentation with various machine learning models to determine the most accurate yield predictions based on mean

square error, root mean squared error, and R-squared values. The major aim of these experiment is to test which regression model handles crop dataset efficiently and what happens when we expand the feature set which is already preprocessed through domain knowledge selection technique.

The selection of the regression models like Random Forest, Linear Regression, and Gradient Boosting models is supported by the strengths of each model for the dataset. With a lower risk of overfitting, Random Forest is excellent at managing big feature sets and non-linear interactions. As demonstrated by its greater performance in the research studies, Linear Regression is easy to understand and captures linear trends. Gradient Boosting is used because it may improve predictions iteratively, reducing bias and variation. They allow stakeholders to understand the relationship between input features and crop yield, aiding decision-making. Regression models also facilitate feature importance analysis and can handle mixed feature types present in agricultural datasets. Evaluation metrics like MSE and $R^2$ assess prediction accuracy. Their simplicity, efficiency, and scalability suit large-scale datasets and real-time decision-making. Regression models provide insights into environmental-crop interactions, promoting sustainable agriculture and food security.

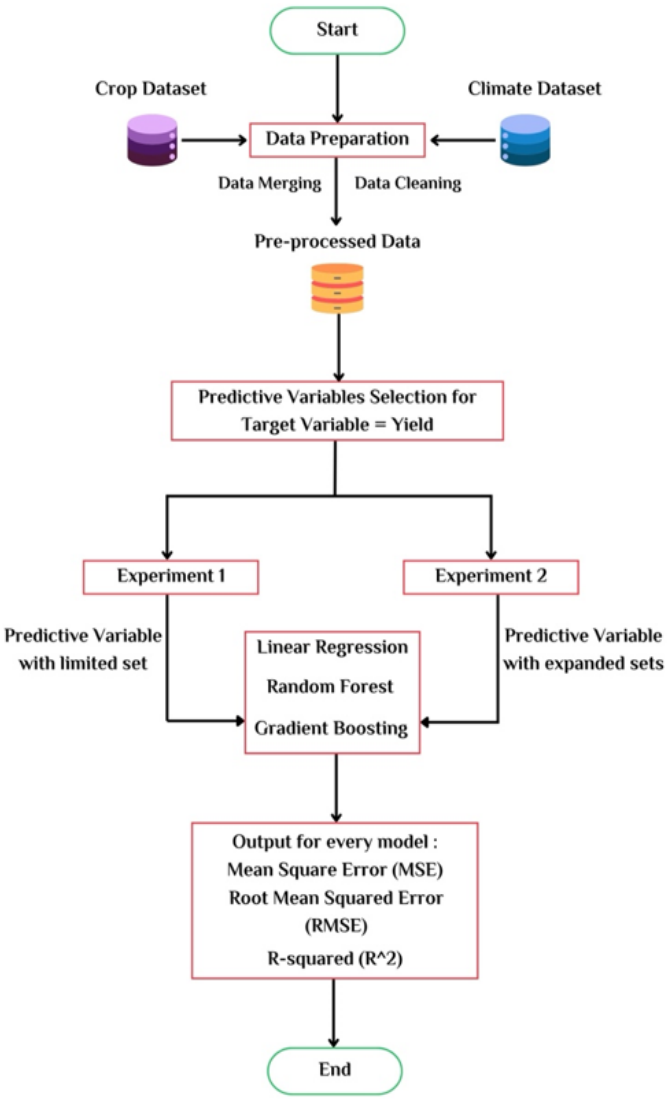



**Figure 1.** Yield forecasting model evaluation flowchart

### 2.1 Data collection and preprocessing

The dataset for this study was compiled from publicly accessible agricultural data sources, focusing on wheat production metrics using the wheat crop dataset from ICRISAT–TCI (Tata Cornell Institute of Agriculture and Nutrition) [25]. Key variables include 'Area' (hectares under cultivation), 'Production' (total produced quantity in metric tons), 'Yield' (production per unit area), along with climate factors such as 'Rainfall' and temperature metrics ('Temp(Max)', 'Temp(Mean)', 'Temp(Min)') from ERA5-Land climate dataset [26] which were merged into single dataset during data pre-processing performed through domain knowledge-based feature selection method. The dataset spans records from 1966 to 2017, ensuring relevance and applicability to current agricultural practices.

Area: The total hectares allocated for wheat cultivation. Production: The total quantity of wheat produced, measured in metric tons. Yield: Calculated as the ratio of Production to Area, representing the efficiency of production per unit area.

### 2.2 Experimental setup

The study was structured around two primary experiments, each employing different feature sets to train and test the models:

Experiment 1: Limited Feature Set Model
Features: 'Area' and 'Production'
Target: 'Yield'
Regression Models: Random Forest, Linear Regression, Gradient Boosting

Experiment 2: Expanded Feature Set Model
Features: 'Area', 'Production', 'Rainfall', 'Temp(Max)', 'Temp(Mean)', and 'Temp(Min)'
Target: 'Yield'
Regression Models: Random Forest, Linear Regression, Gradient Boosting

For each experiment, the dataset was randomly split into training (80%) and testing (20%) sets, ensuring a balanced representation of data across both. This split facilitated the evaluation of model generalizability and predictive accuracy on unseen data.

