European Journal of Agronomy 157 (2024) 127193

Contents lists available at ScienceDirect

European Journal of Agronomy

journal homepage: www.elsevier.com/locate/eja

Evaluating machine learning models and identifying key factors influencing spatial maize yield predictions in data intensive farm management

S. Maseko [a,*], M. van der Laan [a,b], E.H. Tesfamariam [a], M. Delport [c], H. Otterman [c]

[a] Department of Plant and Soil Sciences, University of Pretoria, Private Bag X20 Hatfield, Pretoria 0028, RSA
[b] Agricultural Research Council – Natural Resources and Engineering, Private Bag X79, Pretoria 0001, RSA
[c] Bureau for Food and Agricultural Policy (BFAP), Agri Hub Office Park, RSA

ARTICLE INFO

Keywords:
Precision agriculture
Random Forest
Yield limiting factors
Yield variability

ABSTRACT

Understanding the relationships between crop yields, soil properties, weather patterns and input applications is important for optimizing agricultural production. Data variation analysis using statistical and machine learning (ML) approaches can help identify and understand the practices that optimize yield. The objectives of this study were (i) to evaluate the predictive accuracy of selected ML models for estimating grain yields in on-farm maize (*Zea mays* L.) trials with different combinations of seeding and fertilizer rates in a commercial field, and (ii) to investigate the ability of ML models to assist in identifying yield-limiting factors in the same field. Multiple linear regression, multilayer perceptron, decision tree, and random forest (RF) ML models were trained and tested using crop management and soil from a data-intensive farm management (DIFM) trial and remotely sensed data. The dataset consisted of multiple subplot treatment observations of crop management, soil properties and normalized difference vegetation index (NDVI), linked to final grain yield for the 2019/2020 and 2020/2021 seasons. The RF had the best combination of high correlation (R^2 = 0.69 and 0.80) and low error (MAPE = 5.4 and 8.4% and RMSE = 0.69 and 0.95 t ha^{-1}) when compared to other models for both seasons. Feature importance analysis revealed that urea application was consistently the most critical variable and explained yield variations to the greatest extent, whereas soil phosphorus (P), plant population, and sodium in 2020, and soil P, soil pH, clay content, and plant population in 2021 emerged as the most influential factors for explaining yields. This study concluded that the RF model was the best for spatial yield predictions using DIFM trial datasets. There was also variability between seasons in yield limiting factors resulting from temporal variations in growing conditions. To effectively apply insights from yield prediction models, it is crucial that the variables incorporated into these models have a significant connection to yield and the findings can be translated into actionable management decisions. The DIFM trials combined with ML can play an important role in advancing the field of precision agriculture by providing valuable insights into the complex interactions between crops, soils, and management practices, and identifying new opportunities for improving crop yields and environmental sustainability.

1. Introduction

Yield prediction research is essential in current agricultural systems since it serves as a key point of reference for farm management during planning, agrotechnological investment intervention, and preharvest procedures. Optimizing crop yields in every part of the field is a key objective of precision agriculture (PA). The ability to predict crop yield is a valuable tool for making informed management decisions and implementing PA (Taylor et al., 2007, Bishop et al., 2015). However, providing reliable predictions and recommendations can be challenging due to the multitude of factors that influence the required inputs for a field during a given growing season. To achieve maximum yield through PA management practices (Kaspar et al., 2004), it is imperative to first recognize the dominant spatial factors and comprehend their interdependent connections. Although the influence of weather conditions, such as precipitation, on crop yield can be more simply determined, the impact of spatial factors, such as soil and terrain diversity, is often more challenging to pinpoint and quantify (Kravchenko and Bullock, 2000, Jones et al., 2022).

One viable solution to this challenge entails conducting a thorough and targeted analysis of data sourced from on-farm research trials. The advent of advanced technologies, such as variable-rate planters and

* Corresponding author.

https://doi.org/10.1016/j.eja.2024.127193
Received 10 November 2023; Received in revised form 19 April 2024; Accepted 19 April 2024
Available online 27 April 2024

applicators, has made site-specific data acquisition more cost-effective and has opened up avenues for the development of new decision support systems that can handle more intricate and data-intensive tasks than the conventional systems currently in use (Nyéki et al., 2017). The last two decades have seen a rapid expansion of on-farm research, especially in developing countries, owing to the increased adoption and use of PA technologies (Kyveryga, 2019). On-farm precision experimentation is a type of on-farm experimentation (OFE) that enables the collection of large amounts of crop and soil data in a relatively short period of time, and can be of special interest to large-scale farmers aiming to improve site-specific crop input management (Bullock et al., 2019). A multidisciplinary research project initiative called data-intensive farm management (DIFM) (Bullock et al., 2019) enables researchers to develop data-intensive, site-specific input management advice and collaborate with farmers to provide guidance on how to make OFE systems maximise their return on investment. This data collection may result in increased cost-effectiveness, and if adequately analyzed, can have a huge potential to refine the current knowledge of agricultural systems.

To fully utilize big data analytics in the field of agriculture, it is necessary to advance scientific methodologies. Thus, the application of artificial intelligence (AI), particularly machine learning (ML) techniques, is highly relevant. As the amount of large geo-referenced on-farm data becomes increasingly available, there is a need for analytical AI frameworks that can provide crop management recommendations and yield predictions. With the help of large datasets, it is now possible to conduct inductive research methodologies and investigate the complex interactions between crop management practices, environmental factors, and yield. This approach offers a practical and effective method to conduct large-scale agronomic research (Silva et al., 2020). By leveraging AI-powered data analysis, forecasting, and prediction techniques such as those outlined by Basso et al. (2016), farmers can make informed decisions that increase productivity and ensure the success of their crops.

Over the last decade, ML approaches have become increasingly prevalent in agriculture because of their ability to effectively address complex agricultural problems and nonlinear relationships leading to more accurate results (Pantazi et al., 2016, Tantalaki et al., 2019). One area that has seen particular growth is the use of ML to forecast crop yields, although the research community still debates the most effective techniques for various data types and situations (Ransom et al., 2019, Van Klompenburg et al., 2020). Precision agriculture data, combined with ML techniques has proven to be particularly helpful in estimating crop biomass and yield (Näsi et al., 2018, Li et al., 2020), thanks to the ability of ML to handle large datasets with numerous variables, such as those created using PA equipment with data collection capabilities (Li et al., 2022). The advancement of Al applications has led to a broad range of applications for ML in agriculture, benefiting data gathering and selection to improve agricultural practices (Nawar et al., 2017).

Studies that have examined the predictive accuracy of regression and ML models have primarily focused on meteorological and farm management attributes (Basso and Liu, 2019, Nayak et al., 2022), soil nutrients, and topographic data (Burdett and Wellen, 2022). It has been found that the creation of a yield prediction model using regression models can be complicated by the complex interactions between maize yields and spatially variable soil and management parameters as predictive variables (Jaynes and Colvin, 1997, Jones et al., 2022). The AI applications in the agricultural sector are still in the developmental stage, and the challenges and implications of different ML methods remain unclear (Lassoued et al., 2021).

The continued success of agricultural production relies heavily on farmers implementing advanced techniques at every level of crop production to increase yield per unit area. To aid in this endeavor, farmers can be assisted by an accurate model for crop yield prediction, which enables them to make informed decisions regarding when and how to produce certain crops. By predicting the yield of a specific site, farmers can also adjust the application of farm inputs, such as fertilizers, based on anticipated crop and soil needs. The objectives of this study were to use ML algorithms to: (i) assess the predictive ability of ML models using DIFM datasets and determine the best-performing ML model, and (ii) investigate the potential of ML models to identify optimal input rates and limiting factors in a spatially variable field. We hypothesized that the performance of ML models in spatial predictions would be enhanced by utilizing spatially representative DIFM datasets for training and testing. The findings of this study will significantly contribute to the understanding of the predictive capabilities of various ML models and the importance of soil and agronomic management attributes in predicting maize yields.

2. Materials and Methods

2.1. Experimental site

The study was conducted in Henneman, in the Free State province of South Africa (2751'16"S, 2701'15"E, 1 412 m.a.s.l.). This region is characterized by commercial medium- to large-scale farming of crops and livestock. It has a cold semi-arid climate, with hot and wet summer days, cooler, dry winters, an average temperature of approximately 18°C, and an annual average rainfall of approximately 600 mm yr^{-1}. Seasonal rainfall usually starts in October and ends in April, with more than 80% of rainfall occurring from December to March.

