Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 2 (2025) 225 - 238 e-ISSN: 2580-0760

Comparison of Sugarcane Drought Stress Based on Climatology Data Using Machine Learning Regression Model in East Java

Aries Suharso[1*], Yeni Herdiyeni[2], Suria Darma Tarigan[3], Yandra Arkeman[4]
[1,2]Computer Science, School of Data Science, Mathematics and Informatics, IPB University, Bogor, Indonesia
[1]Informatics, Faculty of Computer Science, Singaperbangsa Karawang University, Karawang, Indonesia
[3]Soil Science and Land Resources, Faculty of Agriculture, IPB University, Bogor, Indonesia
[4]Agricultural Industrial Technology, Faculty of Agriculture, IPB University, Bogor, Indonesia
[1]aries.suharso@unsika.ac.id , [2]yeni.herdiyeni@apps.ipb.ac.id , [3]sdtarigan@apps.ipb.ac.id , [4]yandra@apps.ipb.ac.id

Abstract

Crop Water Stress Index (CWSI), derived from vegetation features (NDVI) and canopy thermal temperature (LST), is an effective method to evaluate sugarcane sensitivity to drought using satellite data. However, obtaining CWSI values is complicated. This study introduces a novel approach to estimate CWSI using climatological data, including average air temperature, humidity, rainfall, sunshine duration, and wind speed features obtained from the local weather station BMKG Malang City, East Java, for the period 2021-2023. Before estimating CWSI, we analyzed sugarcane water stress phenology, examined the strength of the correlation between climatological features and CWSI, and looked at the potential for adding lag features. Our proposed prediction model uses climatological features with additional Lag features in a machine learning regression approach and 5-fold cross-validation of the training-testing data split with the help of optimization using hyperparameters. Different machine learning regression models are implemented and compared. The evaluation results showed that the prediction performance of the SVR model achieved the best accuracy with R2 = 90.45% and MAPE = 9.55%, which outperformed other models. These findings indicate that climatological features with lag effects can effectively predict water stress conditions in rainfed sugarcane if using an appropriate prediction model. The main contribution of this study is the utilization of local climatological data, which is easier to obtain and collect than sophisticated satellite data, to estimate CWSI. The application of the results shows that climatological data with lag effects can accurately estimate water stress conditions in rainfed sugarcane. In drought-prone areas, this strategy can help sugarcane farmers make better choices about land management and irrigation.

Keywords: crop water stress index; climatological data; machine learning regression; sugarcane

How to Cite: Aries Suharso, Yeni Herdiyeni, Suria Darma Tarigan, and Yandra Arkeman, "Comparison of Sugarcane Drought Stress Based on Climatology Data using Machine Learning Regression Model in East Java", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 2, pp. 225 - 238, Mar. 2025.
DOI: https://doi.org/10.29207/resti.v9i2.6159

1. Introduction

A primary factor contributing to the decline in sugarcane productivity in Indonesia is the conversion of wet or paddy fields to dry land, driven by competition with rice as the predominant food crop [1], [2]. This transition necessitates a shift in the sugarcane's irrigation regime, compelling it to rely exclusively on rain-fed irrigation methods. This stands in contrast to the technical irrigation systems employed in wetland regions. The annual sugarcane planting cycle is synchronized with the harvest season, during which there is minimal rainfall, to ensure a high sucrose yield quality. Consequently, the commencement of the sugarcane planting season follows this pattern. Consequently, sugarcane frequently experiences water stress, particularly during the germination and tillering stages of early growth. Water stress during this critical phase has been shown to inhibit growth and significantly impact biomass weight at harvest [3]-[5].
Sugarcane (Saccharum officinarum) is an anisohydric plant, meaning it is able to absorb and store large amounts of water from the soil, even when its own water content is low. During the dry season, sugarcane stomata tend to remain open for extended periods, facilitating photosynthesis and maintaining biomass production. As a result, sugarcane can thrive in

Received: 31-10-2025 | Accepted: 05-03-2025 | Published Online: 18-03-2025

environments with limited water availability; however, the plant becomes more susceptible to physiological damage when subjected to prolonged drought. In response to these challenges, several initiatives have been undertaken, including the integration of water-efficient drip irrigation systems and the development of drought-tolerant sugarcane varieties. [6], [7]. Nevertheless, the efficacy of these solutions is contingent upon the consideration of the initial sugarcane planting schedule in the context of the shifting patterns of the rainy and dry seasons, as influenced by climate change [8]-[10]. The report was received from the State Plantation Company (PTPN-X), the largest sugarcane plantation operator in East Java, Indonesia in Figure 1. The report shows a decline in productivity from the first sugarcane planting season (ratoon1) 2021-2022 to the second sugarcane planting season (ratoon2) 2022-2023. The decline in productivity per hectare showed variability across planting areas (G30-G36), with the lowest recorded decline of 20% and the highest of 37% in G33. Given the findings of this report, we hypothesize that the significant decline in productivity in G33 is due to water stress conditions. Therefore, this study will prioritize the evaluation of the water stress index for rainfed sugarcane land in G33.

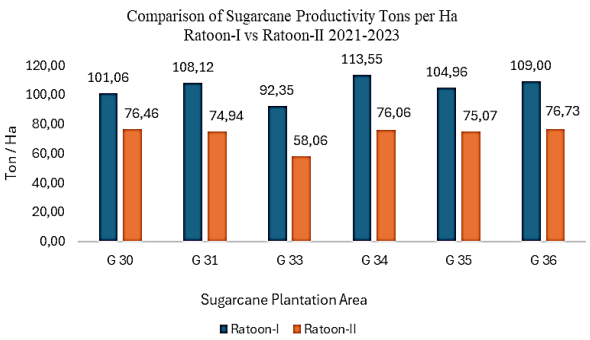


Figure 1. Sugarcane productivity 2021 - 2023

The assessment of plant responses to water stress serves as an early indicator of drought, a task of significant value. A comprehensive review of contemporary plant water stress monitoring methodologies reveals four predominant approaches: first, gravimetry and soil moisture sensors, which are based on measuring soil water content; second, the soil water balance approach; third, measurement based on plant parts (e.g., stomatal conductance, leaf water potential, sap flow, stem and fruit diameter); and fourth, measurement based on remote sensing (e.g., infrared thermometry and vegetation spectral index) [11], [12].

The Crop Water Stress Index (CWSI) is one of the infrared thermometry methods in remote sensing that calculates stomatal closure to plant water deficit based on the difference between plant canopy temperature and air temperature to measure plant water stress. However, this CWSI calculation has a weakness where the vegetation spectral index derived from satellite image extraction is affected by cloud cover noise, which means that the availability of clean data is certainly small, then requires a different reference baseline for wet (Twet) and dry (Tdry) plant surface temperatures for different crops [13], and other studies state that determining LST requires water vapor data that has a very large resolution so it tends to be inaccurate [14]. Recent empirical CWSI sensitivity research has been conducted on a combination of climatological data input features, namely air temperature (Ta), and relative humidity (RH) with field data of canopy temperature (Tc). However, the research results still show a fairly large error in CWSI predictions, namely 52%. This shows that data quality is very important for research related to CWSI for irrigation scheduling, especially in humid climate conditions.

In this study, we use time series data from vegetation spectral feature extraction on Landsat 8 imagery as the basis of truth, especially vegetation properties, from Landsat imagery [15] that can reveal plant reactions to water stress. We propose a new approach to predict sugarcane CWSI using climatological time series data, including air temperature, humidity, rainfall, sunshine duration, and wind speed with Machine learning regression model approach, using a selection of algorithms Ada Boost Regressor (ABR), Decision Tree Regressor (DTR), k-Nearest Neighbors Regressor (KNNR), Light GBM Regressor (LGBMR), Random Forest Regressor (RFR), Support Vector Regressor (SVR), and XGBoost Regressor. Model performance is evaluated using the coefficient of determination (R2) and mean absolute percentage error (MAPE).

2. Research Methods

This study to estimate CWSI using climatological data includes limitations related to the study location, data, and specific machine learning regression approaches to be used. In addition, external factors such as soil quality, agricultural practices, and plant genetics are not discussed in this study. Likewise, the irrigation regime, according to field information, adheres to a rainfed system, which relies on natural rainfall as the main source of water for sugarcane growth without using an artificial irrigation system.

2.1 Research location

The observed research location is an agro-industrial sugarcane plantation (G33), as shown in Figure 2(a), which is marked in red on the sugarcane plantation plot in Figure 2(b), managed by the State Plantation PTPN Plosoklaten, Kediri Regency, East Java, Republic of Indonesia.

The coordinates of the observation location are at Latitude 112.16467404367307 and Longitude -7.904521068941556. The area of the observed plantation plot is 11,234 Ha of the total sugarcane plantation area of 4,900 Ha2.