### 2.3 Model training and evaluation

To train and evaluate the predictive performance of machine learning regression models for crop yield prediction, we followed a systematic approach. Initially, we prepared the dataset by handling missing values, encoding categorical variables, and splitting it into training and testing sets to enable robust model evaluation. Subsequently, we selected the regression algorithms, including Random Forest, Linear Regression and Gradient Boosting. These models were trained using the training dataset to learn the underlying patterns between input features related to wheat production and the target variable, crop yield.

Following the model trainings, we evaluated the performance of each trained model using the testing dataset. Key performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared ($R^2$) were computed to assess predictive accuracy. Also, the study has followed cross-validation techniques to validate the models' robustness and mitigate overfitting risks. Through the

hyperparameter tuning process, the best models were identified for both Random Forest and Gradient Boosting based on their mean squared error (MSE) performance. By achieving a reduced MSE, Random Forest's optimized model proved to be successful in identifying relationships within the dataset. Despite being optimized, Gradient Boosting produced a higher MSE than Random Forest. Both models demonstrated how adjusting important parameters like max_depth, learning_rate, and n_estimators may have a big impact on how well they predict outcomes. By systematically training and evaluating machine learning regression models, we aimed to identify the most effective algorithm for crop yield prediction, facilitating informed decision-making in agricultural planning and management.

### 2.4 Evaluation of model

Model performance was evaluated using the following metrics:

Mean Squared Error (MSE): Measures the average of the squares of the errors between the actual and predicted values.

MSE is calculated using the formula in Eq. (1):

$$MSE = \frac{1}{p}\sum_{i=1}^{p}(m_i - \hat{m}_i) \quad (1)$$

where, $m_i$ is the actual value, $\hat{m}_i$ is the predicted value, and $p$ is the no. of observations.

Root Mean Squared Error (RMSE): It is the square root of Mean Square Error, that provides an error metric in the same units as the target variable. RMSE calculated using the formula in Eq. (2):

$$RMSE = \sqrt{MSE} \quad (2)$$

R-squared ($R^2$): It's the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated using the formula in Eq. (3):

$$R^2 = 1 - \frac{\sum_{i=1}^{p}(m_i - \hat{m}_i)^2}{\sum_{i=1}^{p}(m_i - \bar{m}_i)^2} \quad (3)$$

where, $\bar{m}_i$ is the mean of the actual values.

These metrics are widely used in regression analysis to assess model accuracy and fit [27].

## 3. RESULTS

The analysis conducted in this study meticulously evaluated the predictive performance of three regression models—Linear Regression, Random Forest Regression, and Gradient Boosting Regression—under two distinct conditions determined by the complexity of the feature set. The primary aim was to ascertain the extent to which the number and nature of features influence model accuracy in predicting crop yields. The experiments utilized a basic feature set comprising 'Area' and 'Production' and an expanded feature set that additionally included environmental variables such as 'Rainfall', 'Temp(Max)', 'Temp(Mean)', and 'Temp(Min)'.

### 3.1 Experiment 1 with the limited feature set

Linear Regression with the limited feature set, outperformed the more complex models, exhibiting the lowest Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), alongside the highest R-squared ($R^2$), indicating superior predictive accuracy and model fit as seen in Table 1.

**Table 1.** Model performance with limited feature set

| Regression Model | MSE Score | RMSE Score | $R^2$ Score |
|---|---|---|---|
| Linear Regression | 4,316 | 65.7 | 0.98 |
| Random Forest Regression | 38,728 | 196.89 | 0.89 |
| Gradient Boosting Regression | 30,026 | 173.28 | 0.91 |

