Water Stress Prediction Model in Rainfed Sugarcane Fields Using Temporal Landsat Data Based on Random Forest Regressor

Aries Suharso, *Member, IAENG,* Yeni Herdiyeni, Suria Darma Tarigan, and Yandra Arkeman

***Abstract*— Sugarcane drought (water stress) at the critical germination and tillering phases results in a decrease in milled sugarcane production every year. Using temporal data from Landsat 8 satellite images and local meteorological data, this study attempts to assess sugarcane water stress using upper canopy reflectance (LST). The difficulty in obtaining when calculating canopy temperature is quite complicated, we propose a sugarcane water stress indices prediction model based on the Random Forest Regressor machine learning algorithm with other multivariate vegetative feature data NDVI, NDWI, NDDI, OSAVI, LSWI combined with local daily climate data in the form of air temperature, humidity, rainfall, sunshine hours and wind speed. This data is relatively easier to obtain compared to LST. The research study was carried out at the Indonesian Sugar Research Center's sugarcane plantation in Djengkol Kediri, East Java, Indonesia. Observations were focused on the sugarcane plantation area that was suspected of experiencing the most severe decline in milled sugarcane yields (G33). Initial analysis showed an increase in water stress phenology during the germination - tillering phase between June - October in 2021 and 2022. Through cross-correlation tests, and time lag effect tests, we compiled the dataset. The prediction performance results of our proposed Random Forest Regressor Model achieved the best performance with the vegetation dataset without LST achieving an accuracy of $R2$ = 91.08% and MAPE = 8.93%. These results emphasize that the multi-feature method, excluding LST, was effective in forecasting variations in the sugarcane water stress index. Consequently, this approach is anticipated to mitigate potential losses in future sugarcane milling productivity.**

***Index Terms*—Sugarcane, water stress, vegetative indices, prediction model.**

I. INTRODUCTION

Most of Indonesia's sugarcane fields are rainfed, making them highly susceptible to the country's climate and environmental variations. This susceptibility poses a significant challenge for sugarcane productivity in Indonesia. Prolonged dry seasons create critical conditions by reducing rainfall and water availability for soil absorption during the vegetative development phase [1] - [3]. According to the PTPN study, as illustrated in Figure 1, the milled sugarcane production (tons per hectare) for the sugarcane plantation areas G30, G31, G32, G33, G34, G35, and G36 decreased during the 2021–2022 ratoon1 planting season as opposed to the 2022–2023 ratoon2 planting season.

The worst production decline occurred in the G33 sugarcane land, reaching a 37% production loss. Therefore, we will focus our research on the G33 sugarcane land area. Field officers also provided additional information indicating evidence of stunted sugarcane stem growth. Based on the manager's report and field findings, we suspect that the location is experiencing water stress conditions during the critical phase of sugarcane growth. Understanding sugarcane's response to water stress is crucial, especially during the germination and tillering phases, as it affects growth and productivity [4] - [6].

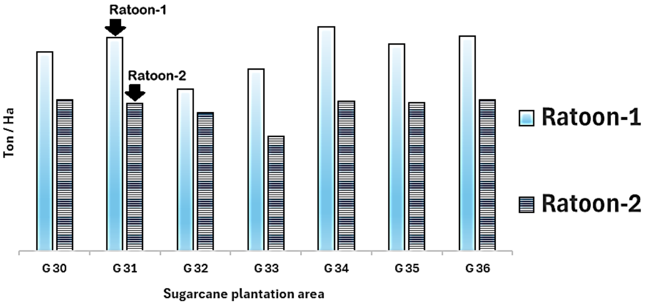


Fig. 1. Comparison of milled sugarcane production Tons per hectare of sugarcane fields G30, G31, G32, G33, G34, G35 and G36 planting season period ratoon1 vs ratoon2

Understanding sugarcane's response to water stress is crucial, especially during the germination and tillering phases, as it affects growth and productivity [6]. Plant phenology, which studies vegetation development throughout the year, can be observed through field observations and remote sensing [7], [8]. Instead of using time-consuming and costly field observations of species-specific phenological responses, we conducted sugarcane phenology observations using Landsat 8 satellite products and local daily climate data. This research evaluates drought stress in plants by assessing plant health, soil moisture levels, surface water availability, and plant canopy temperature, derived from remote sensing feature extraction using Landsat 8 satellite data [9]. The remote sensing data from Landsat 8 satellite imagery was obtained from USGS, and daily climate data was sourced from the nearest Meteorology, Climatology, and Geophysics Agency. Other vegetation indices used include NDVI, NDWI, NDDI, LSWI, OSAVI, and LST.

The study has three main objectives: (1) to calculate the sugarcane water stress index (CWSI); (2) to analyze changes in sugarcane phenology during the germination and tillering periods using harmonic models and time lag cross-correlation; and (3) to develop a prediction model for the

Manuscript received July 1, 2024; revised February 18, 2025.

Suharso. A. is a PhD candidate of the Graduate School of Computer Science, IPB University, Bogor, Indonesia. (correspondence author to provide phone: +62 812 807 222 69; email: arso220377aries@apps.ipb.ac.id, aries.suharso@unsika.ac.id).

Herdiyeni. Y. is a professor of Computer Science Department of Mathematics and Natural Sciences Faculty, IPB University, Bogor, Indonesia (email: yeni.herdiyeni@apps.ipb.ac.id).

Tarigan. S.D. is a professor of Soil and Land Resources Science Department of Agriculture Faculty, IPB University, Bogor, Indonesia (email: sdtarigan@apps.ipb.ac.id).

Arkeman. Y. is a professor of the Agricultural Industrial Technology Department of IPB University, Bogor, Indonesia (email: yandra@apps.ipb.ac.id).

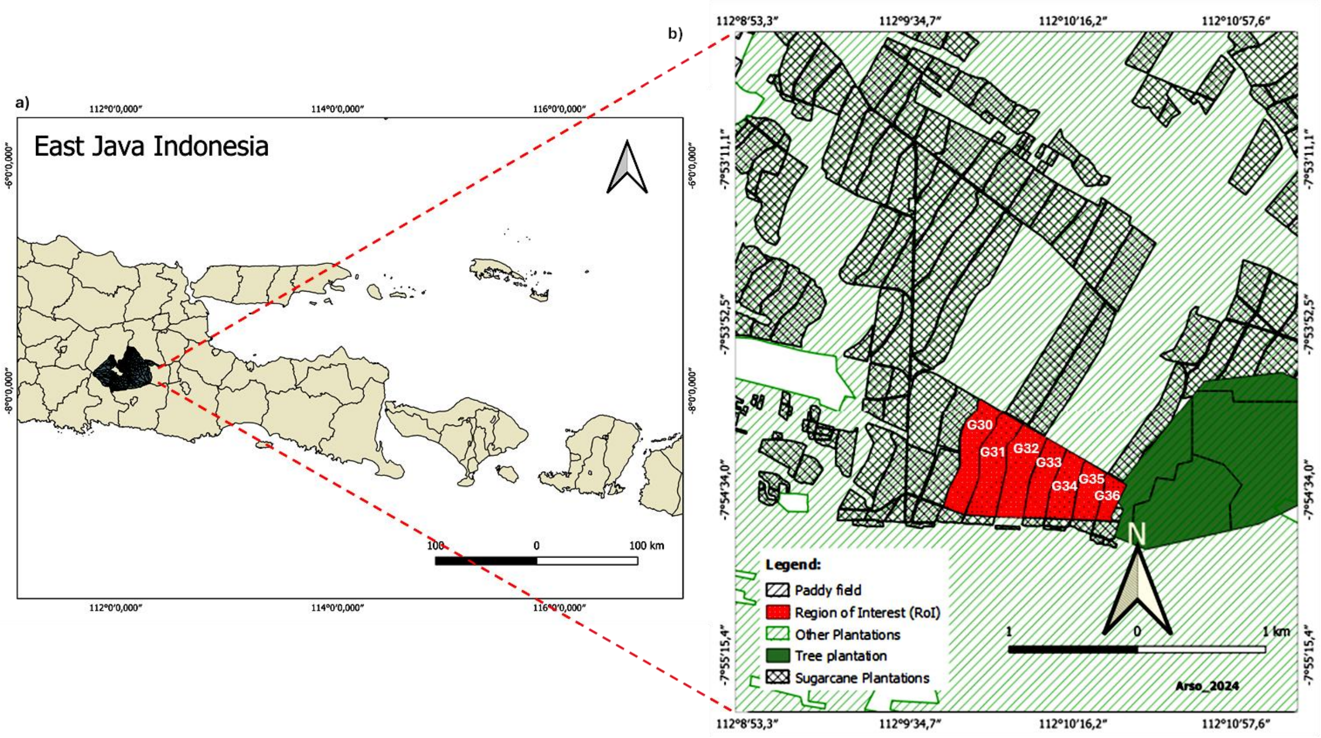


Fig. 2. Research location a) Indonesia, b) Sugarcane Plantation Area of PTPN-X East Java Province, red is the Region of Interest Sugarcane Field (G30, G31, G32, G33, G34, G35 and G36)

sugarcane water stress index using the random forest regressor (RFR) algorithm.

II. MATERIAL AND METHODS

A. *Research location*

Based on Figure 2 highlighted in red, the research location consists of seven agro-industrial sugarcane plantation plots (G30, G31, G32, G33, G34, G35, and G36) under the auspices of PTPN Plosoklaten in Kediri Regency, East Java. The coordinates are Lat 112.16467404367307, Long 7.904521068941556. The area of the sugarcane plantation is 4,900 Ha2. The nature of the Regosol soil is grayish brown. The topography at the foot of Mount Kelud is mostly undulating and hilly. Category Flat land at an altitude of between 292 and 323 meters above sea level. The slope of the land surface is relatively flat between 1% and 4%. The soil in this area has light surface erosion, moderate surface flow, rather slow permeability, and moderate drainage.

B. *Research data*

This study collected data on the sugarcane planting schedule and temporal spectral vegetation data from the Landsat 8 satellite. Focusing on G33 sugarcane land, as shown in Table I.

