Optimized machine learning framework for crop yield prediction using climate and emission data

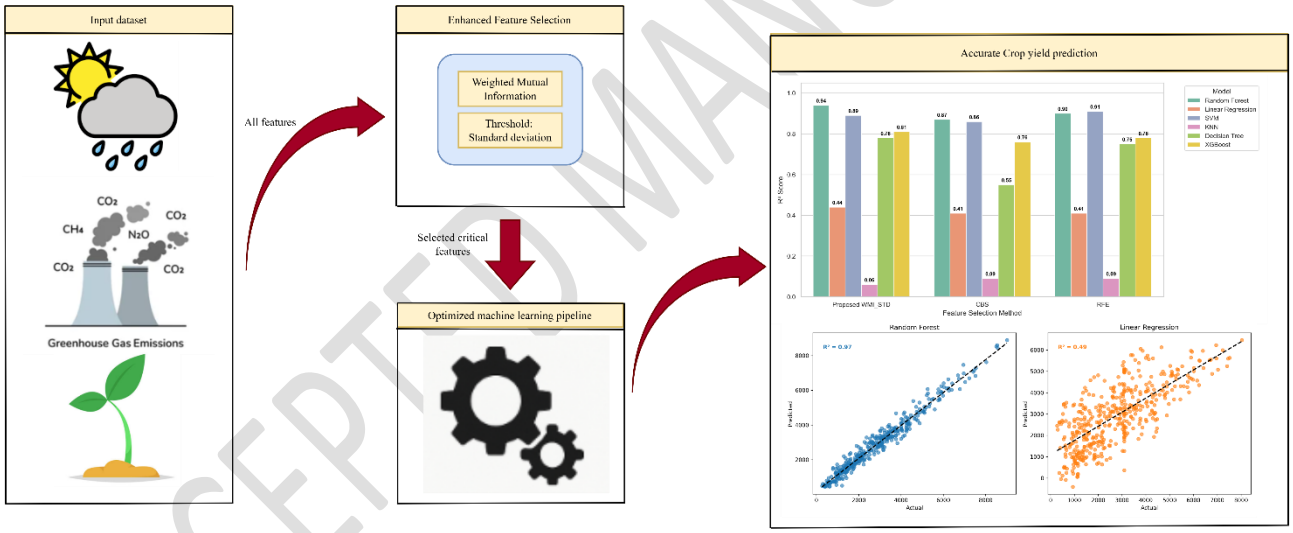
Nivethitha Krishnadoss[1], LokeshKumar Ramasamy[2,*]

[1] Research Scholar, School of computer science and engineering, Vellore institute of technology, Vellore 632 014, Tamil Nadu, India.

[2] Professor, School of computer science and engineering, Vellore institute of technology, Vellore 632 014, Tamil Nadu, India.

* to whom all correspondence should be addressed: e-mail: lokeshkumar.r@vit.ac.in

Graphical Abstract



ACCEPTED MANUSCRIPT

Abstract

The global food security challenge poses a significant risk due to climate change and rapidly growing greenhouse gas (GHG) emissions. In this study, an ensemble learning (EL) framework for global crop yield prediction was developed using key climate variables and two major GHG emissions: carbon dioxide (CO_2) and nitrous oxide (N_2O). The key contribution lies in the proposed Weighted Mutual Information with Standard Deviation Method (WMI_SDM), a feature selection to handle high-dimensional dataset. The method estimates a mutually independent feature dependence score for each feature relative to crop yield, normalizes these scores in feature weights and finds a standard deviation-based threshold to identify the most significant features. Features with weights exceeding this adaptive threshold are retained for model training. Using these selected features, several machine learning (ML) models - including Linear Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and Extended Gradient Boost (XGBoost) were trained and evaluated. The experimental results confirmed that the proposed approach outperformed other models, achieving a coefficient of determination (R^2) of 0.9673, MAE of 226.71 kg ha^{-1}, and MAPE of 11.33%. Furthermore, the proposed method showed remarkable computation efficiency, consuming only 281.99 MiB of memory while completing in 0.55 seconds.

Keywords: Climate data; Greenhouse gas emissions; Feature selection; Mutual information; Ensemble learning; Standard deviation thresholding; Crop yield prediction.

1. Introduction

The prediction of global crop yield is critical for efficient food security strategies and agricultural planning (*Varma et al. 2023, Karthikeyan et al. 2023*). The need for agricultural predictive modeling has grown as a result of issues such as population expansion, environmental change, and scarcity of arable land (*Yewle et al. 2025* , *Iniyan et al. 2023*). Classical crop yield forecast methodologies, which generally rely on historical averages or expert judgment, are no longer

ACCEPTED MANUSCRIPT

effective in reflecting the unpredictability and complexity of present agricultural systems (*Lokeshwari et al. 2024*). Crop development and final yields are significantly influenced by agricultural inputs such as precipitation, fertilizer use, and arable land (*Abdulla et al. 2015*). Furthermore, GHG emissions have an indirect effect on crop performance by affecting the long-term quality of the soil and air (*Richa et al. 2015*). Prediction models have significantly improved as a result of including these parameters (*Gong et al. 2021* , *Sharafi et al. 2023*).

There are many reasons why redundant inputs can negatively affect the model performance. Thus, the present state of ill-selected features is a major problem for most ML models (*Fraino et al. 2023*). This prevents the generalization of new data and has the potential to lead to overfitting. Based on recent research, more advanced and reliable approaches are needed to find and preserve only the most important features of large and complicated models (*Kamangir et al. 2024*).

Mutual Information (MI) is a good FS method capable of finding linear and non-linear associations between features and a target variable (*Zhou et al. 2022*). WMI is another method that weighs features based on their relevance scores. In this method, threshold a decision point defined by a significance set is used to select features with a higher confidence in high-dimensional datasets (*Abdel et al. 2024*). If the FS problem does not yield the desired results using other methods, a solution to this problem could be provided by using an approach called SDM thresholding. The FS algorithm can then eliminate all features with less information than the specified variability threshold to preserve only stable and statistically significant characteristics. By reducing the number of redundant input features, WMI_SDM can facilitate successful characterization of crop yields. Such strategies can be coupled with EL techniques to attain a high degree of accuracy and reduce the number of errors (*Sah et al. 2022* , *Singh et al. 2025*).