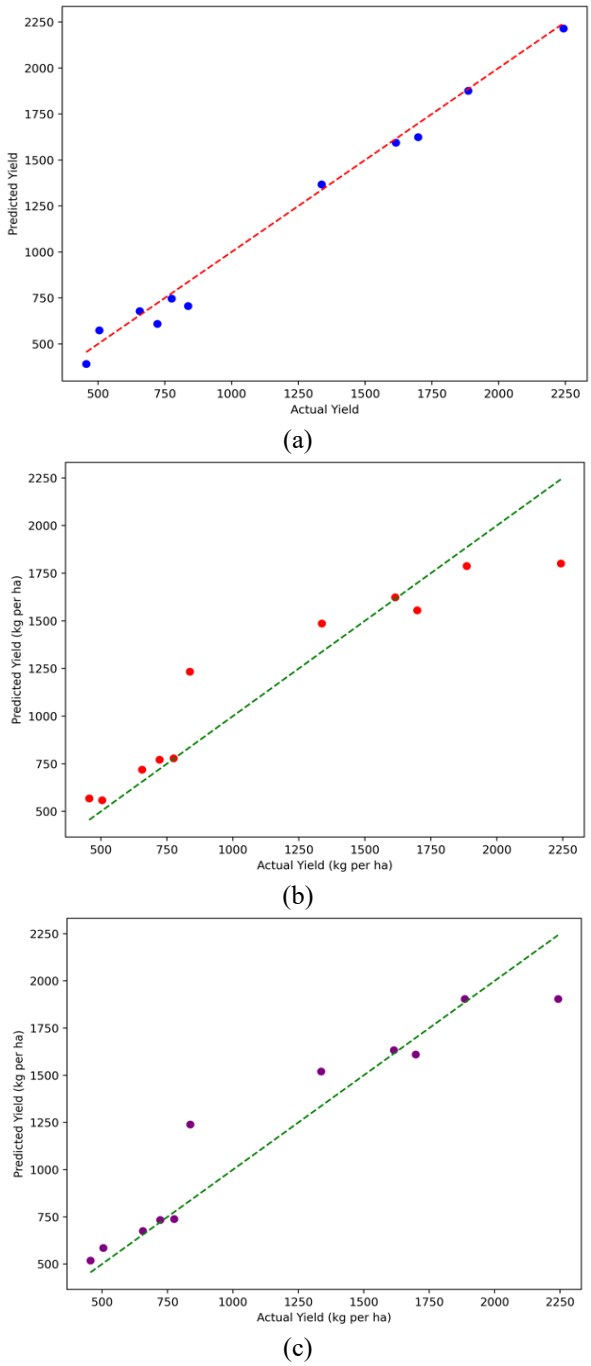
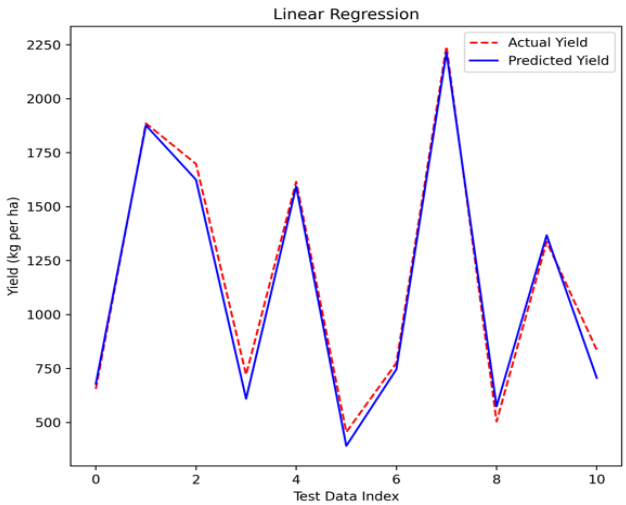


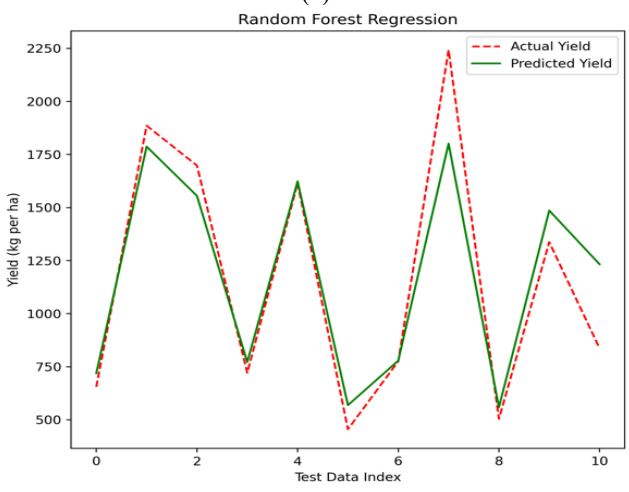

**Figure 2.** (a) Linear Regression - Predicted vs actual crop yield with a limited feature set, (b) Random Forest - Predicted vs actual crop yield with a limited feature set, (c) Gradient Boosting - Predicted vs actual crop yield with a limited feature set

These results highlight Linear Regression's effectiveness in capturing the linear relationship between the selected features and crop yield, suggesting a direct and strong correlation between these variables and the target outcome which can be re-validated from Figure 2(a).