TABLE I
RATOON SUGAR CANE PLANTING SCHEDULE

Period of Ratoon	Germ	Till	GG	MR	Harvest
	(0 - 45)	(45 - 120)	(120 - 250)	(250 - 365)	(365 -)
2021 - 2022	06A - 08A	08A - 10A	10A - 02A	02A - 06A	06A - 07A
2022 - 2023	07B - 09B	09B - 11B	11B - 03B	03B - 07B	07B - 08B

Notes: A is the first 2 weeks of the month, B is the last 2 weeks of the month.

The ratoon sugarcane planting seasons occurred from 2021 to 2022 (ratoon1) and 2022 to 2023 (ratoon2). Each planting season was divided into five phases: germination, tillering, grand growth, mature-ripening, and harvesting. Temporal spectral data from Landsat-8, sourced from the USGS.gov.id site, was processed on the Google Earth Engine platform with a maximum cloud cover limitation of 30%, resulting in 174 images (Fig. 3). The following spectral selections were used: Long Wave Infrared1 Band 10 (LWIR1), Panchromatic Wave Band 8, Short Wave Infrared1 Band 6 (SWIR1), Short Wave Infrared2 Band 7 (SWIR2), Near Infrared (NIR) Wave Band 5, Blue Wave Band 2, Green Wave Band 3, and Red Wave Band 4. Climate variable data, such as average air temperature, air humidity, rainfall, sunshine length, and average wind speed, are available from the regional Meteorology, Climatology, and Geophysics Agency for the same time period as the observed sugarcane growing season.

1) Normalized Difference Vegetation Index (NDVI)

NDVI is an index used to measure photosynthetic activity and vegetation health conditions in a particular area or region [10], [11].

$$NDVI = (NIR - Red) / (NIR + Red) \qquad (1)$$

NIR stands for near-infrared light reflectance (Band 5), and Red for red light reflectance (Band 4). NDVI values range from -1 to +1. Healthy, abundant vegetation is indicated by a high NDVI value (closer to +1), while non-vegetated locations such as buildings or water are indicated by low values (closer to -1). Water surfaces and other non-vegetated objects typically have a negative NDVI rating.

2) Normalized Difference Water Index (NDWI)

The NDWI is employed to determine the presence of water or moisture content in soil, vegetation, or surface water within a specific area [12], [13]. NDWI is computed by contrasting the light reflectance at two

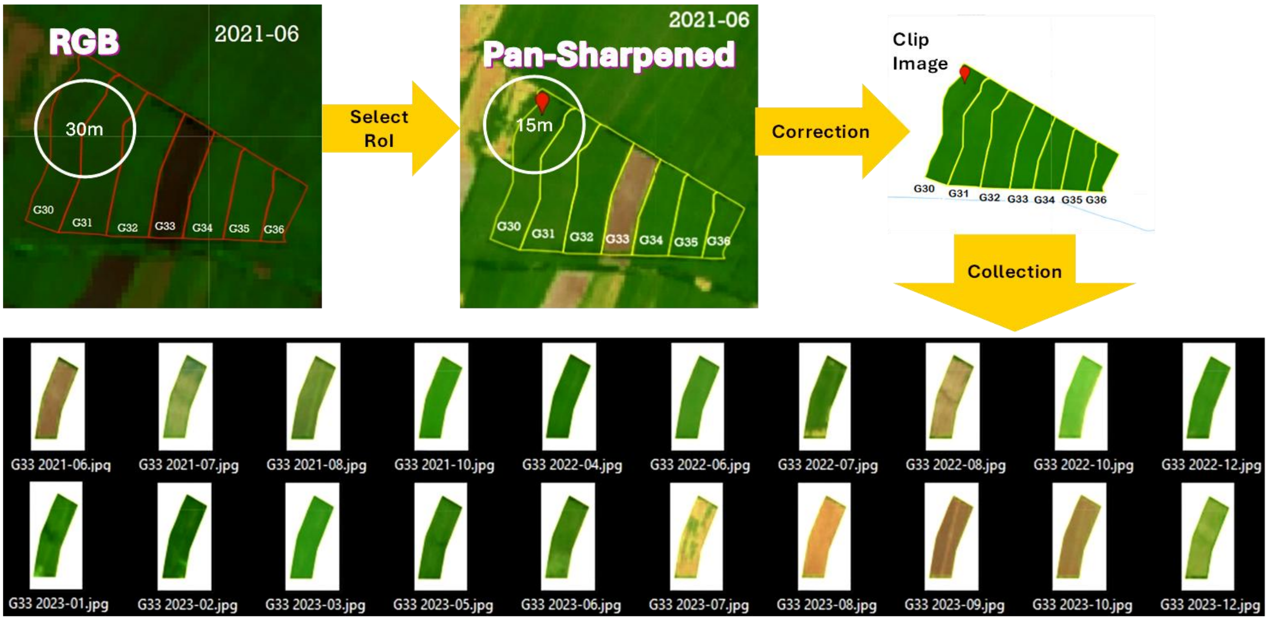


Fig. 3. Satellite Imagery Data Collection Acquisition Process

distinct wavelengths, specifically in the near infrared (NIR) region and the shortwave infrared (SWIR) region.

$$NDWI = (NIR - SWIR1) / (NIR + SWIR1) \qquad (2)$$

Near infrared light reflectance is denoted by NIR (Band 5), and far infrared light reflectance is denoted by SWIR1 (Band 6). The range of NDWI is -1 to +1. Lower NDWI values (near to -1) suggest dry or water-scarce areas, whereas higher values (closer to +1) indicate places with abundant water.

3) Normalized Difference Drought Index (NDDI)

According to Gu et al. (2007), the NDDI is a drought index that combines the NDVI and NDWI algorithms. This combination aids in determining the amount of moisture in the soil and the existence of robust vegetation [14], [15].

$$NDDI = (NDVI - NDWI)/(NDVI + NDWI) \qquad (3)$$

NDDI values range from -1 to +1. Interpretation of NDDI has a useful classification: if the NDDI value is between 0.01 and 0.15, the region is experiencing mild drought, while between 0.15 and 0.25, the region is experiencing moderate drought. NDDI values between 0.25 and 1 indicate severe drought. When the NDDI exceeds 1, the region is experiencing very severe drought.

4) Land Surface Water Index (LSWI)

LSWI is an index that evaluates the water content in plant leaves. Changes in LSWI can provide an indication of the level of water stress in plants. LSWI is calculated by comparing the reflectance of light at two different wavelengths, namely in the NIR region and in the SWIR2 (Shortwave Infrared 2) region [16].

$$LSWI = (NIR - SWIR2) \backslash (NIR + SWIR2) \qquad (4)$$

With NIR: near infrared light reflectance, value (Band 5). SWIR2: short infrared light reflectance (Band 7). The range of values for LSWI is -1 to +1. While low values (near -1) suggest drought conditions or soil devoid of water, high values (near +1) indicate the presence of ample surface water.

5) Optimized Soil Adjusted Vegetation Index (OSAVI).

The OSAVI offers an improved assessment of plant conditions in relation to soil moisture by making specific adjustments to light reflectance in the NIR and Red light regions [17].

$$OSAVI = (NIR - Red) / (NIR + Red + L) \qquad (5)$$

With NIR: near infrared light reflectance, Red: red light reflectance, and L: adjustment factor usually in the range (0.16 to 0.2) to compensate for the influence of ground conditions.

6) Land Surface Temperature (LST)

LST is the temperature measured without accounting for air factors, straight from the Earth's surface. LST is essential for environmental science, hydrology, agriculture, and climate monitoring, among other disciplines [18] - [20].

$$LST = [BT / (1 + L\lambda(BT/p) * ln(\varepsilon\lambda)] \qquad (6)$$

With

$$BT = [K2 / ln\ (K1/ L\lambda) + 1] - 273.15 \qquad (7)$$

$$L\lambda = ML\ .\ Qcal + AL \qquad (8)$$

$$Pv = \left[\frac{(NDVI - NDVImin)}{(NDVImax - NDVImin)}\right]^2 \qquad (9)$$

$$e\lambda = \varepsilon v\lambda + Pv + \varepsilon s\lambda(1\text{-}Pv) + C\lambda \qquad (10)$$

$$p = h(c/\lambda) = 1.438\ x\ 10\text{^-}2\ mK \qquad (11)$$

The following describes the variables that are utilized in the LST formula: BT = ToA Brightness Temperature (ºC); Lλ = ToA Radiant Spectral Value; ML = Radiance Multiplicative Band; AL = Radiance Add

Band; Qcal = Quantized and calibrated standard product pixel value (DN); K1 = Thermal conversion constant 1; K2 = Thermal conversion constant 2;Pv = Vegetation Fraction; NDVImax = Maximum NDVI value; NDVImin = Minimum NDVI value; ελ = Land surface emissivity; C = Surface roughness (c = 0 for a homogeneous flat surface); εsλ = 0.996 (if NDVI is between 0 – 0.2), 0.973 (if NDVI value is greater than 0.5); p = radiation function (1.438x10-2 mK); h = Planck's constant (6.26x10-34 J sec); c = Speed of light (2.998 x108 m sec-1); λ = Stefan Boltzman constant (1.38x10-23 JK-1); ε = emissivity of the object.

7) Brightness Temperature (BT)
BT is the temperature recorded by a satellite sensor without considering the effects of the atmosphere. This is the temperature that would be observed if the object was heated only by electromagnetic radiation and does not take into account the effects of atmospheric absorption, scattering, and emission. Brightness Temperature (BT) was first introduced by Sir Arthur W. S. Taylor in the context of remote sensing and satellite image processing in 1978. The formula used to calculate BT from TIRS Landsat-8 is:

$$BT = K2 / \ln((K1/L\lambda) + 1) \qquad (12)$$

BT stands for brightness temperature in Kelvin (K), and K1 and K2 are TIRS-specific calibration constants. Lλ is the brightness measured by Landsat-8 at a specific wavelength. Instead of being a direct digital number from the satellite image, the value of (Lλ) is expressed as radiance (Watts per square meter per steradian per micrometer) [21].