The soil category is flat regosol 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.

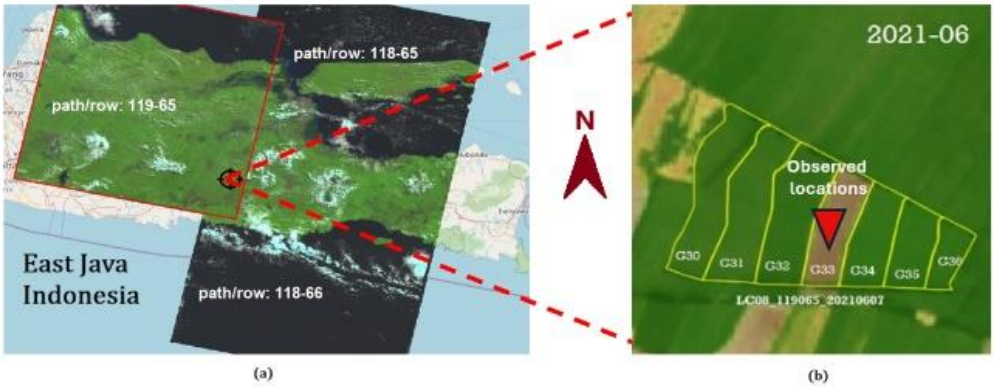


Figure 2. The observed research locations are (a) Map of East Java with three Landsat satellite image capture lines, (b) Plan of sugarcane plantation area owned by the National Plantation in Kediri Regency, East Java Province, Indonesia.

2.2 Research Data

The report of the results of observations by officers in the field showed that there was stunted growth in the sugarcane stem segments, which indicated water stress at the beginning of growth. Other data provided were in the form of a vector map (shp) of sugarcane land at PTPN-X and the Sugarcane Planting Schedule (PTPN-X) for the 2021-2023 Period, as shown in Table 1.

Table 1. Sugarcane Planting Schedule for G33 2021-2023

Ratoon Period	Ger (0 - 45)	Till (45 - 120)	GG (120 - 250)	MR (250 - 365)	H (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

The description Ger is the germination phase, Till is the tillering phase, GG is the sugarcane's peak growth period, MR is the mature and ripening phase, and H is the harvesting phase. With schedule notation, A represents the first two weeks of each month, and B represents the latter two weeks of each month.

Landsat 8 satellite imagery for the 2021-2023 period is sourced from the USGS.gov.id site with properties in Table 2, then extracted on the Google Earth Engine platform, which obtains indications of vegetation features. Meanwhile, data processing, analysis, and predictive modeling are carried out on the Google Colab platform with Python 3.10.

Table 2. Landsat Data Properties

Attribute	Detail
Landsat 8 Landsat 9	LANDSAT/LC08/C02/T1_TOA LANDSAT/LC09/C02/T1_TOA
Period	2021 to 2023
Region of Interest (ROI)	PTPN X Kediri East Jawa Indonesia Latitude: 112.16467404367307, Longitude: -7.904521068941556
Scale	30m to 15m Cloud cover free 30%
Path/Row	LC08-(118-65, 118-66, 119-65)
Bands	Bands (2,3,4,5,6,8,10)

Climatology data is used to predict water stress features as CWSI targets. Climatology data, including air temperature, air humidity, rainfall, sunshine duration, and wind speed features, were obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG) of Malang City, East Java, Indonesia, for the 2019-2023 period.

2.3 Research Method

We apply the concept that plant water stress is the impact of the influence of abiotic environmental variables [11]. We partially observe and test changes in climate features as part of abiotic environmental variables on their contribution to sugarcane water stress specifically. Our research limitations are determined according to the observed sugarcane planting cycle [8], [9]. Climatology data from BMKG is used according to the observation period within a 2-year span, from 2021 to 2023. Assuming 1 year is 365 days, the number of daily data for 2 years is around 730. However, this number is then reduced again according to the available satellite imagery data, which is only 102 satellite imagery data that are clean from maximum cloud cover of 30%. It should be noted that Landsat 8 satellite data is obtained with a frequency of once every 16 days in the same region of interest (roi), even if the conditions are without cloud cover [14], [16].

Therefore, the temporal granularity of our climatology data is adjusted to the monthly frequency of vegetation

index data. This was chosen because it is more appropriate for long-term trends that affect the model's ability to capture temporal patterns, such as seasonal trends or short-term fluctuations. The potential impact of cloud cover filtering (30%) on data quality can result in the loss of important data on certain days, especially during the rainy season or in areas with high cloud cover. Therefore, time identity is very important in this spatial data so as not to cause bias towards other periods so that the analysis carried out may not represent the actual conditions in the field.

The research stages are explained in Figure 3. Starting from the first step, collecting data and information from various sources regarding the research object. The second step is data preparation, which consists of data cleaning and feature engineering. The third stage is sugarcane phenology analysis. The fourth stage is comparing the intensity of climatological variables with the sugarcane water stress index through cross-correlation. The fifth step is to create a prediction model based on machine learning regression. The last stage is to determine the optimal prediction model.

The sugarcane water stress prediction model proposed in Figure 4 is based on machine learning regression. The selected models include Ada Boost Regressor (ABR), Decision Tree Regressor (DTR), k-Nearest Neighbors Regressor (KNNR), Light GBM Regressor (LGBMR), Random Forest Regressor (RFR), Support Vector Regressor (SVR), and XGBoost Regressor.

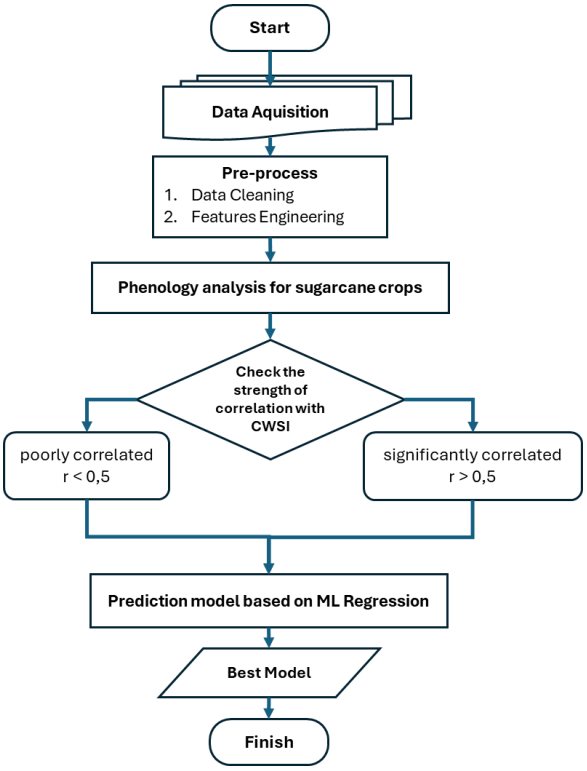


Figure 3. Proposed workflow method

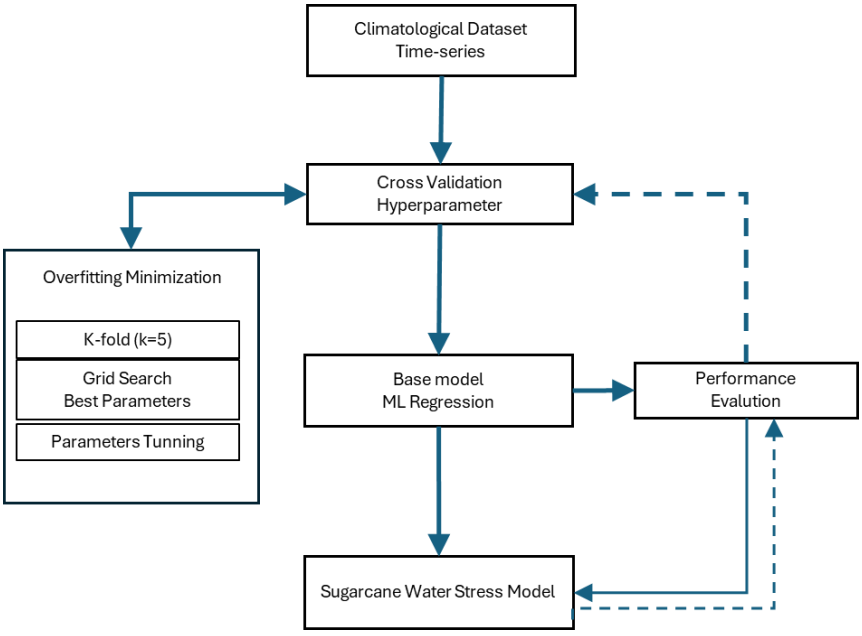


Figure 4. The proposed prediction model

2.4 Pre-Process

The preprocessing steps performed include the following aspects of data cleaning and feature engineering.

