Field-level sugarcane yield estimation utilizing Sentinel-2 time-series and machine learning

Rekha B. U.[1], Veena V. Desai[1], Suresh Kuri[1], Pratijnya S Ajawan[1], Sunil Kumar Jha[2], V. C. Patil[3]
[1]Department of Electronics and Communication Engineering, KLS Gogte Institute of Technology,
Affiliated to Visvesvaraya Technological University, Belagavi, India
[2]Applied Agricultural Remote Sensing Centre, University of New England, Armidale, Australia
[3]K J Somaiya Institute of Applied Agricultural Research, Bagalkot, India

Article Info

Article history:

Received Apr 16, 2024
Revised Sep 10, 2024
Accepted Sep 29, 2024

Keywords:

Machine learning
Red edge spectral bands
Remote sensing
Sugarcane yield
Vegetation indices time series

ABSTRACT

This work focused on developing a methodology for using machine learning (ML) approaches to establish a pre-harvest yield prediction model for sugarcane at field level by integrating time-series remote sensing imagery data with ML techniques. Ground truth agro data and thirty-one spectral vegetation indices were extracted from Sentinel-2 imagery and were considered for yield modeling. A two-level feature selection technique was used to determine the most significant variables that best correlated with sugarcane yield to predict yield in advance. Seven ML algorithms, including those based on regularization, decision trees, and ensemble methods like boosting, were used to predict yield. The approach achieved the highest R^2 score of 0.73 and the lowest root mean squared error (RMSE) of 13.45 t/ha with random forest (RF) among the seven ML models tested. Furthermore, all feature selection procedures identified normalized difference red edge (NDRE), red edge chlorophyll index (RECI), and ratio vegetation index (RVI) as major yield-driving variables. The experiments during feature selection demonstrated the potential of red edge spectral bands in development of a reliable sugarcane-yield prediction approach. The RF model obtained using the proposed methodology outperforms the two baseline models developed using NDVI and GNDVI indices, with an improved RMSE of 16-18%.

Corresponding Author:

Rekha B. U.
Department of Electronics and Communication Engineering, KLS Gogte Institute of Technology
Affiliated to Visvesvaraya Technological University
Belagavi, 590018, India
Email: dhdrekha@gmail.com

1. INTRODUCTION

Sugarcane is a multifunctional crop mainly utilized for extracting sugar and within a diversified portfolio. Precise early-season also known as “within-the-season” yield prediction of sugarcane has an important role in the management of agroecosystems [1]. The conventional methods of sugarcane yield estimation before harvest involve manual scouting through established sugarcane fields and crop cutting experiments (CCE). Walking through the dense sugarcane crops that are as tall as 6 m probe challenge to the employees. As a result the conventional methods are prone to inaccuracies and often biased due to human errors. Since the establishment of commercial cultivation of sugarcane, in the recent past a few sugar mills are exploring the possibilities of estimating sugarcane yield using global positioning system (GPS) technology through manual survey of field perimeters in their mill catchment area promoting intervention of engineering technologies in sugarcane crop yield prediction via field-based studies [2]. Highly detailed remote-sensing data with precise temporal and spatial resolution can offer valuable insights into the development and yield of sugarcane crops [3]-[6]. The application of remote-sensing technologies also facilitates non-invasive and unbiased estimations of crop production. Furthermore, vegetation indices are numerical indicators that indicate the level of vegetation vitality. Several vegetation indicators, including NDVI [7], [8], RVI [9], GNDVI [10], EVI, NDWI, and SAVI [11], are currently utilized in the development of statistical methods for sugarcane yield prediction. When it comes to predicting sugarcane productivity on individual farms in Iran, the green vegetation index (GVI) is more accurate than both GNDVI and NDVI [12]. In an effort to estimate the sugarcane yield at the plot level using machine learning (ML) considering unmanned aerial vehicle (UAV) multi-spectral images, an investigation was conducted in Bundaberg, Australia. Among the twenty-three identified indicators, five specific indices, notably NDRE, GRNDVI, PNDVI, CCCI, and WDRVI, demonstrated a notable degree of precision in the prediction of sugarcane yield. According to an investigation in [13], metrics obtained from standardized NDVI time-series collected by the MODIS sensor along with an ensemble approach of artificial neural networks (ANNs) were used for predicting the yield of sugarcane in São Paulo State, Brazil, three months prior to harvesting.

Predicting sugarcane yield at the regional scale (district-wise) in Uttar Pradesh, India was done using the readily available MODIS satellite data. A pilot-scale research was conducted in [14] to evaluate the cultivation of sugarcane across four sugar processing plants located in the two Indian states. The investigation was conducted throughout the period between 2017 and 2019 and utilized NDVI, water scalar (WS), and land surface water index (LSWI) multi-date multi-spectral information gathered from LISS III, LISS IV of Resourcesat-2 and 2A, in addition to ground data and GPS information. Additionally, CCE was utilized as part of the assessment process. A comparable investigation conducted similar to [14], was presented in [15] where similar sugar processing plants in two Indian states were considered and utilized an ensemble approach of optical information from both synthetic aperture radar (SAR) and Sentinel-2 information collected through the sugarcane harvest cycle's major stage of development for yield prediction. A number of factors, including a combination of NDVI, NDRE, EVI, MTVI, and WS, according to information gathered by Sentinel-2, were used to determine the crop yield. On the contrary, VV and VH polarizations, in addition to the cross pol ratio (VV/VH), consisted of the parameters derived from SAR information that were used to drive yield modeling. The researchers ultimately employed three factors, namely EVI, WS, and VH, at different stages of development in order to obtain a precise estimation of sugarcane yield. Factors like less cultivating region, a hybrid cropping structure, crop varietal shifts, variation in the date of planting, soil variants, and irrigation methods contributed to the difficulty of accurately predicting yields at a spatial scale. ML has been widely recognized for its potential to achieve higher predicting accuracy levels in agricultural yield at the field level. This is mainly due to ML's capability of capturing the complicated connection among crop yield alongside the various features that are utilized for yield predictions. Therefore, the utilization of satellite imagery when combined with ML approach enhances the precision of sugarcane yield predictions. Several ML algorithms, such as multiple linear regression (MLR), gradient boosting tree (GBR), support vector machine (SVM), random forest (RF), ANNs, extreme gradient boosting tree (XGB), and Bayesian approach, have been widely employed for predicting sugarcane crop yield at the field to regional or mill-level employing agrometeorological and satellite imagery [15]-[18].

One of the most important steps in developing any kind of crop yield prediction approach is choosing the right ML model and incorporating the right feature parameters. In spite of a known correlation between sugarcane yield and different vegetation indices, it is crucial to weigh their suitability for the intended application at a given location and growth stages of sugarcane. A thorough examination of vegetation indices in [19] indicated that various factors can significantly impact the accuracy and reliability of vegetation indices. These factors include environmental variables, sensor measurements, sensor observing conditions, sun's lighting geometry, soil moisture levels, brightness, and color. Xue and Su [20], examined over one hundred vegetation indices. They have analyzed their usefulness in relation to the specific vegetation of interest and the surrounding environment. While there are numerous indices mentioned in the literature, only a few of them have been essentially utilized or assessed for predicting sugarcane yield. According to previous research, the NDVI has been identified as the most commonly utilized index for predicting crop yield [21]-[26]. The use of RS and ML for sugarcane yield prediction is picking the momentum in recent years after satellite data providers like USGS and ESA changed their policies to make the data free and open to users. Hence it remains a site-specific open research problem the solutions of which can lead to a universal model gradually with minimal changes to its operation. From our understanding of recent studies to the best of our knowledge, a set of vegetation indices to be included for yield modeling and a detailed ML based data driven approach to model sugarcane crop yield at field level was unknown in our study area. There is not a specific vegetation index or a preset set of indices that is universally applicable to sugarcane crop modeling. Thus, the choice of one vegetation index over the other, for an application, is quite delicate to make. Therefore, the present work aims to propose a methodology to evaluate the relationship of ground data variables and spectral vegetation indices derived from hyperspectral imagery with yield at the field level and develop a ML based sugarcane yield prediction model for a tropical region like India particularly in our study area while targeting the following objectives:

- Identify and evaluate the sensitivity of indices calculated from narrowband and red edge bands with sugarcane yield.
- Determine the optimal time for sugarcane yield prediction.
- Identify the significant vegetation indices that best explain the yield at the field level.
- Create and evaluate various ML techniques for predicting sugarcane yield using Sentinel-2 data in an environment with sparse ground data.