8) Crops Water Stress Index (CWSI)
The CWSI is utilized to assess and monitor drought levels in crops or agricultural plants. Jackson et al. first presented the CWSI in 1981. The fundamental formula for calculating CWSI has been refined through recent advancements [22] is:

$$CWSI = (Ts - Tcold) /(Thot - Tcold) \qquad (13)$$

Where Ts is the leaf temperature converted to LST; Tcold is the ambient air temperature converted to LSTmin; Thot is the maximum temperature that can be achieved by the plant in a non-drought converted to LSTmax.

A number of levels can be distinguished in the CWSI threshold for sugarcane based on the modified check and balance results of [23], [24]: The CWSI value can be classified as follows: No Water Stress if it is less than 0.2, Low Water Stress if it is between 0.2 and 0.4, Moderate Water Stress if it is between 0.4 and 0.6, and High Water Stress (Drought) if it is between 0.6 and 0.8. The area is considered to be bare land when the CWSI value is greater than 0.8 and almost 1.0, which is the most severe condition.

C. Imputation and Filling in Gaps

Imputation and gap-filling techniques are applied to Vegetation Index (VI) time series to address missing information due to cloud cover, which causes temporal and geographic data gaps and biases in future image processing and application. We use state-of-the-art temporal-based methodologies or temporal "gap-filling" techniques, such as data fusion, pixel blending, data interpolation, or best pixel selection, to create gap-filled satellite imagery to map land cover in the observation zone [23], [24]. The following is the equation (14) for the linear interpolation function.

$$f(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{(x_1 - x_0)}(x - x_0) \qquad (14)$$

For each value *x* of the independent variable, the value of the dependent variable is expressed by the function *f(x)*. In this case, the independent variable is *x*, and the known values of the independent variable are x_1 and x_0. Plant phenology trends can be assessed, and noise can be reduced through time series smoothing techniques [25], [26]. In this study, the central point in the time series data of the Vegetation Index feature is replaced with an equation-based average of all points within a fixed-size moving window, utilizing an interpolation method and a moving average (MA) filter. This technique has the advantage of preserving the peaks of the seasonal curve.

D. Smoothing with Harmonic series

Figure 4, Shows Sinusoidal harmonic series, which can express a function with periodicity as a sum of sinusoidal waves with frequencies different from the primary frequency, are used in methods to smooth time series data [27].

Harmonic series can also help reduce interference. This is especially true if the data noise is essentially random and does not have a periodic structure. Harmonic series methods can help "filter out" random noise by representing periodic trends and patterns, which will help make patterns and trends easier to see. The harmonic series equation (15) that we use in this study is as follows [28], [29].

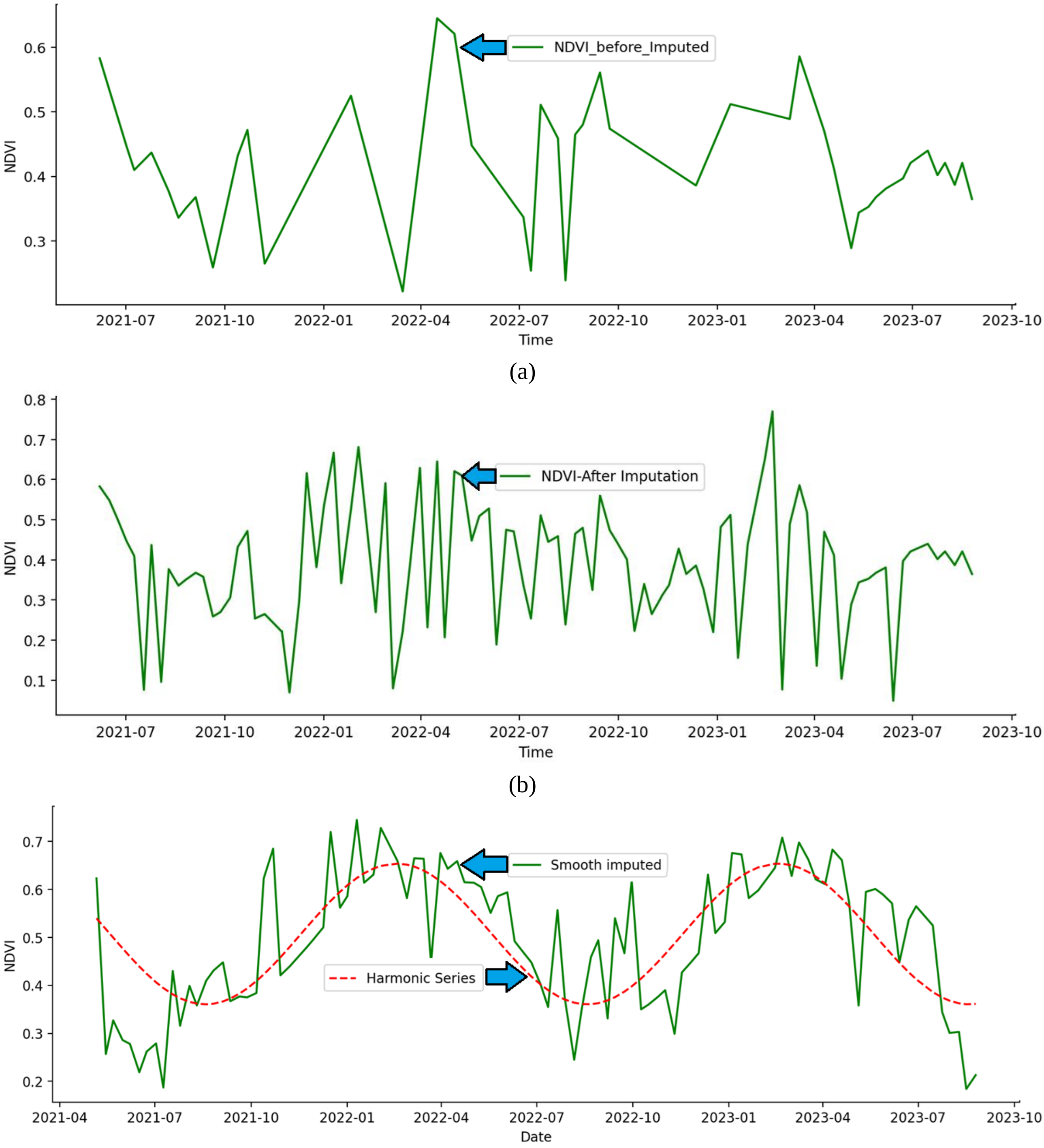
$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n cos\ (n\omega t)) + b_n\ sin\ (n\omega t) \qquad (15)$$

The frequency, or period value, of the sinusoidal wave in the series is denotedby *f(t)*. Periodic, on the other hand, is the cycle length that denotes the separation between the function's two repeating points. The unit of measurement known as "omega" (*ω*) is radians per unit time (2*π*) of a regular frequency rotation. It is commonly utilized in relation to Fourier and harmonic series. In the meanwhile, the series amplitude of sinusoidal waves at various frequencies is given by the coefficients a_0, a_n, and b_n.

E. Development water stress prediction model

The stages carried out in creating a prediction model, as in Figure 5. The following are the steps related to data analysis and development of a water stress prediction model:

1) *Dataset*: This stage is the starting point of the process produced through the collection of raw data from various sources (USGS for Landsat8 ToA imagery, PTPN-X HGU management data for the 2019-2023 sugarcane planting schedule, vector maps of sugarcane land area).
2) *Data repair*: at this stage, repair of missing data or disturbed data is carried out through imputation techniques



(a)

(b)

(c)

Fig. 4. Time series data correction for vegetation indices (a) Raw data series, (b) Imputation missing value with linear interpolation, (c) Trend smoothing with harmonic series method

and smoothing of data distribution patterns through harmonic techniques.

3) *Data preprocessing*: includes Data Scale Standardization, data structuring based on date, formation of data feature variations using Lag and Rolling techniques, then data cleaning with NaN values.

4) *Handling overfitting*: Cross Validation is carried out to separate Training Data and Test Data, overfitting control through k-fold cross validation, and Grid search technique evaluates all combinations of hyperparameters given with tuning parameters (nest = n-estimator, md = maximum depth, mf = maximum features, msl = minimum sample leaf, and mss = minimum sample split). The best hyperparameter combination is selected based on the average score of cross-validation [30], [31].

5) *Model application*: Training data is entered into the basic Random Forest Regressor (RFR) algorithm. which generates its final prediction by averaging the predictions of all the individual decision trees (16).

$$\hat{Y} = argmax_{y\in Y}\frac{1}{T}\sum_{t=1}^{T}P(y\,|x,P_t) \qquad (16)$$

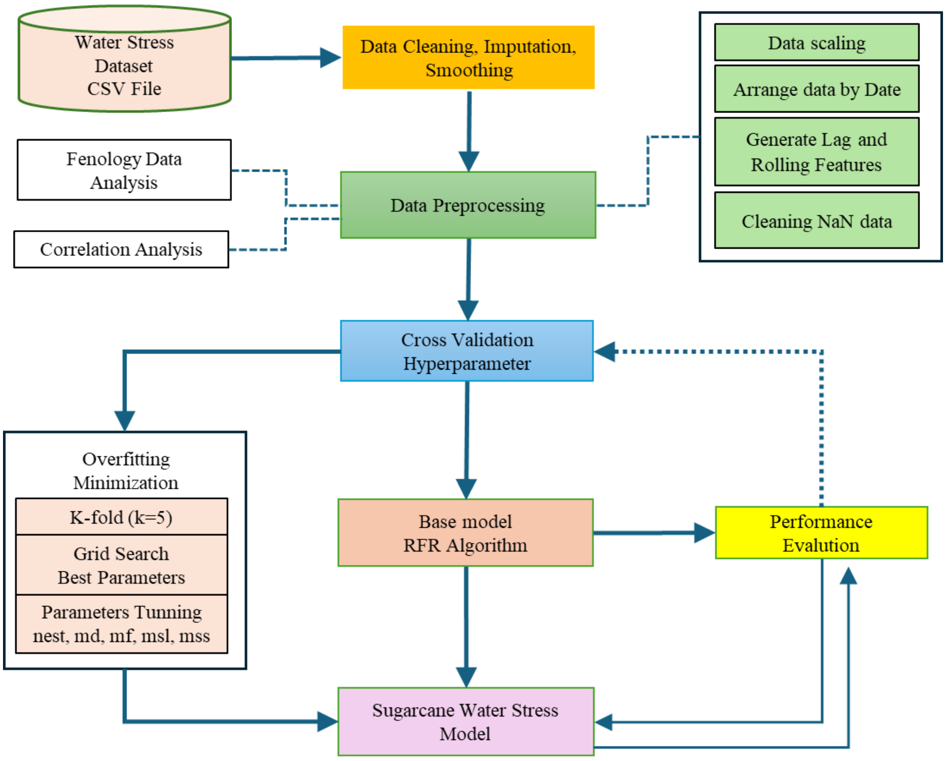


Fig. 5. Workflow of the proposed sugarcane water stress model

In the context of the Random Forest Regressor, *Pt* is the partition forecast that tree *t* introduced, and *T* is the number of trees in the forest. By maximizing the posterior, the model forecasts the duration and position of the relevant mistake for every new case, with $\hat{Y}$ is the input *x*'s eventual forecast. Bootstrap samples are acquired from the original training data in order to create random subsets of the data required to train each decision tree in the Random Forest Regressor process. To make each decision tree more unique, Random Feature Selection only takes into account a random subset of features at each split [32], [33].