This study aims to devise a precise and understandable framework for forecasting crop yield by integrating climate and emission data, applying a FS technique based on WMI, enhanced using SDM, and tested on a strong ensemble of ML models.

ACCEPTED MANUSCRIPT

2. Dataset and pre-processing

The dataset used for this study was a full set of variables from a publicly accessible repository from the World Bank (http: //data. worldbank. org/indicator? tab=all), covering all indicators related to crop production, socioeconomic and environmental conditions. The World Bank data are heavily validated and up-to-date, providing high quality and credible data, which makes them more robust in terms of the generalizability of research findings, especially in regards to the analysis of global trends in development (*Fernández et al. 2025* , *Heydari et al. 2025*). There are 20 variables in the dataset, and the number of observations is 3,724 for different countries and years. Table. 1 shows a detailed description of global crop yield prediction dataset.

Table 1. Data set description

Variable name	Description
Avg_Perc	Average precipitation in depth (mm per year)
Annu_Fres_Wat_with	Annual freshwater withdrawals, total (% of internal resources).
For_Area	Forest area (% of land area)
Fert_Consump	Fertilizer consumption (kilograms per hectare of arable land)
Agri_Irri_Land	Agricultural irrigated land (% of total agricultural land)
Agri_Land	Agricultural land (% of land area)
Ara_Land	Arable land (% of land area)
CO2_emis_agri	Agricultural CO_2 emissions (Mt CO2e)
N2O_emission	Nitrous Oxide emissions (Mt)
Ene_Use	Energy use (kg of oil equivalent per capita)
Ele_Pow_Consmp	Electric power consumption (kWh per capita)
Ren_Elec_OP	Renewable electricity output (% of total electricity output)
Ren_Ene_Consp	Renewable energy consumption (% of total energy consumption)
Acc_To_Elect	Access to electricity (% of population)

ACCEPTED MANUSCRIPT

GDP_Per_Capita	GDP per capita (current US$)
Pop_Growth	Population growth (annual %)
Country Name	Name of the country.
Country Code	Country code.
Year	Year of observation.

The dataset was pre-processed, that is, only values in the fields were provided, missing data were replaced with column means, and the feature distribution was normalized to ensure there was a clean and well-structured dataset that could be used in the model training.

3. Methodology

This section describes the overall framework of the proposed crop yield prediction model, including WMI_SDM to filter weak features, and EM design to enhance prediction accuracy. This process has been validated using well-established evaluation metrics for multiple regression models.

3.1.Overview of proposed crop yield model using WMI_SDM

The objectives of this study were to predict global crop yields with an EL strategy using WMI_SDM method, which provides both optimal selection of features and improved robustness of model predictions. The overall workflow of the proposed crop yield model using WMI SDM method is as shown in Figure 1. This study presents the WMI_SDM model as a feature selection algorithm designed to eliminate less informative predictors in high-dimensional data on crop yield, weather conditions, and GHG emissions by integrating information-theoretic and statistical features. The ML process was divided into four major stages: (i) data preprocessing, (ii) feature selection, (iii) model building, and (iv) evaluation.

- **Phase 1 - Data preprocessing**: Phase 1 was the numeric attribute filtering step, missing value replacement with the mean calculation step, and elimination of non-numeric or irrelevant data to maintain the integrity of the dataset. A PowerTransformer was subsequently used to normalize

ACCEPTED MANUSCRIPT

the variance and make the distribution more Gaussian, and normalization was then used to transform all variables to a uniform distribution to eliminate the effects of high-magnitude variables. The data were divided in the relationship of 80:20 to be used as training and testing. Finally, the outlier was identified by the absolute errors of prediction by absolute thresholds after first training the model; an observation that surpassed the defined limit was eliminated to enhance the strength and reliability of the further model analysis.

- **Phase 2 - Feature selection**: The proposed WMI SDM method was used to identify features based on the algorithm described below. The reason for computing the MI between each predictor and the target variable such as crop yield was to capture both linear and nonlinear dependencies. The resulting MI scores were then normalized to produce feature weights, and an adaptive threshold, which is defined as the standard deviation of the normalized MI values, was used to eliminate redundant and less informative predictors. Such an adaptive process also eliminates manual hyperparameter tuning and dynamically adapts to the underlying data distribution.
- **Phase 3 - Model building**: Phase three was implemented where six regression algorithms, including RF, SVR, KNN, DT, XGBoost and LR, were trained on 80 percent of the data and evaluated on the remaining 20 percent using default scikit-learn parameter settings to provide a fair comparison.
- **Phase 4 - Evaluation:** Phase four focused on predictive and computational performance. R^2, MAE, RMSE, and MAPE were used to evaluate the predictive accuracy, and processing time and peak memory consumption were used to evaluate the computational efficiency of each feature-selection tool. This holistic evaluation includes accuracy, scalability, and efficiency.