2. MATERIALS AND METHODS

2.1. Study area

The investigation was carried out in the operational region of Godavari Bio-Refineries Ltd. (GBL), a sugar mill located in Sameerwadi, which is positioned within the Bagalkot and Belagavi districts of Karnataka, India. The geographical coordinates of the study area are 16.3898 °N and 75.0371 °E. The research region is depicted in Figure 1. The region was characterized by a humid subtropical environment, with the monsoon period from June to September. The region was dry and sparse with temperatures ranging from 16.20 °C to 39.00 °C. The mean precipitation in Belagavi and Bagalkot was 545 millimeters (mm) and 808 mm respectively. Sugarcane cultivation in the study area observes three planting seasons: seasonal (between Jan-Mar and Jan-Feb), pre-seasonal (between Oct-Nov and Dec-Feb), and adsali (between July-Aug and Oct-Dec). The common sugarcane varieties grown here include CO86032 and CO91010. Apart from sugarcane, other crops grown include maize, turmeric, green leaves, seasonal fruits, and vegetables.

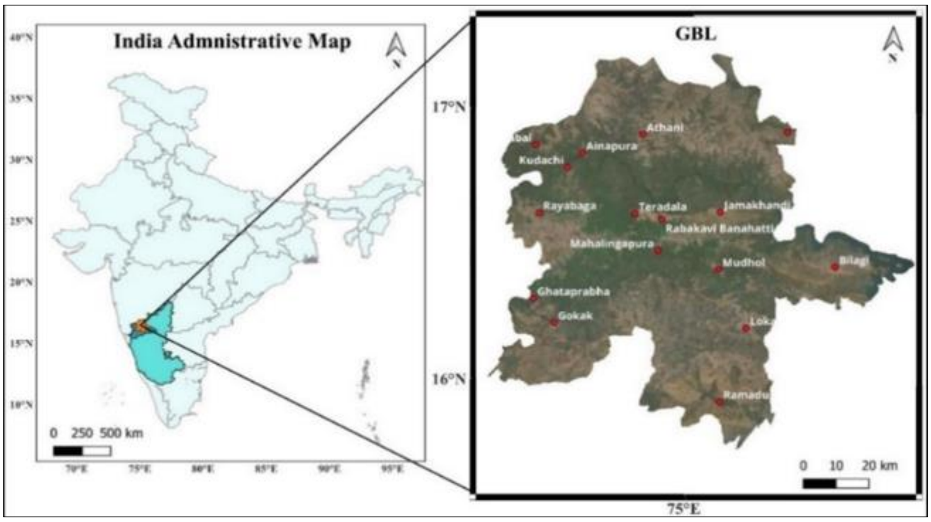


Figure 1. Location map of the study area

2.2. Field data

A GPS survey of sugarcane fields registered with GBL for crop yield modeling was performed between January and April 2019. The field boundary geo-coordinates were collected using a handheld Garmin Montana 680 GPS device. The GPS tracks were digitized into KML files using Google Earth Pro and combined in QGIS to create a single-shape file representing the surveyed farms in 2D. During digitization, we ensured to carefully map the field boundary coordinates and avoid any intersection of resulting polygons. A random subset of fields from different villages in the study area were chosen for the survey. This was done to ensure the overall sample reflected the typical growing conditions and hence, diverse crop vigor categories in our sampling strata. While we acknowledge the possibility of certain sources of bias, such as variation in farm management practices, soil fertility, or weather conditions, though we made efforts to mitigate these factors. Agronomic data for 196 surveyed sugarcane fields, including crop variety, cane type, and yield for the surveyed fields were obtained from the GBL agro database for three consecutive growing seasons 2017-

2018, 2018-2019, and 2019-2020. The field level sugarcane yield modeling was done for the growing season 2017-2018. The 2018-2019 and 2019-2020 season data were used to validate the final model.

2.3. Satellite imagery data

The Sentinel-2 top-of-atmosphere level-1C (L1C) collection in Google earth engine (GEE) is used for the study. The Sentinel-2 surface reflectance Level-2A collection in GEE does not extend back to 2017. So we used the L1C collection. L1C data is also reported to provide good results without the need for atmospheric correction [27]-[29]. The effect of clouds was reduced in two levels. First, Sentinel-2 tiles with a cloud cover of less than 20% were filtered by the "cloud pixel percentage" (CPP) metadata field. Second, at the pixel level clouds were masked using the "QA60" quality band [27]. The 20 m spatial resolutions of the red edge and SWIR bands were downscaled to 10 m using bilinear interpolation [28] with the original projection maintained, to improve interpretation and achieve common differentiability.

2.4. Spectral vegetation indices

31 vegetation indices listed in Table 1, specific to crop yield were selected for the study. In Table 1, the notations are as follows N - NIR, R - Red, G - Green, B - Blue, RE1 - Red Edge 1, RE2 - Red Edge 2, RE3 - Red Edge 3 band, N2 - Red Edge 4 band. The selection of these indices was based on the knowledge that the NIR, Green, and Red spectral bands play a crucial role in representing the dense canopy surface such as that of a sugarcane crop. In addition, recent studies for rice yield estimation found narrow bands to be useful in improving the yield model accuracy [29]. To examine the potential of narrow bands in sugarcane yield prediction, a set of indices based on Narrow NIR and red edge bands was also included.