Data Cleaning includes Linear Interpolation, Rolling Average, and Row or Column Deletion. Linear Interpolation is done as an attempt to fill in missing values, this step cleans the data because the focus is on fixing or completing missing data to maintain the integrity of the data set. Use of Rolling Average (for missing data imputation). This rolling average is used to fill in missing values without adding new information. Row or Column Deletion for many missing values is a data-cleaning step because it is a

way to handle data that cannot be used in the configuration to be analyzed.

Feature engineering at the preprocessing stage of model development is very important, and it includes normalization or standardization, rolling average, harmonic series, and lag feature engineering. Normalization or Standardization functions as a standardization of vegetation and climatology indices on a range scale (-1, 1) without changing the actual value, so that it is more relevant in improving model performance. In this study, we standardize the climatology feature values. Rolling averages are usually used to smooth the distribution of time series data and also support the creation of new features (not just to fill in missing values). We apply the Harmonic Series to maintain seasonal patterns in vegetation indices. This is also expected to capture water stress patterns in sugarcane. Lag Feature Engineering applied to climatology Features show the time delay in influencing water stress conditions. For example, rainfall on the previous day (lag 1) can affect soil moisture levels on the following day, which in turn affects water stress levels [17].

The delay pattern for climatology variables varies. For example, air temperature may have an immediate effect (0-1 day lag), while variables such as rainfall and humidity have a slower effect (2-5 day lag). Including too much lag data can increase dimensionality and cause overfitting, so only significant lags (2-5 days) are considered to maintain a balance between model accuracy and efficiency. The lag test results strongly suggest that climatological features play an important role in influencing plant water stress conditions. Therefore, lag engineering and rolling mean features are essential to capture these temporal relationships in modeling, which improves the predictive ability of the model in reflecting water stress indices.

The Normalized Difference Vegetation Index (NDVI) is an index used to measure photosynthetic activity and vegetation health conditions in a particular area or region as shown in Equation 1 [18], [19].

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

NIR: near-infrared light reflectance (Band 5), and Red: red light reflectance (Band 4). NDVI has a range of values between -1 and +1. High NDVI values (close to +1) indicate the presence of healthy and abundant vegetation, while low values(close to -1) indicate non-vegetation areas such as water or buildings. Negative NDVI values usually occur on water surfaces or other non-vegetation objects.

Land Surface Temperature (LST) is the temperature measured directly from the ground surface, without taking into account the influence of the atmosphere. LST is very important in climate monitoring, hydrology, agriculture, and environmental science, as shown in Equation 2 [20], [21].

$$LST = [BT / (1 + L\lambda(BT/p) * \ln(\varepsilon\lambda)] \quad (2)$$

BT variable is Top of Atmosphere (ToA) Brightness Temperature (°C); Lλ is ToA Radiant Spectral Value; ελ is the Emissivity of the ground surface and p is the radiation function (1.438x10-2 mK).

The Crops Water Stress Index (CWSI) is used to measure and monitor the level of drought in crops or agricultural plants. The Crop Water Stress Index (CWSI) was first introduced by Jackson et al. in 1981 [22]. The general formula for CWSI is based on the latest developments by Veysi et al. In 2017, as shown in Equation 3 [13].

$$CWSI = (Ts - Tcold) / (Thot - Tcold) \quad (3)$$

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

Meanwhile, data cleaning on climatology data comprises screening for abnormalities and eliminating layers that do not fit the typical baseline. Also, check for missing values and remove non-numeric data (NaN). The feature engineering process for climatology data involves arranging the data by date, standardizing the value scale, filling in missing values with linear interpolation, preserving seasonal patterns with harmonic sine series (4), and smoothing the time series distribution with rolling window statistical techniques.

The harmonic series Equation 4 that we use in this study is as follows [23].

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos(n\omega t)) + b_n \sin(n\omega t) \quad (4)$$

The frequency, or period value, of the sinusoidal wave in the series is denoted by 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 (2π) of a regular frequency rotation. It is commonly utilized in relation to Fourier and harmonic series. Meanwhile, the series amplitude of sinusoidal waves at various frequencies is given by the coefficients a_0, an, and bn.

2.5 Analyzing Sugarcane Phenology

Phenology analysis helps us to understand the state of vegetation [24]. Matching vegetation characteristic data to planting schedule time in observed sugarcane fields is known as phenology space. In this study, the vegetation characteristics evaluated to estimate water stress conditions are CWSI and NDVI, which reflect the current vegetation.

The time-series statistical line created in the phenology space between CWSI and NDVI is particularly visible during the early growth phase, which is considered a key period of the sugarcane plant against water stress circumstances. With a typical baseline for NDVI ranging from 0.4 to +1, the baseline for CWSI is more

than 0.5, indicating that the vegetation is suffering water drought stress [25].

2.6 Check the strength of correlation

At this stage, the pattern consistency in each characteristic, as well as the strength of the association between climatological factors and sugarcane water stress indicators as goals, will be investigated. The cross-correlation test is a statistical method used to measure the extent to which two time-series variables are related to each other. It is very useful for finding relationships between two data sets that may not be obvious in the initial analysis.

After further development after conducting the cross-correlation test, we measure the extent to which climate features affect the CWSI by shifting one-time series to real-time. The time unit we use is t days, with data at time t+k in the second series, where k is the lag. In this study, we apply k values according to the observation results on Lag Cross-Correlation.

2.7 Prediction Model Based on Machine Learning Regression

The selection of the model used is adjusted to the type of data processed and the objectives to be achieved, where the data processed is in the form of a numerical time series as a representation of spatial-temporal data that has a strong seasonal pattern. The purpose of this modeling is to estimate the value of the vegetation index based on climatological data. Therefore, the selection of the model is directed at a combination of regression-based, ensemble, and non-linear methods.

There are several candidate algorithm options, including *Ada Boost Regressor (ABR), Decision Tree Regressor (DTR), k-nearest Neighbors Regressor (KNNR), Light GBM Regressor (LGBMR), Random Forest Regressor (RFR), Support Vector Regressor (SVR), and XGBoost Regressor (XGBR)*. The machine learning regression model is used to predict the sugarcane worker stress index because these models are relevant for predicting time series data trends. On the other hand, we did not choose the *CatBoost* algorithm because of its advantages for categorical data, so it is not relevant to the numerical dataset used and its application to time series.

Next, each basic algorithm is given a component in the form of hyperparameters to achieve optimal prediction results (see Table 3). Choosing the right method and configuring hyperparameters are essential for developing an efficient prediction model. Each machine learning regression algorithm offers advantages for processing temporal data, such as less overfitting, competitive results, and faster processing time.

AdaBoost Regressor (ABR) This model is more focused on the prior prediction inaccuracy, with the goal of fixing it in the next model [26]. Hyperparameters that can be configured in (ABR): n_estimators (nested) Learning_rate specifies how many estimators (weak models) will be added and how much each estimator contributes to the final model. Low values enable the model to learn slowly but steadily. The loss function chosen, such as 'linear', 'quadratic', or 'exponential', might impact error management.

Table 3. Models Architecture for Prediction

No.	Machine Learning Regression	Hyperparameter Tuning
1	Ada Boost Regressor (ABR)	learning_rate: 0.001, nest: 150
2	Decision Tree Regressor (DTR)	Max_depth: 2, min_samples_leaf: 2, min_samples_split: 2
3	k-Nearest Neighbors Regressor (KNNR)	Algorithm: auto, neighbors: 9, weights: uniform
4	Light GBM Regressor (LGBMR)	Auto-choosing-row-wise multi-threading
5	Random Forest Regressor (RFR)	Max_depth: 25, max_features: 5 min_samples_leaf: 3, min_samples_split: 3, n_estimators: 200
6	Support Vector Regressor (SVR)	C: 1, epsilon: 0.1, kernel: rbf
7	XGBoost Regressor (XGBR)	learning_rate: 0.001, max_depth: 25, n_estimators: 750

Decision Tree Regressor (DTR) predicts target values using a tree structure [27]. Each node splits the input based on the most valuable feature. The following hyperparameters can be set: The maximum depth of the tree is denoted by max_depth (md), while min_samples_split (mss) is the minimal number of samples necessary to split a node. min_samples_leaf (msl) is the minimum number of samples necessary to become a leaf node. The max_features (mf) defines the maximum number of features utilized in each split.

k-Nearest Neighbors Regressor (kNNR) forecasts a target value using the average of the k nearest neighbors [28]. This model is built on the closeness of data points in the feature space. Hyperparameters that may be set include n_neighbors, which is the number of nearest neighbors evaluated, and weight, which is the weight given to the neighbors, such as 'uniform' (for all data to be considered identically) or 'distance' (based on data distance). The notation 'p' shows the kind of distance used to calculate the distance between neighbors (for example, 1 for Manhattan distance and 2 for Euclidean distance).