2.2. Trial design and management

This study was based on datasets collected in a DIFM maize field trial conducted in the 2019/2020 and 2020/2021 growing seasons. The DIFM trials are designed to generate data for localized crop responses to site-specific input factors (Bullock et al., 2019). The experiment had two management input factors: seeding rate (S) and nitrogen fertilizer application rate (urea). The treatments were set up in a completely randomized factorial design, with nine seeding rate factors (10 000, 15 000, 18 000, 21 500, 27 000, 32 000, 38 000, 44 000 and 50 000 seeds ha^{-1}) for each of the two seasons. There were eight levels of urea fertilizer rates (90, 120, 150, 170, 200, 225, 250, and 270 kg ha^{-1} urea in 2019/2020, and 105, 120, 150, 170, 200, 225, 250, 270, and 300 kg ha^{-1} urea in 2020/2021) assigned randomly throughout the field in each of the two seasons (Fig. 1). The procedure used for fertilizer application was an initial uniform 200 kg ha^{-1} 15.10.6 (31) NPK mixture applied throughout the field at planting. The N variation treatments were then implemented by applying different urea rates, banded 15 cm offset to the row and 10 cm deep. The standard practice of the farmer was to apply 18 000 seeds ha^{-1} and 224 kg ha^{-1} of urea. The plots were designed to be 15 m wide and 73 m long. The plot width is typically determined by the width of the planter used, such that every plot hosted one pass of the planter and one pass of the fertilizer applicator, the 73 m length was determined by the distance it takes for the planter to change input application rates as determined by the trial design. All treatments were implemented in the field using variable-rate-enabled seed planters and fertilizer applicators. A medium-season cultivar from the seed company Dekalb (DKC 78–77 BR) was used consistently throughout the field over the two seasons. A rate of 18 000 seeds ha^{-1} was assigned to a buffer zone around the perimeter of the trial, and observations from the buffer zone were not included in the subsequent analysis.

2.3. Data collection and processing

Before the input application from the precision planter and yield data from the yield monitor were used for training the ML algorithms, data-cleaning procedures as described below were applied to the raw data. The final dataset used consisted of maize grain yield (20% moisture content) as a dependent variable and 24 georeferenced management,

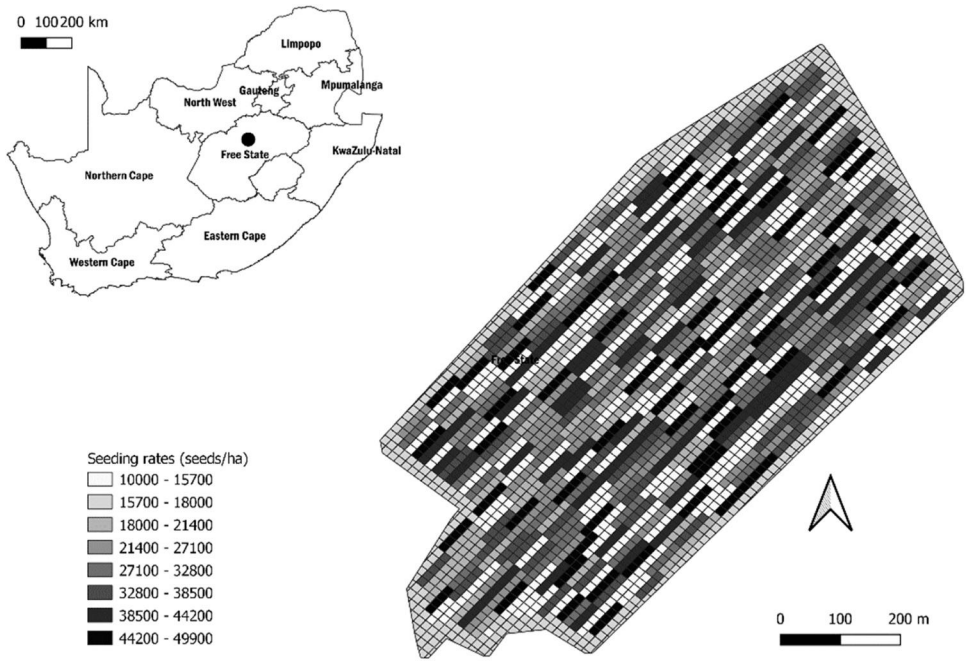

Fig. 1. The maize (*Zea mays* L.) field in which the study was conducted, located in Henneman, Free State province, South Africa. The figure illustrates the various seeding rate treatments used in the experiment.

soil, and remotely sensed data as independent variables (see Table 1 for all list). The data were processed to represent multiple small plots across the field, with each plot having a unique yield value linked to management, soil, and remotely sensed data.

2.3.1. Yield data

The yield data were collected using a calibrated yield-monitoring system mounted on the combined harvester, which recorded the yield data every two seconds during the harvesting process. The farmer utilized a strategic harvesting technique in which, during every other pass, the combined harvester traversed through the center of the plot. As a result, yield data were gathered solely from the middle 50% of each plot.

Table 1
The agronomic management, soil and remotely sensed variables used in model development.

	Variable name	Description	Units
Agronomic	Plant_pop	Plant population	seeds ha^{-1}
	Urea	Urea application	kg ha^{-1}
Soil	pH_top	Soil pH in topsoil	-
	Bray_top	Phosphorus in topsoil	mg kg^{-1}
	K_top	Potassium in topsoil	mg kg^{-1}
	Mg_top	Magnesium in topsoil	mg kg^{-1}
	Na_top	Sodium in topsoil	mg kg^{-1}
	S_top	Sulphur in topsoil	mg kg^{-1}
	Clay_top	Clay content in topsoil	%
	Bray_sub	Phosphorus in sub soil	mg kg^{-1}
	K_sub	Potassium in sub soil	mg kg^{-1}
	Mg_sub	Magnesium in sub soil	mg kg^{-1}
	Na_sub	Sodium in sub soil	mg kg^{-1}
	S_sub	Sulphur in sub soil	mg kg^{-1}
	Clay_sub	Clay content in sub soil	%
	Soil_d	soil depth	m
Remotely sensed	11DAE_ndvi	NDVI at 11 DAE	
	25DAE_ndvi	NDVI at 25 DAE	
	60DAE_ndvi	NDVI at 60 DAE	
	85DAE_ndvi	NDVI at 85 DAE	unitless
	100DAE_ndvi	NDVI at 100 DAE	
	110DAE_ndvi	NDVI at 110 DAE	
	120DAE_ndvi	NDVI at 120 DAE	
	135DAE_ndvi	NDVI at 135 DAE	

The data cleaning procedures for maize yield in a DIFM trial executed in this study were discussed in detail by Bullock et al. (2019). Briefly, raw 'as applied' and harvest data were retrieved directly from the variable rate applicators and yield monitors. Raw data were cleaned to remove observations with extreme yields or applied rates ('outliers'). Additionally, data points were excluded from the headlands due to varying sun exposure, fluctuations in machinery driving speed, and the possibility of application overlaps, which made the data less reliable. The DIFM strategy also involved the placement of about 10 m 'transitional buffer zones' at the end of each plot where the planter could be changing from one application rate to another. The distance between points, swath width, and headings recorded in the raw yield data were used to create yield polygons, and subplots were created by combining yield polygons with similar N rates into groups (yield polygons combined to form subplots approximately 12 m in length). The average value of all the yield points within each subplot was calculated and used in the analysis as a single observational unit. After data cleaning, 5 748 and 3 409 observational units were analyzed for the respective seasons. The first season yielded more data points than the second due to more heterogeneous yields throughout the field in the 2020/2021 season, and some plots had yields that were too low for meaningful analysis in the second.

2.3.2. Soil data

Data on the soil physical and chemical properties were collected by the Omnia fertilizer company for soil analysis, which was conducted on-site before the start of each planting season. The variables measured at each location for both the topsoil (0–0.3 m) and subsoil (0.3–0.6 m) included physical properties such as clay percentage and soil depth as well as chemical properties such as soil pH, potassium (K), Bray P, calcium (Ca), magnesium (Mg), and sodium (Na). The soil was sampled at 62 locations across the field on a 100 m grid (Fig. 2), and five samples were taken at each sampling point approximately 3 m apart. Of the 24 variables used for yield predictions, 14 were primarily focused on the physical and chemical properties of both topsoil and subsoil.