Table 1. The vegetation indices used in this study

Index	Band combination	Reference
Red edge band greenness indices		
Chlorophyll carotenoid index (CCCI)	(N - RE1)*(N + R)/(N + RE1)*(N - R)	[20]
Modified Chlorophyll Absorption in Reflectance Index (MCARI)	((RE1 - R) - 0.2 * (RE1 - G)) * (RE1/R)	[20]
Modified red edge NDVI (MRENDVI)	(RE1 - N)/(RE1 + N - (2 * B))	[30]
Normalized difference red edge index (NDRE)	(N - RE1)/(N + RE1)	[30]
Red edge chlorophyll index (RECI)	(N/RE1) - 1	[30]
Sentinel-2 red edge position (S2REP)	705.0 + 35.0 * ((((RE3 + R)/2.0) - RE1)/(RE2 - RE1))	[31]
Atmospherically resistant vegetation index (ARVI)	(N2 - (R - gamma * (R - B)))/(N2 + (R - gamma * (R - B)))	[32]
Broadband and narrowband greenness indices		
Blue normalized difference vegetation index (BNDVI)	(N - B)/(N + B)	[30]
Chlorophyll vegetation index (CVI)	(N * R)/(G ** 2.0)	[30]
Difference vegetation index (DVI)	N– R	[20]
Enhanced vegetation index (EVI)	g * (N - R)/(N + C1 * R - C2 * B + L)	[20]
Green-blue normalized difference vegetation index (GBNDVI)	(N - (G + B))/(N + (G + B))	[30]
Green chlorophyll vegetation index (GCI)	(N/G) – 1	[30]
Global environment monitoring index (GEMI)	((2.0*((N ** 2.0) - (R ** 2.0)) + 1.5*N + 0.5*R)/(N + R + 0.5))*(1.0 - 0.25*((2.0 * ((N ** 2.0) - (R ** 2)) + 1.5 * N + 0.5 * R)/(N + R + 0.5)))-((R - 0.125)/(1 - R))	[20]
Green leaf index (GLI)	(2.0 * G - R - B)/(2.0 * G + R + B)	[20]
Green normalized difference vegetation index (GNDVI)	(N - G)/(N + G)	[20]
Green red normalized difference vegetation index (GRNDVI)	(N - (G + R))/(N + (G + R))	[30]
Green ratio vegetation index (GRVI)	N/G	[20]
Modified soil-adjusted vegetation index (MSAVI)	0.5 * (2.0 * N + 1 - (((2 * N + 1) ** 2) - 8 * (N - R)) ** 0.5)	[20]
Modified simple ratio (MSR)	(N / R - 1)/((N / R + 1) ** 0.5)	[20]
Modified triangular vegetation index (MTVI)	1.2 * (1.2 * (N - G) - 2.5 * (R - G))	[20]
Normalized difference vegetation index (NDVI)	(N - R)/(N + R)	[20]
Normalized difference water index (NDWI)	(G - N)/(G + N)	[20]
Optimized soil-adjusted vegetation index (OSAVI)	(N - R)/(N + R + 0.16)	[20]
Pan NDVI (PNDVI)	(N - (G + R + B))/(N + (G + R + B))	[30]
Red blue NDVI (RBNDVI)	N - (R + B)/N + (R + B)	[30]
Renormalized difference vegetation index (RDVI)	((N – R)/(N + R))$^{1/2}$	[20]
Ratio vegetation index (RVI)	N/R	[20]
Soil-adjusted vegetation index (SAVI)	(1.0 + L) * (N - R)/(N + R + L)	[20]
Transformed vegetation index (TVI)	(((N - R)/(N + R)) + 0.5) ** 0.5	[20]
Wide dynamic range vegetation index (WDRVI)	(alpha * N - R)/(alpha * N + R)	[20]

2.5. Feature engineering

We started constructing a time series for all the selected 31 vegetation indices using the equations described in Table 1 from January 2017 to March 2018 by filtering the Sentinel-2 L1C collection and applying a cloud-masking function in GEE [33]. The masked cloudy pixels were replaced by the temporal gap-filling technique with linear interpolation [34]. Following that, with the zonal statistics tool, we calculated the mean value of all pixels within each polygon for every spectral index band, to derive the respective vegetation index for a field on each day an image was available for it. Since the fields are spread geographically, they were captured by multiple tiles. Hence, all the calculated vegetation indices were not available at the same temporal resolution for each field. To harmonize this data we modeled the time series indices to be available at a regular interval i.e., at every 5-day time-step. We used a time window size of 21 days to look for an unmasked pixel in the time series. This helps in visualizing more gradual changes in the temporal profiles of indices compared to the original ones. The effect of interpolation and smoothing on the temporal profile of NDVI can be seen in Figure 2.

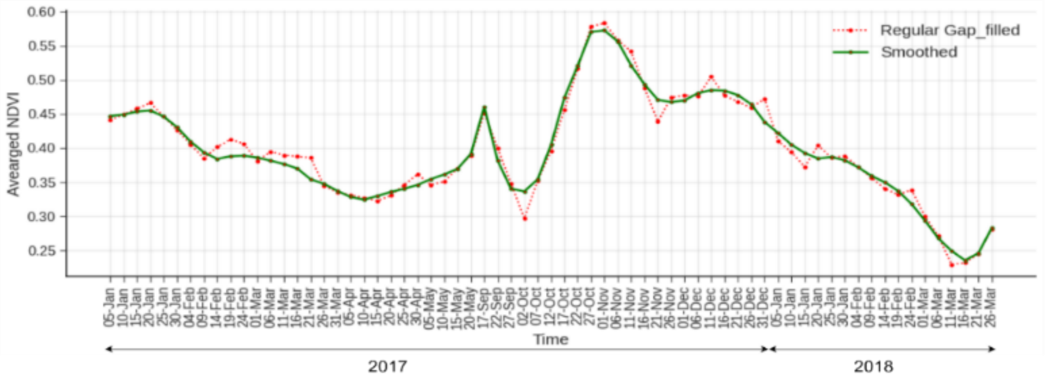


Figure 2. The averaged 5-day NDVI temporal profile across all fields. *The rise in NDVI between May - Sep could be due to intercropping farming practices in the study area

2.6. Optimum time for yield prediction

The monthly average values of estimated time-series vegetation indices were utilized. However, due to the cloud pixel percentage threshold used to filter cloudy images, no indices were available between June and September 2017, even with interpolation. The correlation between each feature variable and observed sugarcane yield was evaluated using Pearson's correlation coefficient [1]. The correlation coefficients were computed for grand growth (120 - 270 days after sowing) and maturity phase (270-360 days after sowing) except during the cloudy period.

2.7. Feature selection

A total of 37 features for the yield model were engineered from the combination of ground truth (6 categorical variables that include cane variety, crop type-plantation, ratoon 1, ratoon 2, ratoon 3, mix of ratoon and plantation) and time-series remote sensing data (31 variables) for every field. The performance of any ML model is largely based on the features it is trained on. Therefore, to select only the most suitable or significant features for predicting sugarcane yield we performed feature selection in two phases using Pearson's correlation coefficient (r) [15] and recursive feature elimination with cross-validation (RFECV) [35], [36].

2.8. Model evaluation and validation

The sugarcane yield prediction models were developed using the yield data for 160 field samples. As part of the modeling strategy, randomly 70% data were selected for model training and rest 30% data were held out as blind data for model testing. Seven ML algorithms namely, MLR, Ridge, Elasticnet, support vector regression (SVR), RF, GBR, and XGB were fitted over the training data. The reliability of the final model was validated using the 2018-2019 and 2019-2020 growing season sugarcane yield data. The framework for the proposed methodology is presented in Figure 3.