6) *Evaluation*: The performance testing of the CWSI prediction model with dataset features is determined based on Accuracy (17), mean error (18) and MAPE (19), the ratio of test data: training data used is 20%: 80% of the original dataset. To avoid overfitting, the Grid Search data grouping technique is used with 5-fold cross validation. The accuracy value is obtained by the following calculation:

$$Accuracy = 100 - MAPE \qquad (17)$$

$$Error = |Prediction_i - data\ test_i| \qquad (18)$$

$$MAPE = 100 \times \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Error}{Data\ Test_i} \right) \qquad (19)$$

$$Prediction = model\ prediction\ (feature\ test) \qquad (20)$$

III. RESULT AND DISCUSSION

A. *Phenological Analysis*

After going through a series of pre-processing stages starting from collecting clean satellite imagery, extracting vegetation index feature values (NDVI, NDWI, NDDI, OSAVI, LSWI, LST and CWSI), to compiling vegetation index feature values on a monthly average according to the planting schedule (ratoon1 and ratoon2) in Figure 6. Furthermore, the data distribution pattern formed in each vegetation index throughout the planting year (ratoon1 and ratoon2) can be observed. In Figure 6, there are two gray areas representing the germination and tillering phases in the first planting season (ratoon1) of 2021-2022 and the second planting season (ratoon2) of 2022-2023, which are critical phases of sugarcane plants against water stress conditions [34]. The green and red lines are the NDVI and CWSI character labels, respectively. The results of observations in the grayscale area show that the

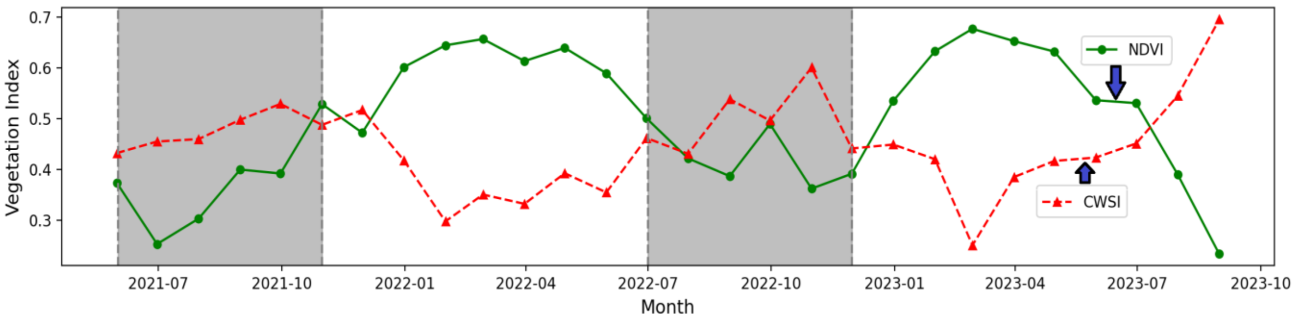


Fig. 6. Phenology of sugarcane based on NDVI and CWSI

CWSI value is higher than the NDVI. The decreasing NDVI value indicates that the water capacity of plants in the leaves and soil surface is in deficit due to the absence of irrigation or rain. This is due to the very low rainfall in the dry season between June - October 2021 and July - November 2022. It is natural that the NDVI index decreases and the CWSI increases, which is a representative response to sugarcane water stress conditions. During ratoon1, the CWSI value < 0.5 is in a mild stress condition, while during ratoon2, the CWSI value can be seen in the Figure 6 increasing to 0.6, which means it is in a moderate stress condition.

B. *Cross Correlation Test of CWSI with Vegetation Features*

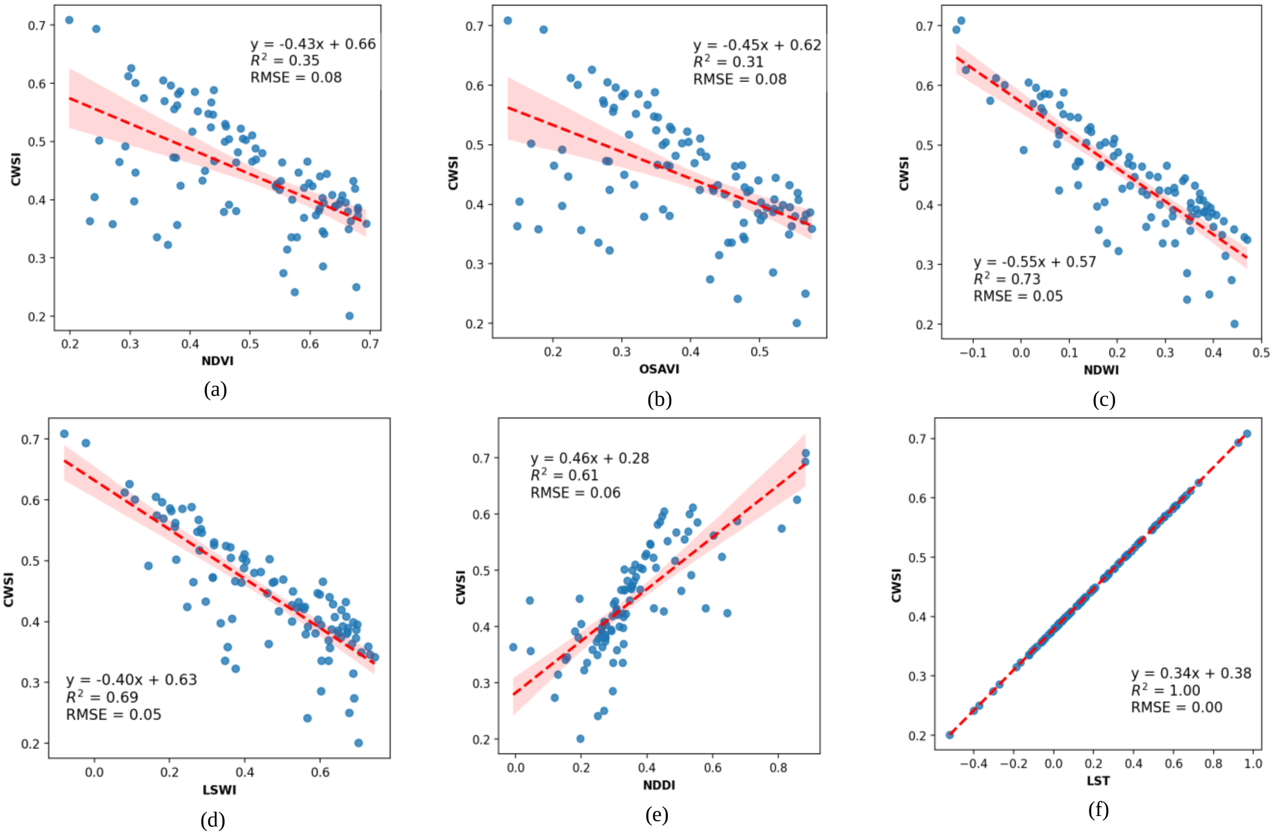
We conducted a cross-correlation test to better understand the relation between CWSI and various vegetation spectrum features: NDVI (Figure 7(a)), OSAVI (Figure 7(b)), NDWI (Figure 7(c)), LSWI (Figure 7(d)), NDDI (Figure 7(e)), and LST (Figure 7(f)). As illustrated in Figure 7, With R2 values of 0.37 and 0.31, respectively, the data show a slight negative connection between NDVI and OSAVI and CWSI. However, with R2 values of 0.77 and 0.73, respectively, NDWI and LSWI show a high negative connection with CWSI. On the other hand, NDDI and CWSI have a moderately positive connection (R2 = 0.60). While LST is strongly positively correlated with CWSI, with an R² value of 1.00.

The error low RMSE for NDVI is 0.07 (7%), OSAVI is 0.08 (8%), NDWI is 0.05 (5%), LSWI is 0.05 (5%), NDDI is 0.06 (6%), and LST is very strongly positively correlated with CWSI, with R2 = 1.00 and RMSE of 0.00, indicating no error. These findings highlight the varying degrees of correlation between various vegetation features and CWSI, with LST showing the strongest relationship.

Furthermore, as shown in Figure 8, a cross-correlation test was conducted between the CWSI and a number of climatology features, such as rainfall (RR), solar radiation level (ss), minimum air temperature (Tn), maximum air temperature (Tx), average air temperature (Tavg), average air humidity (RH_avg), and average wind speed (ff_avg). All climatic features exhibit a modest connection with CWSI, according to the linear regression model, as indicated by the R2 determinant coefficient values for each feature being less than 0.50 or 50%. With an average RMSE error value of 0.09, or 9%, for all climatic features, the model does, however, show a comparatively low error rate.