Light GBM Regressor (LGBMR) The Light GBM Regressor (LGBMR) is a gradient-based boosting strategy that values speed and efficiency [29]. Light GBM uses histogram binning to swiftly process vast quantities of data. Hyperparameters can be configured as follows: num_leaves is the maximum number of leaves in each tree; learning_rate determines the step size while updating the model; n_estimators (nest) is the number of trees to create; max_depth (md) is the

maximum depth of each tree; and boosting_type, which determines the kind of boost.

Random Forest Regressor (RFR) is an ensemble approach that generates many decision trees and combines their predictions to improve accuracy while reducing overfitting [30], [31]. The RFR hyperparameters that may be specified are the same as for the Decision Tree, including n_estimators (nest), which is the number of trees in the forest, max_depth (md), and min_samples_split (mss), as well as max_features (mf), which is the maximum amount of features used for each split.

The *Support Vector Regressor* (SVR) is a regression-specific form of SVM in which the model looks for the best-fit hyperplane among margins [32]. The SVR hyperparameters that can be set include the notation 'C', which governs the trade-off between margin size and error, the type of kernel used (linear, polynomial, RBF, etc.), and epsilon, which is the margin around the hyperplane used to set the threshold of tolerated error. In addition to classification, the application of reduced kernel tricks for regression and dimensionality reduction in the MapReduce framework has good potential to solve large-scale nonlinear support vector machines that are reduced to an important technique in the Big Data era [33].

XGBoost Regressor (XGBR) is a high-performance gradient-boosting algorithm [34]. To avoid overfitting, this approach utilizes early stopping and L1/L2 regularization. The following hyperparameters can be set: n_estimators (nest), which is the number of trees in the model, learning_rate, which is the learning stride to update the weights, max_depth (ms), which is the maximum depth of the tree, gamma, which is an early pruning controller to reduce model complexity, and lambda (λ) and alpha (α), which are L2 and L1 regularization features to prevent overfitting.

2.8. The Optimal Prediction Model

The process of finding the best prediction model begins with assessing the performance of the CWSI prediction model with the k-fold cross-validation approach. K-fold validation is used to enhance model accuracy [35]. The data set's time series is separated into three sets: training, validation, and testing. Training and validation take up 80% of the data set, while testing takes 20%. Equation 5 evaluates the performance of the machine learning regression prediction model using numerous evaluation measures typically used for regression issues. R-squared (R^2) (5) indicates the model's capacity to explain the variability of the target variable. R-squared values vary from 0 to 100%, with higher values indicating greater model performance [36].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (5)$$

n represents the number of samples, *y*i is the observed value from the *i*-th sample, *ŷ*i represents the expected value for the i-th sample, and $\bar{y}_i$ is the average of observed values. Meanwhile, the performance error assessment was evaluated using the absolute error percentage (MAPE) (6) to confirm the prediction model's correctness [37].

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{E_t - A_t}{A_t}\right| \quad (6)$$

In Equation 6, Et represents the predicted value for t-th, whereas At represents the actual value for t-th.

3. Results and Discussions

This section contains the results and discussion according to the proposed process, from data collection to performance assessment of the most appropriate regression machine learning model for predicting sugarcane water stress.

3.1 Analyzing Sugarcane Phenology

After obtaining the NDVI, LST and CWSI time series data as reference data or ground truth. Furthermore, phenology analysis can be carried out and the results are shown in Figure 5. Cross-correlation of the CWSI water stress index (red dotted line) with the NDVI feature (green solid line) and LST sugarcane canopy temperature (blue dotted line). The highlighted phase is the gray shaded area depicting the germination-processing phase also called the beginning of the planting season (SoS) with a pattern of two planting seasons ratoon-1 and ratoon-2 which are susceptible to dry conditions. Based on Figure 5, the gray area shows that the NDVI value is lower than the CWSI, with an index range of < 0.4 while the LST and CWSI values are > 0.5 together. These data indicate that sugarcane vegetation is experiencing moderate water stress [13]. Figure 6 shows a comparison of the seasonal time series data pattern between the actual monthly mean CWSI values represented by the red dotted line with the climate features (a) air temperature, (b) air humidity, (c) rainfall, (d) sunshine, and (e) wind speed represented by the solid blue line.

This provides a good insight into the pattern of climate influence on sugarcane water stress. The shaded areas in Figure 6 represent the beginning of growing season 1 (ratoon-1) and growing season 2 (ratoon-2).

3.2 Check the strength of the correlation

Figure 7 presents the results of the pairwise correlation test to determine the strength of the relationship between the CWSI target features and the climate data predictor features. The distribution of cross-correlation data in Figures 7(a), 7(b), and 7(c) shows that the average air temperature, air humidity, and rainfall are weakly negatively correlated with CWSI. Figures 7(d) and 7(e) show that solar radiation and wind speed are weakly positively correlated with CWSI.