2.3.3. Remotely sensed data

The normalized difference vegetation index (NDVI), which provides

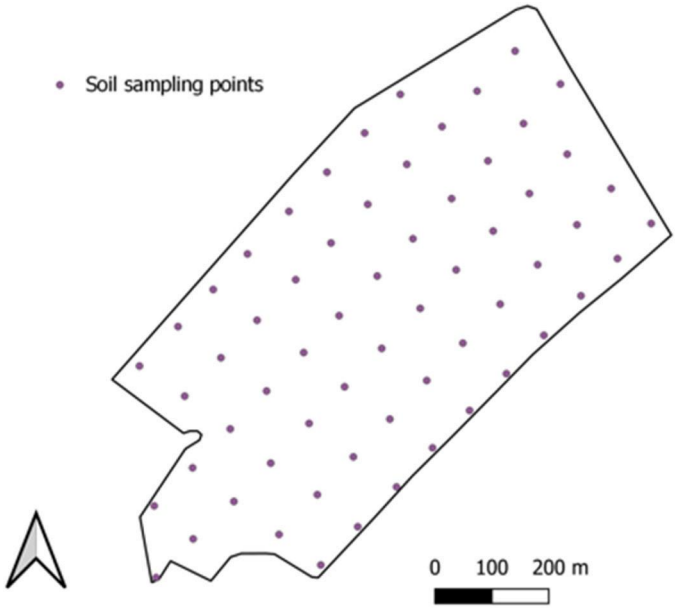

Fig. 2. A map illustrating the locations of the soil sampling points used in the maize (*Zea mays* L.) trial study conducted in Henneman in the Free State province of South Africa.

an indication of canopy development and health, is a highly effective tool for assessing crop yield potential. To incorporate real-time data into our model training and testing, NDVI data were used to track crop growth at various stages during the growing season. NDVI values were calculated using Sentinel 2 A images with a 10 m resolution in QGIS, creating raster files from which the NDVI was calculated for each pixel. These images were captured between 1 November 2019 and 30 April 2020 as well as 20 November 2020 and 30 April 2021, and downloaded through the Corpernicus hub (https://scihub.copernicus.eu). The centroids for each plot from the yield data shapefile were used to sample the NDVI values. The NDVI data were extracted at seven different intervals, beginning 11 days after emergence (DAE) and continuing until the crop reached physiological maturity (135 DAE). The images were carefully selected to focus on the area of interest (the maize field), and only those with less than 5% cloud cover were included in the analysis. Finally, the raster files were sampled using centroids from each yield polygon to extract the NDVI time series for the corresponding yield points from emergence to harvest.

2.4. Machine learning maize yield predictions

The ML models were built using Python Keras libraries in a Google Colaboratory cloud computing environment. The ML algorithms were implemented using a multistep process. The data file included attributes of the soil, agronomic management practices, and remotely sensed data as independent variables, with maize yield as the dependent variable. The processed data were utilized for both model training and testing purposes. The data used for model training and testing were from 2019/2020, 2020/2021, and a merged dataset of the two seasons. The data points used for model training and testing were 4 025, 2 387 and 6 410 observational units for the 2019/2020, 2020/2021 and merged datasets for both seasons, respectively. An 80/20 ratio data split was used for model training and testing for each of the four ML models.

Four ML algorithms (MLR, MLP, DT, and RF) were investigated. The MLR models the relationship between two or more explanatory variables and a dependent variable, assuming a linear relationship. For predicting crop yields, MLR has been a popular technique (Drummond et al., 2003, Van Klompenburg et al., 2020). Multiple linear regression uses least-squares optimization to determine the dependent variable that best fits each independent variable (measured yield). It assumes normality, homoscedasticity, no multicollinearity, and the presence of a linear relationship between the predictors and response variables (James et al., 2013). The MLR model was developed according to Eq. (1).

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots + \beta_n X_n + \varepsilon_i \quad (1)$$

where y_i is the grain yield, β_0 represents bias, $\beta_1 - \beta_n$ are the coefficients of regression, $X_1 - X_n$ are the input variables, and ε_i is the error associated with the ith observation. Multicollinearity was evaluated in the training and testing of the MLR model, which refers to a situation in which two or more explanatory variables in the regression model were highly linearly related. This was evaluated using the variance inflation factor (VIF). A commonly used rule is that if the VIF is less than 5, there is low multicollinearity; between 5 and 10, there is high multicollinearity; and more than 10, the multicollinearity is too high (Kutner et al., 2005). The explanatory variables with the highest VIF were deleted, one at a time, and the model was refitted. This procedure was repeated until all the VIFs were below 5.

The MLP model was created using Keras, a deep learning application programming interface (API) implemented in Python. Because the models in Keras are described as a series of layers, a sequential model was initially created, and then four additional layers were added using the Rectified Linear Unit (ReLu) activation function. An Adams gradient descent optimizer was chosen with default hyperparameters, as tests have shown that this is a good optimizer when used with adaptive learning rates (Ruder, 2016). We implemented a mean squared error (MSE) loss function and a maximum of 500 epochs.

The decision tree (DT) is a supervised learning model that can be used for both classification and regression tasks. It can select an outcome from a tree of potential decisions (Maimon and Rokach, 2014, Perez-Alonso et al., 2017). The tree structure resembles a flowchart and is used to evaluate issues by considering numerous features and attributes. In this study, the Scikit-learn Python module "DecisionTreeRegressor" class was applied, with a maximum depth of 30 trees.

Random Forest is a tree-based ensemble model built on the concept of bagging, which averages final predictions from different training subsets made by sample training data with replacement in an effort to reduce prediction variation (Breiman, 2001). Random forest adds a new feature to bagging by randomly selecting a set of features, building a tree with those features, repeating this process numerous times, and then averaging all the predictions made by the trees (Shahhosseini et al., 2021). The Gini index was used to identify the key characteristics that significantly influenced yield based on various independent variables. This feature selection process is crucial for identifying significant variables that explain yields and could highlight limiting factors in agronomic terms.

2.5. Model evaluation

Performance evaluation measures were used to assess the accuracy of each prediction model and to select the most suitable algorithm for supervised learning regression exercises. To evaluate our ML models, we used the root mean square error (RMSE) and mean absolute error percentage (MAPE). The degree of correlation between predicted and actual values was determined using the coefficient of determination (R^2). These metrics can be used for both regression and classification tasks (Naser and Alavi, 2020). It is generally accepted that the model with the smallest estimation error is the best.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \cdot 100\% \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} \quad (4)$$

where y_i is the measured value, $\hat{y}$ represents the predicted values, $\bar{y}$ mean value of y_i, and n represents the number of observations.

3. Results

3.1. Descriptive statistics

The descriptive statistics of seeding rates, urea application, and grain yield during the 2019/2020 and 2020/2021 seasons are presented in Table 2. Yields varied from 6.8–12.4 t ha^{-1} and 2–13.7 t ha^{-1} in the 2019/2020 and 2020/2021 seasons, respectively. The 2019/2020 season had a higher average yield (9.7 t ha^{-1}) and lower standard deviation (1.2 t ha^{-1}) than the 2020/2021 season (8.6 t ha^{-1} average yield and 2.1 t ha^{-1} standard deviation). Table 2 also shows the descriptive statistics for the soil physical and chemical properties over the two seasons. The pH, K, and Mg levels in the soil were slightly higher during the 2020/2021 season, whereas the P, Na, and S levels were lower than those in the 2019/2020 season. Although these small changes were noticeable, they were statistically insignificant. The chemical properties of the subsoil remained consistent and did not vary between the 2019/2020 and 2020/2021 seasons.

The two seasons were also characterized by differences in the total rainfall received from planting to harvesting, which was 674 mm and 439 mm for the 2019/2020 and 2020/2021 seasons, respectively (Fig. 3). There was a high incidence of tillering in the planted cultivar in various treatments, especially in the combination of low plant populations and high fertilizer rate treatments.