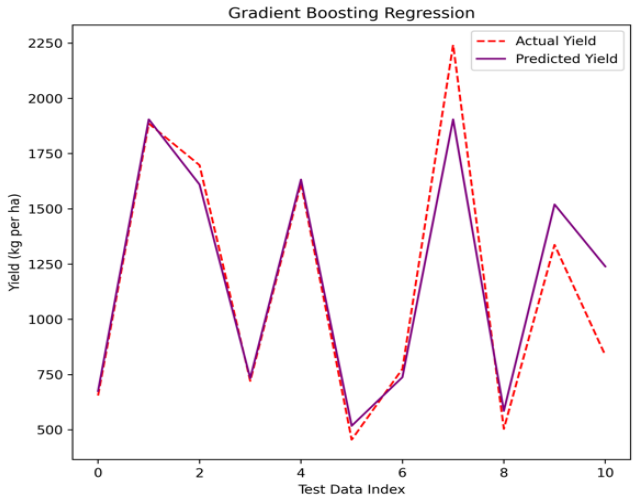
The Linear Regression scatterplot shows a very strong linear relationship between the actual and predicted yields, with the data points closely aligned along the dashed line. This indicates that the Linear Regression model has performed quite well, with predictions that are very close to the actual values. The model seems to accurately capture the underlying pattern in the data without significant overfitting or underfitting. The Random Forest Regression plot from Figure 2(b) also indicates a positive relationship between actual and predicted yields; however, the data points are more spread out around the dashed line compared to the Linear Regression graph. This spread suggests that while the Random Forest model is capturing the general trend, its predictions are less accurate than those of the Linear Regression model. There is greater variability in the results, which might be due to the model capturing more complex patterns in the data or possibly overfitting. From Figure 2(c), the Gradient Boosting Regression plot shows a pattern similar to the Random Forest model, with data points spread out around the dashed line but still following the trend. This indicates that Gradient Boosting is performing similarly to Random Forest in terms of prediction accuracy. The spread of points suggests some degree of prediction error, but the model still captures the general trend of the data.




(a)




(b)




(c)

**Figure 3.** (a) Linear Regression - Line plot of predicted vs actual crop yield with a limited feature set, (b) Random Forest - Line plot of predicted vs actual crop yield with a limited feature set, (c) Gradient Boosting - Line plot of predicted vs actual crop yield with a limited feature set

Figure 3(a) clearly shows that Linear Regression has a good alignment of actual and predicted values of yield using the test data index which refers to the numerical labels assigned to the samples within the test dataset, essentially marking their order or position.

The Random Forest and Gradient Boosting from Figure 3(b) and Figure 3(c), respectively, both have a slight difference between the predicted values and actual values of yield when we have a limited feature set.

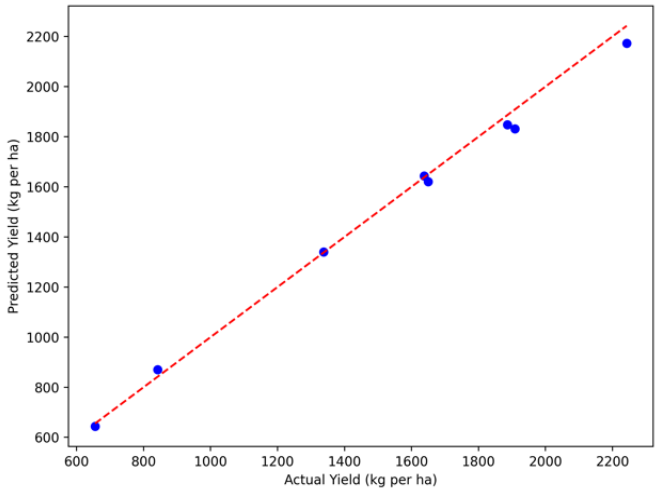
### 3.2 Experiment 2 with expanded feature set

The introduction of the expanded feature set marked a significant shift in model performance, especially for Linear Regression, which saw substantial improvements across all metrics. Table 2 highlights the model performance with expanded feature sets.

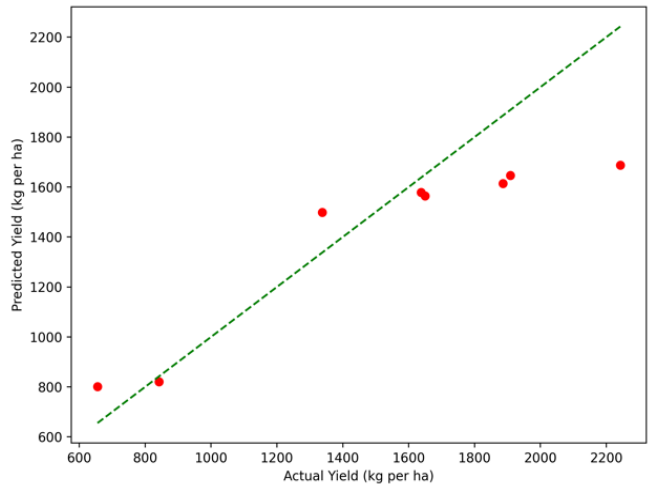
**Table 2.** Model performance with expanded feature set

| Regression Model | MSE Score | RMSE Score | $R^2$ Score |
|---|---|---|---|
| Linear Regression | 1,757 | 41.92 | 0.99 |
| Random Forest Regression | 63,564 | 252.12 | 0.76 |
| Gradient Boosting Regression | 71,630 | 267.64 | 0.72 |