ACCEPTED MANUSCRIPT

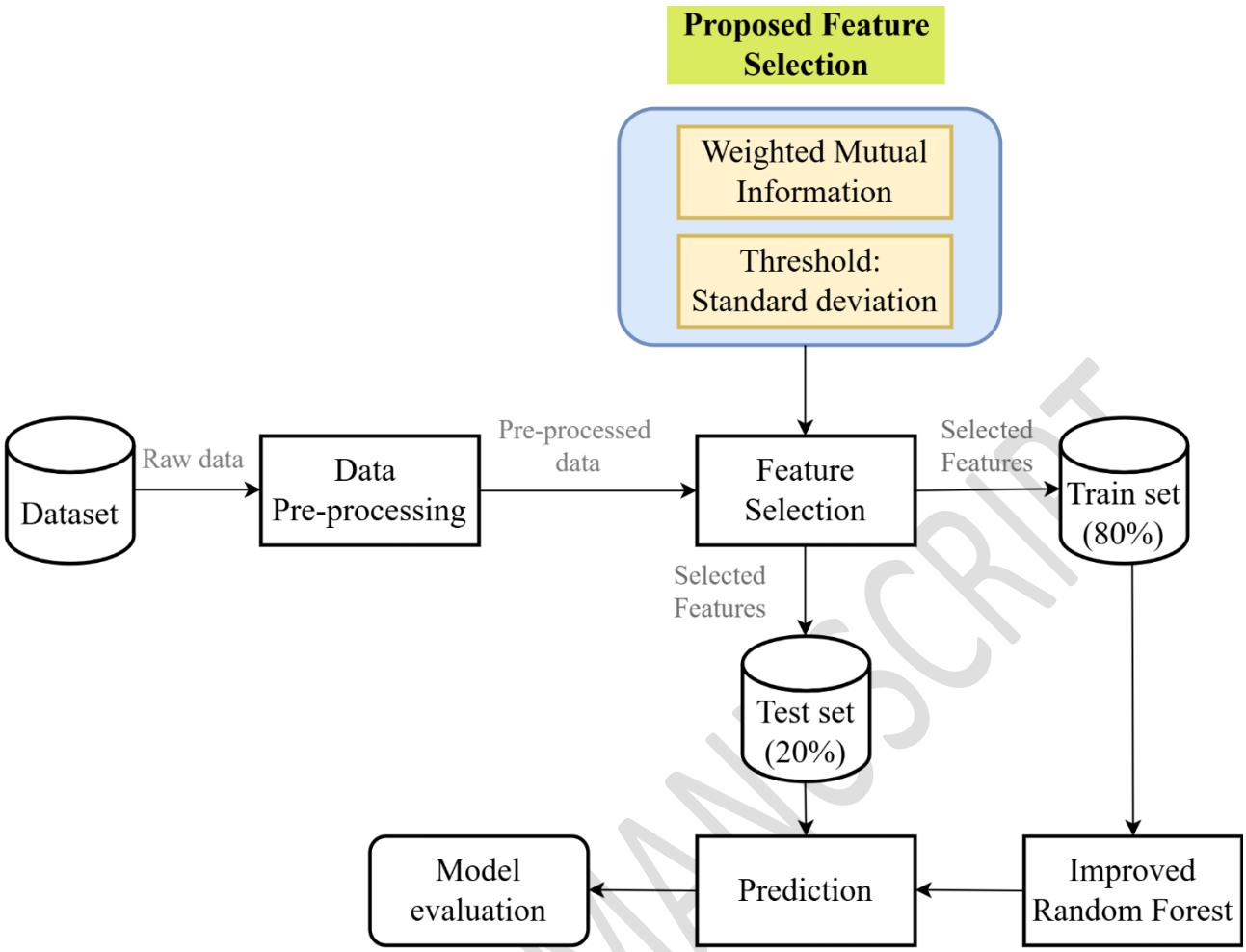


Figure 1. Overall workflow of the proposed crop yield prediction model using the WMI_SDM feature selection approach. The process includes data preprocessing, feature selection using WMI with SD-based thresholding, model training, prediction, and evaluation.

ALGORITHM: Proposed Feature selection using WMI_SDM

Input: Pre-processed Dataset $D = \{X, y\}$ with features $X = \{X_1, X_2, \ldots\ldots, X_n\}$ and crop yield y as the target.

Output: Critical features $X' \subseteq X$.

Step 1: Compute $\mathrm{MI}(\mathrm{X_i}, \mathrm{y}) = \sum_{\mathrm{X_i}} \sum_{y} \mathrm{P}(\mathrm{X_i}, \mathrm{y}) \log \frac{\mathrm{P}(\mathrm{X_i}, \mathrm{y})}{\mathrm{P}(\mathrm{X_i})\mathrm{P}(\mathrm{y})}$, where $\mathrm{P}(\mathrm{X_i}, \mathrm{y})$ is the joint probability distribution of features $\mathrm{X_i}$ and y, and $\mathrm{P}(\mathrm{X_i})$, $\mathrm{P}(\mathrm{y})$ are the marginal distributions.

ACCEPTED MANUSCRIPT

Step 2: Normalize MI score with $\mathrm{w_i} = \frac{\mathrm{MI(X_i,y)}}{\sum_j \mathrm{MI(X_j,y)}}$.

Where $w_i$represents the WMI score for feature X_i, ensuring that $\sum_i \mathrm{w_i} = 1$.

Step 3: Compute the standard deviation threshold using $S = \sqrt{\frac{n\sum_{i=1}^{n} w_i^2 - \left(\sum_{i=1}^{n} w_i\right)^2}{n(n-1)}}$.

Where n is the total number of samples.

Step 4: Select critical features based on threshold (t):

- Define threshold (t): $t = S$.
- Select all features whose importance weights exceed or equal the threshold ($w_i \geq t$):

$$\mathrm{X'} = \{\mathrm{X_i} \mid \mathrm{w_i} \geq \mathrm{t}\}.$$

For instance, if the computed normalized MI score for features $[X_1, X_2, X_3, X_4, X_5,] = [0.12, 0.28, 0.31, 0.05, 0.24]$ and the adaptive threshold $S = 0.09$, then the features satisfying $w_i \geq 0.09$ that is $X' = [X_1, X_2, X_3, X_5,]$ will be selected as the critical features for model training.

Step 5: Use the selected features $X'$for model training.

3.2.WMI based FS

MI is a measure of the degree of informational exchange between the input features and the target variable, in this case, cereal yield (*Cheng et al. 2022* , *Saleh et al. 2018*). In this study, we transform MI by assigning weights based on relevance scores and obtain WMI. The aim of this approach is to prioritize features with large informational value when predicting yield while not overlying them. We computed the MI and ranked the features according to their scores. This provides a logical basis for isolating the most informative predictors from high-dimensional input space.

ACCEPTED MANUSCRIPT

3.3.SDM based thresholding

The final FS strategy involves a thresholding algorithm based on the SD, and features with MI scores below one SD from the mean are deleted, assuming they have a low predictive value or noise, to produce a FS strategy that only produces robust and consistently informative features. The resulting optimization decreases overfitting and improves model generalization. In this study, the SD of the normalized MI score is used as an improvised threshold whereby the most influential features are identified. The SD is used to show how much the feature relevance scores are scattered around; thus, a large SD would imply more variability of feature importance. Setting the threshold to the SD keeps only the features whose weighted MI scores exceed the natural variability that the data possesses.