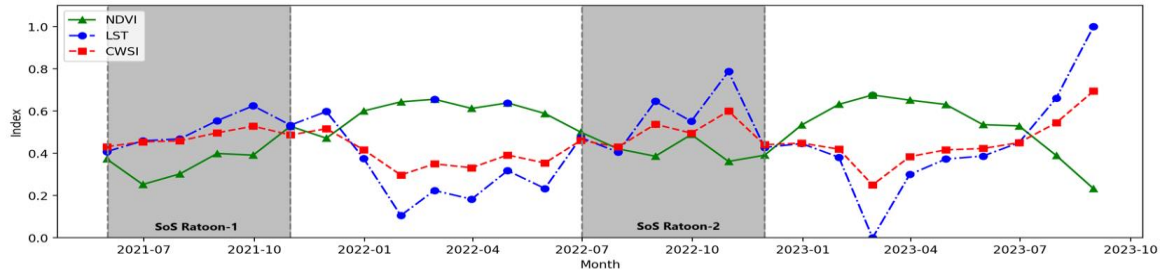


Figure 5. Sugarcane Phenology with NDVI, LST and CWS

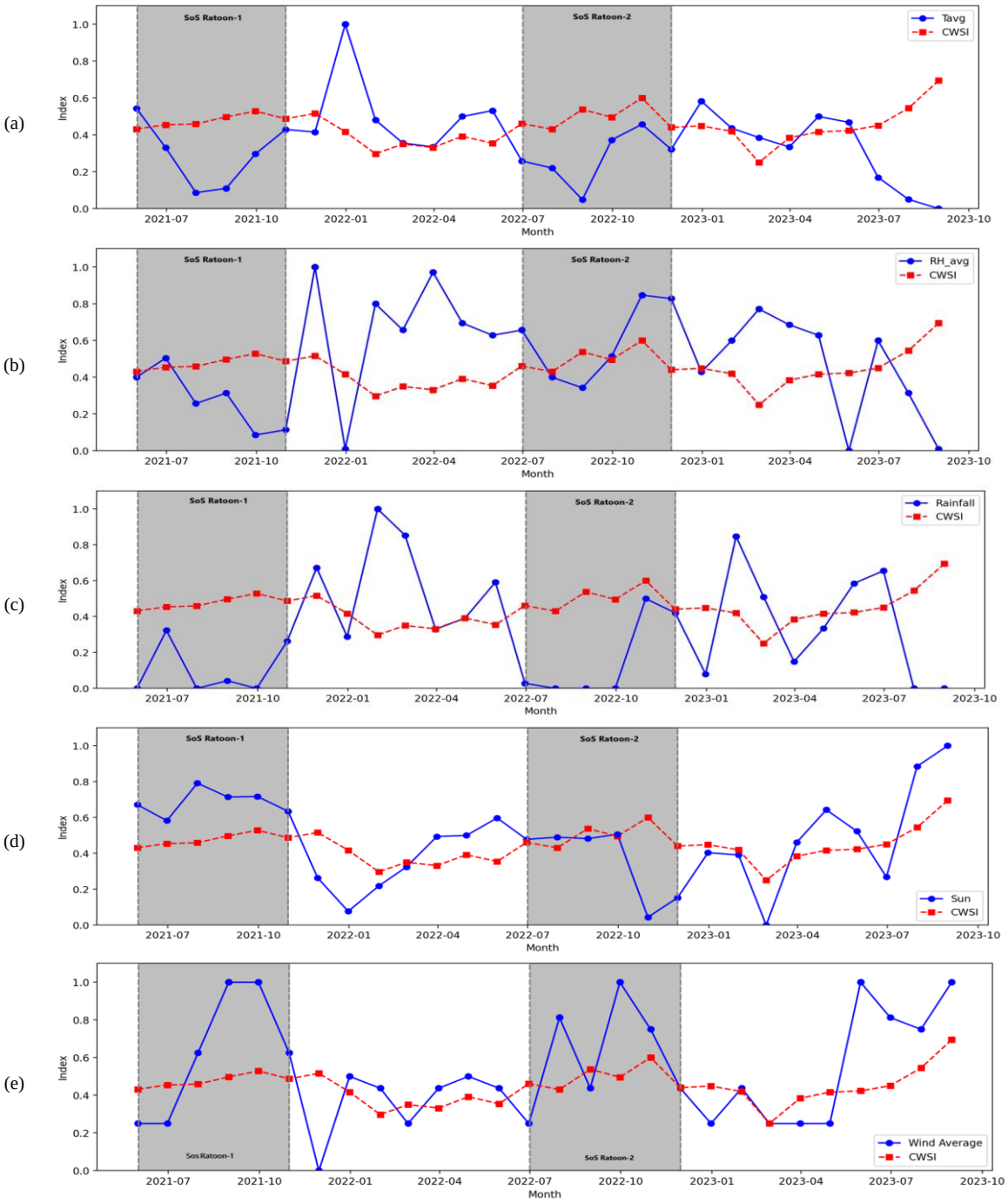


Figure 6. Sugarcane phenology CWSI with climatological features: (a) Air-Temperature, (b) Humidity, (c) Rainfall, (d) Solar-Radiation, (e) Wind-Speed

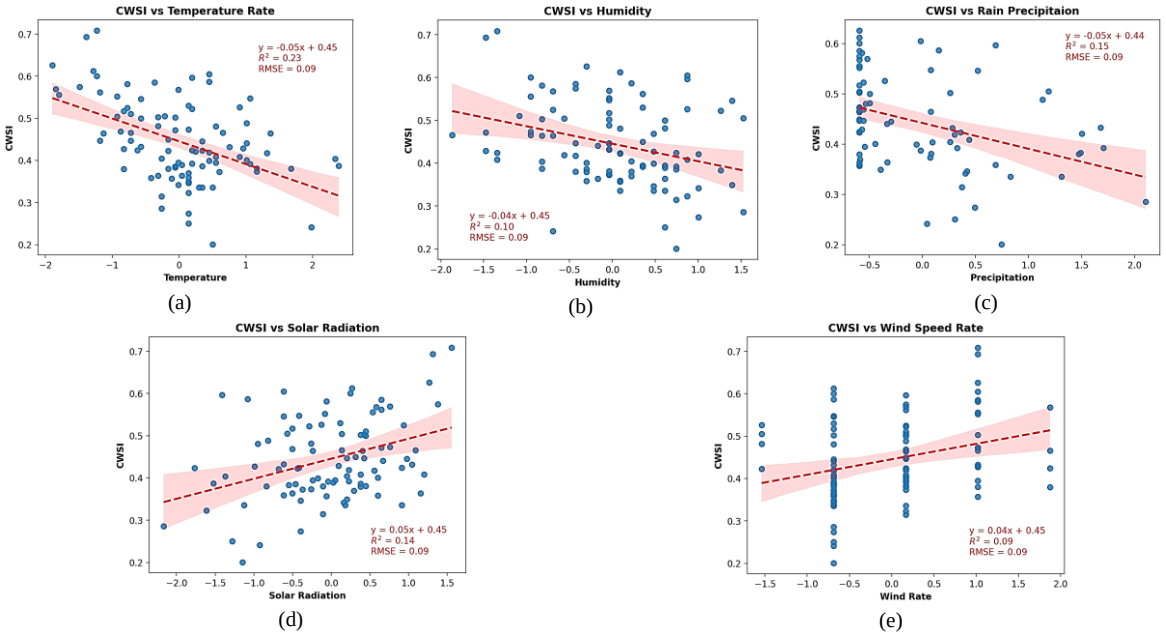


Figure 7. Cross-correlation of CWSI versus Climate features in the scattering distribution: (a) Air Temperature (b), Humidity, (c) Precipitation, (d) Solar Radiation, (e) Wind Speed.

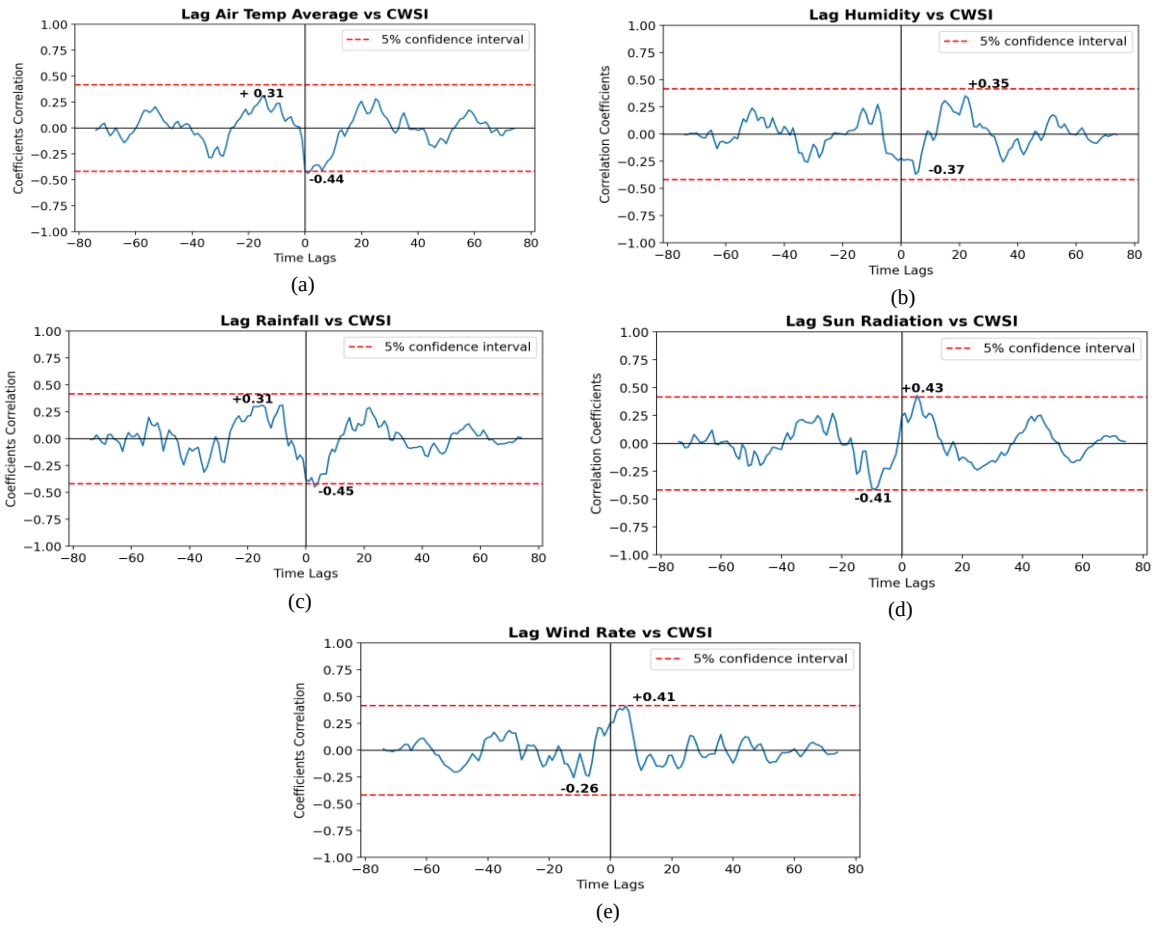


Figure 8. Cross-Lag-correlation of CWSI versus Climate features in Time Lags distribution: (a) Air Temperature (b), Humidity, (c) Rainfall, (d) Solar Radiation, (e) Wind Speed.

The weak relationship is evidenced by the correlation coefficient value *R*2 below 0.5 (50%). However, the value of the correlation determinant coefficient R2 (5) shows a weak value, far below 0.5. Therefore, we took the initiative to form a new variation of climatological characteristics with a time lag effect ranging from 1 to 5 days in the future affecting the CWSI.