C. *Time Lag Cross Correlation Test*

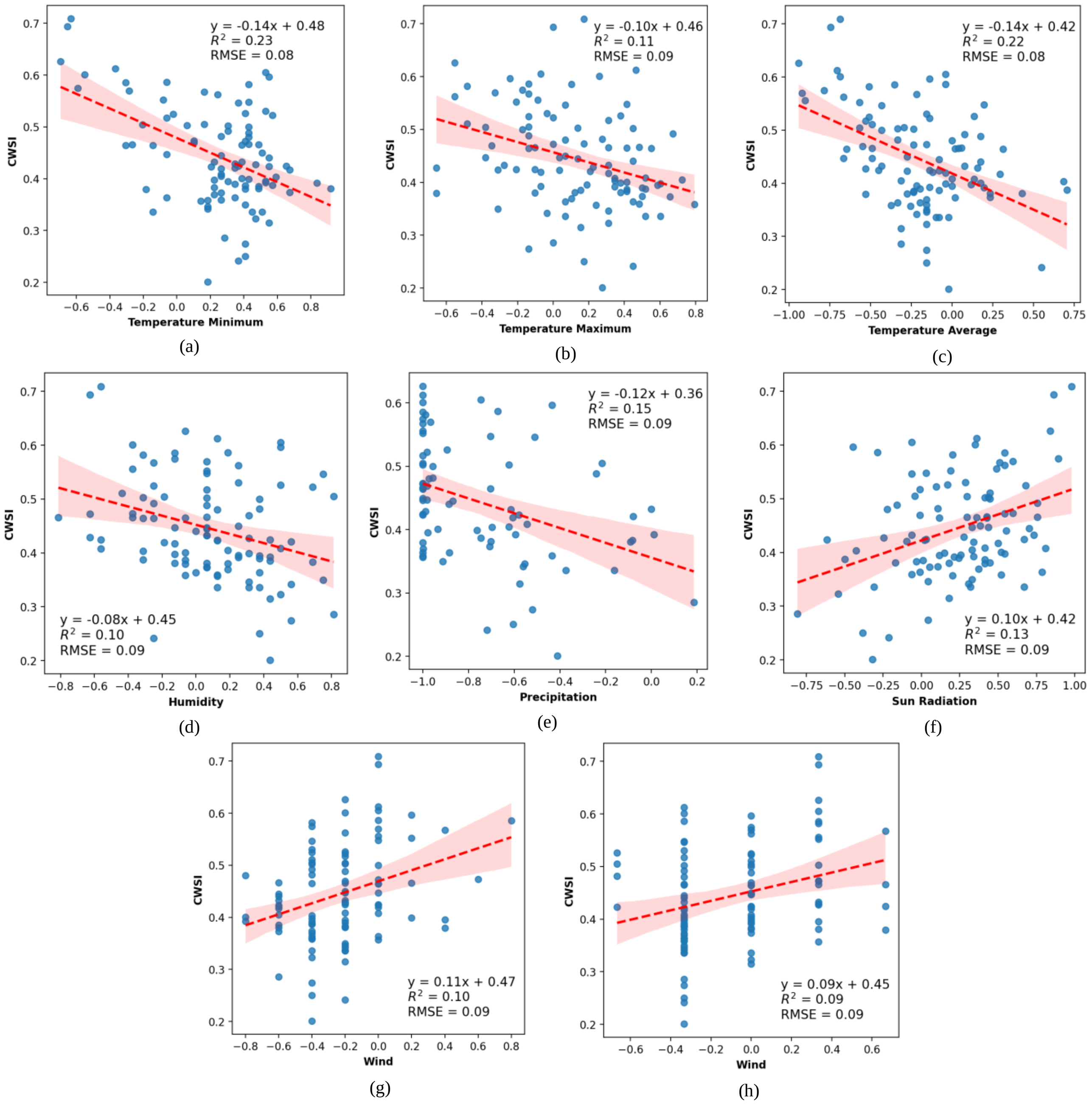
Cross-correlation test with time lag is applied to examine the potential influence of time lag of vegetation spectral features on changes in CWSI, as shown in Figure 9. The peak cross-correlation value between CWSI and all vegetation features occurs at zero time lag, indicating no influence of time lag. Meanwhile, changes in vegetation feature values directly affect the sugarcane CWSI in real time. There is a significant negative association between the NDVI and CWSI, as shown by the correlation coefficient value of -0.61 in Figure 9(a). This suggests that when the NDVI rises, the CWSI falls, and vice versa. Similarly, Figure 9(b) shows a negative correlation for OSAVI-CWSI with a coefficient of -0.57, for NDWI with a coefficient of -0.88, and for LSWI with a coefficient of -0.86. As with NDVI, this indicates that increases in OSAVI, NDWI, and LSWI values lead to decreases in CWSI values, and vice versa.



(a) (b) (c)

(d) (e) (f)

Fig. 7. Regression Correlation between CWSI with features: (a) NDVI, (b) OSAVI, (c) NDWI, (d) LSWI, (e) NDDI, and (f) LST



(a) (b) (c) (d) (e) (f) (g) (h)

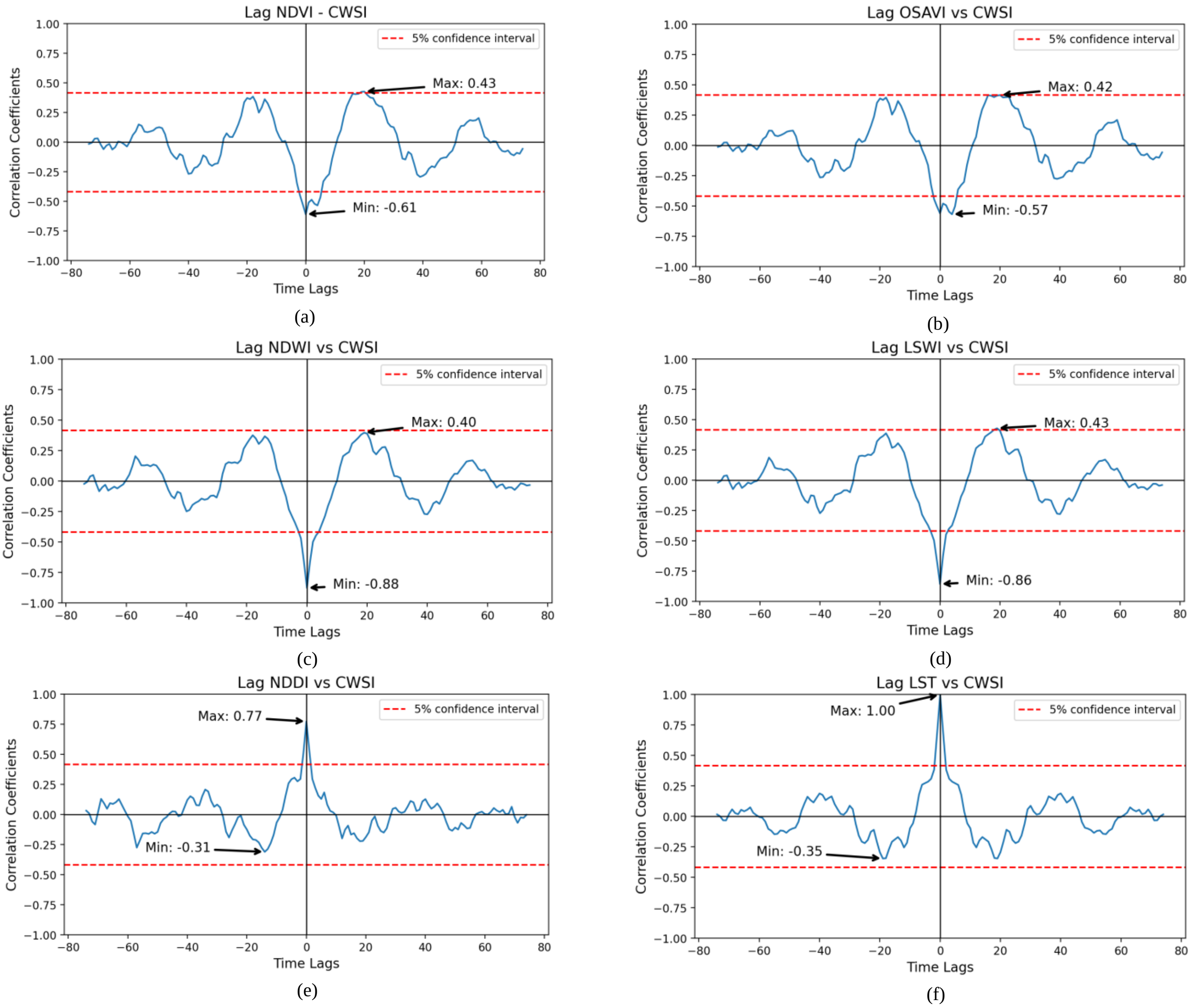
Fig. 8. Regression analysis of the correlation between CWSI and climatological factors, including: (a) minimum air temperature, (b) maximum air temperature, (c) average air temperature, (d) average air humidity, (e) rainfall, (f) solar radiation level, (g) maximum wind speed, and (h) average wind speed.

On the other hand, with coefficients of 0.77 and 1.00, respectively, NDDI and LST show a significantly positive connection with CWSI. Since LST, which stands for canopy temperature, records thermal waves picked up by the Landsat 8 satellite sensor, its substantial impact on CWSI is clear. Thus, changes in the values of vegetation features directly affect the sugarcane water stress index (CWSI) in real time..

Conversely, as depicted in Figure 10, the peak value of the cross-correlation coefficient between the water stress feature and the climatology features reveals that most features exhibit a time lag effect, except for maximum air temperature (b) and average air temperature (c), which have a direct impact on sugarcane water stress. Other features, including minimum air temperature (a), air humidity (d), rainfall (e), sunlight exposure (f), maximum wind speed (g), and average wind speed (h), demonstrate a time lag effect ranging from 1 to 5 days.