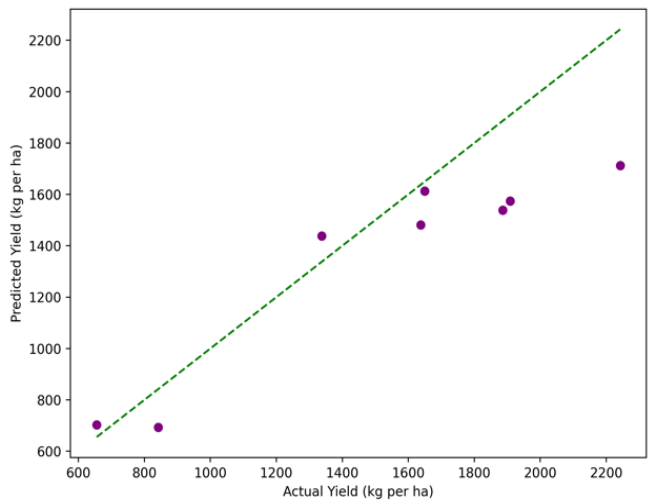
Linear Regression not only maintained its lead but also enhanced its predictive accuracy and model fit with the addition of climate variables, underscoring its adaptability and efficiency in utilizing a broader array of features. In contrast, both Random Forest and Gradient Boosting Regression models experienced a decline in performance, as indicated by increased MSE and RMSE values and reduced $R^2$. This suggests that the incorporation of additional features may have introduced complexities that these models could not as effectively manage, possibly due to overfitting or the dilution of relevant information among a larger set of variables. Figure 3(a), Figure 3(b) and Figure 3(c) also indicated the similar findings from the scatterplots generated with expanded features of datasets.




(a)




(b)



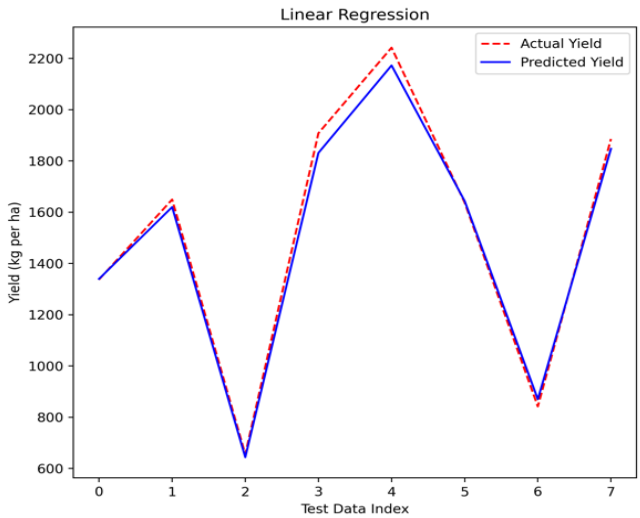

(c)

**Figure 4.** (a) Linear Regression - Predicted vs actual crop yield with expanded feature set, (b) Random Forest - Predicted vs actual crop yield with expanded feature set, (c) Gradient Boosting - Predicted vs actual crop yield with expanded feature set

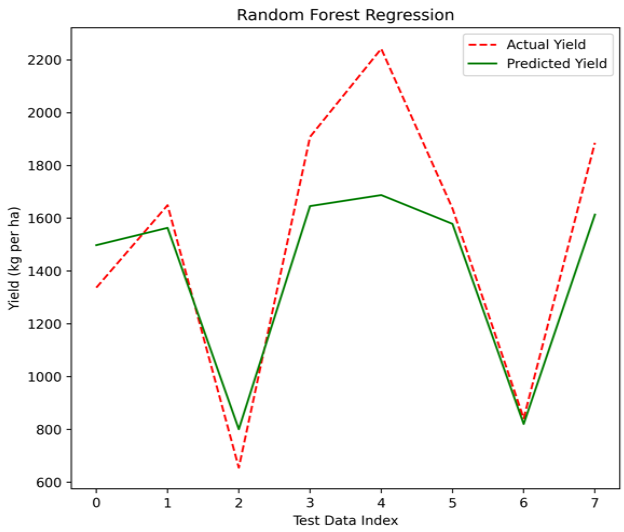
The Linear Regression plot in Figure 4(a) shows data points that are very close to the dashed line of perfect prediction, indicating that the model's predictions are highly accurate. The tight clustering of points around this line suggests that Linear Regression has effectively utilized the additional features to provide precise yield estimates. The model has responded well to the complexity added by the expanded feature set, maintaining a strong linear relationship between predicted and actual yields. Figure 4(b) shows the Random Forest Regression plot shows that the data points are more dispersed around the dashed line compared to the Linear Regression plot. This dispersion implies that the predictions made by the Random Forest model are less accurate than those of the Linear Regression model. The spread of points could be a sign of the model's struggle to fully capitalize on the additional information provided by the expanded feature set, leading to greater variability in the predicted values. Similarly, the Gradient Boosting Regression plot on Figure 4(c) indicates that the model has not performed as well as Linear Regression. The data points are scattered further from the line of perfect prediction, suggesting that the model is less precise in its predictions. This might be due to overfitting, where the model has become too closely fitted to the training data, affecting its generalization capability.

Figure 5(a) clearly shows that Linear Regression has a good alignment of actual and predicted values of yield using the test data index.