Based on the correlation coefficient value and the 5% confidence interval limit in Figure 8, it can be seen that the results of the cross-correlation test with a time lag indicate that climatological characteristics have a less strong influence, either positive or negative, on sugarcane water stress (CWSI). This is as evidenced by the peak value of the cross-correlation coefficient which is less than the confidence limit range (+/- 0.5). The position of the peak value of the correlation coefficient on the time lag axis indicates when the climatological characteristics affect the CWSI, either in the past (-), present (0), or future (+) direction. Figure 8(a) shows that the average air temperature feature has a weak negative correlation with the CWSI, with a correlation coefficient value of (-0.44) at a time lag of zero (0) days, which indicates a direct influence of the average air temperature feature on the CWSI. This is then the specific reason for the average air temperature feature, the addition of the lag feature is not proposed. Figure 8(b) illustrates that humidity is weakly negatively correlated with a correlation coefficient value of (-0.37) and the influence of a lag of around +5 days in the future affects the CWSI. Figure 8(c) shows a weak negative relationship of rainfall feature, with a correlation coefficient value (-0.45) affecting CWSI after 2 to 3 days. Figure 8(d) shows a positive correlation of solar radiation to CWSI with a correlation coefficient value (+0.43) and a lag time effect of around +5 days. Figure 8(e) illustrates a positive correlation for wind speed with a correlation coefficient value of (+0.41), and a lag time effect of around 5 days on CWSI. The results of the Cross-Correlation Lag Test observations show the potential for 26 new features that can be formed from the lag time effect process which includes 20 lag features: lag_cwsi, lag_rh, lag_rr, lag_ss, lag_ffavg and 5 new rolling_mean features and 1 'month' time feature from the Lag and Rolling means processes as seen in Table 4.

Table 4. New features parameter from time Lag climate effect

New Parameter	Description
lag_cwsi1	cwsi value on the previous 1-day time shift
lag_cwsi2	cwsi value on the previous 2-day time shift
lag_cwsi3	cwsi value on the previous 3-day time shift
lag_cwsi4	cwsi value on the previous 4-day time shift
month	calculation internal range
lag_rh1	humidity value on the previous 1-day time shift
lag_rh2	humidity value on the previous 2-day time shift
lag_rh3	humidity value on the previous 3-day time shift
lag_rh4	humidity value on the previous 4-day time shift
lag_rr1	rainfall value on the previous 1-day time shift
lag_rr2	rainfall value on the previous 2-day time shift
lag_rr3	rainfall value on the previous 3-day time shift
lag_rr4	rainfall value on the previous 4-day time shift
lag_ss1	duration of sunlight value on the previous 1-day time shift
lag_ss2	duration of sunlight value on the previous 2-day time shift
lag_ss3	duration of sunlight value on the previous 3-day time shift
lag_ss4	duration of sunlight value on the previous 4-day time shift
lag_ffavg1	win speed average value on the previous 1-day time shift
lag_ffavg2	win speed average value on the previous 2-day time shift
lag_ffavg3	win speed average value on the previous 3-day time shift
lag_ffavg4	win speed average value on the previous 4-day time shift
rolling_mean_rh	the average value of all lags in humidity
rolling_mean_rr	the average value of all lags in rainfall
rolling_mean_ss	the average value of all lags in the duration of sunlight
rolling_mean_ffavg	the average value of all lags in wind speed

3.3 Prediction Models

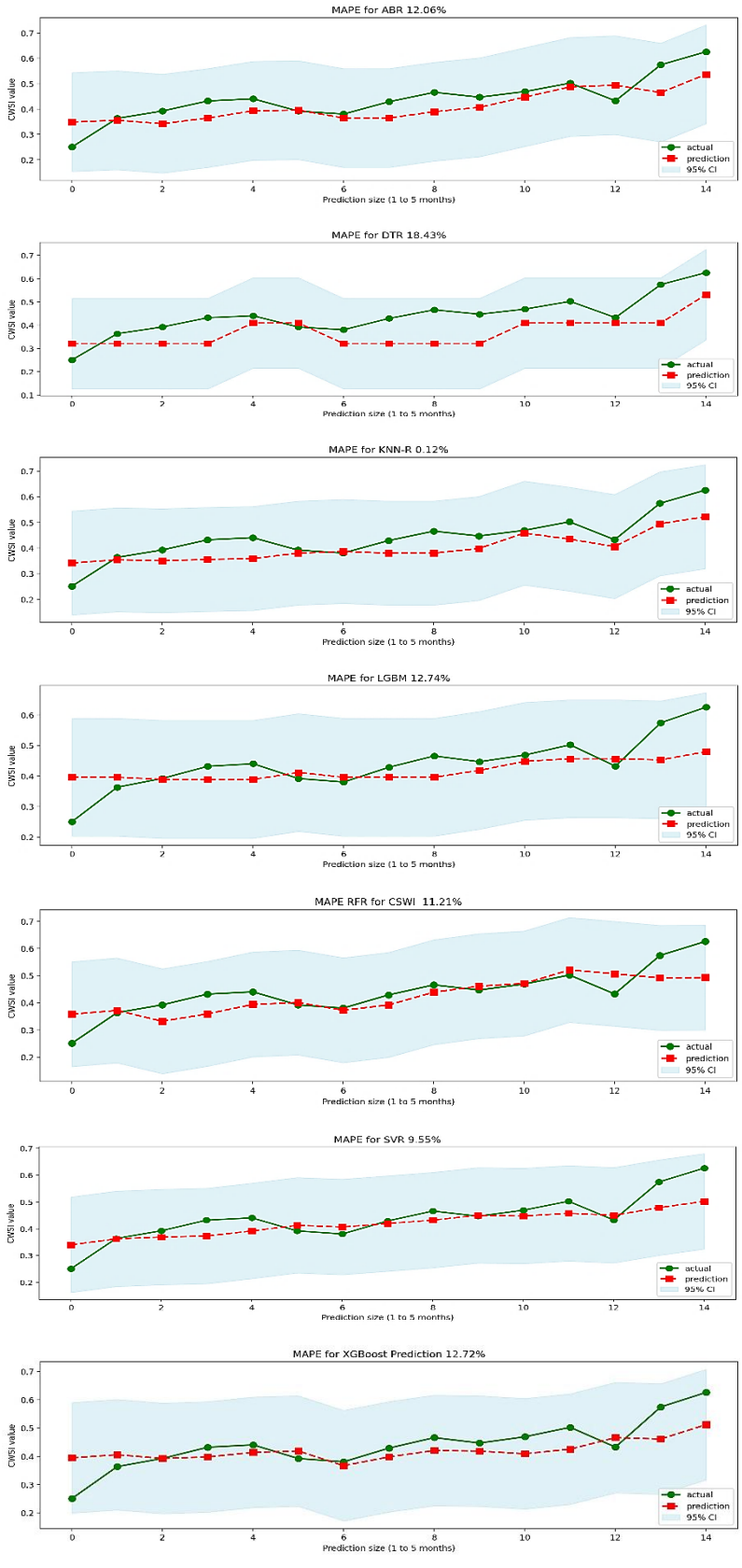
The sugarcane stress prediction models will use the following machine learning regression algorithms: Ada Boost Regressor (ABR), Decision Tree Regressor (DTR), k-Nearest Neighbors Regressor (KNNR), Light GBM Regressor (LGBMR), Random Forest Regressor (RFR), Support Vector Regressor (SVR), and XGBoost Regressor (XGBR). Both the base models and models with hyperparameter tuning were tested.

Table 5 shows the performance of models with and without hyperparameter tuning (BM only and BM + Hyperparameter) for seven machine learning algorithms. Several evaluation metrics are used, such as accuracy correlation determinant coefficient R2 (5), which is the percentage of correct predictions. Avg_Error: Average prediction error. MAPE (Mean Absolute Percentage Error) (6): Average absolute error percentage. Model Improvement: Performance improvement after hyperparameter tuning compared to the base model.

3.4 Error Analysis

Based on Table 5, the results of each machine learning regression model approach can be seen, which shows that the SVR model has the most significant performance improvement after hyperparameter tuning, followed by DTR. The RFR and SVR models have the best performance in terms of MAPE and Avg_Error. On the other hand, models such as LGBMR and XGBR show that tuning does not have a significant impact on improving their performance. In Figure 9 the time-series data plot the actual CWSI values are represented by the solid green line and the predicted CWSI values are represented by the dashed red line which provides

good insight into the model performance on each model approach.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Figure 9. Visualization for performance Sugarcane CWSI prediction based on climatological data model using the regression learning machine: (a) ABR, (b) DTR, (c) KNNR, (d) LGBMR, (e) RFR, (f) SVR, (g) XGBR

Table 5. Model Performance for Prediction

No.	Base Model (BM) Prediction	BM only		BM + Hyperparameter		MAPE	Model improvement
		Accuracy	Avg_Error	Accuracy	Avg_Error		
1	ABR	87,90	0,0511	87,94	0,0511	12,06	0,05
2	DTR	78,89	0,0893	81,57	0,0814	18,43	3,40
3	KNNR	87,09	0,0563	87,75	0,0528	12,00	0,76
4	LGBMR	86,90	0,0539	87,26	0,0533	12,74	0,42
5	RFR	88,87	0,0468	89,19	0,0456	11,21	0,36
6	SVR	85,75	0,0636	90,45	0,0413	9,55	5,48
7	XGBR	87,26	0,0525	87,28	0,0527	12,72	0,02

3.5 Feature importance ranking for each model

Furthermore, to complete the explanation and add insight into the prediction results of each approach model, the Feature Importance is sought for each model. The visualization in Figure 10 shows which features in the SVR contribute the most to influencing the prediction of the vegetation water stress index, for other models, we present them in Table 6.