Additionally, Figure 10 shows that the correlation coefficients for maximum air temperature and average air temperature are 0.37, while rainfall (RR) has a coefficient of -0.45. In contrast, positive correlations are observed between the water stress feature and solar radiation level (ss) with a coefficient of 0.43, maximum wind speed (ff_x) with a coefficient of 0.36, and average wind speed with a coefficient of 0.41.-0.37; and (e) rainfall (RR) of -0.45. On the other hand, a positive correlation occurs between the water stress feature and the features (f) solar radiation level (ss) of 0.43; (g) maximum wind speed (ff_x) of 0.36 and feature (h) average wind speed of 0.41.

This finding is significant in creating new features using the Lag and Rolling Windows methods. These new features are labeled as follows: lag1 for a 1-day lag, lag2 for a 2-day



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 9. Time-lag cross correlation between water stress features (CWSI) and vegetation features: (a) NDVI, (b) OSAVI, (c) NDWI, (d) LSWI, (e) NDDI, (f) LST

lag, lag3 for a 3-day lag, lag4 for a 4-day lag, and lag5 for a 5-day lag. These labels are applied to each climatological feature with a time lag effect, including average humidity (RH), rainfall (RR), solar radiation level (ss), maximum wind speed (ff_x), and average wind speed (ff_avg), along with one time feature, 'month,' and the rolling_mean feature, which represents the average value of each feature with a lag effect.

We constructed the dataframe based on the cross-correlation and cross-lag correlation of the standard features. However, the analysis results revealed that the LST vegetation index feature has a very strong positive correlation with CWSI and to avoid bias against the model built, we separated the LST feature from the dataset. This was done in accordance with the objectives of this study which sought another easier approach without involving the complicated procedure of finding LST feature values for CWSI.

The dataset consists of 42 features: 5 vegetation features (NDVI, OSAVI, NDWI, LSWI, NDDI), 8 climate features (minimum air temperature, maximum air temperature, mean air temperature, mean humidity, precipitation, solar radiation, maximum wind speed, and mean wind speed), 24 new lag features (lag_cwsi1, lag_cwsi2, lag_cwsi3, lag_cwsi4, lag_rh1, lag_rh2, lag_rh3, lag_rh4, lag_rr1, lag_rr2, lag_rr3, lag_rr4, lag_ss1, lag_ss2, lag_ss3, lag_ss4, lag_ffx1, lag_ffx2, lag_ffx3, lag_ffx4, lag_ffavg1, lag_ffavg2, lag_ffavg3, lag_ffavg4), 1 time feature 'month,' and 5 rolling_mean features representing the average value of each feature with a lag effect.

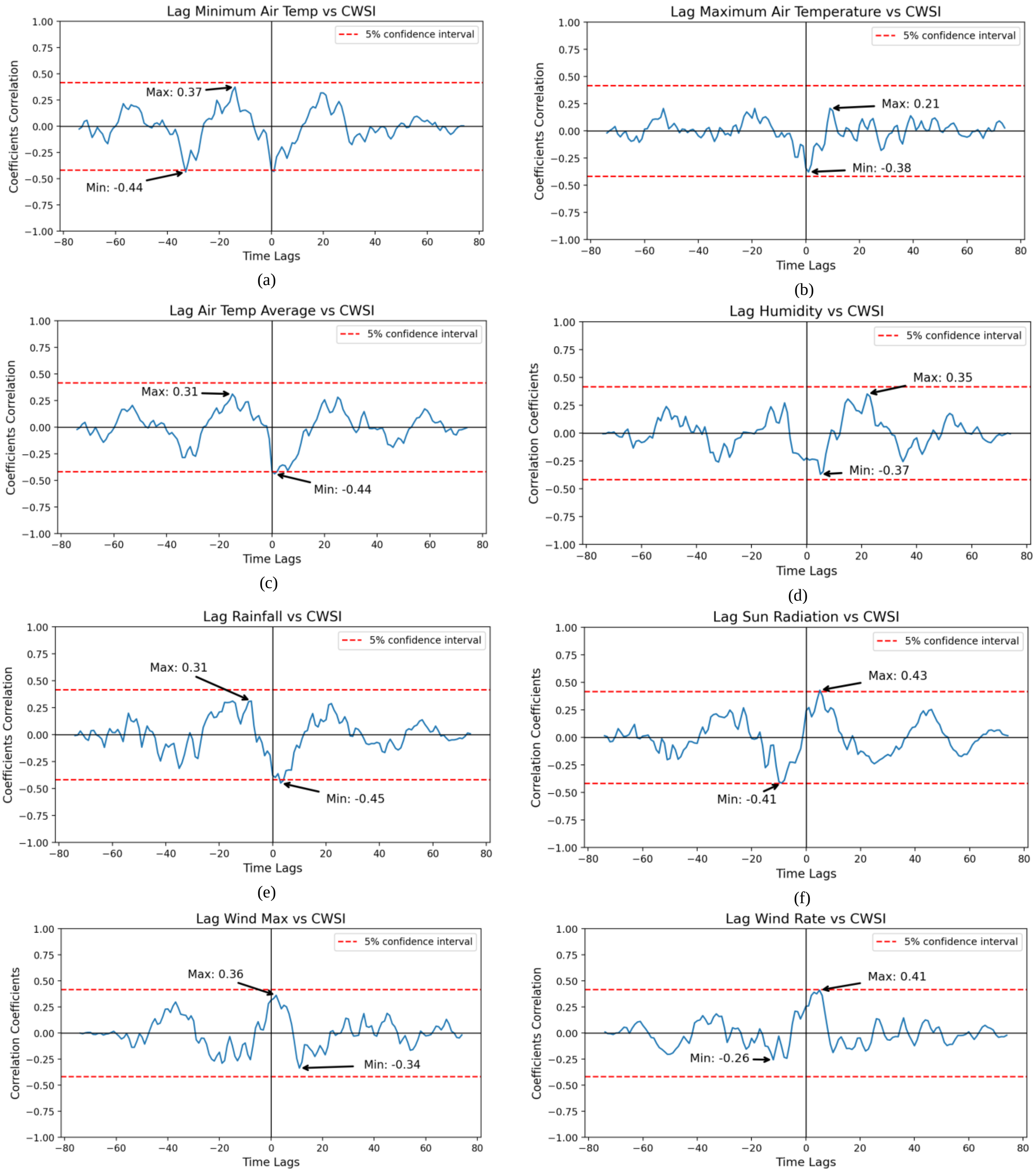
D. Validation and Evaluation Prediction Model

The performance of the prediction shown in Table II.

TABLE II
PERFORMANCE EVALUATION OF PREDICTION MODELS

Dataset Scheme	Accuracy (R2)		MAPE
	RFR	RFR + Hyper	
Vegetation Data	90,96	91,08	8,93
Climate Data	89,10	89,12	10,61
Vegetation + Climate	89,67	89,98	10,04

The Random Forest Regressor (RFR) baseline model scenario was used to test the model, and Grid Search and k-fold cross-validation were used to perform the hyperparameter-adjusted RFR scenario. performance evaluation results of the



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 10. Time-lag cross correlation between water stress features (CWSI) and vegetation features: (a) minimum air temperature, (b) maximum air temperature, (c) average air temperature, (d) average air humidity, (e) rainfall, (f) solar radiation level, (g) maximum wind speed, and (h) average wind speed.

prediction model, as outlined in Table II, indicate that the dataset scheme with vegetation features achieved the highest performance, with an R² accuracy of 91.08% and a MAPE of 8.93%. The combination of vegetation and climatology features, optimized through hyperparameter tuning, ranked second with an R² accuracy of 89.98% and a MAPE of 10.04%. The prediction model based solely on climatology data features obtained an R² accuracy of approximately 89.12% and a MAPE of 10.61%. These three results surpassed those from previous studies: an R² accuracy of 0.74 using the ANOVA technique [35], and an R² accuracy of 0.70 using the Random Forest technique [36].

This predictive model is built using data from 2021 to 2023. The entire time period is divided into training and testing sets. The validation method used is k-fold validation