3.4.EM framework

The basis of this study is a set of baseline learners (RF, SVM, KNN, DT, XGBoost, LR) selected for their ability to capture linear, nonlinear, and hierarchical dynamics of the data; although each model is evaluated separately, the ensemble potential can be exploited by means of fusion strategies. This work compensates for the performance weaknesses of individual models and provides a more balanced output, with RF and XGBoost used to mimic bagging and boosting behaviors, respectively.

Overall, the proposed WMI_SDM method was implemented in Python using the scikit-learn library to ensure transparency, reproducibility, and clear parameterization. The method resulted in the selection of fifteen critical predictors: Avg_Perc, Fert_Consump, Agri_Land, Ara_Land, For_Area, Acc_To_Elect, Ren_Elec_OP, Ren_Ene_Consp, Ele_Pow_Consmp, Ene_Use, CO2_emis_agri, Annu_Fres_Wat_with, GDP_Per_Capita, N2O_emission, and Pop_Growth. Because WMI_SDM is a statistical threshold-based feature selection technique, it does not require hyperparameter tuning, ensuring consistent and interpretable results across different datasets.

ACCEPTED MANUSCRIPT

4. Evaluation metrics

In this study, four widely used regression evaluation metrics were used to assess the proposed crop yield model: Coefficient of Determination (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). R^2 specifies the percentage of variance in the observed yield values, and a value close to one indicates a better fit between the predicted and ground truth values (*Anand et al. 2025*). The MAE was obtained by calculating the mean of predicted and ground truth values (*Garai et al. 2024*). RMSE is the square root of the average squared difference between the predicted and ground truth values (*Panigrahi et al. 2023*). MAPE describes prediction accuracy as a percentage (*Singh et al. 2025*).

5. Results and Discussion

Table. 2 shows the performance comparison of different baseline (LR, RF, SVM, KNN, DT,and XGBoost) without FS. As shown, the RF model is superior to all other regression models. Table. 3 presents a comparison of baseline models with three different FS approaches before removing outliers: proposed WMI_SDM, correlation-based selection (CBS), and Recursive Feature Elimination (RFE). The proposed WMI_SDM technique consistently outperformed with baseline methods across most models.

Table 2. Performance comparison of various models without Feature selection

Model	R^2 (0-1)	MAE (kg ha^{-1})	RMSE (kg ha^{-1})	MAPE (%)
LR	0.4496	1036.7347	1418.3994	58.29%
RF	**0.8848**	**424.4335**	**648.8668**	**19.13%**
SVM	0.0540	1370.8190	1859.4363	82.24%
KNN	0.8568	424.6861	723.5230	18.78%
DT	0.7809	647.3546	894.8645	29.85%

ACCEPTED MANUSCRIPT

XGBoost	0.7800	659.8946	896.6243	35.23%	

Table 3. Model performance across different Feature selection methods before removing outliers

Model	Feature selection method	No.of features	R^2 (0 – 1)	MAE (kg ha^{-1})	RMSE (kg ha^{-1})	MAPE (%)
LR	WMI_SDM	15	**0.4428**	**1043.8166**	**1427.0756**	**58.86%**
	CBS	8	0.4059	1087.7437	1473.5539	63.92%
	RFE	5	0.4098	1092.1618	1468.7116	62.83%
RF	WMI_SDM	15	**0.9433**	**281.4370**	**455.0473**	**12.97%**
	CBS	8	0.8737	392.2563	679.4332	18.91%
	RFE	5	0.8983	346.1991	609.6327	15.11%
SVM	WMI_SDM	15	**0.0621**	**1362.7810**	**1851.4962**	**81.61%**
	CBS	8	0.0875	1338.0578	1826.2969	79.77%
	RFE	5	0.0861	1338.9177	1827.6542	79.26%
KNN	WMI_SDM	15	**0.8935**	**359.6666**	**623.9856**	**15.89%**
	CBS	8	0.8568	411.2501	723.4210	19.42%
	RFE	5	0.9066	384.8964	584.2081	17.37%
DT	WMI_SDM	15	**0.7809**	**647.3546**	**894.8645**	**29.85%**
	CBS	8	0.5475	804.9802	1286.0917	41.12%
	RFE	5	0.7461	682.7632	963.3275	31.27%
XGBoost	WMI_SDM	15	**0.8095**	**624.7687**	**834.4900**	**34.95%**
	CBS	8	0.7577	706.7919	941.0804	39.66%
	RFE	5	0.7790	662.6792	898.7520	35.14%

ACCEPTED MANUSCRIPT

As shown in Figure 2, the R^2 results obtained from various ML models using three feature selection techniques: Proposed WMI_STD, CBS, and RFE are illustrated. Out of all methods, proposed WMI_STD has the upper hand in R^2 values on most models.