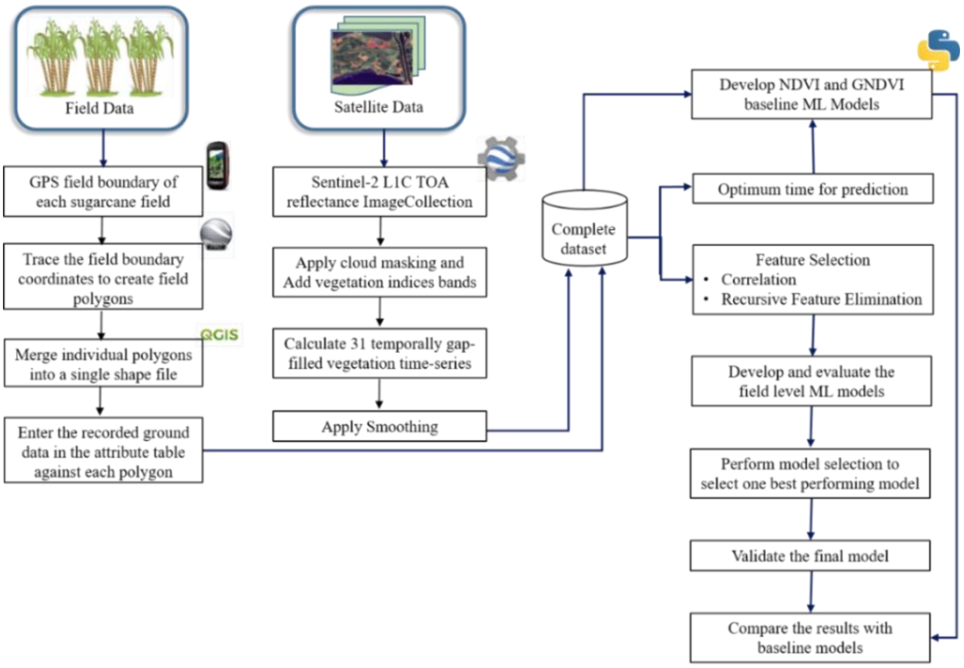

Field Data
Satellite Data
GPS field boundary of each sugarcane field
Trace the field boundary coordinates to create field polygons
QGIS
Merge individual polygons into a single shape file
Enter the recorded ground data in the attribute table against each polygon
Sentinel-2 L1C TOA reflectance ImageCollection
Apply cloud masking and Add vegetation indices bands
Calculate 31 temporally gap-filled vegetation time-series
Apply Smoothing
Complete dataset
Develop NDVI and GNDVI baseline ML Models
Optimum time for prediction
Feature Selection
• Correlation
• Recursive Feature Elimination
Develop and evaluate the field level ML models
Perform model selection to select one best performing model
Validate the final model
Compare the results with baseline models

Figure 3. Framework for the study

3. RESULTS AND DISCUSSION

The vegetation time series of all indices vary over time as a function of the standing cane's developmental phases. This temporal variation is analyzed to identify the optimum time for yield prediction using RS data in the study. The correlation between each feature variable and observed sugarcane yield was evaluated using Pearson's correlation coefficient. The 'r' values for ground truth variables ranged between -0.30 to 0.18, indicating a negligible correlation with yield. The correlation of vegetation indices started to see an increasing trend after May, attained a peak in the first fortnight of November, and then decreased in December. November where the highest correlation was observed for the majority of the indices, corresponds to the month when most of the sugarcane is towards the end of the grand growth phase and the start of the maturity phase, i.e., only two months before harvesting [37]. This trend in correlation lead us to identify November month as the right time for yield prediction. The variation in 'r' for vegetation indices with yield is shown in Figure 4. Among all optical variables, MRENDVI (r_{min}=-0.13, r_{max}=0.17, r_{nov}=-0.13) and S2REP (r_{min}=-0.11, r_{max}=0.21, r_{nov}=0.21) indicated low correlation with yield throughout the sugarcane crop cycle. Hence, these may not be the best indicators of sugarcane yield.

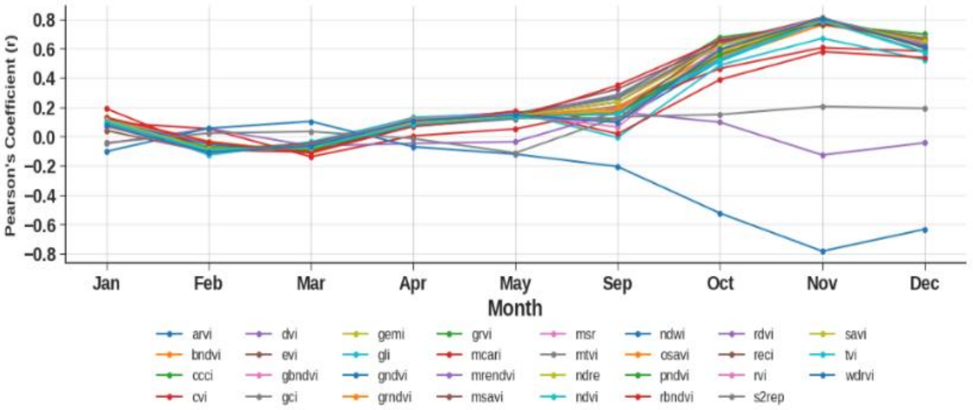
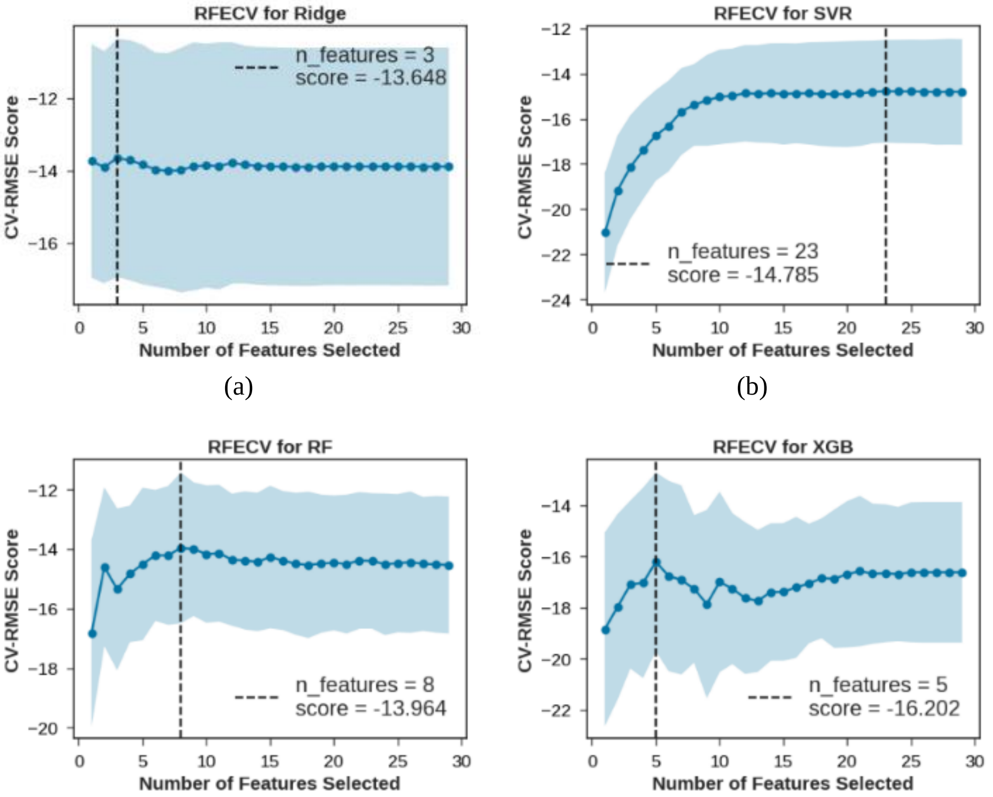


Figure 4. The variation in correlation coefficient for each VI with target yield between January to December

Since the correlation coefficient quantified the relationship between the features and yield, features with r≥±0.5 were selected for further analysis. This resulted in 29 out of 37 features we started with. Interestingly, all 29 features were the spectral vegetation indices having p-value<0.05, indicating their statistical significance with yield. Despite having a strong association with sugarcane yield, the 29 spectral vegetation indices that were chosen had moderate to high correlations with one another, causing feature redundancy. To minimize feature redundancy in the model, we decided to retain only those vegetation indices that were of the highest importance in predicting sugarcane yield using RFECV. The exact number of features that are important for yield modeling is unknown in advance. For this reason, we used RFECV over traditional RFE method. RFECV automatically determines the optimal number of features using cross-validation with RFE. This feature selection method's sensitivity to estimator choice is its limitation. In other words, a subset of features selected by one estimator used with RFECV algorithm may or might not be significant to another estimator. To mitigate this limitation, we evaluated the RFECV with 4 different estimators. The RFECV results for various estimators are presented in Figure 5, including Figures 5(a) to 5(d).