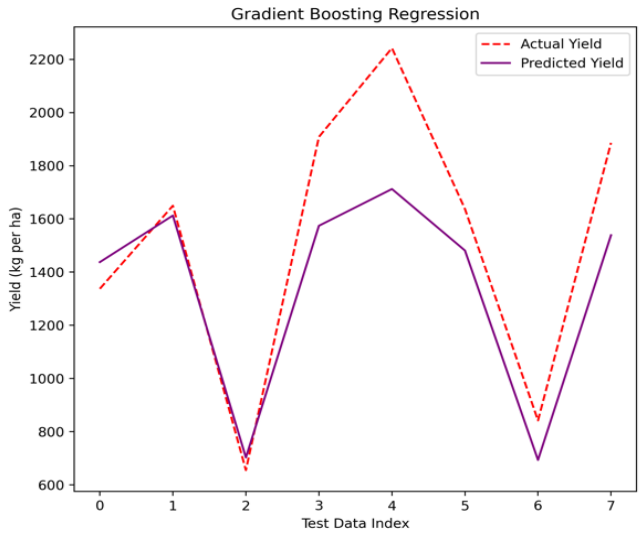
While the Random Forest and Gradient Boosting in Figure 5(b) and Figure 5(c) both have a significant difference between the predicted values and actual values of yield when we have an expanded features set.




(a)




(b)




(c)

**Figure 5.** (a) Linear Regression - Line plot of predicted vs actual crop yield with expanded feature set, (b) Random Forest - Line plot of predicted vs actual crop yield with expanded feature set, (c) Gradient Boosting - Line plot of predicted vs actual crop yield with expanded feature set

## 4. DISCUSSION

The varying responses of regression models namely Random Forest, Linear Regression, and Gradient Boosting to different feature sets highlight the importance of aligning model selection with feature selection strategies. The superior performance of the Linear Regression model across both feature sets underscores the effectiveness of simpler models. The notable improvement in predictive accuracy observed with the addition of environmental variables to Linear Regression emphasizes the model's adaptability and its potential to leverage a diverse range of features. On the other hand, the declined performance of the Gradient Boosting and Random Forest models with the expanded feature set serves as a caution, suggesting that adding more features to such complex model does not always improve their performance. Despite their theoretical capability to model intricate relationships, these models showed signs of overfitting and struggled to incorporate added complexity effectively, as indicated by increased MSE and RMSE, and decreased $R^2$.

Linear Regression, with its simplicity, may avoid overfitting to less important or noisy features, which can sometimes hinder more advanced models. This could suggest that the fundamental patterns in the dataset are simple enough for Linear Regression to accurately identify. However, models like Random Forest or Gradient Boosting may become less predictive due to their increased complexity, which could lead them to concentrate on irrelevant details. This shows that more complicated models may struggle to maintain robust and dependable performance in scenarios when feature interactions are weak or noisy.

This highlights the critical role of thoughtful feature selection and model tuning, challenging the notion that more complex models are inherently superior. The contrasting performances also underscore the importance of aligning model complexity with the dataset's characteristics and the need for a balanced approach to feature selection. These insights not only inform strategic considerations for model selection and feature engineering in agricultural yield prediction but also contribute to broader discussions on optimizing model performance in diverse domains. Moving forward, continued research into tailored feature selection methods is crucial for enhancing model effectiveness in fields where accurate prediction is paramount. Many feature selection strategies can be used to enhance the performance of complex models. These include Recursive Feature Elimination (RFE), which reduces complexity and improves accuracy by methodically eliminating less significant characteristics from the model to keep it focused on the most relevant predictors. Another helpful method is Principal Component Analysis (PCA), which lowers the dataset's dimensionality, simplifies feature interactions, and improves model generalization by concentrating on important components. Furthermore, Domain Knowledge-Based Selection helps minimize unnecessary data, reduce noise, and improve predictive power by using expert insights to choose parameters proven to have the biggest impact on crop productivity. By using these techniques, complicated models may handle big feature sets more effectively, which improves performance in the end.

### 4.1 Analysis of feature set's influence

The contrasting responses of the models to the expanded feature set reveal the subtle influence of feature selection on predictive performance. Linear Regression's ability to significantly improve with the addition of climate variables demonstrates its capacity to leverage a wider data spectrum for enhancing yield predictions. This improvement highlights the model's flexibility and the nature of the relationships between both limited and expanded features and target variable crop yield. The diminished performance of Gradient Boosting and Random Forest models with the expanded feature set may reflect the challenges associated with integrating a diverse range of predictors, where the addition of features does not necessarily translate into better model performance and may, in fact, detract from the models' ability to focus on the most predictive variables.

When the feature sets are expanded, it becomes clear that the additional information can significantly influence model performance. The Linear Regression model demonstrates robustness and appears to benefit from the expanded feature set, showing a strong predictive performance. On the other hand, both Random Forest and Gradient Boosting Regression models do not show the same level of improvement with the expanded features and display a higher level of prediction error. These results indicate that the choice of model is crucial when dealing with different feature sets, and the complexity of the model does not always correlate with improved performance. For this dataset, Linear Regression is able to make the most of the additional features provided, while the more complex models do not necessarily translate the added feature information into more accurate predictions. This could be due to the data's nature and underlying patterns, which a linear model might capture more suitably than models designed for complex, non-linear interactions.