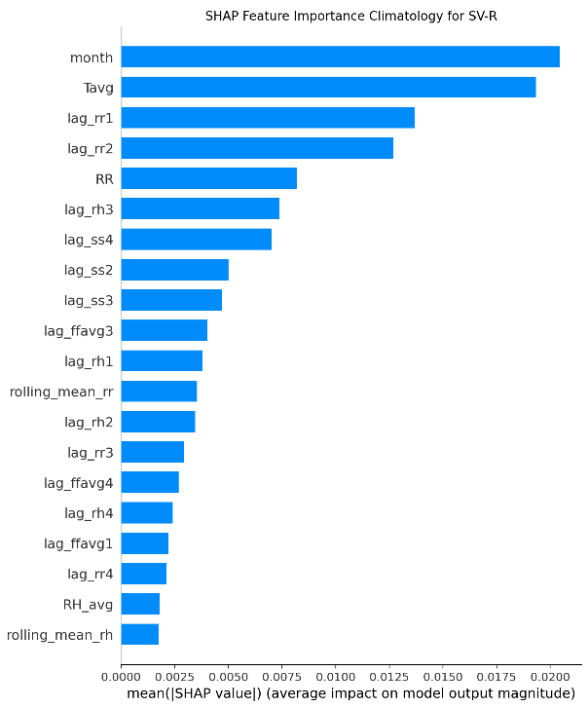


Figure 10. Features importance of prediction model using SVR

The Importance Feature can be explained by which features contribute significantly to the prediction of the water stress index (CWSI) in sugarcane plants. The contribution of features that have a strong influence with an important threshold value limit > 0.1 is generally based on the default attributes in ensemble algorithms such as ABR, DTR, LGBM-R, RF-R, XGBoost (feature_importances_) or SHAP (Shapley Additive Explanations) with an important threshold value limit > 0.004 to display the contribution of features to the KNN-R and SV-R prediction models.

3.6 Contribution of feature

Based on Table 6, the "month" feature shows that seasonality has a big influence on predictions, using time to identify seasonal trends. The "lag_cwsi1" feature (previous CWSI value) highlights the importance of historical data in predicting current values. Rainfall-related features ("rolling_mean_rr" and "lag_rr3") show that accumulated rainfall significantly impacts plant water stress. Sunshine duration features ("lag_ss4" and "ss") indicate that sunlight affects plant stress, but plants are generally safe from drought. The "Tavg" feature (average temperature) has a moderate impact, with extreme temperatures causing more water stress, though it's less involved in model construction due to immediate effects with no lag variation.

3.7 Computational Cost Analysis Based on Model Complexity

In the context of energy efficiency, at the end of this study, we also review the computational costs for training and implementing the model, especially for settings related to resource constraints, presented in Table 7.

Based on the results of model performance in Table 5 and the results of the computational cost analysis in Table 7, the selected models for sugarcane water stress prediction, such as Random Forest or SVR, have high training costs. This can be overcome by limiting or reducing the number of trees or limiting the number of kernel iterations as an effort to reduce computing time.

3.8 Comparison Between Models

SVR (Support Vector Regression) shows the best performance with the highest accuracy (90.45%), lowest average error (0.0413), and lowest MAPE (9.55), benefiting greatly from tuning. ABR (AdaBoost Regression) and RFR (Random Forest Regression) are the most stable models, showing consistent performance with or without hyperparameter tuning. SVR and DTR (Decision Tree Regression) show significant improvements after tuning, indicating high dependence on hyperparameters. While complex models like SVR and RFR provide excellent performance, they require more computational resources than simpler models like DTR. SVR is recommended for its ability to capture complex patterns

and handle non-linear relationships [33]. For model stability without intensive tuning, ABR and RFR are suitable options. Additionally, adding more data and model combinations like ensembles can further improve performance. However, if model stability is required without intensive tuning, ABR and RFR can be considered. The tuning function is more suitable for improving performance for models such as SVR and DTR. On the other hand, there is still the possibility of adding additional data and model combinations, such as ensembles, to further reduce MAPE.

Table 6. Features Importance of Each Approach Model

No.	Machine Learning Regression	Feature Importance	n-Features
1	Ada Boost Regressor (ABR)	month, lag_cwsi1, rolling_mean_rr, lag_rr3, lag_rr2, lag_ss4, Tavg, lag_rr4, ss	9
2	Decession Tree Regressor (DTR)	month, rolling_mean_rr, lag_cwsi2, lag_ss4, lag_rr2	5
3	k-Nearest Neighbors Regressor (KNNR)	month, lag_rr1, RR, Tavg, lag_rr2, lag_rh1, lag_ss4, lag_rh2, lag_rh4, ff_avg, ss, RH_avg, lag_rr4, lag_rh3, lag_rr3, lag_ss1, lag_ss2, lag_ffavg1, lag_ffavg2, lag_ffavg3	20
4	Light GBM Regressor (LGBMR)	month, rolling_mean_rr, lag_rr3	3
5	Random Forest Regressor (RFR)	rolling_mean_rr, month, lag_cwsi1, lag_rr3, lag_ss4, lag_rr2, lag_rr1, Tavg, rolling_mean_rh, rolling_mean_ss, RR, lag_cwsi4	12
6	Support Vector Regressor (SVR)	month, Tavg, lag_rr1, lag_rr2, RR, lag_rh3, lagss4, lag_ss2, lag_ss3, lag_ffavg3, lag_rh1, rolling_mean_rr, lag_rh2, lag_rr3, lag_ffavg4, lag_rh4, lag_ffavg1, lag_rr4, RH_avg, rolling_mean_rh	20
7	XGBoost Regressor (XGBR)	month, rolling_mean_rr, lag_ss4, lag_rr2, lag_rr4, RR, lag_cwsi2, lag_ffavg1, lag_ss1, ff_avg	10

Table 7. Computational Cost Analysis Based on Model Complexity

Model	Training Complexity	Implementation Complexity	Computational Cost
AdaBoost Regressor (ABR)	Moderate	Low	Requires many iterations for training but is quite efficient when predicting.
Decision Tree Regressor (DTR)	Low	Low	Fast in training and deployment, suitable for limited resource settings.
k-Nearest Neighbors (KNNR)	Low	High	Requires a lot of memory to store data and is expensive when searching for nearest neighbors.
LightGBM Regressor (LGBMR)	Moderate	Low	Designed for efficiency, training and prediction is relatively resource-saving.
Random Forest Regressor (RFR)	High	Moderate	It takes a long time due to the large number of trees, but the implementation is quite efficient.
Support Vector Regressor (SVR)	High	High	Using kernels makes it computationally expensive for training and prediction.
XGBoost Regressor (XGBR)	High	Low	Very efficient for distribution setup, but requires large resources initially.

4. Conclusions

The difficulty in measuring the water stress index in sugarcane led to using climatological data with time lag features for estimation. Although the correlation analysis was not very strong, phenological observations suggested seasonal patterns of water stress related to climatological features like temperature, humidity, and rainfall. Machine learning regression models, including SVR, ABR, DTR, KNNR, LGBMR, RFR, and XGBR, were used for estimation. SVR showed the best performance, significantly improving after hyperparameter tuning. The "month" feature had the highest contribution, indicating the significant influence of seasonality. These findings suggest that local climatological data can be a viable alternative to satellite data for predicting water stress, emphasizing the importance of local data in agricultural research and water resource management. Further challenges include developing model scales for wider areas and integrating climatological variables into larger datasets.

Acknowledgements

We would like to thank the PTPN-X Research Center and HGU for their essential assistance and permission to conduct research on the sugarcane fields they administer. We'd also like to thank the United States Geological Survey (USGS) and Google Earth Engine (GEE) for offering free satellite data and cloud-based data processing engines. The authors would also like to express their profound thanks to IPB University and Singaperbangsa University Karawang for offering the chance to pursue a Doctoral degree and supporting this paper.

References

[1] A. Yusara, H. Handoko, and B. Budianto, “Water Demand Analysis of Sugarcane Based on Crop Simulation Model (Case Study: Kediri Regency, East Java),” *Agromet*, vol. 33, no. 1, pp. 30–40, 2019, doi: 10.29244/j.agromet.33.1.30-40.