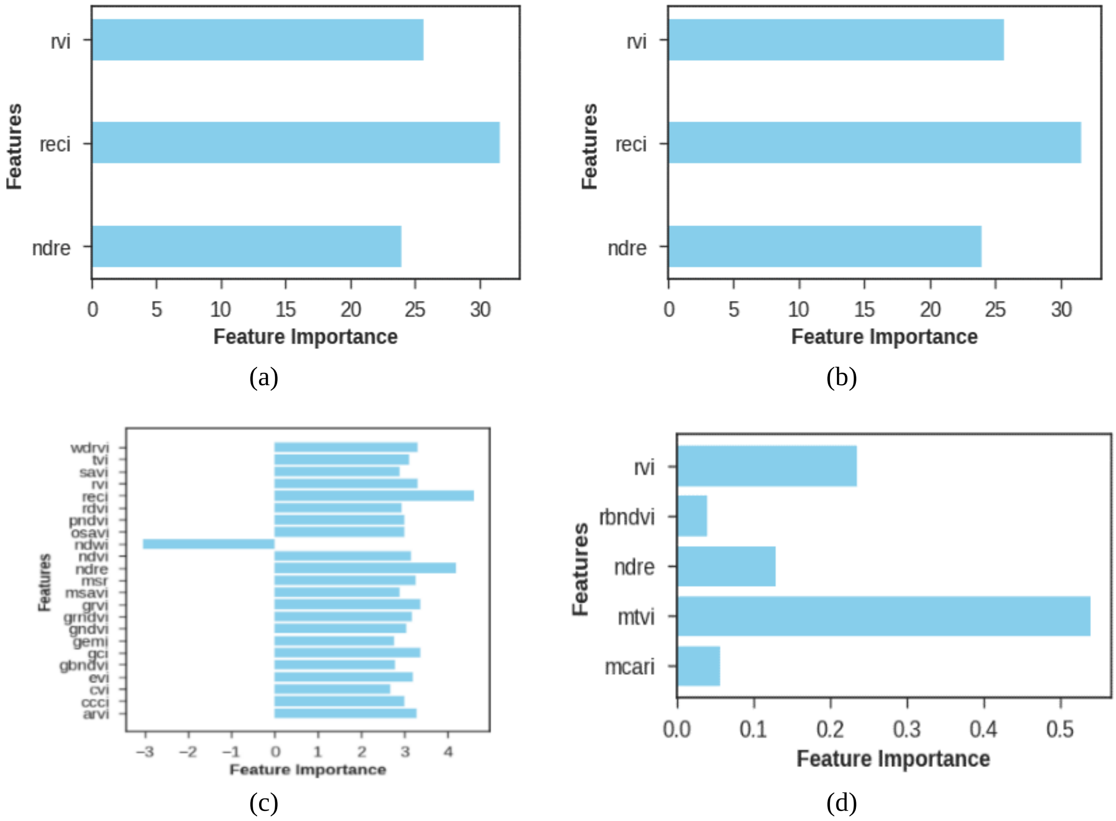


(a) (b)

(c) (d)

Figure 5. Applying: (a) Ridge-based RFECV, (b) SVR-based RFECV, (c) RFR-based RFECV, and (d) XGBoost regressor-based RFECV to the initial feature space

As expected, the optimum features obtained during the feature selection process varied depending on the estimator used, with RFECV. The feature subset chosen by each RFECV-estimator (Ridge, SVR, RF, and XGB) combination had notable differences. RFECV-Ridge and RFECV-RF chose fewer features with comparatively better RMSE due to their ability to reduce redundancy and model complexity. During our experimentation with different estimators, we evaluated the feature importance computed by the RFECV methods for the selected features. It can be noticed, that NDRE and RVI are deemed as significant by all 4 RFECV estimators with consistent high feature importance value. And RECI was chosen by 3 out of 4 estimators tested. The other indices were shortlisted either once or twice only. Thus, these 3 indices play a prominent role in predicting sugarcane yield. The optimum features selected by each RFECV estimator along with the importance values of the features are presented in Figure 6, including Figure 6(a) to 6(d).



(a) (b)

(c) (d)

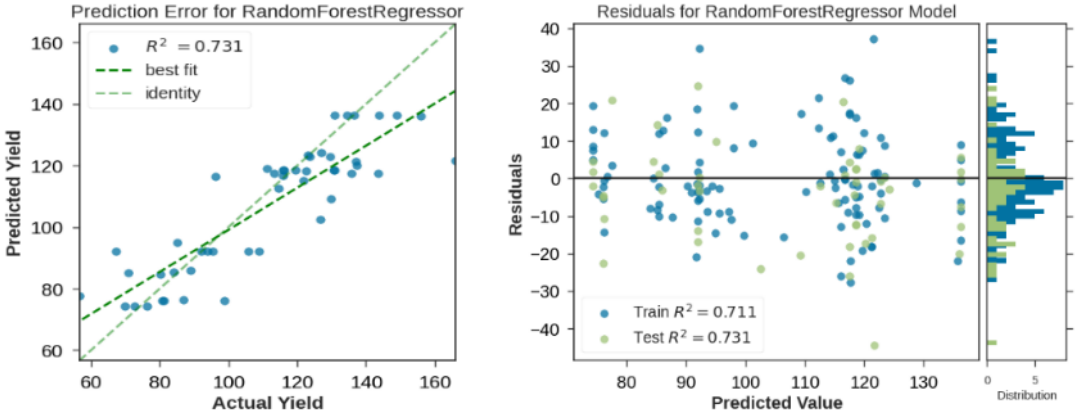
Figure 6. Feature importance of predictors estimated by the algorithms; (a) RFECV-Ridge, (b) RFECV-SVR, (c) RFECV-RF, and (d) RFECV-XGB

Though, the choice of 3 variables by RFECV-Ridge, seemed an ideal feature subset, choosing the best estimator for RFECV can depend on various factors, including the performance metrics, computational resources, and the specific characteristics of our data. To eliminate the possible trade-off between the number of optimum features and model performance, all seven ML models chosen for this study were trained using the optimal features from both RFECV-Ridge and RFECV-RF methods. Hyperparameter optimization and 10-fold cross-validation were also employed to mitigate the chances of underfitting or overfitting the model. This methodology resulted in fourteen sugarcane yield models. Based on the RMSE test score, the top 3 models are RF trained with RFECV-Ridge feature subset (R^2=0.731), RF trained with RFECV-RF feature subset (R^2=0.713), and Ridge trained with RFECV-RF feature subset (R^2=0.698). The accuracy metrics of all seven ML models considered for yield modeling are tabulated below in Table 2. A K-fold cross-validation procedure with k=10 was used to select one best model from the top three performing models stated above.