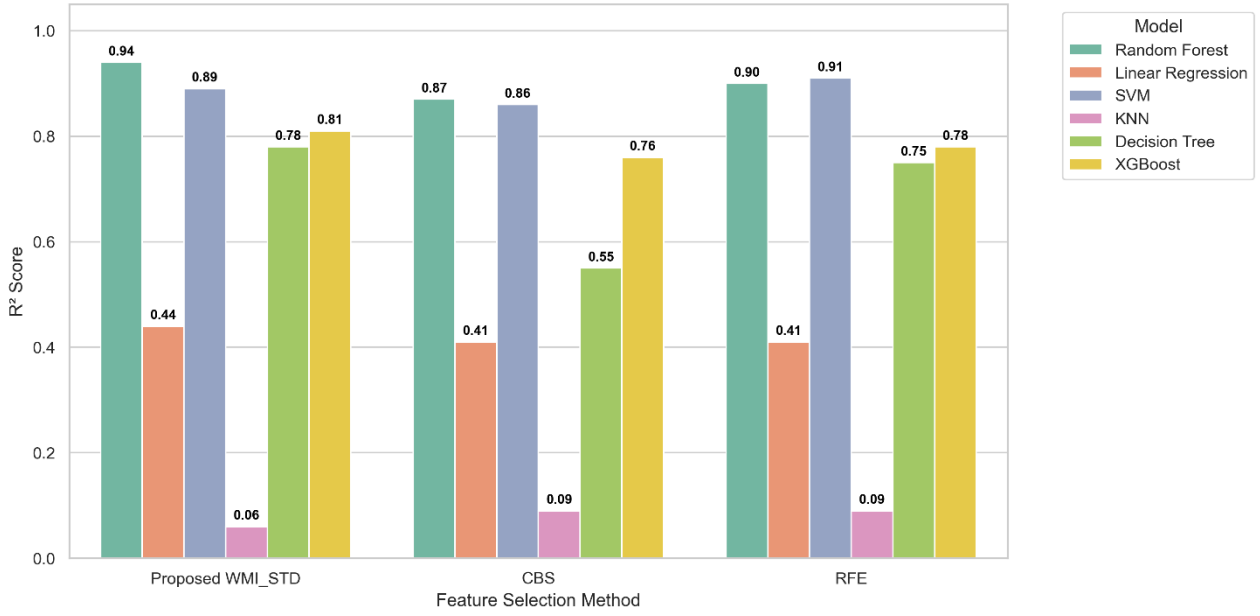


Figure 2. Model performance comparison in terms of R^2 scores across different feature selection methods, showing that the proposed WMI_SDM method achieves superior predictive accuracy compared to CBS and RFE techniques.

Table. 4 shows the performance of regression models was assessed after the proposed WMI_ SDM method and after removal of outliers from the data set. As shown in table, the RF achieved the best overall performance with an R^2 score of 0.9673, a low MAE of 226.7131 kg ha^{-1}, a RMSE of 306.1670 kg ha^{-1} and a MAPE of 11.33%.

Table4. Performance comparison of models with proposed WMI_SDM method after removing outliers

Model	**R^2 (0-1)**	**MAE (kg ha^{-1})**	**RMSE (kg ha^{-1})**	**MAPE (%)**
LR	0.4855	918.8378	1151.6394	56.06%

ACCEPTED MA

RF	**0.9673**	**226.7113**	**306.1670**	**11.33%**
SVM	0.0663	1183.0345	1428.3965	82.45%
KNN	0.9468	297.6387	404.4728	14.82%
DT	0.8118	540.0511	699.6213	26.94%
XGBoost	0.8297	562.9566	684.3553	34.01%

Based on the results in Table 4, the RF model demonstrated the highest predictive performance (R^2 = 0.9673, MAE = 226.71 kg ha^{-1}, RMSE = 306.17 kg ha^{-1}, and MAPE = 11.33%), outperforming other regression models when coupled with the proposed WMI_SDM feature selection method. Consequently, RF was selected for further evaluation using k-fold cross-validation (k = 5 and k = 10) to assess the robustness and reliability of the model's performance. The corresponding results, summarized in Table 5, include mean ± SD values to account for model uncertainty and variability across folds. The proposed WMI_SDM method achieved the highest predictive accuracy, with an R^2 of 0.884 ± 0.088, RMSE of 811.68 ± 419.70 kg ha^{-1}, MAE of 336.60 ± 59.47 kg ha^{-1}, and MAPE of 13.41 ± 3.01% under 10-fold cross-validation. The improvement from 5-fold (R^2 = 0.858 ± 0.071) to 10-fold validation indicates enhanced model stability and better generalization with increased data variability. Compared to CBS and RFE, the proposed approach consistently yielded superior accuracy and lower uncertainty in error metrics, confirming its robustness, consistency, and efficiency in identifying the most informative features for crop yield prediction.

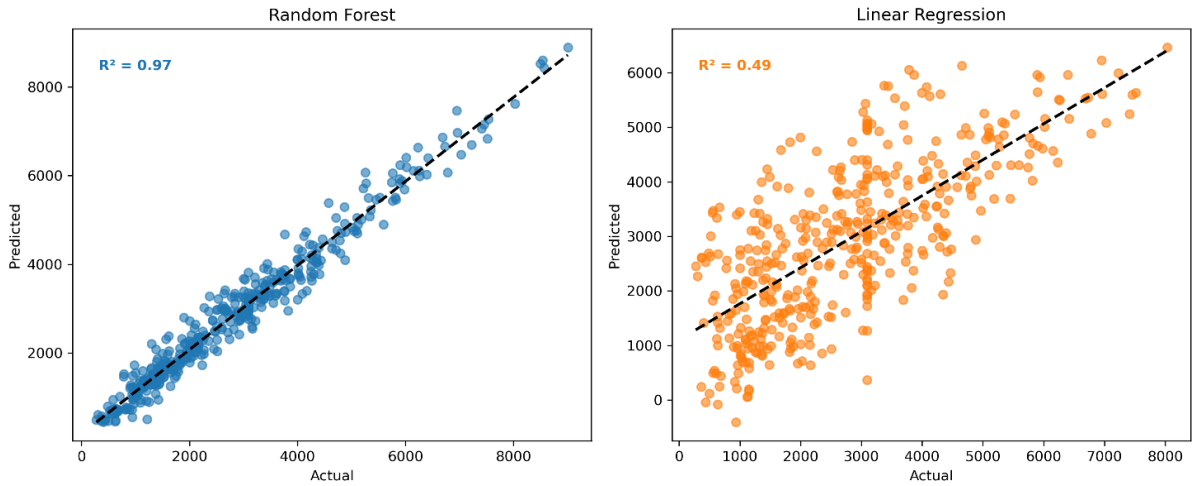
Table5. Performance comparison of the proposed WMI_STD method for Random Forest and baseline feature selection approaches using 5-fold and 10-fold cross-validation.