3.2. Correlation analysis

The Pearson correlation analysis results for all the datasets are presented in Fig. 4. The results indicated that the extent of the relationship between yield and individual yield-influencing attributes varied between the seasons. Despite seasonal variations, a positive correlation between crop yield and agronomic management, as well as NDVI at all stages of growth, was evident in both seasons. Urea application had a stronger relationship with yield than the plant population for the two individual seasons and when the two seasons were merged. Plant population and urea application both helped explain the yield differences between the two seasons, with urea application having a higher correlation with yields in the 2020/2021 season than in the 2019/2020 season, whereas the opposite was observed for the plant population, with a higher correlation in the 2019/2020 season than in the 2020/2021 season. The NDVI at 110 and 85 DAE had the highest correlation with yield in the 2019/2020 and 2020/2021 seasons, respectively. Further analysis of the results showed that during the 2019/2020 season, neither Na_sub nor Clay_sub was significant in explaining the yield variation, and the same was true for Plant_pop, K_top, Mg_sub, and Na_sub during the 2020/2021 season ($P > 0.05$). However, all other attributes were found to be significant ($P < 0.05$) in explaining yields in each of the two seasons. The extent to which some soil attributes were related to yield could be negative (pH_top, and S_sub), positive (Bray_top, Bray_sub, K_top, K_sub and clay_sub), or both (Mg_top, Mg_sub, clay_top and Na_sub) over the two seasons. There were no clear positive or negative interpretations of these variables and yields, suggesting that the relationships were nonlinear or that there was no relationship.

After merging the data from both seasons into a single dataset and incorporating monthly rainfall as variables, the results indicated that urea application was more influential than plant population in explaining the yield. Other significant factors contributing to yield were S_top, Bray_top, and pH_top, which demonstrated a negative correlation. Meanwhile, S_sub and Na_sub showed the weakest correlation with the yield. Only Na_sub did not significantly explain the yield variation ($P > 0.05$); all other attributes significantly explained the yield variation ($P < 0.05$) when the two seasons were combined into one dataset.

3.3. Evaluation of ML algorithms for spatial maize yield predictions

The initial model training and testing on data with NDVI included 24 explanatory variables for all four models. However, Bray_sub, 100DAE_ndvi, and 120DAE_ndvi were dropped from the MLR model training in the VIF evaluation step to correct the effects of multicollinearity. There was no variable removed for MLR training and testing without NDVI because all VIF values were below 10. The R^2 values between the actual and predicted yields for the four ML models using datasets with and without NDVI are shown in Fig. 5. An analysis of the overall trend revealed that when NDVI data were factored into the ML models, improved predictive accuracy was achieved in both seasons. The incorporation of NDVI strengthened the linear associations between the yield and yield prediction attributes within the dataset, which were effectively captured by the models in most cases. In contrast, when comparing model performance with data without NDVI, the MLR, DT, and RF exhibited either increased or equal prediction accuracy from the 2019/2020–2020/2021 seasons, while MLP showed a decline in prediction accuracy from the same period when trained and tested without NDVI data. The results of the model predictions using NDVI data showed that MLP had a higher prediction accuracy than MLR in the first season

Table 2
The descriptive statistics of input application, maize yields, top and subsoil physical and chemical properties for 2019/2020 and 2020/2021 seasons.

		2019/2020 season				2020/2021 season			
Variable		**std**	**min**	**mean**	**max**	**std**	**min**	**mean**	**max**
Seeding rate	seeds ha^{-1}	10527	5613	30434	49881	9594	13676	28862	49381
Urea rate	kg ha^{-1}	55	94	181	276	62	104	204	308
Yields	t ha^{-1}	1.2	6.8	9.7	12.4	2.1	2	8.6	13.7
pH_top	-	0	4	4	6	0	4	6	7
Bray_top	mg kg^{-1}	11	25	48	100	10	25	44	92
K_top		29	95	217	360	43	170	277	373
Mg_top		12	43	80	151	10	57	94	127
Na_top		2	4	9	23	4	4	7	36
S_top		3	1	12	31	1	1	2	7
Clay_top	%	2	15	19	25	2	15	19	29
Bray_sub	mg kg^{-1}	8	5	16	39	11	5	16	39
K_sub		25	103	155	247	44	103	154	248
Mg_sub		37	84	141	274	54	84	143	274
Na_sub		1	4	6	12	2	4	6	12
S_sub		4	2	16	26	6	2	16	26
Clay_sub	%	2	20	28	35	3	20	28	35

std: standard deviation, **min**: minimum, **max**: maximum

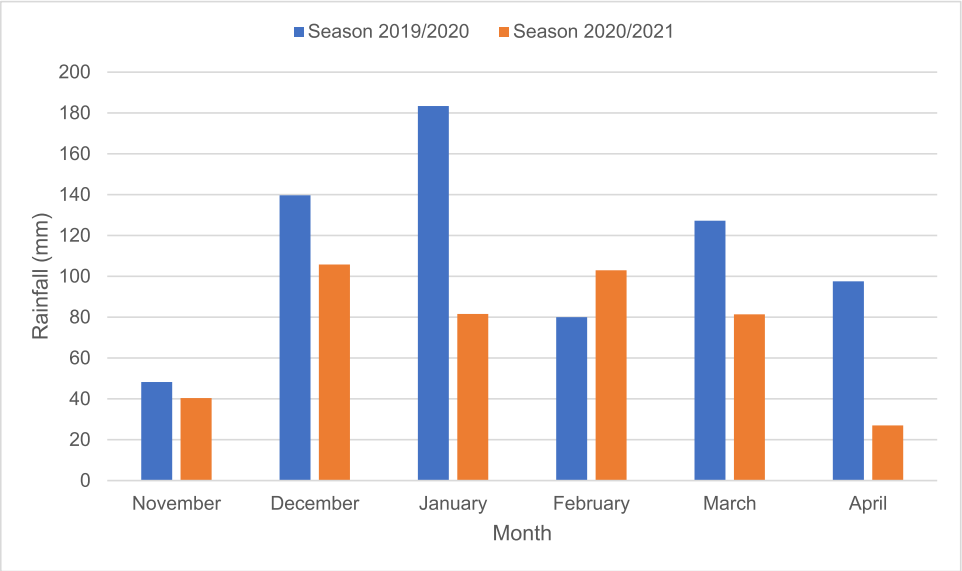

Fig. 3. Rainfall distribution for the two seasons (2019/20 and 2020/21) covering the period from planting to harvesting.

($R^2 = 0.47$ and 0.5 for MLR and MLP), but lower accuracy in the second season ($R^2 = 0.71$ and 0.65 for MLR and MLP) and when the two seasons' data were combined ($R^2 = 0.65$ and 0.62 for MLR and MLP).

The model performance was further evaluated using the MAPE and RMSE, as shown in Table 3. Similar to the R^2 analysis, the inclusion of NDVI reduced the error of the ML predictions, as indicated by the low MAPE and RMSE values with NDVI. The MAPE and RMSE accuracy trends displayed inconsistencies during the 2019/2020 season, with lower values observed for the MLP and RF models without the NDVI. Although model accuracy varied depending on the ML model and dataset used, DT was the least accurate model, with MLP and MLR scoring reasonable accuracies, and RF was the most accurate model when NDVI was included in the training and testing data. The MLR was the least accurate, DT and MLP were reasonably accurate, and RF was the most accurate ML model without NDVI data included in the training and testing datasets.

3.3.1. The important factors for maize yield predictions

The RF model was utilized not only to assess the predictive abilities of the models but also to evaluate the interaction between seeding and fertilizer rate combinations and various attributes within each plot to ascertain the most influential attributes for yield prediction (Fig. 6). As the RF model has emerged as the top performer, this section delves into the implications of the important variables extracted from that model. Although the relative impact of a solitary variable cannot be measured independently of other variables, a measure for assessing the relative importance of factors on prediction outcomes was provided.

The variable importance generated from the RF model on data without NDVI indicated that urea application and plant population, which are variables that the farmer can most easily control, explained 24, 40 and 27% of the yield variation in the first, second, and combined seasons, respectively. The inclusion of NDVI data in model training resulted in urea application and plant population explaining 15, 13 and 14% of the yield variation, respectively, for the three datasets used. Despite the changes in the variables (with and without NDVI) used for model development, soil pH and clay content in the topsoil consistently explained 5–6% and 2–3% of the yield variation, respectively. The NDVI variables had a stronger impact on maize yield predictions than the soil and management variables. This explains the improved model performance when NDVI data were included, as indicated by the higher prediction accuracies in all ML models used. Specific NDVI measurements on different days after emergence demonstrated a significant explanatory power. For example, NDVI at 110 DAE in the 2019/2020 season explained 38% of the yield variation, whereas NDVI at 85 DAE in both the 2020/2021 season and the combined seasons explained 55% and 45% of the yield variation, respectively. The feature importance showed that urea application, plant population, soil pH, and clay content in the topsoil were agronomic management and soil attributes that led to larger information gains on yield variability, whereas subsoil chemical properties explained yield variability the least in the low rainfall season or combined season analysis. In the wet season, however, urea application, plant population, and subsoil properties (subsoil P, Na, K, and clay) explained the yield variability the most. The overall feature importance showed that topsoil attributes dominated the first half, while the subsoil attributes dominated the latter when the two-season dataset was merged.