### 4.2 Practical implications for agricultural decision making

According to the results, decision-making in agriculture can benefit greatly from the use of simpler models like Linear Regression, especially when practicality, efficiency, and

interpretability are crucial. Stakeholders in agriculture, such as farmers, legislators, and resource managers, frequently require easily understood and applied actionable insights. Because they are easier to understand, simpler models clearly demonstrate how variables such as production, area, and environmental conditions directly affect crop output. Making educated decisions about risk management, crop planning, and resource allocation depends on this openness.

Simpler models can also be used in areas with limited access to powerful computer equipment because they are computationally efficient. They provide fast and accurate forecasts that have a direct impact on daily agricultural operations when used in real-time decision support systems without consuming a large amount of computing resources. However, sophisticated models, requiring more data, adjustments, and resources, may not always be practical or necessary, especially when a straightforward linear relationship suffices for reliable forecasts. Therefore, in many agricultural contexts, simpler models provide a practical and affordable approach that minimizes computational and interpretive hurdles while yielding dependable results.

## 5. CONCLUSION

The findings of this research shed light on the intricate relationship between feature selection and predictive accuracy in the realm of crop yield prediction using machine learning regression algorithms. Through meticulous experimentation with Random Forest, Linear Regression, and Gradient Boosting models across varying feature sets, we gained valuable insights into the trade-offs between model complexity and predictive performance in agricultural data analysis. Our study challenges the common assumption that complex models inherently outperform simpler ones. Surprisingly, Linear Regression models consistently surpassed their more intricate counterparts in predictive accuracy, particularly when provided with a comprehensive feature set. This underscores the importance of judicious feature selection and highlights the efficacy of linear models in capturing essential patterns in agricultural data.

The study emphasizes how important feature selection is when utilizing machine learning algorithms to estimate crop productivity. The results of the study show that, even with its simplicity, Linear Regression consistently performed better than the more intricate models of Gradient Boosting and Random Forest, especially when a large feature set was used. The model's steady performance can be attributed to the dataset's predominantly linear relationships, which the model handles effectively without overfitting or being impacted by additional complexity.

In Experiment 1, where limited features such as 'Area' and 'Production' were considered, Linear Regression exhibited superior predictive accuracy compared to Random Forest and Gradient Boosting models, that is justified by its higher R-squared ($R^2$) value. Specifically, Linear Regression achieved an $R^2$ of 0.988, indicating that approximately 98.8% of the variance in crop yield could be explained by the model using only these basic features. In contrast, all three regression models achieved lower $R^2$ values, suggesting a weaker ability to capture the underlying dynamics of crop yield with the limited feature set. The subsequent expansion of the feature set in Experiment 2, incorporating climate variables such as rainfall and temperature metrics, further accentuated the prowess of Linear Regression. The model achieved an even higher $R^2$ value of 0.993, signifying an improved ability to explain approximately 99.3% of the variance in crop yield with the additional features. This substantial increase in $R^2$ reinforces the effectiveness of Linear Regression in harnessing and effectively utilizing additional information for crop yield prediction. While the decline of performance of complex models suggests challenges in adapting to increased feature complexity and points out the importance of feature relevance in agricultural data.

The MSE, RMSE and $R^2$ values, obtained from both experiments quantitatively illustrate the superiority of Linear Regression in capturing crop yield dynamics, particularly when provided with a comprehensive feature set. The significant improvements in $R^2$ observed with the expanded feature set underscore the effectiveness of Linear Regression in harnessing additional information. Conversely, the declining $R^2$ values of Random Forest and Gradient Boosting models emphasize the challenges associated with increased feature complexity. The study underscores the critical role of feature selection in shaping the predictive accuracy of regression models for crop yield prediction. While complex ensemble methods offer versatility, the simplicity and interpretability of Linear Regression models make them a compelling choice for agricultural data analysis.

The practical implications of using simpler models like Linear Regression in agricultural yield prediction suggest that, in many cases, simplicity and interpretability can outweigh complexity. The application of less sophisticated models, such as Linear Regression, to agricultural yield prediction suggests that simplicity and interpretability might often be more important than complexity. Less complex models are simpler to use, involve less processing power, and provide simpler tools for decision-making. These qualities are especially important in agricultural environments where immediate and useful insights are required. The study emphasizes how important feature selection is to model performance and how it may help make simpler models perform on par with or even better than more complicated ones provided relevant features are carefully chosen. These results highlight for practitioners how crucial it is to match model selection to the type of data and the issue at hand.

## REFERENCES


[1] Outlook, O.F.A. (2013). OECD/Food and Agriculture Organization of United Nations.

[2] Shah, A., Dubey, A., Hemnani, V., Gala, D., Kalbande, D.R. (2018). Smart farming system: Crop yield prediction using regression techniques. In Proceedings of International Conference on Wireless Communication: ICWiCom 2017, pp. 49-56. https://doi.org/10.1007/978-981-10-8339-6_6

[3] Shyamala, K., Rajeshwar, I. (2020). Enhanced Gradient Boosting regression tree for crop yield prediction. International Journal of Scientific and Technological Research, 9(3): 1651-1654.