Model	**Folds**	**R^2 (Mean ± SD)**	**RMSE (kg ha^{-1}) ± SD**	**MAE (kg ha^{-1}) ± SD**	**MAPE (%) ± SD**
WMI_STD	5	0.858 ± 0.071	916.64 ± 23.44	360.18 ± 47.64	14.44 ± 1.98
WMI_STD	**10**	**0.884 ± 0.088**	**811.68 ± 19.70**	**336.60 ± 59.47**	**13.41 ± 3.01**

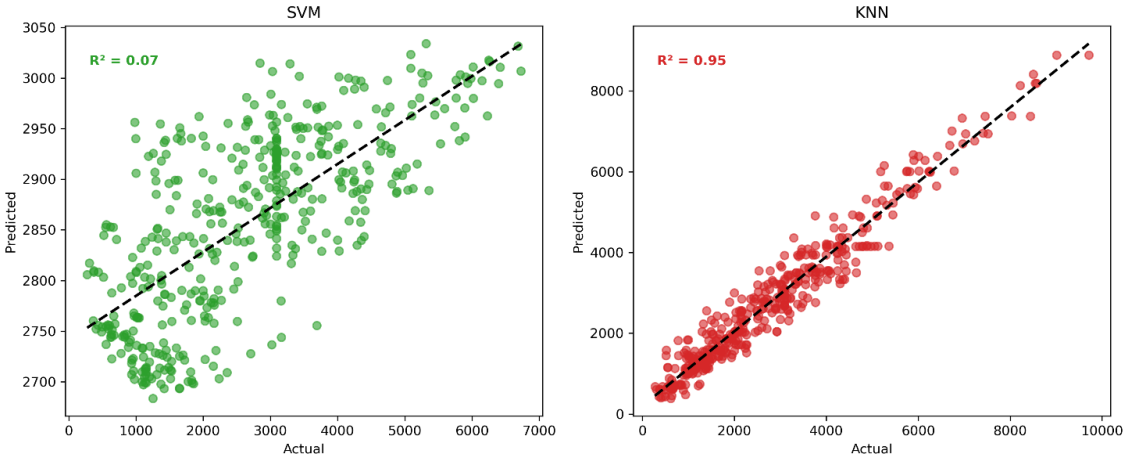
ACCEPTED MANUSCRIPT

CBS	10	0.869 ± 0.103	849.46 ± 36.62	382.68 ± 71.62	16.45 ± 3.62
RFE	10	0.879 ± 0.078	848.88 ± 390.87	387.08 ± 57.19	16.55 ± 3.54

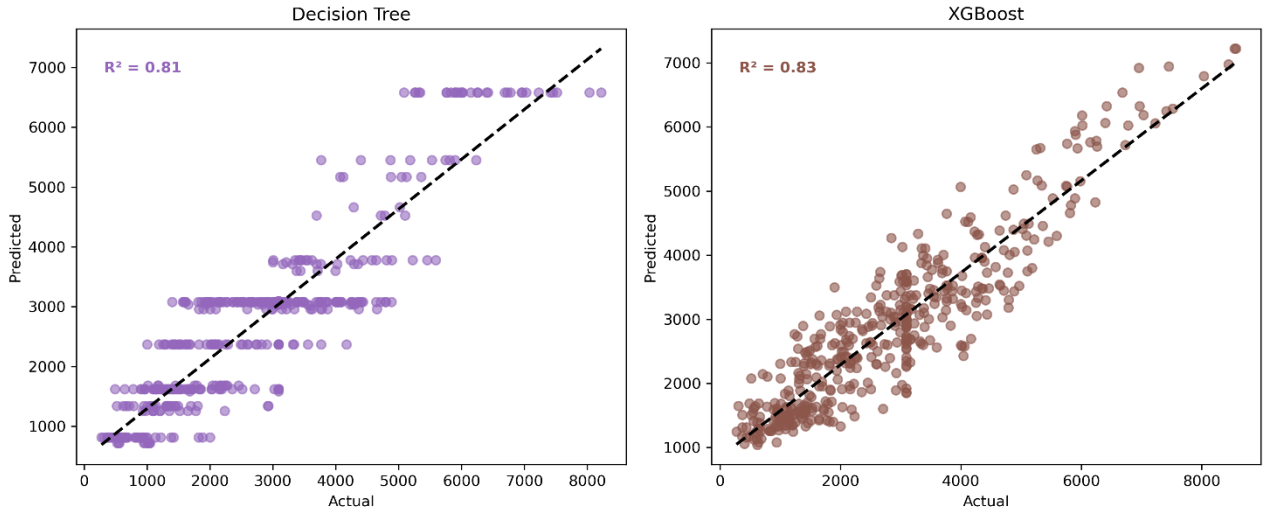
Figure 3 (a) – (f) shows a scatter plot of the relationship between an observed ground truth and predicted crop yield for different models using the proposed WMI_SDM after removing outlier. Each subplot illustrates the performance of different models along with the corresponding R^2 value. The dashed diagonal line represents the ideal fit where predicted values equal actual values. Models like RF ($R^2 = 0.97$) and KNN ($R^2 = 0.95$) demonstrate superior predictive accuracy, while LR ($R^2 = 0.49$) shows relatively lower performance.



(a) (b)



(c) (d)

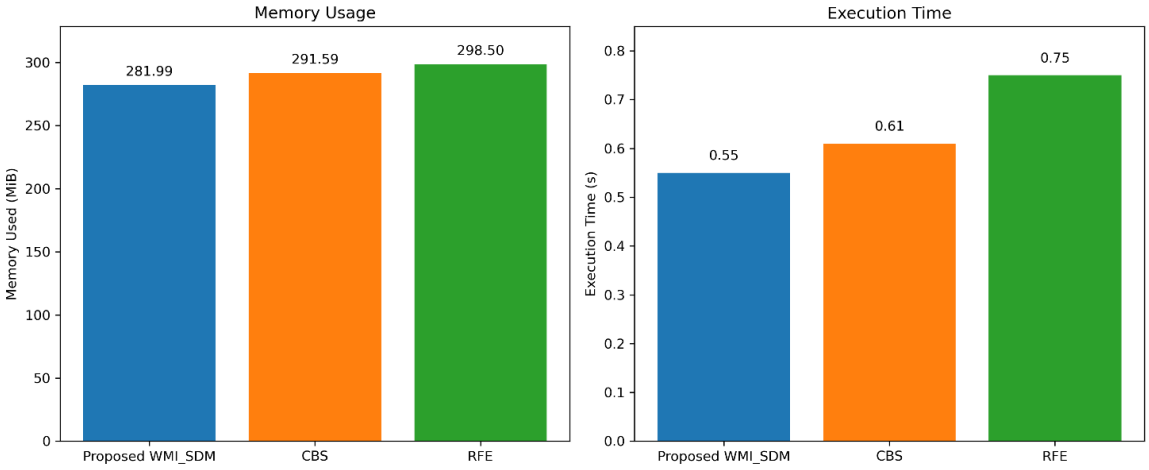


(e) (f)

