Anusandhanvallari
Vol 2026, No.1
July 2026
ISSN 2229-3388

Data-Driven Crop Yield Forecasting Using Regression, Classification Methods: A Systematic Approach

Dr. S. Anusuya

MCA, M. Phil., Ph.D., Assistant Professor, Department of Computer Science, Sri Vasavi College, Self Finance Wing, Erode

Abstract

Accurate prediction of Crop Yield Forecasting (CYF) is essential for improving Agricultural productivity, optimizing resource utilization, and supporting data-driven decision-making in precision farming. Traditional yield estimation methods often fail to capture the complex nonlinear relationships among environmental, soil, and cultivation factors. This study proposes CYF-CSVR, CYF-WKNN a Machine Learning-based Paddy Yield Forecasting and Classification model to enhance prediction accuracy. The first analysis is on predicting the yield and the second analysis is to do classification on the attribute 'Yield Status' -a binomial attribute which is created at the time of implementation. The experimental dataset was obtained from the UCI Machine Learning Repository Paddy dataset, which contains 45 agricultural attributes related to soil properties, weather conditions, crop management, and growth parameters. The dataset was preprocessed through missing value handling, normalization, and feature selection to improve model efficiency. Performance evaluation was conducted using standard regression metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Error(RE), and Correlation (r). In addition for the method CYF-WKNN, the performance is assessed with the classification metrics, Accuracy and Kappa Statistics. The results from Rapidminer demonstrate that the proposed model CYF-CSVR and CYF-WKNN effectively gives strong generalization capability. The proposed approach contributes to precision agriculture by enabling early yield forecasting, better crop management strategies, and ultimately supporting sustainable paddy cultivation.

Keywords – CrayFish Optimization, Feature Selection, Support Vector Regression, Weighted KNN.

I. Introduction

Agriculture crop yield forecasting is the process of estimating the expected production of crops before harvest using historical and current agricultural data. [1] It plays a crucial role in improving farming efficiency, resource management, and food security. Yield forecasting considers multiple factors such as soil quality, rainfall, temperature, humidity, irrigation, fertilizer usage, and pest conditions. Traditional forecasting methods often rely on manual observations, whereas modern approaches use machine learning and regression analysis to identify patterns and predict yield more accurately. Accurate crop yield forecasting helps farmers, researchers, and policymakers make better decisions for sustainable agriculture and production planning.

Agriculture Crop Prediction Using Machine Learning

Agricultural crop prediction is a data-driven process that estimates future crop production based on historical and real-time farming data.[2] The increasing availability of agricultural datasets from weather stations, soil sensors, drones, satellite imagery, and IoT devices has significantly enhanced predictive agriculture. Crop prediction involves collecting relevant attributes such as soil pH, nitrogen, phosphorus, potassium, rainfall, humidity, temperature, sunlight exposure, and irrigation frequency, and then applying machine learning models to analyze their influence on crop growth. Machine learning improves agricultural forecasting by identifying hidden patterns and nonlinear dependencies that traditional statistical methods may overlook. Accurate crop prediction helps optimize resource utilization, reduce production risks, improve food supply planning, and increase economic returns for farmers. In paddy cultivation, for example, predicting yield before harvest enables better supply chain management and government policy planning. Machine learning-based crop prediction supports sustainable agriculture by reducing wastage of water, fertilizer, and labor while maximizing productivity.

Role of Various Machine Learning Algorithms in Crop Prediction

[3]Different machine learning algorithms play different roles in crop yield prediction depending on the data characteristics, complexity, and prediction requirements. Some algorithms are simple and interpretable, while others provide higher accuracy for nonlinear datasets.

Table I. Role of Various Machine Learning Algorithms in Crop Prediction

Algorithm	Type	Role in Crop Prediction	Advantages
Linear Regression	Regression	Models linear relationships between agricultural features and yield	Simple, interpretable
Support Vector Regression	Regression	Handles nonlinear crop-environment relationships	High accuracy, robust
Decision Tree	Regression/Classification	Splits agricultural data into rule-based structures	Easy interpretation
Random Forest	Ensemble Regression	Uses multiple trees for prediction	Reduces overfitting
XGBoost	Ensemble Regression	Optimized boosting for high-dimensional agricultural data	Fast and accurate
K-Nearest Neighbors	Regression/Classification	Predicts yield based on similar past farms	Simple and effective
Artificial Neural Network	Deep Learning	Learns complex patterns in crop growth	Handles nonlinear data well
Long Short-Term Memory	Deep Learning	Predicts yield using time-series agricultural data	Good for seasonal patterns
Convolutional Neural Network	Deep Learning	Uses satellite or image-based crop monitoring	Strong image analysis

Applications of Research Analysis in Agriculture

1. Crop Yield Prediction Analysis – [4]Crop yield prediction analysis focuses on estimating the expected paddy production using agricultural parameters such as soil nutrients, rainfall, temperature, humidity, irrigation, and fertilizer usage. This analysis helps identify the relationship between input features and yield output. Regression algorithms like Support Vector Regression and Random Forest can accurately model these relationships. The outcome supports farmers in forecasting production before harvest.