[2] P. D. dan S. I. Pertanian, “Outlook Tebu,” *Pusat Data dan Sistem Informasi Pertanain*, 2016.

[3] N. Qin, Q. Lu, G. Fu, J. Wang, K. Fei, and L. Gao, “Assessing the drought impact on sugarcane yield based on crop water requirements and standardized precipitation evapotranspiration index,” *Agric Water Manag*, vol. 275, no.

July 2022, p. 108037, 2023, doi: 10.1016/j.agwat.2022.108037.

[4] J. A. O. Reyes, D. E. Casas, J. L. Gandia, and E. F. Delfin, *Drought impact on sugarcane production*, vol. 35, no. June. 2021.

[5] L. C. Santos *et al.*, "Influence of deficit irrigation on accumulation and partitioning of sugarcane biomass under drip irrigation in commercial varieties," *Agric Water Manag*, vol. 221, no. June 2018, pp. 322–333, 2019, doi: 10.1016/j.agwat.2019.05.013.

[6] A. Narayanamoorthy, "Impact assessment of drip irrigation in India: The case of sugarcane," *Development Policy Review*, vol. 22, no. 4, pp. 443–462, 2004, doi: 10.1111/j.1467-7679.2004.00259.x.

[7] P. J. Dlamini, "Drought stress tolerance mechanisms and breeding effort in sugarcane: A review of progress and constraints in South Africa," *Plant Stress*, vol. 2, no. December 2020, p. 100027, 2021, doi: 10.1016/j.stress.2021.100027.

[8] G. Sonkar, N. Singh, R. K. Mall, K. K. Singh, and A. Gupta, "Simulating the Impacts of Climate Change on Sugarcane in Diverse Agro-climatic Zones of Northern India Using CANEGRO-Sugarcane Model," *Sugar Tech*, vol. 22, no. 3, pp. 460–472, 2020, doi: 10.1007/s12355-019-00787-w.

[9] D. Zhao and Y. R. Li, "Climate Change and Sugarcane Production: Potential Impact and Mitigation Strategies," *International Journal of Agronomy*, vol. 2015, 2015, doi: 10.1155/2015/547386.

[10] S. O. Ihuoma and C. A. Madramootoo, "Recent advances in crop water stress detection," *Comput Electron Agric*, vol. 141, pp. 267–275, 2017, doi: 10.1016/j.compag.2017.07.026.

[11] H. G. Jones and R. A. Vaughan, *Remote sensing of vegetation: principles, techniques, and applications*. Oxford University Press, USA, 2010.

[12] Jackson, "Thermal Crop Water Stress Indices," *Stress: The International Journal on the Biology of Stress*, pp. 1–12, 1982.

[13] 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.

[14] R. Triadi, Y. Herdiyeni, and S. D. Tarigan, "Estimating crop water stress of sugarcane in indonesia using landsat 8," *2020 International Conference on Computer Science and Its Application in Agriculture, ICOSICA 2020*, pp. 8–11, 2020, doi: 10.1109/ICOSICA49951.2020.9243255.

[15] S. Y. J. Prasetyo, K. D. Hartomo, and M. C. Paseleng, "Satellite imagery and machine learning for identification of aridity risk in central Java Indonesia," *PeerJ Comput Sci*, vol. 7, pp. 1–21, 2021, doi: 10.7717/PEERJ-CS.415.

[16] 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.

[17] 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.

[18] R. Calderón, J. A. Navas-Cortés, C. Lucena, and P. J. Zarco-Tejada, "High-resolution airborne hyperspectral and thermal imagery for early detection of Verticillium wilt of olive using fluorescence, temperature and narrow-band spectral indices," *Remote Sens Environ*, vol. 139, pp. 231–245, 2013, doi: 10.1016/j.rse.2013.07.031.

[19] 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.

[20] J. E. Nichol and S. Abbas, "Integration of remote sensing datasets for local scale assessment and prediction of drought," *Science of the Total Environment*, vol. 505, pp. 503–507, 2015, doi: 10.1016/j.scitotenv.2014.09.099.

[21] P. S. Käfer, S. B. A. Rolim, L. R. Diaz, N. S. da Rocha, M. L. Iglesias, and F. E. Rex, "Comparative Analysis of Split-Window and Single-Channel Algorithms for Land Surface Temperature Retrieval of a Pseudo-Invariant Target," *Boletim de Ciências Geodésicas*, vol. 26, no. 2, pp. 1–17, 2020, doi: 10.1590/s1982-21702020000200008.

[22] R. D. Jackson, S. B. Idso, R. J. Reginato, and P. J. Pinter, "Canopy temperature as a crop water stress indicator," *Water Resour Res*, vol. 17, no. 4, pp. 1133–1138, 1981, doi: 10.1029/WR017i004p01133.

[23] I. Pesenson, Q. Thong, L. Gia, A. Mayeli, H. Mhaskar, and D. Zhou, *Novel Methods in Harmonic Analysis, Volume 2*, vol. 2.

[24] Y. Herdiyeni, M. F. Mumtaz, G. F. Laxmi, Y. Setiawan, L. B. Prasetyo, and T. R. Febbiyanti, "Analysis and prediction of rubber tree phenological changes during Pestalotiopsis infection using Sentinel-2 imagery and random forest," *J Appl Remote Sens*, vol. 18, no. 1, p. 14524, 2024, doi: 10.1117/1.JRS.18.014524.

[25] S. Irmak, D. Z. Haman, and R. Bastug, "Determination of crop water stress index for irrigation timing and yield estimation of corn," *Agron J*, vol. 92, no. 6, pp. 1221–1227, 2000, doi: 10.2134/agronj2000.9261221x.

[26] W. Wang and D. Sun, "The improved AdaBoost algorithms for imbalanced data classification," *Inf Sci (N Y)*, vol. 563, pp. 358–374, 2021, doi: 10.1016/j.ins.2021.03.042.

[27] N. Desai and V. Patel, "Linear Decision Tree Regressor: Decision Tree Regressor Combined with Linear Regressor," no. July, 2021.

[28] J. Luo, Y. Wang, Y. Ou, B. He, and B. Li, "Neighbor-Based Label Distribution Learning to Model Label Ambiguity for Aerial Scene Classification," *Remote Sens (Basel)*, vol. 13, no. 4, p. 755, 2021, doi: 10.3390/rs13040755.

[29] R. P. Sheridan, A. Liaw, M. Tudor, and S. Chemistry, "Light Gradient Boosting Machine as a Regression Method for Quantitative Structure- Activity Relationships."

[30] Y. Ao, H. Li, L. Zhu, S. Ali, and Z. Yang, "The Linear Random Forest Algorithm and its Advantages in Machine Learning assisted Logging Regression Modeling," *J Pet Sci Eng*, 2018, doi: https://doi.org/10.1016/j.petrol.2018.11.067.

[31] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata Journal*, vol. 20, no. 1, pp. 3–29, 2020, doi: 10.1177/1536867X20909688.

[32] Y. Forghani, R. S. Tabrizi, H. S. Yazdi, and M. R. Akbarzadeh-T, "Fuzzy support vector regression," *2011 1st International eConference on Computer and Knowledge Engineering, ICCKE 2011*, no. Vc, pp. 28–33, 2011, doi: 10.1109/ICCKE.2011.6413319.

[33] Y.-R. Yeh, S.-Y. Huang, H.-K. Pao, and Y.-J. Lee, "a Review of Reduced Kernel Trick," *Journal of the Chinese Statistical Association*, vol. 52, pp. 85–114, 2014, [Online]. Available: http://jupiter.math.nctu.edu.tw/~yuhjye/assets/file/publications/journal_papers/J4_A Review of Reduced Kernel Trick in Machine Learning.pdf

[34] N. Uzir, S. Raman, S. Banerjee, and R. S. Nishant Uzir Sunil R, "Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets Experimenting XGBoost Algorithm for Prediction and Classifi cation of Different Datasets," *International Journal of Control Theory and Applications*, vol. 9, no. July, 2016.

[35] Y. Jung, "Multiple predicting K-fold cross-validation for model selection," *J Nonparametr Stat*, vol. 30, no. 1, pp. 197–215, 2018, doi: 10.1080/10485252.2017.1404598.

[36] A. C. Cameron and F. A. G. Windmeijer, "An R-squared measure of goodness of fit for some common nonlinear regression models," *J Econom*, vol. 77, no. 2, pp. 329–342, 1997, doi: 10.1016/s0304-4076(96)01818-0.

[37] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean Absolute Percentage Error for regression models," *Neurocomputing*, vol. 192, pp. 38–48, 2016, doi: 10.1016/j.neucom.2015.12.114.