4. Discussion

This study focused on comparing ML techniques to define the relationship between soil, agronomic management, and remotely sensed data for spatial maize yield predictions, and the identification of influential attributes explaining yields. Gaining insights from data through the application of statistical techniques is important for effectively training ML algorithms. The descriptive statistics of the measured yield revealed that there was a variation in the sub-plot yields both within a single season and across the two seasons. Although the rich dataset presented a wide range of yield values, the spatial structure of the variability was still unclear. A yield distribution heterogeneity similar to that of this field has been reported in other DIFM fields (Trevisan et al., 2021). Bullock et al. (2019) stated that crop yields are a product of natural processes influenced by input management choices (x), field characteristics (c), and weather (z), which can be represented mathematically as $y = f(x, c, z)$. The yield variation between the two seasons can be explained by the differences in the total rainfall received and the overall rainfall distribution from planting to harvesting. The precipitation pattern during the 2019/2020 season was more favourable than that of the 2020/2021 season, with a higher total amount and better-distributed precipitation from planting to harvest. Most months in 2019/2020 recorded higher monthly rainfall, except for February. This discrepancy in February rainfall explains the negative correlation between February rainfall and the crop yield.

While it is relatively simple to determine the effect of weather factors such as precipitation, on crop yield, it is more difficult to accurately

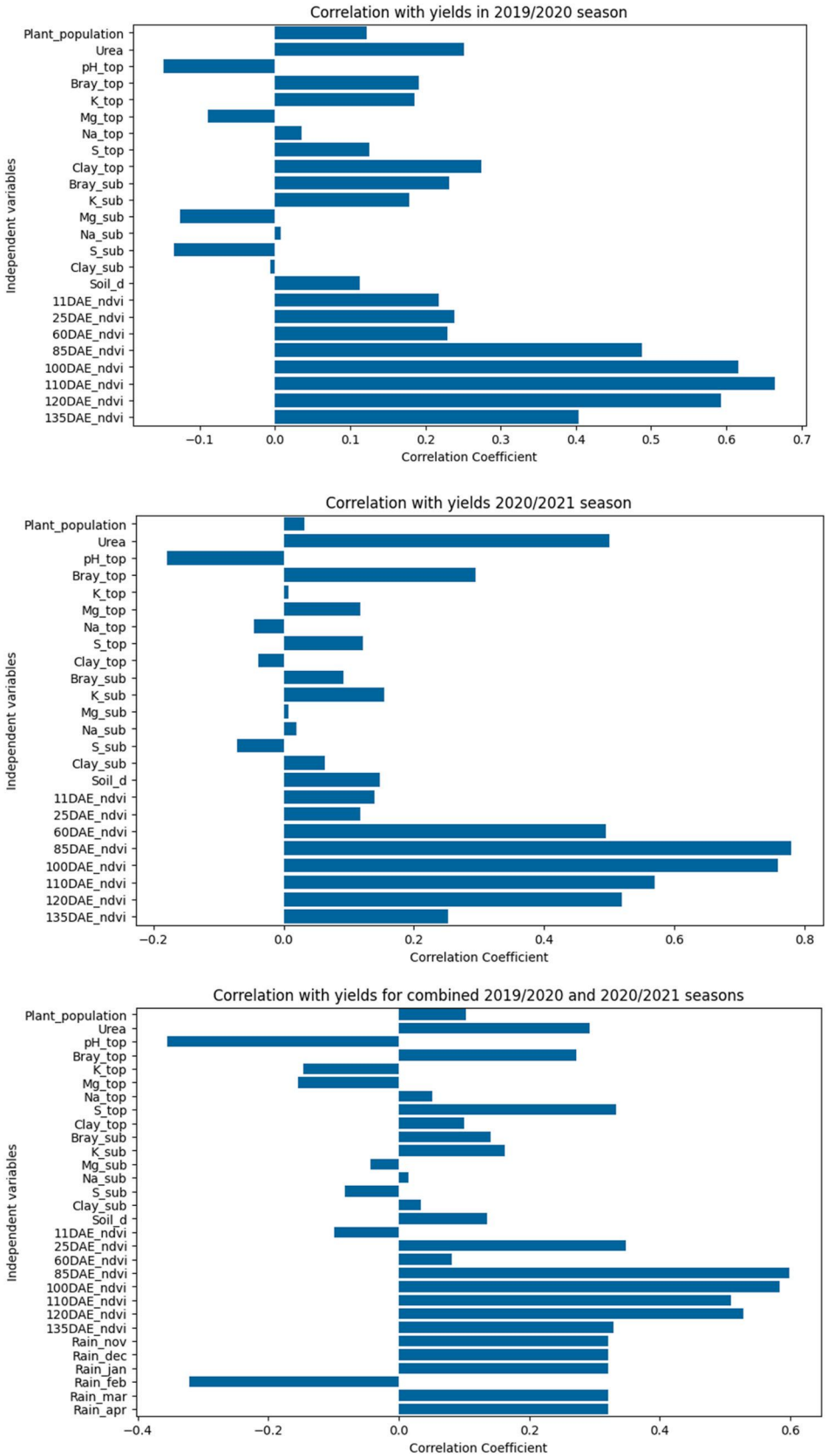

Fig. 4. Pearson correlation analysis for the relationship between agronomic management, soil, remotely sensed, and weather data and maize yields for the 2019/2020 and 2020/2021 seasons, and combined data for the seasons. (Plant_pop: plant population, Urea: urea application, ph_top: soil pH in topsoil, bray_top: phosphorus in topsoil, K_top: potassium in topsoil, Mg_top: magnesium in topsoil, Na_top: sodium in topsoil, S_top: sulphur in topsoil, Clay_top: clay content in topsoil, Bray_sub: phosphorus in sub soil, K_sub: potassium in sub soil, Mg_sub: magnesium in sub soil, Na_sub: sodium in sub soil, S_sub: Sulphur in sub soil, Clay_sub: Clay content in sub soil, Soil_d: soil depth).

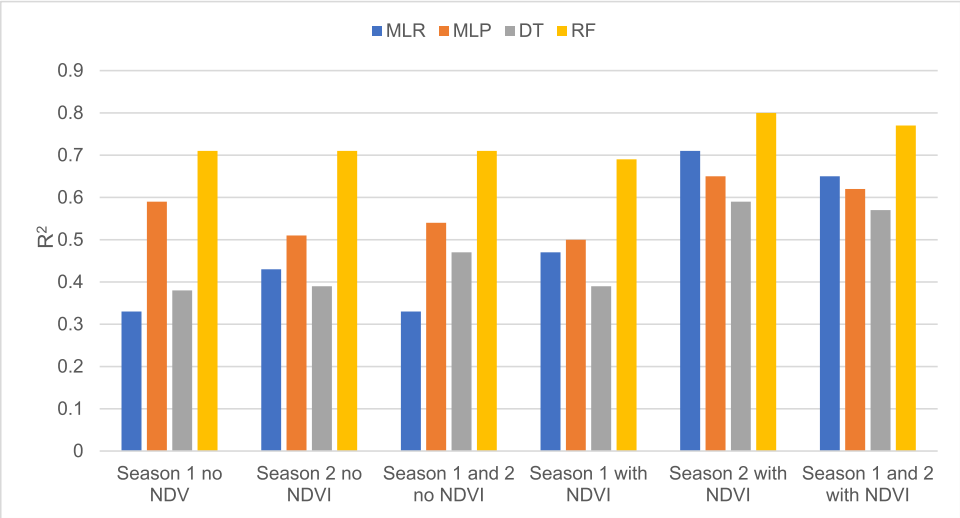

Fig. 5. Comparison of the performance of machine learning algorithms for season 1 (2019/2020) and season 2 (2020/2021) and combining the data from the two seasons with and without NDVI (MLR: multiple linear regression, MLP: multilayer perceptron, DT: decision tree, RF: random forest).

Table 3
Statistical analysis comparison of machine learning regression models on the DIFM trial maize field for 2019/2020, 2020/201, and the combined dataset with and without NDVI evaluated using the 80/20 training and testing analysis (MAPE: mean absolute percentage error, RMSE: root mean square error).