Table 2. The predictive performance of the ML models on training and testing datasets

Model	RFECV estimator: Ridge				RFECV estimator: RF			
	MAE (t/ha)		RMSE (t/ha)		MAE (t/ha)		RMSE (t/ha)	
	Train	Test	Train	Test	Train	Test	Train	Test
LR	11.250	10.855	13.789	14.433	11.348	11.704	14.525	15.169
Ridge	10.829	10.934	13.560	14.503	11.019	10.751	13.851	14.283
Elasticnet	15.459	18.385	18.528	21.547	14.062	15.818	17.024	19.134
SVR	10.811	10.889	13.497	14.320	11.186	10.981	13.931	14.463
RF	11.306	9.734	14.088	13.448	11.159	10.046	14.061	13.644
GBR	11.619	10.409	14.401	14.328	11.696	10.481	14.275	13.999
XGB	11.703	11.735	14.297	15.353	11.855	11.743	14.591	15.266

The RF model trained with features selected by RFECV-Ridge emerged as the best-fitted model with an RMSE=13.79 t/ha during the model selection procedure. In addition, an R^2=0.731 reflected that the model was able to explain the variance in yield reasonably well. The sugarcane yield prediction based on the RF model has higher accuracy than others. This could be attributed to the RF's ability to handle non-linear and dynamic input data, such as environmental impacts, and cropping systems. The RF model trained with RFECV-RF features subset was the second-best model with a closer RMSE score. This combination is computationally expensive, particularly for large samples at the production level. In our analysis [38] we achieved an R^2=0.71 with GNDVI alone as the yield predictor using Polynomial regression. However, from our current methodology, we see that a combination of indices is giving us an increased R^2=0.731, reflecting the fact that a combination of selected features are able to improve the yield accuracy. In addition, the current study, performed model validation on 2018-2019 and 2019-2020 growing season data. The best model i.e., RF, used NDRE, RECI, and RVI as the predictor variables as shown in Figure 7. The model exhibits a decent fit on training data as shown in Figure 7(a). From the residual plot in Figure 7(b), we see a fairly random and uniform distribution of the residuals against the target yield. From the histogram, we can also see that the error is normally distributed around zero, which generally indicates a well-fitted model.



(a) (b)

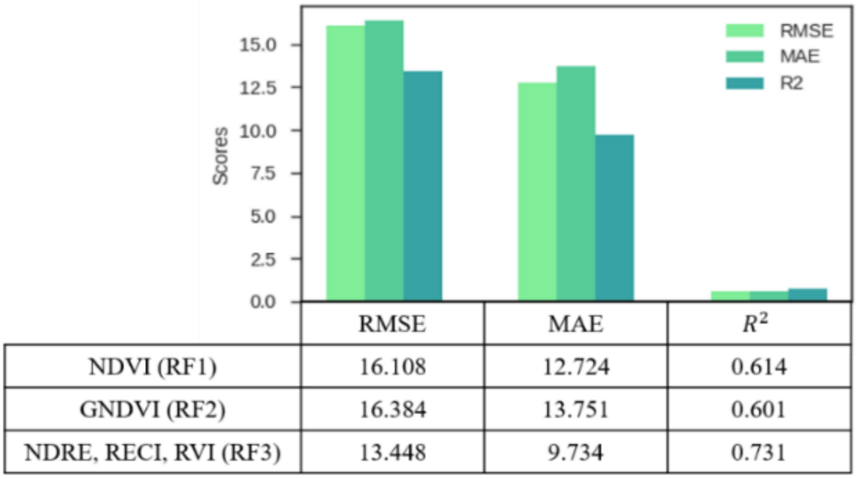
Figure 7. Plot of (a) a prediction error with actual yield plotted against predicted yield values by model with best-fit line and (b) residual showing the distribution of the difference between the actual yield and the predicted yield

In our study, the choice of indices - NDRE, RECI, and RVI for sugarcane yield prediction in the study area is completely data-driven. From all the experimentations, we could note that indices derived using red edge bands are favored compared to other indices. Hence the introduction of indices in the red edge bands can play a significant role in improving the sugarcane yield prediction model. While NDRE and RVI are a combination of the NIR band and red edge range between visible Red and NIR, RVI uses the NIR and Red band of Sentinel-2. Given agricultural monitoring, the roles of NDVI and NDRE are complementary. NDRE offers advantages in certain situations, particularly for deeper canopy measurements and areas where NDVI may reach saturation [39], [40]. Sugarcane is also a perennial crop. This makes NDRE particularly useful for deeper canopy assessment of sugarcane crops at maturity. Also at pixel level, when NDVI saturates and detects a consistent +1, NDRE can occasionally offer a more accurate assessment of variability in those regions [29], [41]. At maturity, sugarcane is associated with high canopy density making RVI a viable index for this study [20]. In November, while some cane has reached a stage of maturity, a certain portion of it is still in the mid-late maturity stage due to differences in the date of planting. In such a scenario, RECI helps in assessing the crop at a stage of active vegetation development [3]. The final model is validated for 20 fields and 16 fields for the growing season 2018-2019 and 2019-2020 respectively. The data of these fields was not exposed to the model during the training process. The validation results are presented in Table 3.

Table 3. The predictive performance of the final ML model on validation datasets

Model	2018 – 2019		2019 – 2020	
	MAE (t/ha)	RMSE (t/ha)	MAE (t/ha)	RMSE (t/ha)
RandomForestRegressor (n_estimators=100, max_depth=2, min_samples_leaf=2, random_state=42)	10.065	12.531	11.634	13.539

We also performed an extensive feature selection on the final RF model. Here, all possible feature combinations were tested without eliminating any feature. We could see that the combination of NDRE and RECI had RMSE=14.00 t/ha with a standard deviation of 3.00 t/ha compared to RMSE=13.49 t/ha and a standard deviation of 3.15 t/ha for NDRE and RVI. This shows that a combination of two or more indices as yield predictors is effective in sugarcane yield modeling compared to models with one yield predictor. In addition, red edge bands have the potential to lead to improved accuracy in yield prediction. Based on the literature, it was found NDVI and GNDVI are the most prevalent vegetation indices used for sugarcane yield prediction in advance of harvest. With an objective to have a baseline result against which the results of our approach could be compared, we implemented two RF models, one trained with NDVI (RF1) and another trained with GNDVI (RF2) feature as baseline models. The results of these models when compared to the RF results obtained through the proposed methodology (RF3) validate and highlight the usefulness of the approach proposed which has better accuracy comparatively. RF3 has shown improved accuracy by 16.5 % and 17.9% compared to RF1 and RF2 respectively as shown in Figure 8.



	RMSE	MAE	R^2
NDVI (RF1)	16.108	12.724	0.614
GNDVI (RF2)	16.384	13.751	0.601
NDRE, RECI, RVI (RF3)	13.448	9.734	0.731

Figure 8. Comparison of the proposed model with baseline models

4. CONCLUSION

The study aimed to establish a method to develop a sugarcane yield prediction model using ML algorithms and remote sensing data. For this purpose, we investigated the suitability of 31 vegetation indices including both red edge and narrowband greenness indices to explain sugarcane yield in the study area. We find that RF trained with NDRE, RECI and RVI indices as yield predictors efficiently provides a decent estimate of predicted yield two months in advance to harvest in the study area. Another prominent finding is that, due to its high canopy density at maturity, sugarcane yield may be more accurately estimated by vegetation indices based on wavelengths located in the red edge region of electromagnetic spectrum. Through our experimentations based on most commonly used indices for sugarcane yield, it is found that a combination of NDRE, RECI, and RVI indices provide higher accuracy than NDVI and GNDVI based ML models. Since sugarcane crop growth varies between localities with different growing weather conditions, precipitation, and environmental variables that in turn drive sugarcane yield, the study reflects on the fact that indices that explain yield have to be chosen with caution after their significance is tested to explain yield in the study area. The model can be improved by exploring more red edge band based vegetation indices from hyperspectral imageries. Also, verification of the method’s consistency through additional growing seasons is suggested.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support extended for this work by Shri Samir Somaiya (Chairman and MD, Somaiya group), sugarcane farmers of the region and Godavari Bio-refineries Ltd.