Figure 3 (a) – (f). Scatter plots showing the relationship between actual and predicted crop yield values for different baseline models (RF, LR, SVM, KNN, DT, and XGBoost) using the proposed WMI_SDM feature selection method after removing outliers. The R^2 scores indicate that Random Forest and KNN achieved the highest predictive accuracy.

Figure 4 (a) & (b) gives a comparative assessment of the memory footprint and execution time for WMI_SDM, CBS, and RFE feature selection techniques. The WMI_SDM technique appears to work the best, needing the least amount of memory (281.99 MiB) and the least amount of execution time (0.55 seconds).

ACCEPTED MANUSCRIPT



(a) (b)

Figure 4 a & b. Comparison of feature selection methods in terms of (a) memory usage and (b) execution time, showing that the proposed WMI_SDM method achieves lower computational cost and faster execution compared to CBS and RFE.

Table 6 presents an overview of the proposed model and its competitive approaches with respect to their feature selection techniques, ML models, and R^2 values. The proposed model has significantly better predictive performance compared to the existing approaches, not only for the predictive performance but also in terms of the robustness of the feature selection method employed. The high R^2 value for the WMI_SDM method confirms the proven ability of the WMI_SDM method to find the most relevant and non-redundant features to predict crop yields in data-driven agro-environmental systems.

Table6. Comparison of proposed and existing methods

Study	Feature selection method	ML model used	R^2
Proposed Method	WMI_SDM	RF	0.9673
Iniyan et al. 2022	Mutual Information Feature Selection (MIFS)	Stacked ensemble	0.9242
Ingio et al. 2024	RFE	Multiple Linear	0.9031

ACCEPTED MANUSCRIPT

		Regression	
Li et al. 2023	Pearson correlation coefficient and random forest importance	RNN	0.6670
Fan et al. 2024	Correlation analysis	RF	0.8501

Overall, the key findings presented in this study highlight the benefits of the proposed WMI_SDM model in terms of computational efficiency and predictive validity, thus making it one of the most attractive solutions for large-scale or resource-intensive agricultural systems. The ability of the framework to handle high-dimensional data and simultaneously reduce the time and memory required to run the framework is a testimony to its functionality in real-time decision-making. However, this research is limited by the fact that it relies on region-specific climatic and emission data, which can reduce its applicability in other geographic settings. To increase generalizability, the framework may be retrained or fine-tuned using external data based on different regions and crops, thus helping to evaluate the strength and flexibility of the framework in out-of-sample conditions. Despite such restrictions, the suggested strategy provides an effective and scalable basis for further use in the framework of working systems of agricultural prediction, where fast and correct yield prediction is one of the key requirements.

6. Conclusion

This paper presents a crop yield prediction model under changing climate conditions based on both climate and emission data. A WMI_SDM technique was proposed to reduce redundancy and noise in large, complex datasets. Extensive experimental analysis was performed to test the robustness and generalizability of the method, such as performance comparisons of regressors without FS, with FS methods prior to outlier removal, and with the proposed WMI_SDM method after outlier removal. The FS methods WMI_SDM, CBS, and RFE were also systematically compared using

ACCEPTED MANUSCRIPT

memory utilization and execution time (speed). The experiments showed that the proposed model can achieve high accuracy predictions with R^2 = 0.9673, MAE = 226.7113 kg ha^{-1}, and MAPE = 11.33%, which is reasonably high compared with other baseline models. FS using the WMI_SDM method was found to consistently outperform traditional approaches, ensuring stable and relevant feature selection in the presence of outliers. Notably, the proposed WMI_SDM method demonstrated superior efficiency, using only 281.99 MiB of memory and executing in just 0.55 s, which is significantly lower than other FS techniques. In this work, EL was used as a base for reliable forecasting of crop yield and the experiments demonstrated that the EL framework and several ML algorithms produced very promising results. Our model ensures the stability, relevance, and efficiency of selected features giving a scalable, data-efficient approach that policymakers and agricultural planners can use to tackle the challenges of climate change and food security.

Conflict of interest

The authors declare that there is no conflict of interest.

References

Abdel-salam, M., Kumar, N., & Mahajan, S. (2024). A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning. *Neural Computing and Applications*, *36*(33), 20723-20750.

Abdulla, A., & Ouki, S. (2015). The potential of wastewater reuse for agricultural irrigation in Libya: Tobruk as a case study. *Global NEST Journal*, 17(2), 357-369.

Anand, P., Singh, S. D., Bhowmik, P. N., & Kontoni, D. P. N. (2025). Optimizing concrete mix proportions with zeolite, GGBS, and CDW: a data-driven approach integrating experimental analysis and machine learning models. *Engineering Research Express*, *7*(1), 015105.

ACCEPTED MANUSCRIPT

Cheng, J., Sun, J., Yao, K., Xu, M., & Cao, Y. (2022). A variable selection method based on mutual information and variance inflation factor. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *268*, 120652.

Fan, L., Fang, S., Fan, J., Wang, Y., Zhan, L., & He, Y. (2024). Rice Yield Estimation Using Machine Learning and Feature Selection in Hilly and Mountainous Chongqing, China. Agriculture, 14(9), 1615.

Fernández-Olmos, M., Fleta-Asín, J., Gómez-Aguas, T., Muñoz, F., & Sáenz-Royo, C. (2025). Improved database of public-private partnerships from World Bank with imputed economic, institutional and conflict data. *Data in Brief*, *60*, 111457.