Season	ML algorithm	MAPE (%)	RMSE (t ha^{-1})
With NDVI data			
2019/2020	MLR	7.3	0.89
	MLP	6.5	0.85
	DT	7.4	0.96
	RF	5.4	0.69
2020/2021	MLR	9.8	1.08
	MLP	10.5	1.22
	DT	12.0	1.35
	RF	8.4	0.95
Combined seasons	MLR	8.4	0.95
	MLP	7.7	0.93
	DT	8.7	1.10
	RF	6.6	0.81
Without NDVI data			
2019/2020	MLR	8.4	1.01
	MLP	6.2	0.79
	DT	7.4	0.97
	RF	5.3	0.66
2020/2021	MLR	14.4	1.59
	MLP	12.4	1.40
	DT	13.3	1.65
	RF	10.0	1.14
Combined seasons	MLR	11.6	1.36
	MLP	7.9	0.98
	DT	8.9	1.21
	RF	7.0	0.89

identify and quantify the influence of spatial elements, such as soil attributes (Kravchenko and Bullock, 2000, Jones et al., 2022). Through correlation analysis, it was evident that there was a complex relationship between maize yield and different yield attributes during and between the seasons. Urea application had a stronger relationship with yield than the plant population for each of the two seasons and when seasonal data were combined. The varying correlations between yield and yield-influencing attributes across seasons reflect the complexity of the agricultural systems. Crop yield is a function of the interaction between spatial and temporal changes in variables (Bullock et al., 2019), and crop yield prediction is affected by several of these variables. Varying weather conditions and the interaction between multiple yield-influencing variables can contribute to varying correlations, such as for soil properties, which may have a more positive effect on yield in a wet season than in a subsequent drier season. Site-specific spatial soil and environmental variability has been reported to contribute towards yield variability in previous studies (Ma and Herath, 2016, Wen et al., 2021). The negative correlation between soil pH and yield indicates that, in this field, there may be a compounding effect of other yield-influencing factors associated with an increase in soil pH, which results in certain parts of the field having lower yields at higher soil pH.

4.1. Comparative analysis of ML algorithms for spatial maize yield predictions

Several studies have compared the predictive capabilities of regression and ML models, focusing primarily on climatic factors, farm management, soil nutrients, and topographic attributes (Gonzalez-Sanchez et al., 2014, Burdett and Wellen, 2022). However, the integration of DIFM, leading to thousands of unique data points, with ML algorithms for spatial yield predictions has not been well explored to date. The results of this study indicate that although varying ML techniques are applicable for spatial maize yield predictions, the accuracy of these models can be influenced by the specific datasets used and may vary with temporal changes between successive seasons. The lower accuracy of the MLR model in predicting spatial yields without NDVI data suggests the complexity of capturing the intricate and nonlinear connections present in the dataset or that there may not be a direct relationship. Previous studies have suggested that linear regression analysis is less effective in agronomic studies, because yield data are a result of multiple interacting factors (Kitchen et al., 2003). The integration of NDVI data boosted the initially low prediction accuracy of the MLR model more than that of the other models, suggesting that NDVI data can improve the predictive capability of the MLR model. This was likely a result of the higher correlation between NDVI and yield, as indicated in the descriptive statistics, being able to capture more linear relationships between NDVI and yields in the MLR model, whereas the MLP also considered the nonlinear relationships that lowered the prediction accuracy. For the same dataset without NDVI, the MLP outperformed MLR and DT in terms of prediction accuracy because of its superior capability to effectively manage nonlinear relationships. The MLP architecture is comprised of multiple hidden layers and employs nonlinear activation functions that are adept at handling nonlinear relationships. Although DT can also manage nonlinear relationships, it struggles with irregular patterns and is prone to overfitting (Bramer, 2002).

Although incorporating NDVI data into the training and testing

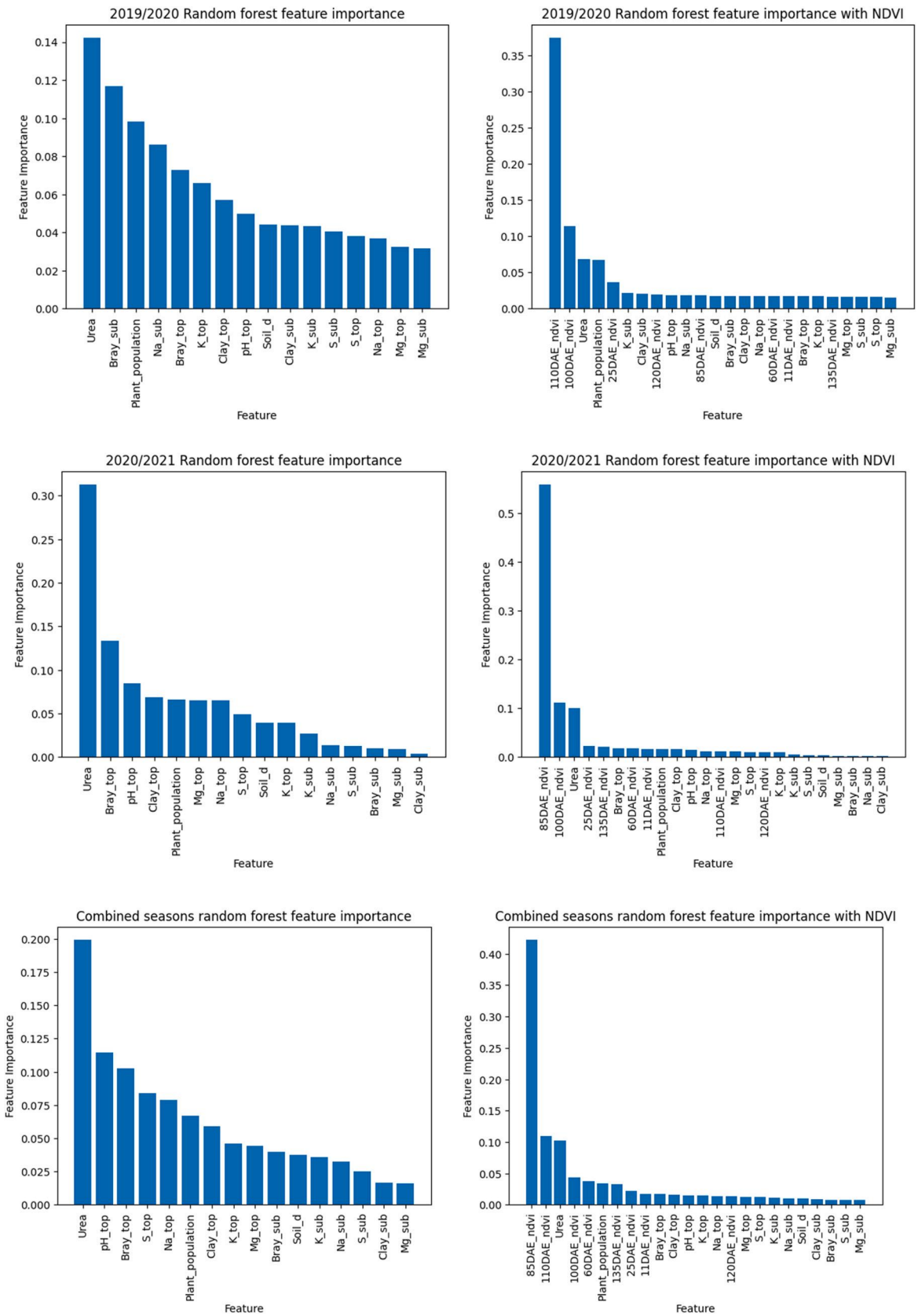

Fig. 6. Feature importance from the random forest for 2019/2022, 2020/2021, and the combination of the two-season data with and without normalized difference vegetation index (NDVI) using the 80/20% training and testing analysis (Plant_pop: plant population, Urea: urea application, ph_top: soil pH in topsoil, bray_top: phosphorus in topsoil, K_top: potassium in topsoil, Mg_top: magnesium in topsoil, Na_top: sodium in topsoil, S_top: sulphur in topsoil, Clay_top: clay content in topsoil, Bray_sub: phosphorus in sub soil, K_sub: potassium in sub soil, Mg_sub: magnesium in sub soil, Na_sub: sodium in sub soil, S_sub: Sulphur in sub soil, Clay_sub: Clay content in sub soil, Soil_d: soil depth).

datasets led to improved model predictions and a decrease in prediction errors for all models, the extent of these improvements varied among the different models and datasets used. The performance of the RF model surpassed that of the MLR, DT, and MLP models for spatial yield predictions, with and without the inclusion of NDVI. Therefore, the RF model effectively captured the relationship between maize yield and various factors representing soil, management, and remotely sensed attributes. In several previous studies (Han et al., 2020, Burdett and Wellen, 2022), RF also outperformed other ML models such as MLR, DT, and MLP. The RF uses the single best variable when splitting responses on each node and averages the predictions of different trees in the forest to create a multi-dimensional function, which gives RF an advantage over the other models when predictor variables are highly correlated. The application of ML models in the two distinct seasons yielded different prediction accuracy results for all models, although there was consistency in the overall comparison between the models. This could be an indication of a shift in the impact of temporal and spatial variables on yield between seasons (Kravchenko and Bullock, 2000), and these interactions are more important in rain-fed cropping systems.