REFERENCES

[1] J. Som-ard, M. D. Hossain, S. Ninsawat, and V. Veerachitt, “Pre-harvest sugarcane yield estimation using UAV-based RGB images and ground observation,” *Sugar Tech*, vol. 20, no. 6, pp. 645–657, Dec. 2018, doi: 10.1007/s12355-018-0601-7.

[2] A. Nihar, N. R. Patel, and A. Danodia, “Machine-learning-based regional yield forecasting for sugarcane crop in Uttar Pradesh, India,” *Journal of the Indian Society of Remote Sensing*, vol. 50, no. 8, pp. 1519–1530, Aug. 2022, doi: 10.1007/s12524-022-01549-0.

[3] S. Akbarian, M. R. Jamnani, C. Xu, W. Wang, and S. Lim, “Plot level sugarcane yield estimation by machine learning on multispectral images: A case study of Bundaberg, Australia,” *Information Processing in Agriculture*, Jun. 2023, doi: 10.1016/j.inpa.2023.06.004.

[4] M. Hajeb, S. Hamzeh, S. K. Alavipanah, L. Neissi, and J. Verrelst, “Simultaneous retrieval of sugarcane variables from Sentinel-2 data using Bayesian regularized neural network,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 116, p. 103168, Feb. 2023, doi: 10.1016/j.jag.2022.103168.

[5] K. Amankulova, N. Farmonov, and L. Mucsi, “Time-series analysis of Sentinel-2 satellite images for sunflower yield estimation,” *Smart Agricultural Technology*, vol. 3, p. 100098, Feb. 2023, doi: 10.1016/j.atech.2022.100098.

[6] K. Amankulova, N. Farmonov, K. Omonov, M. Abdurakhimova, and L. Mucsi, “Integrating the Sentinel-1, Sentinel-2 and topographic data into soybean yield modelling using machine learning,” *Advances in Space Research*, vol. 73, no. 8, pp. 4052–4066, Apr. 2024, doi: 10.1016/j.asr.2024.01.040.

[7] A. Bégué *et al.*, “Spatio-temporal variability of sugarcane fields and recommendations for yield forecast using NDVI,” *International Journal of Remote Sensing*, vol. 31, no. 20, pp. 5391–5407, Oct. 2010, doi: 10.1080/01431160903349057.

[8] H. Jiang *et al.*, “Early season mapping of sugarcane by applying machine learning algorithms to Sentinel-1A/2 time series data: a case study in Zhanjiang City, China,” *Remote Sensing*, vol. 11, no. 7, p. 861, Apr. 2019, doi: 10.3390/RS11070861.

[9] K. Zhang *et al.*, “Predicting rice grain yield based on dynamic changes in vegetation indexes during early to mid-growth stages,” *Remote Sensing*, vol. 11, no. 4, p. 387, Feb. 2019, doi: 10.3390/rs11040387.

[10] M. M. Rahman and A. J. Robson, “A novel approach for sugarcane yield prediction using landsat time series imagery: a case study on Bundaberg Region,” *Advances in Remote Sensing*, vol. 05, no. 02, pp. 93–102, 2016, doi: 10.4236/ars.2016.52008.

[11] M. dos S. Simões, J. V. Rocha, and R. A. C. Lamparelli, “Variáveis espectrais e indicadores de desenvolvimento e produtividade da cana-de-açúcar,” *Scientia Agricola*, vol. 62, no. 3, pp. 199–207, Jun. 2005, doi: 10.1590/S0103-90162005000300001.

[12] M. Rahimi Jamnani, A. Liaghat, and F. Mirzaei, “Optimization of sugarcane harvest using remote sensing,” *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, vol. 42, no. 4/W18, pp. 857–861, Oct. 2019, doi: 10.5194/isprs-archives-XLII-4-W18-857-2019.

[13] J. L. Fernandes, N. F. F. Ebecken, and J. C. D. M. Esquerdo, “Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble,” *International Journal of Remote Sensing*, vol. 38, no. 16, pp. 4631–4644, Aug. 2017, doi: 10.1080/01431161.2017.1325531.

[14] M. Kumar, A. Das, K. N. Chaudhari, S. Dutta, K. K. Dakhore, and B. K. Bhattacharya, “Field-scale assessment of sugarcane for mill-level production forecasting using indian satellite data,” *Journal of the Indian Society of Remote Sensing*, vol. 50, no. 2, pp. 313–329, Feb. 2022, doi: 10.1007/s12524-021-01442-2.

[15] A. Das *et al.*, “Machine learning model ensemble for predicting sugarcane yield through synergy of optical and SAR remote sensing,” *Remote Sensing Applications: Society and Environment*, vol. 30, p. 100962, Apr. 2023, doi: 10.1016/j.rsase.2023.100962.

[16] T. van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: a systematic literature review,” *Computers and Electronics in Agriculture*, vol. 177, p. 105709, Oct. 2020, doi: 10.1016/j.compag.2020.105709.

[17] R. A. Medar, V. S. Rajpurohit, and A. M. Ambekar, “Sugarcane crop yield forecasting model using supervised machine learning,” *International Journal of Intelligent Systems and Applications*, vol. 11, no. 8, pp. 11–20, Aug. 2019, doi: 10.5815/ijisa.2019.08.02.

[18] A. C. dos S. Luciano, M. C. A. Picoli, D. G. Duft, J. V. Rocha, M. R. L. V. Leal, and G. le Maire, “Empirical model for forecasting sugarcane yield on a local scale in Brazil using Landsat imagery and random forest algorithm,” *Computers and Electronics in Agriculture*, vol. 184, p. 106063, May 2021, doi: 10.1016/j.compag.2021.106063.

[19] A. Bannari, D. Morin, F. Bonn, and A. R. Huete, “A review of vegetation indices,” *Remote Sensing Reviews*, vol. 13, no. 1–2, pp. 95–120, Aug. 1995, doi: 10.1080/02757259509532298.

[20] J. Xue and B. Su, “Significant remote sensing vegetation indices: a review of developments and applications,” *Journal of Sensors*, vol. 2017, pp. 1–17, 2017, doi: 10.1155/2017/1353691.

[21] J. Som-Ard, C. Atzberger, E. Izquierdo-Verdiguier, F. Vuolo, and M. Immitzer, “Remote sensing applications in sugarcane cultivation: A review,” *Remote Sensing*, vol. 13, no. 20, p. 4040, Oct. 2021, doi: 10.3390/rs13204040.

[22] A. Agapiou, D. G. Hadjimitsis, and D. D. Alexakis, “Evaluation of broadband and narrowband vegetation indices for the identification of archaeological crop marks,” *Remote Sensing*, vol. 4, no. 12, pp. 3892–3919, Dec. 2012, doi: 10.3390/rs4123892.

[23] D. Haboudane, J. R. Miller, E. Pattey, P. J. Zarco-Tejada, and I. B. Strachan, “Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: modeling and validation in the context of precision agriculture,” *Remote Sensing of Environment*, vol. 90, no. 3, pp. 337–352, Apr. 2004, doi: 10.1016/j.rse.2003.12.013.