Fraino, P. E. (2023). Using principal component analysis to explore multi-variable relationships. *Nature Reviews Earth & Environment*, *4*(5), 294-294.

Garai, S., Paul, R. K., Yeasin, M., Roy, H. S., & Paul, A. K. (2024). Machine learning algorithms for predicting rainfall in India. *Curr. Sci*, *126*, 360-367.

Gong, L., Yu, M., Jiang, S., Cutsuridis, V., & Pearson, S. (2021). Deep learning based prediction on greenhouse crop yield combined TCN and RNN. *Sensors*, *21*(13), 4537.

Heydari, A., Mirzaei, N., Pamucar, D., Niroomand, S., & Nowzari, R. (2025). A Feature Selection Approach Based on Information Theory with Application to the International Monetary Fund and World Bank Economic Datasets. *International Journal of Information Technology & Decision Making*, 1-24.

Ingio, J. A., Nsang, A. S., & Iorliam, A. (2024). Optimizing Rice Production Forecasting Through Integrating Multiple Linear Regression with Recursive Feature Elimination. Journal of Future Artificial Intelligence and Technologies, 1(2), 96-108.

ACCEPTED MANUSCRIPT

Iniyan, S., & Jebakumar, R. (2022). Mutual information feature selection (MIFS) based crop yield prediction on corn and soybean crops using multilayer stacked ensemble regression (MSER). Wireless Personal Communications, 126(3), 1935-1964.

Iniyan, S., Varma, V. A., & Naidu, C. T. (2023). Crop yield prediction using machine learning techniques. *Advances in Engineering Software*, *175*, 103326.

Kamangir, H., Sams, B., Dokoozlian, N., Sanchez, L., & Earles, J. M. (2024). CMAViT: Integrating Climate, Managment, and Remote Sensing Data for Crop Yield Estimation with Multimodel Vision Transformers. *arXiv preprint arXiv:2411.16989*.

Karthikeyan, B., Mohan, V., Chamundeeswari, G., & Ruba, M. (2023). Deep learning driven crop classification and chlorophyll content estimation for the Nexus food higher productions using multispectral remote sensing images. *Global NEST Journal*, 25(3), 164-173.

Li, Z., Zhou, X., Cheng, Q., Zhai, W., Mao, B., Li, Y., & Chen, Z. (2023). An integrated feature selection approach to high water stress yield prediction. Frontiers in Plant Science, 14, 1289692.

Lokeshwari, M., Jha, G. K., Praveen, K. V., & Bharadwaj, A. (2024). Artificial intelligence for crop yield prediction: a bibliometric analysis. *Current Science (00113891)*, *126*(10).

Panigrahi, B., Kathala, K. C. R., & Sujatha, M. (2023). A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models. *Procedia Computer Science*, *218*, 2684-2693.

Richa, A., Douaoui, A., Bettahar, N., Qiang, Z., & Mailhol, J. C. (2015). Assessment and modeling the influence of nitrogen input in the soil on groundwater nitrate pollution: plain of upper-cheliff (north Algeria). *Global Nest Journal*, 17(4), 744-755.

Sah, G., Banerjee, S., & Dutta, M. P. (2022). Ensemble learning algorithms with feature reduction mechanism for intrusion detection system. *International Journal of Information and Computer Security*, *19*(1-2), 88-117.

ACCEPTED MANUSCRIPT

Saleh Al-rimy, B. A., Aizaini Maarof, M., & Shaid, S. Z. M. (2018). Redundancy Coefficient Gradual Up-weighting-based Mutual Information Feature Selection Technique for Crypto-ransomware Early Detection. *arXiv e-prints*, arXiv-1807.

Sharafi, S., Kazemi, A., & Amiri, Z. (2023). Estimating energy consumption and GHG emissions in crop production: A machine learning approach. *Journal of Cleaner Production*, *408*, 137242.

Singh, A. K., Yeasin, M., Paul, R. K., Roy, H. S., Kumar, P., Paul, A. K., & Sarkar, A. (2025). Optimisation-based weighted ensemble algorithm for predicting prices of spices. *Current Science (00113891)*, *128*(8).

Varma, M., Lama, A., Singh, K. N., & Gurung, B. (2023). Evaluating the performance of crop yield forecasting models coupled with feature selection in regression framework. *Curr Sci*, *125*(6), 649.

Yewle, A. D., Mirzayeva, L., & Karakuş, O. (2025). Multi-modal Data Fusion and Deep Ensemble Learning for Accurate Crop Yield Prediction. *arXiv preprint arXiv:2502.06062*.

Zhou, H., Wang, X., & Zhu, R. (2022). Feature selection based on mutual information with correlation coefficient. *Applied intelligence*, *52*(5), 5457-5474.

ACCEPTED MANUSCRIPT

List of tables

Table number	Title
Table 1	Data set description
Table 2	Model performance comparison without Feature selection
Table 3	Model performance across different Feature selection methods before removing outliers
Table 4	Performance comparison of models with proposed WMI_SDM method after removing outliers
Table 5	Performance comparison of the proposed WMI_STD method for Random Forest and baseline feature selection approaches using 5-fold and 10-fold cross-validation.
Table 6	Comparison of proposed and existing methods

ACCEPTED MANUSCRIPT

Figure captions

Figure number	Caption
Figure 1	Overall workflow of proposed crop yield model using WMI_SDM
Figure 2	Model R^2 scores across feature selection methods
Figure 3 (a) – (f)	Scatter plots showing the relationship between actual and predicted crop yield values for different baseline models (RF, LR, SVM, KNN, DT, and XGBoost) using the proposed WMI_SDM feature selection method after removing outliers. The R^2 scores indicate that Random Forest and KNN achieved the highest predictive accuracy.
Figure 4 a & b	Comparison of feature selection methods in terms of (a) memory usage and (b) execution time, showing that the proposed WMI_SDM method achieves lower computational cost and faster execution compared to CBS and RFE.

ACCEPTED MANUSCRIPT