4.2. Feature importance and identification of yield limiting factors

Before adopting a full input application strategy, feature importance can help identify regions with a high possibility of success in modifying management to a site-specific approach. This study showed that the important features explaining yield variability differed between seasons and were influenced by the inclusion of other variables in the dataset. The interaction between soil and management variables makes it challenging to identify the main yield-limiting factors and the extent of the impact of each variable on the final yield. Urea application was consistently the most critical variable in all model training and testing scenarios, although the degree of importance varied with the amount of rainfall received during the season. The influence of N application on yield variability was notably higher during the drier season than during the wet season, thus emphasizing the significance of N availability and adequate rainfall in achieving optimum yields. Additional factors, including soil P, plant population and Na in 2020, as well as soil P, soil pH, clay content, and plant population in 2021, emerged as the most influential factors for explaining yields. Merging the two seasons dataset indicated that topsoil properties explained yield variation more than subsoil properties, highlighting the importance of adequate nutrient availability in the topsoil for optimum yields.

The tillering abilities of the cultivar in low plant populations may have resulted in more cobs and grains, which could be the reason why the low plant population was not as influential as expected in explaining yield variability. This could be important for applications of RF models in evaluating varied input applications, indicating the importance of investing in N fertilization for optimum yields, and possibly not as much on seeding rates. The NDVI data explained yields to a higher degree than did the soil and management variables. The addition of NDVI data can only be crucial for improving the yield prediction accuracy of the models, as this can be an attribute that farmers cannot control. Integrating more within-season data sources, such as NDVI, rainfall distribution in relation to plant growth stage, and the timing of split fertilizer application, should be considered in future research because it can potentially improve spatial yield predictions. Filippi et al. (2019) reported an increase in model predictions in response to an increase in the within-season data used for model predictions. Other studies have highlighted the importance of other variables not included in this study, such as available water content (Kravchenko and Bullock, 2000, Xu et al., 2019, De Souza et al., 2023), soil organic matter (Xu et al., 2021) and topography variables (Lacerda et al., 2022). The vapor pressure deficit and temperature have also been found to be important factors in model training, and testing includes multiple seasons (Xu et al., 2019). Although the best ML model performance could be achieved using DIFM and NDVI data, the feature importance revealed that the most accurate prediction could be achieved at 85–110 DAE, which could be too late for a farmer to take meaningful remedial actions if required, such as applying more fertilizer. It is also crucial that the variables included in the yield prediction models are directly connected to yield and translatable into management choices for the ML model's insights to be implemented effectively.

4.3. Uncertainties in the study

There are still some limitations and uncertainties regarding the use of ML models in yield predictions. Machine learning models typically do not incorporate mechanisms related to crop growth during the model development process (Han et al., 2020), which may contribute to the increased uncertainty in model performance. As a result, it can be challenging to comprehend the reasoning behind these predictions. This makes it challenging to identify and address errors, biases, and uncertainties in a model. Machine learning models are often trained using data from specific locations and times, which can limit their generalizability to other locations and times. This limitation arises because the relationships between input variables and yield may differ depending on factors, such as soil type, in-season weather patterns, and management practices. Having a larger dataset with increased weather pattern ranges included in the training dataset can improve the ML predictions, even under extreme conditions. In future studies, it is recommended that the trained models be retrained and validated on comparable commercial farms from other locations with varying weather and soil conditions.

5. Conclusion

This study showed that despite differences in accuracy, all four ML techniques could be effectively trained and tested on DIFM data to develop models that can be used to estimate spatial maize yields in a highly heterogeneous field. This could be important for helping farmers and agronomists estimate yield profit margins and determine the cost-benefit of agronomic intervention. The RF model was the best for spatial yield predictions using DIFM datasets and the inclusion of NDVI data improved model performance. Although there was variability between seasons in the factors that strongly influenced yield, urea application was consistently the most critical variable, along with other key factors including soil P, plant population, and Na in 2020, as well as soil P, soil pH, clay content, and plant population in 2021. Machine learning models in agricultural systems require the consideration of variations in weather, soil characteristics, and other environmental factors that can vary in space and time. The DIFM trials can play an important role in advancing the field of PA by providing valuable insights into the complex interactions between crops, soils, weather, and management practices, and by identifying new opportunities for improving crop yields and environmental sustainability. Although distinct zones were not identified, the results, when combined with AI technologies, can aid farmers in managing their fields more effectively. The approach used in this study can be enhanced and modified for use in various agricultural settings and incorporating additional vegetative indices that may be associated with crop production.

CRediT authorship contribution statement

Simphiwe Maseko: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Helga Otterman**: Formal analysis, Data curation. **Marion Delport**: Formal analysis, Data curation. **Eyob Habte** Tesfamariam: Writing – review & editing, Supervision, Methodology. **Michael van der Laan**: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Michael van der Laan reports financial support was provided by Water Research Commission. Simphiwe Maseko reports financial support was provided by National Research Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

We gratefully acknowledge the Water Research Commission [Project title "Development And Application Of A Big Data Platform To Improve Agricultural Water Resources Management In South Africa" (C2020/2021–00440)] for funding this research, and National Research Foundation (NRF) for providing us with a partial student Funding for this work to be carried out.

References

Basso, B., Fiorentino, C., Cammarano, D., Schulthess, U., 2016. Variable rate nitrogen fertilizer response in wheat using remote sensing. Precis. Agric. 17, 168–182. https://doi.org/10.1007/s11119-015-9414-9.

Basso, B., Liu, L., 2019. Seasonal crop yield forecast: methods, applications, and accuracies. Adv. Agron. 154, 201–255. https://doi.org/10.1016/bs.agron.2018.11.002.

Bishop, T., Horta, A., Karunaratne, S., 2015. Validation of digital soil maps at different spatial supports. Geoderma 241, 238–249. https://doi.org/10.1016/j.geoderma.2014.11.026.

Bramer, M., 2002. Using J-pruning to reduce overfitting in classification trees. Knowl. -Based Syst. 15, 301–308. https://doi.org/10.1016/S0950-7051(01)00163-0.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/a:1010933404324.

Bullock, D.S., Boerngen, M., Tao, H., Maxwell, B., Luck, J.D., Shiratsuchi, L., Puntel, L., Martin, N.F., 2019. The data-intensive farm management project: changing agronomic research through on-farm precision experimentation. Agron. J. 111, 2736–2746. https://doi.org/10.2134/agronj2019.03.0165.

Burdett, H., Wellen, C., 2022. Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. Precis. Agric. 23, 1553–1574. https://doi.org/10.1007/s11119-022-09897-0.

De Souza, P.V.D., De Rezende, L.P., Duarte, A.P., Miranda, G.V., 2023. Maize yield prediction using artificial neural networks based on a trial network dataset. Eng. Technol. Appl. Sci. Res. 13, 10338–10346. https://doi.org/10.48084/etasr.5664.

Drummond, S.T., Sudduth, K.A., Joshi, A., Birrell, S.J., Kitchen, N.R., 2003. Statistical and neural methods for site–specific yield prediction. Trans. ASAE 46, 5.

Filippi, P., Jones, E.J., Wimalathunge, N.S., Somarathna, P.D., Pozza, L.E., Ugbaje, S.U., Jephcott, T.G., Paterson, S.E., Whelan, B.M., Bishop, T.F., 2019. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precis. Agric. 20, 1015–1029. https://doi.org/10.1007/s11119-018-09628-4.