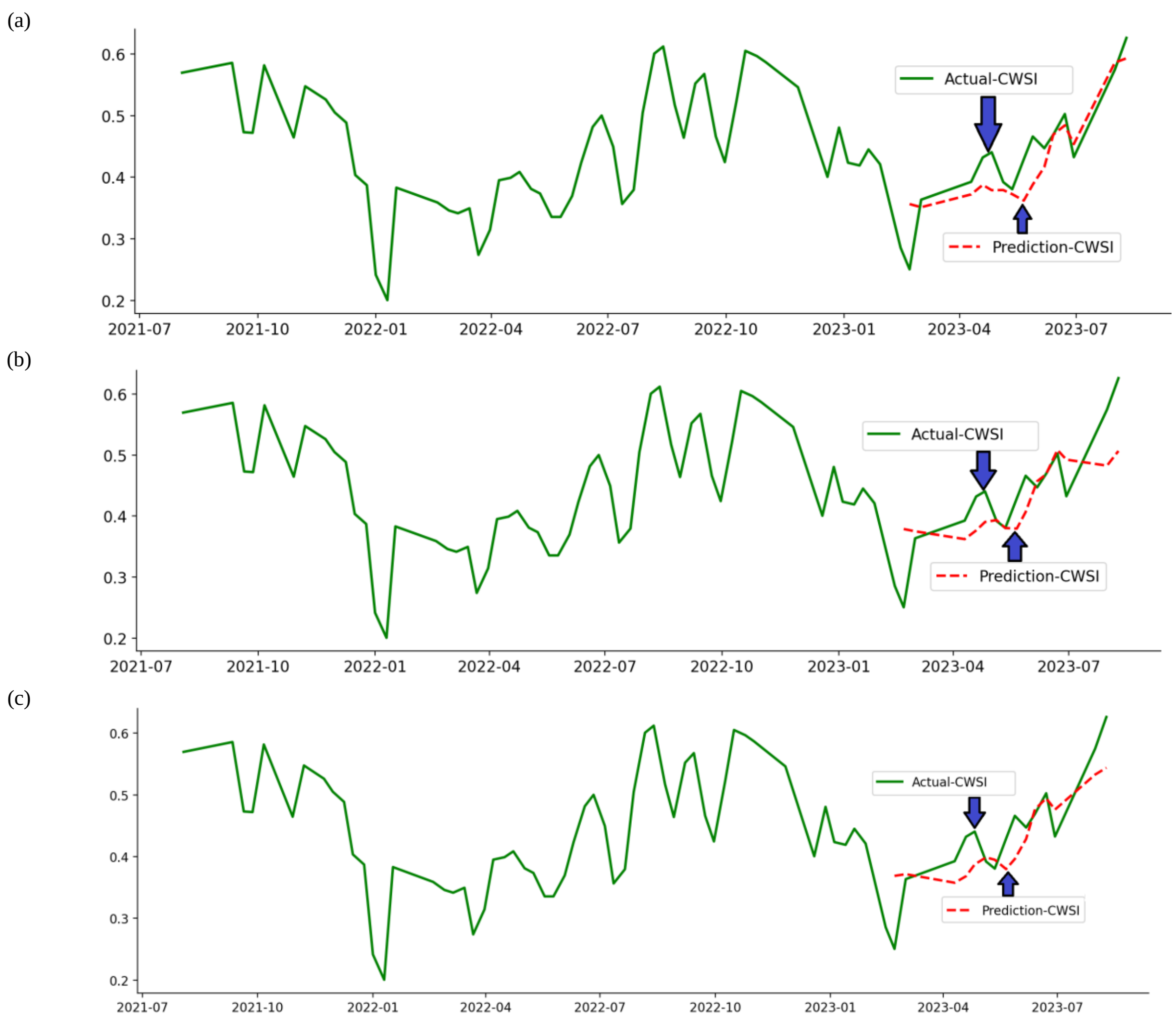
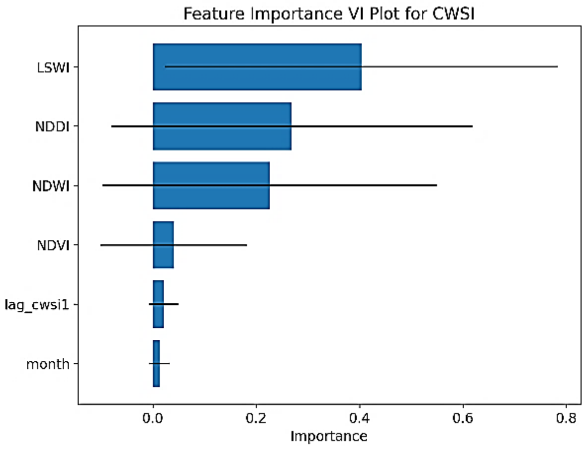


Fig. 11. CWSI prediction results with three different dataset schemes: (a) vegetation dataset, (b) climatology dataset, (c) mixed vegetation and climatology dataset without LST

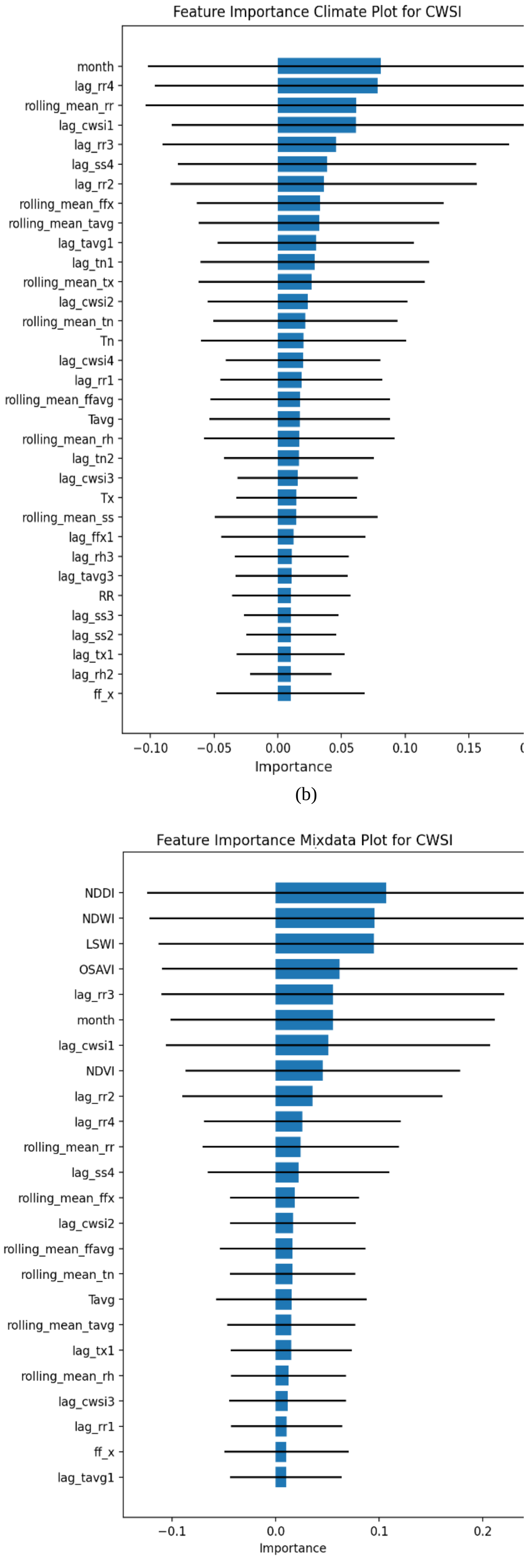
with a five-fold time series cross-validation window with a forecast interval varying between 1 and 5 months. As the cross-validation process progresses, the training data expands to cover all previous data, while the test data size remains constant by running a random search for each fold in the time series. Cross-validation is optimized from the gird search space to ensure unbiased tuning that prevents overfitting. Hyperparameter tuning is set at maximal depth = 25; maximal features = 5; minimal samples leaf = 3; minimal samples split = 3; and number of estimators = 200. The visualization of the model performance is shown in Figure 11, which shows the success of the RFR model in applying the water stress approach (CWSI) for sugarcane fields. According to Figure 11 shows the prediction results, with a time span ranging from 1 to 5 months. All datasets used in the modeling process do not use LST features.

E. Feature Importance

To see what features play a major role in each data set schema, see Figure 12.



(a)



(b)

(c)

Fig. 12. List of important features of CWSI different dataset schemes: (a) vegetation dataset, (b) climatology dataset, (c) mixed vegetation and climatology dataset.

The figure shows data with three different data set schemes: (a) vegetation data set, (b) climatology data set, (c) mixed vegetation and climatology data set. Based on the information in Figure 12(a), it is known that the important features for scheme vegetation feature dataset are LSWI, NDDI, NDWI, NDVI, lag_cwsi1, and month. While in the Figure 12(b) order of important features for scheme 2 climatology dataset is month, lag_rr4, rolling_mean_rr, lag_cwsi1, lag_rr3, lag_ss4, lag_rr2, rolling_mean_ffx, rolling_mean_tavg, lag_tavg1, lag_tn1, rolling_mean_tx, lag_cwsi2, rolling_mean_tn, Tn and so on. Then for scheme 3 mixed dataset between vegetation and climatology features, it is known that the order of important features is NDDI, NDWI, LSWI, OSAVI, lag_rr3, month, lag_cwsi1, NDVI, lag_rr2, lag_rr4, rolling_mean_rr, lag_cwsi2, rolling_mean_ffavg, rolling_mean_tn, Tavg, rolling_mean_tavg, lag_tx1, rolling_mean_rh and so on.

The three schemes demonstrate that the combination of random data from vegetation features, climatology data, and historical data effectively predicts future CWSI values. Vegetation features other than LST, such as NDWI, LSWI, NDDI, and NDVI, significantly influence changes in CWSI index values. Additionally, climatology features with time lag effects, like the rainfall time lag feature (lag_rr2) and the solar radiation time lag feature (lag_ss4), also play a crucial role in this predictive collaboration.

IV. CONCLUSION

This study illustrates the intricate process of determining the crop water stress index (CWSI), beginning with the extraction of vegetation features from Landsat 8 satellite data and converting them into CWSI values through the reduction of NDVI and LST. The most challenging part of this process is the complex and time-consuming calculation of LST, which is a crucial feature that significantly impacts CWSI. In this research, we propose an alternative method to determine the CWSI value using a multi-correlation prediction model that utilizes vegetative and climatological features without incorporating the LST feature. Before inputting the data into the model, we conducted cross-correlation and time lag tests to assess the strength of relationships, seasonal patterns, and potential time lag effects among the features. The results indicated that vegetation features have strong correlations, whereas climatological features have weaker correlations. The time lag test showed that air temperature has no time lag effect, while other climatological features do. Based on these findings, we designed three dataset schemes: dataset1 with vegetation only, dataset2 with climatology only, and dataset3 with a combination of vegetation and climatology, all excluding LST features. The best model performance was achieved with the vegetation index dataset scheme, which yielded an R^2 accuracy of 91.08% and a MAPE of 8.93% with hyperparameter adjustment. This suggests that it is feasible to estimate CWSI values using a multi-feature random data approach without involving LST. Additionally, we found that excluding the canopy temperature (LST) feature from the determination of the sugarcane water stress value allows other features such as LSWI, NDDI, NDWI for vegetation, and rainfall and sunshine duration for climatology (with a lag effect of 2 days for rainfall and 4 days for sunshine) to become significant influencers of CWSI values.

V. FUTURE WORK

In future research, we aim to implement this model into a specialized application capable of dynamically processing data to provide recommendations for determining the planting schedule for rainfed sugarcane fields in response to water stress conditions. This research serves as an initial step in developing a smart irrigation system to safeguard plants from water stress.

ACKNOWLEDGMENT

We extend our heartfelt thanks to the Indonesian Sugar Research Center of PT. Perkebunan Nusantara-X and HGU for their invaluable support and services in granting us access to research data on the sugarcane fields they manage. Our appreciation also goes to the United States Geological Survey (USGS) and Google Earth Engine (GEE) for providing free satellite data access and cloud-based data processing platforms. The author is profoundly grateful to IPB University and Singaperbangsa University Karawang for providing the opportunity to pursue doctoral education and for funding the writing of this article.