[4] Charoen-Ung, P., Mittrapiyanuruk, P. (2018). Sugarcane yield grade prediction using random forest and Gradient Boosting tree techniques. In 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE), Nakhonpathom, Thailand, pp. 1-6. https://doi.org/10.1109/JCSSE.2018.8457391

[5] Puligudla, P., Karthik, K.S., Kumar, K.N., Thirugnanam, M. (2020). Prediction of crop yield using Gradient Boosting. Journal of Xi'an University of Architecture & Technology, 12(XI): 1006-7930.

[6] Yasaswy, M.K., Manimegalai, T., Somasundaram, J. (2022). Crop Yield Prediction in Agriculture Using Gradient Boosting Algorithm Compared with Random Forest. In 2022 International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates, pp. 1-4. https://doi.org/10.1109/ICCR56254.2022.9995829

[7] Huber, F., Yushchenko, A., Stratmann, B., Steinhage, V. (2022). Extreme gradient boosting for yield estimation compared with deep learning approaches. Computers and Electronics in Agriculture, 202: 107346. https://doi.org/10.1016/j.compag.2022.107346

[8] Liang, H., Sun, X., Sun, Y., Gao, Y. (2017). Text feature extraction based on deep learning: A review. EURASIP Journal on Wireless Communications and Networking, 2017: 211. https://doi.org/10.1186/s13638-017-0993-1

[9] Miao, J., Niu, L. (2016). A survey on feature selection. Procedia Computer Science, 91: 919-926. https://doi.org/10.1016/j.procs.2016.07.111

[10] Bengio, Y., Courville, A., Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8): 1798-1828. https://doi.org/10.1109/TPAMI.2013.50

[11] Deng, X., Li, Y., Weng, J., Zhang, J. (2019). Feature selection for text classification: A review. Multimedia Tools and Applications, 78(3): 3797-3816. https://doi.org/10.1007/s11042-018-6083-5

[12] Cai, J., Luo, J., Wang, S., Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300: 70-79. https://doi.org/10.1016/j.neucom.2017.11.077

[13] Chen, R.C., Dewi, C., Huang, S.W., Caraka, R.E. (2020). Selecting critical features for data classification based on machine learning methods. Journal of Big Data, 7(1): 52. https://doi.org/10.1186/s40537-020-00327-4

[14] Babatunde, O.H., Armstrong, L., Leng, J., Diepeveen, D. (2014). A genetic algorithm-based feature selection. International Journal of Electronics Communication and Computer Engineering, 5(4): 899-905.

[15] Kumar, R., Singh, M.P., Kumar, P., Singh, J.P. (2015). Crop selection method to maximize crop yield rate using machine learning technique. In 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Avadi, India, pp. 138-145. https://doi.org/10.1109/ICSTM.2015.7225403

[16] Shekoofa, A., Emam, Y., Shekoufa, N., Ebrahimi, M., Ebrahimie, E. (2014). Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithms: A new avenue in intelligent agriculture. PloS ONE, 9(5): e97288. https://doi.org/10.1371/journal.pone.0097288

[17] Rosenberg, M.S. (2010). A generalized formula for converting chi-square tests to effect sizes for meta-analysis. PloS ONE, 5(4): e10059. https://doi.org/10.1371/journal.pone.0010059

[18] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H. (2017). Feature selection: A data perspective. ACM Computing Surveys (CSUR), 50(6): 1-45. https://doi.org/10.1145/3136625

[19] Chandrashekar, G., Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1): 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[20] Montgomery, D.C., Peck, E.A., Vining, G.G. (2021). Introduction to Linear Regression analysis. John Wiley & Sons.

[21] Liu, Y., Wang, Y., Zhang, J. (2012). New machine learning algorithm: Random forest. In Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, pp. 246-252. https://doi.org/10.1007/978-3-642-34062-8_32

[22] Ali, Z.A., Abduljabbar, Z.H., Taher, H.A., Sallow, A.B., Almufti, S.M. (2023). Exploring the power of eXtreme Gradient Boosting algorithm in machine learning: A review. Academic Journal of Nawroz University, 12(2): 320-334.

[23] Miralles-Garcia, J.L. (2023). Challenges and opportunities in managing peri-urban agriculture: A case study of L'Horta de València, Spain. International Journal of Environmental Impacts, 6(3): 89-99. https://doi.org/10.18280/ijei.060301

[24] Caka, F. (2022). Prospects for mainstreaming urban agriculture in Kosovo in support of sustainable urban development. International Journal of Environmental Impacts, 5(1): 23-37. https://doi.org/10.2495/EI-V5-N1-23-37

[25] ICRISAT. (2024). District level data for indian agriculture by ICRISAT 2024. http://doi.org/10.17616/R3D90M

[26] Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., et al. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth System Science Data, 13(9): 4349-4383. https://doi.org/10.5194/essd-13-4349-2021

[27] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning. New York: Springer. https://doi.org/10.1007/978-3-031-38747-0