Gonzalez-Sanchez, A., Frausto-Solis, J., Ojeda-Bustamante, W., 2014. Predictive ability of machine learning methods for massive crop yield prediction. Span. J. Agric. Res. 12, 313–328.

Han, J., Zhang, Z., Cao, J., Luo, Y., Zhang, L., Li, Z., Zhang, J., 2020. Prediction of winter wheat yield based on multi-source data and machine learning in China. Remote Sens. 12, 236. https://doi.org/10.3390/rs12020236.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer, New York vol. 112.

Jaynes, D.B., Colvin, T.S., 1997. Spatiotemporal variability of corn and soybean yield. Agron. J. 89, 30–37. https://doi.org/10.2134/agronj1997.00021962008900010005x.

Jones, E.J., Bishop, T.F., Malone, B.P., Hulme, P.J., Whelan, B.M., Filippi, P., 2022. Identifying causes of crop yield variability with interpretive machine learning. Comput. Electron. Agric. 192, 106632 https://doi.org/10.1016/j.compag.2021.106632.

Kaspar, T.C., Pulido, D., Fenton, T., Colvin, T., Karlen, D., Jaynes, D., Meek, D., 2004. Relationship of corn and soybean yield to soil and terrain properties. Agron. J. 96, 700–709. https://doi.org/10.2134/agronj2004.0700.

Kitchen, N., Drummond, S., Lund, E., Sudduth, K., Buchleiter, G., 2003. Soil electrical conductivity and topography related to yield for three contrasting soil–crop systems. Agron. J. 95, 483–495. https://doi.org/10.2134/agronj2003.4830.

Kravchenko, A.N., Bullock, D.G., 2000. Correlation of corn and soybean grain yield with topography and soil properties. Agron. J. 92, 75–83. https://doi.org/10.2134/agronj2000.92175x.

Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W., 2005. Applied Linear Statistical Models. McGraw-hill.

Kyveryga, P.M., 2019. On-farm research: experimental approaches, analytical frameworks, case studies, and impact. Agron. J. 111, 2633–2635. https://doi.org/10.2134/agronj2019.11.0001.

Lacerda L.N., Miao Y., Mizuta K., Stueve K. 2022. Identifying Key Factors Influencing Yield Spatial Pattern and Temporal Stability for Management Zone Delineation.

Lassoued, R., Macall, D.M., Smyth, S.J., Phillips, P.W., Hesseln, H., 2021. Expert insights on the impacts of, and potential for, agricultural big data. Sustainability 13, 2521. https://doi.org/10.3390/su13052521.

Li, C., Ma, C., Pei, H., Feng, H., Shi, J., Wang, Y., Chen, W., Li, Y., Feng, X., Shi, Y., 2020. Estimation of Potato Biomass and Yield Based on Machine Learning from Hyperspectral Remote Sensing Data. J. Agric. Sci. Technol. B 10, 195–213. https://doi.org/10.17265/2161-6264/2020.04.001.

Li, K.-Y., Sampaio De Lima, R., Burnside, N.G., Vahtmäe, E., Kutser, T., Sepp, K., Cabral Pinheiro, V.H., Yang, M.-D., Vain, A., Sepp, K., 2022. Toward automated machine learning-based hyperspectral image analysis in crop yield and biomass estimation. Remote Sens. 14, 1114. https://doi.org/10.3390/rs14051114.

Ma, B., Herath, A., 2016. Timing and rates of nitrogen fertiliser application on seed yield, quality and nitrogen-use efficiency of canola. Crop Pasture Sci. 67, 167–180. https://doi.org/10.1071/CP15069.

Maimon, O.Z., Rokach, L., 2014. Data Mining with Decision Trees: Theory and Applications. World scientific vol. 81.

Naser M., Alavi A. 2020. Insights Into Performance Fitness and Error Metrics for Machine Learning. arXiv preprint arXiv:2006.00887.

Näsi, R., Viljanen, N., Kaivosoja, J., Alhonoja, K., Hakala, T., Markelin, L., Honkavaara, E., 2018. Estimating biomass and nitrogen amount of barley and grass using UAV and aircraft based spectral and photogrammetric 3D features. Remote Sens. 10, 1082. https://doi.org/10.3390/rs10071082.

Nawar, S., Corstanje, R., Halcro, G., Mulla, D., Mouazen, A.M., 2017. Delineation of soil management zones for variable-rate fertilization: a review. Adv. Agron. 143, 175–245. https://doi.org/10.1016/bs.agron.2017.01.003.

Nayak, H.S., Silva, J.V., Parihar, C.M., Krupnik, T.J., Sena, D.R., Kakraliya, S.K., Jat, H.S., Sidhu, H.S., Sharma, P.C., Jat, M.L., 2022. Interpretable machine learning methods to explain on-farm yield variability of high productivity wheat in Northwest India. Field Crops Res. 287, 108640 https://doi.org/10.1016/j.fcr.2022.10864.

Nyéki, A., Milics, G., Kovács, A., Neményi, M., 2017. Effects of soil compaction on cereal yield: a review. Cereal Res. Commun. 45, 1–22. https://doi.org/10.1556/0806.44.2016.056.

Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R.L., Mouazen, A.M., 2016. Wheat yield prediction using machine learning and advanced sensing techniques. Comput. Electron. Agric. 121, 57–65. https://doi.org/10.1016/j.compag.2015.11.018.

Perez-Alonso, D., Peña-Tejedor, S., Navarro, M., Rad, C., Arnaiz-González, Á., Díez-Pastor, J.-F., 2017. Decision Trees for the prediction of environmental and agronomic effects of the use of Compost of Sewage Slugde (CSS). Sustain. Prod. Consum. 12, 119–133. https://doi.org/10.1016/j.spc.2017.07.001.

Ransom, C.J., Kitchen, N.R., Camberato, J.J., Carter, P.R., Ferguson, R.B., Fernández, F. G., Franzen, D.W., Laboski, C.A., Myers, D.B., Nafziger, E.D., 2019. Statistical and machine learning methods evaluated for incorporating soil and weather into corn nitrogen recommendations. Comput. Electron. Agric. 164, 104872 https://doi.org/10.1016/j.compag.2019.104872.

Ruder S. 2016. An Overview of Gradient Descent Optimization Algorithms. arXiv preprint arXiv:1609.04747.

Shahhosseini, M., Hu, G., Huber, I., Archontoulis, S.V., 2021. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Sci. Rep. 11, 1–15. https://doi.org/10.1038/s41598-020-80820-1.

Silva, J.V., Tenreiro, T.R., Spätjens, L., Anten, N.P., Van Ittersum, M.K., Reidsma, P., 2020. Can big data explain yield variability and water productivity in intensive cropping systems? Field Crops Res. 255, 107828 https://doi.org/10.1016/j.fcr.2020.107828.

Tantalaki, N., Souravlas, S., Roumeliotis, M., 2019. Data-driven decision making in precision agriculture: the rise of big data in agricultural systems. J. Agric. Food Inf. 20, 344–380. https://doi.org/10.1080/10496505.2019.1638264.

Taylor, J., Mcbratney, A., Whelan, B., 2007. Establishing management classes for broadacre agricultural production. Agron. J. 99, 1366–1376. https://doi.org/10.2134/agronj2007.0070.

Trevisan, R., Bullock, D., Martin, N., 2021. Spatial variability of crop responses to agronomic inputs in on-farm precision experimentation. Precis. Agric. 22, 342–363. https://doi.org/10.1007/s11119-020-09720-8.

Van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: a systematic literature review. Comput. Electron. Agric. 177, 105709 https://doi.org/10.1016/j.compag.2020.105709.

Wen, G., Ma, B.-L., Vanasse, A., Caldwell, C.D., Earl, H.J., Smith, D.L., 2021. Machine learning-based canola yield prediction for site-specific nitrogen recommendations. Nutr. Cycl. Agroecosystems 121, 241–256. https://doi.org/10.1007/s10705-021-10170-5.

Xu, X., Gao, P., Zhu, X., Guo, W., Ding, J., Li, C., Zhu, M., Wu, X., 2019. Design of an integrated climatic assessment indicator (ICAI) for wheat production: a case study in Jiangsu Province, China. Ecol. Indic. 101, 943–953.

Xu, T., Guan, K., Peng, B., Wei, S., Zhao, L., 2021. Machine learning-based modeling of spatio-temporally varying responses of rainfed corn yield to climate, soil, and management in the US Corn Belt. Front. Artif. Intell. 4, 647999.