[24] J. Brinkhoff, R. Houborg, and B. W. Dunn, “Rice ponding date detection in Australia using Sentinel-2 and Planet Fusion imagery,” *Agricultural Water Management*, vol. 273, p. 107907, Nov. 2022, doi: 10.1016/j.agwat.2022.107907.

[25] R. Ni *et al.*, “An enhanced pixel-based phenological feature for accurate paddy rice mapping with Sentinel-2 imagery in Google Earth Engine,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 178, pp. 282–296, Aug. 2021, doi: 10.1016/j.isprsjprs.2021.06.018.

[26] S. Wolters, M. Söderström, K. Piikki, H. Reese, and M. Stenberg, “Upscaling proximal sensor N-uptake predictions in winter wheat (Triticum aestivum L.) with Sentinel-2 satellite data for use in a decision support system,” *Precision Agriculture*, vol. 22, no. 4, pp. 1263–1283, Aug. 2021, doi: 10.1007/s11119-020-09783-7.

[27] Y. Zheng, A. C. Dos Santos Luciano, J. Dong, and W. Yuan, “High-resolution map of sugarcane cultivation in Brazil using a phenology-based method,” *Earth System Science Data*, vol. 14, no. 4, pp. 2065–2080, 2022, doi: 10.5194/essd-14-2065-2022.

[28] D. Agrafiotis, “Video error concealment,” in *Academic Press Library in Signal Processing*, 2014, pp. 295–321.

[29] Y. Kanke, B. Tubaña, M. Dalen, and D. Harrell, “Evaluation of red and red-edge reflectance-based vegetation indices for rice biomass and grain yield prediction models in paddy fields,” *Precision Agriculture*, vol. 17, no. 5, pp. 507–530, Oct. 2016, doi: 10.1007/s11119-016-9433-1.

[30] S. Akbarian, C. Xu, W. Wang, S. Ginns, and S. Lim, “Sugarcane yields prediction at the row level using a novel cross-validation approach to multi-year multispectral images,” *Computers and Electronics in Agriculture*, vol. 198, p. 107024, Jul. 2022, doi: 10.1016/j.compag.2022.107024.

[31] W. J. Frampton, J. Dash, G. Watmough, and E. J. Milton, "Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 82, pp. 83–92, Aug. 2013, doi: 10.1016/j.isprsjprs.2013.04.007.

[32] D. Tanré, B. N. Holben, and Y. J. Kaufman, "Atmospheric correction algorithm for NOAA-AVHRR products: theory and application," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 30, no. 2, pp. 231–248, 1992, doi: 10.1109/36.134074.

[33] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, vol. 202, pp. 18–27, Dec. 2017, doi: 10.1016/j.rse.2017.06.031.

[34] U. Gandhi, "End-To-End Google Earth Engine – Spatial Thoughts," *Spatialthoughts*, 2006. https://spatialthoughts.com/courses/google-earth-engine/.

[35] S. P. Raja, B. Sawicka, Z. Stamenkovic, and G. Mariammal, "Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers," *IEEE Access*, vol. 10, pp. 23625–23641, 2022, doi: 10.1109/ACCESS.2022.3154350.

[36] M. Yoosefzadeh-Najafabadi, H. J. Earl, D. Tulpan, J. Sulik, and M. Eskandari, "Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in Soybean," *Frontiers in Plant Science*, vol. 11, Jan. 2021, doi: 10.3389/fpls.2020.624273.

[37] A. K. Verma, P. K. Garg, and K. S. Hari Prasad, "Sugarcane crop identification from LISS IV data using ISODATA, MLC, and indices based decision tree approach," *Arabian Journal of Geosciences*, vol. 10, no. 1, p. 16, Jan. 2017, doi: 10.1007/s12517-016-2815-x.

[38] S. K. Jha *et al.*, "Sugarcane yield prediction using vegetation indices in Northern Karnataka, India," *Universal Journal of Agricultural Research*, vol. 10, no. 6, pp. 699–721, Dec. 2022, doi: 10.13189/ujar.2022.100611.

[39] N. S. Naguib and S. Daliman, "Analysis of NDVI and NDRE indices using satellite images for crop identification at Kelantan," *IOP Conference Series: Earth and Environmental Science*, vol. 1102, no. 1, p. 012054, Nov. 2022, doi: 10.1088/1755-1315/1102/1/012054.

[40] O. Mutanga and A. K. Skidmore, "Narrow band vegetation indices overcome the saturation problem in biomass estimation," *International Journal of Remote Sensing*, vol. 25, no. 19, pp. 3999–4014, Oct. 2004, doi: 10.1080/01431160310001654923.

[41] B. Boiarskii, "Comparison of NDVI and NDRE indices to detect differences in vegetation and chlorophyll content," *Journal of Mechanics of Continua and Mathematical Sciences*, vol. spl1, no. 4, Nov. 2019, doi: 10.26782/jmcms.spl.4/2019.11.00003.

BIOGRAPHIES OF AUTHORS

Rekha B. U. pursuing Ph.D. at Department of Electronics and Communication Engineering, KLS Gogte Institute of Technology, Belagavi, Karnataka, India, in the domain of Machine Learning and Remote Sensing with applications in agriculture and crop production. Has 5 years of teaching experience as assistant professor and 3 years of experience at EdTech industry in curriculum design and development. She can be contacted at email: rekha553@gmail.com.

Dr. Veena V. Desai holds a Ph.D. in Electronics, a Masters in Computer Network Engineering, a Bachelor's degree in Electronics & Communication Engineering, and a Post Graduate Diploma in Cyber laws from the Asian School of Cyber Laws, Mumbai University. Has nearly 30 years of teaching experience as an Academician. Areas of research include computational intelligence applications, network security, biomedical signal processing, and issues concerning agriculture. With over 50 indexed and peer-reviewed research publications, she has served as chair and reviewer at several international conferences and journals. She is a Member of IEEE, a Life member of ISTE, the Computer Society of India, IETE, and the Cryptographic Research Society of India, CRSI. She can be contacted at email: veenades@gmail.com.

Dr. Suresh Kuri working as Associate Professor, Department of Electronics and Communication Engineering, KLS Gogte Institute of Technology, Belagavi, Karnataka, India. Has over 25 years of teaching experience as an Academician. His expertise includes digital electronics, signal processing, digital image watermarking, and neural networks. He can be contacted at email: sureshkuri@git.edu.

Pratijnya S. Ajawan working as Assistant Professor at Department of Electronics and Communication Engineering, KLS Gogte Institute of Technology, Belagavi Karnataka, India. A research scholar with interests in machine learning, artificial intelligence, and natural language processing domain. She can be contacted at email: psajawan@git.edu.

Sunil Kumar Jha is pursuing Ph.D. at Applied Agricultural Remote Sensing Centre, University of New England, Armidale, Australia. He is involved in studying remote monitoring of agricultural dynamics. His research interests include crop identification, crop mapping, yield estimation, rice science, ET mapping at the local and regional scale, crop water stress assessment, chlorophyll mapping, nitrogen mapping, biomass estimation, digital elevation model (DEM), flood mapping using UAV, optical and SAR data. He can be contacted at email: sonu2007.sunil@gmail.com.

Dr. V. C. Patil served as Director at K J Somaiya Institute of Applied Agricultural Research (KIAAR), Sameerwadi, Karnataka, India. With more than 30 years of experience at academic institutions and research organizations, his skills and expertise include plant pathology, plant breeding and plant ecology with particular interests in precision agriculture, precision sugarcane cultivation, sugarcane yield prediction models, and in-situ monitoring of sugarcane crop using remote sensing. He can be contacted at email: vcpatilksu@gmail.com.