REFERENCES

[1] H. Singh and S. S. Bola, “Climate change , its impact and mitigation strategies for sugarcane production : A review,” vol. 9, no. 3, pp. 1288–1294, 2020.

[2] N. Farooq and S. H. Gheewala, “Assessing the impact of climate change on sugarcane and adaptation actions in Pakistan,” *Acta Geophysica*, vol. 68, no. 5, pp. 1489–1503, 2020, doi: 10.1007/s11600-020-00463-8.

[3] S. Flack-Prain *et al.*, “The impact of climate change and climate extremes on sugarcane production,” *GCB Bioenergy*, vol. 13, no. 3, pp. 408–424, 2021, doi: 10.1111/gcbb.12797.

[4] V. Misra *et al.*, “Morphological assessment of water stressed sugarcane: A comparison of waterlogged and drought affected crop,” *Saudi J Biol Sci*, vol. 27, no. 5, pp. 1228–1236, 2020, doi: 10.1016/j.sjbs.2020.02.007.

[5] A. Singels, A. L. Paraskevopoulos, and M. L. Mashabela, “Farm level decision support for sugarcane irrigation management during drought,” *Agric Water Manag*, vol. 222, no. January, pp. 274–285, 2019, doi: 10.1016/j.agwat.2019.05.048.

[6] Y. D. Giroh and A. A. Girei, “Analysis of the Factors affecting Sugarcane (Saccharum officinarum) Production under the Out growers Scheme in Numan Local Government Area Adamawa State, Nigeria,” *Journal of Education and Practice*, vol. 3, no. 8, pp. 195–201, 2012, [Online]. Available: www.iiste.org

[7] S. Sudianto, Y. Herdiyeni, and L. B. Prasetyo, “Early Warning for Sugarcane Growth using Phenology-Based Remote Sensing by Region,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 2, pp. 502–510, 2023, doi: 10.14569/ijacsa.2023.0140259.

[8] Y. Uttaruk and T. Laosuwan, “Drought detection by application of remote sensing technology and vegetation phenology,” *Journal of Ecological Engineering*, vol. 18, no. 6, pp. 115–121, 2017, doi: 10.12911/22998993/76326.

[9] K. C. DeJonge, S. Taghvaeian, T. J. Trout, and L. H. Comas, “Comparison of canopy temperature-based water stress indices for maize,” *Agric Water Manag*, vol. 156, pp. 51–62, 2015, doi: 10.1016/j.agwat.2015.03.023.

[10] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering, “Monitoring Vegetation Systems In The Great Plains With Erts,” *J Agric Food Chem*, vol. 24, no. 1, pp. 24–26, 1974, doi: 10.1021/jf60203a024.

[11] K. L. Hugie, P. J. Bauer, K. C. Stone, E. M. Barnes, D. C. Jones, and B. T. Campbell, “Improving the Precision of NDVI Estimates in Upland Cotton Field Trials,” *The Plant Phenome Journal*, vol. 1, no. 1, pp. 1–9, 2018, doi: 10.2135/tppj2017.09.0009.

[12] P. P. Patil, M. P. Jagtap, N. Khatri, H. Madan, A. A. Vadduri, and T. Patodia, “Exploration and advancement of NDDI leveraging NDVI and NDWI in Indian semi-arid regions: A remote sensing-based study,” *Case Studies in Chemical and Environmental Engineering*, vol. 9, no. December 2023, p. 100573, 2024, doi: 10.1016/j.cscee.2023.100573.

[13] B.-C. Gao, “NDWI A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water From Space,” *Remote Sens. Environ*, vol. 7212, no. April, pp. 257–266, 1996.

[14] E. Tavazohi and M. A. Nadoushan, “Assessment of drought in the Zayandehroud basin during 2000-2015 using NDDI and SPI indices Assessment Of Drought In The Zayandehroud Basin During 2000 – 2015 Using Nddi And Spi Indices,” no. April, 2018.

[15] P. P. Patil, M. P. Jagtap, N. Khatri, H. Madan, A. A. Vadduri, and T. Patodia, “Exploration and advancement of NDDI leveraging NDVI and NDWI in Indian semi-arid regions: A remote sensing-based study,” *Case Studies in Chemical and Environmental Engineering*, vol. 9, no. December 2023, p. 100573, 2024, doi: 10.1016/j.cscee.2023.100573.

[16] K. Chandrasekar, M. V. R. Sesha Sai, P. S. Roy, and R. S. Dwevedi, “Land Surface Water Index (LSWI) response to rainfall and NDVI using the MODIS vegetation index product,” *Int J Remote Sens*, vol. 31, no. 15, pp. 3987–4005, 2010, doi: 10.1080/01431160802575653.

[17] R. R. Fern, E. A. Foxley, A. Bruno, and M. L. Morrison, “Suitability of NDVI and OSAVI as estimators of green biomass and coverage in a semi-arid rangeland,” *Ecol Indic*, vol. 94, no. May, pp. 16–21, 2018, doi: 10.1016/j.ecolind.2018.06.029.

[18] S. Guha, H. Govil, and P. Diwan, “Monitoring LST-NDVI Relationship Using Premonsoon Landsat Datasets,” *Advances in Meteorology*, vol. 2020, no. 1, 2020, doi: 10.1155/2020/4539684.

[19] J. Khan, P. Wang, Y. Xie, L. Wang, and L. Li, “Mapping MODIS LST NDVI Imagery for Drought Monitoring in Punjab Pakistan,” *IEEE Access*, vol. 6, no. c, pp. 19898–19911, 2018, doi: 10.1109/ACCESS.2018.2821717.

[20] M. Arslan, R. Zahid, and B. Ghauri, “Assessing the occurrence of drought based on NDVI, LST and rainfall pattern during 2010-2014,” *International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 2016-Novem, pp. 4233–4236, 2016, doi: 10.1109/IGARSS.2016.7730103.

[21] F. N. Kogan, “Application of vegetation index and brightness temperature for drought detection,” *Advances in Space Research*, vol. 15, no. 11, pp. 91–100, 1995, doi: 10.1016/0273-1177(95)00079-T.

[22] S. Veysi, A. A. Naseri, S. Hamzeh, and H. Bartholomeus, “A satellite based crop water stress index for irrigation scheduling in sugarcane fields,” *Agric Water Manag*, vol. 189, pp. 70–86, 2017, doi: 10.1016/j.agwat.2017.04.016.

[23] C. Labuzzetta, Z. Zhu, X. Chang, and Y. Zhou, “A submonthly surface water classification framework via gap-fill imputation and random forest classifiers of landsat imagery,” *Remote Sens (Basel)*, vol. 13, no. 9, 2021, doi: 10.3390/rs13091742.

[24] S. Belda *et al.*, “DATimeS: A machine learning time series GUI toolbox for gap-filling and vegetation phenology trends detection,” *Environmental Modelling and Software*, vol. 127, May 2020, doi: 10.1016/j.envsoft.2020.104666.

[25] J. Liang *et al.*, “Using Enhanced Gap-Filling and Whittaker Smoothing to Reconstruct High Spatiotemporal Resolution NDVI Time Series Based on Landsat 8, Sentinel-2, and MODIS Imagery,” *ISPRS Int J Geoinf*, vol. 12, no. 6, 2023, doi: 10.3390/ijgi12060214.

[26] Z. Cai, P. Jönsson, H. Jin, and L. Eklundh, “Performance of Smoothing Methods for Reconstructing NDVI Time-Series and Estimating Vegetation Phenology from MODIS Data,” *Remote Sens (Basel)*, vol. 9, no. 12, 2017, doi: 10.3390/rs9121271.

[27] D. P. Roy and L. Yan, “Robust Landsat-based crop time series modelling,” *Remote Sens Environ*, vol. 238, no. June, pp. 0–1, 2020, doi: 10.1016/j.rse.2018.06.038.

[28] M. Jung and E. Chang, “NDVI-based land-cover change detection using harmonic analysis,” *Int J Remote Sens*, vol. 36, no. 4, pp. 1097–1113, 2015, doi: 10.1080/01431161.2015.1007252.

[29] S. K. Padhee and S. Dutta, “Spatio-Temporal Reconstruction of MODIS NDVI by Regional Land Surface Phenology and Harmonic Analysis of Time-Series,” *GIsci Remote Sens*, vol. 56, no. 8, pp. 1261–1288, 2019, doi: 10.1080/15481603.2019.1646977.

[30] M. Becker, L. Schneider, and S. Fischer, *Hyperparameter Optimization*. 2024. doi: 10.1201/9781003402848-4.

[31] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, “Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis,” *Informatics*, vol. 8, no. 4, pp. 1–21, 2021, doi: 10.3390/informatics8040079.

[32] B. Goehry, “Random forests for time-dependent processes,” *ESAIM - Probability and Statistics*, vol. 24, pp. 801–826, 2020, doi: 10.1051/ps/2020015.

[33] Z. El Mrabet, N. Sugunaraj, P. Ranganathan, and S. Abhyankar, “Random Forest Regressor-Based Approach for Detecting Fault Location and Duration in Power Systems,” *Sensors*, vol. 22, no. 2, pp. 1–19, 2022, doi: 10.3390/s22020458.

[34] A. Singels, A. L. Paraskevopoulos, and M. L. Mashabela, “Farm level decision support for sugarcane irrigation management during drought,” *Agric Water Manag*, vol. 222, no. January, pp. 274–285, 2019, doi: 10.1016/j.agwat.2019.05.048.

[35] A. Santillán-Fernández, V. H. Santoyo-Cortés, L. R. García-Chávez, I. Covarrubias-Gutiérrez, and A. Merino, “Influence of drought and irrigation on sugarcane yields in different agroecoregions in Mexico,” *Agric Syst*, vol. 143, no. June, pp. 126–135, 2016, doi: 10.1016/j.agsy.2015.12.013.

[36] T. F. Canata, M. C. F. Wei, L. F. Maldaner, and J. P. Molin, “Sugarcane Yield Mapping Using High-Resolution Imagery Data and Machine Learning Technique,” *Remote Sens (Basel)*, vol. 13, no. 2, p. 232, 2021, doi: 10.3390/rs13020232.