2. Soil Fertility Impact Analysis - Soil fertility impact analysis examines how essential nutrients such as nitrogen, phosphorus, potassium, and soil pH influence paddy growth and yield. This helps in understanding nutrient deficiencies and their effect on productivity. Using correlation and regression techniques, researchers can determine the most influential soil properties. This analysis supports better fertilizer planning and soil health management.

3. Climate Influence Analysis - Climate influence analysis studies the effect of environmental conditions such as rainfall, temperature, humidity, and sunlight on crop yield. Since climate directly affects plant growth, analyzing these factors helps identify favorable and unfavorable conditions for paddy cultivation. Machine learning models can reveal hidden weather patterns that impact production. This analysis improves climate-aware farming decisions.

4. Irrigation Efficiency Analysis - Irrigation efficiency analysis evaluates how water management affects crop productivity. It helps determine whether the current irrigation levels are sufficient, excessive, or inadequate for achieving optimal yield. Regression models can identify the relationship between water supply and crop output. This analysis helps conserve water while maintaining productivity.

5. Fertilizer Optimization Analysis – [5]Fertilizer optimization analysis aims to identify the ideal amount of fertilizer required for maximum crop yield. Excessive fertilizer use increases cost and may harm as well as insufficient reduces productivity. By analyzing nutrient application and yield patterns, machine learning can recommend optimal fertilizer strategies. This supports cost-effective and sustainable agriculture.

6. Pest and Disease Impact Analysis - This analysis investigates how pest attacks and plant diseases affect paddy yield. It helps estimate the extent of crop loss caused by biological stress factors. Machine learning models can detect patterns between pest severity and yield reduction. This supports early intervention and better pest management strategies.

7. Harvest Time Prediction Analysis - Harvest time prediction analysis estimates the number of days required for crop maturity based on growth conditions and environmental factors. Accurate harvest prediction helps farmers plan labor, storage, and transportation efficiently. Regression models can predict the ideal harvesting period. This minimizes post-harvest losses and improves planning.

8. Region-wise Productivity Analysis - Region-wise productivity analysis compares crop yield across different geographical regions or farming zones. It helps identify high-yield and low-yield areas based on environmental and soil variations. This analysis can reveal location-specific farming challenges. It supports regional agricultural planning and policy decisions.

9. Sustainable Agriculture Analysis - Sustainable agriculture analysis focuses on achieving higher crop yield with minimal use of resources such as water, fertilizer, and energy. It studies efficient farming practices that reduce environmental impact while maintaining productivity. Machine learning helps identify balanced input strategies. This promotes long-term agricultural sustainability.

10. Explainable Agriculture Analysis - Explainable agriculture analysis combines prediction models with interpretability methods like SHAP and LIME. It helps explain why a model predicts a particular yield by showing feature contributions. This improves transparency and trust in machine learning systems. Farmers and researchers can better understand the factors affecting crop productivity.

Significance of Regression, Classification Analysis in Crop Yield Forecasting

Regression, Classification analysis is highly significant in crop yield forecasting because it provides quantitative estimates of agricultural production before harvest. This early prediction capability allows farmers to take preventive actions against possible yield reduction. For example, if Regression, Classification analysis predicts low yield due to insufficient rainfall, farmers can improve irrigation planning. [6]Similarly, if soil nutrient deficiency is identified as a major yield-reducing factor, fertilizer strategies can be adjusted. Regression, Classification models also support government agencies and policymakers by helping estimate food production, manage storage facilities, and regulate market prices. In precision agriculture, Regression, Classification serves as the core predictive engine for smart farming applications. It helps optimize resource allocation, reduce operational costs, improve sustainability, and enhance food security. Advanced Regression, Classification methods combined with explainable AI techniques like SHAP and LIME can further reveal how different agricultural factors contribute to yield, increasing trust, and transparency in predictive systems.

Significance of Optimization Algorithm in Crop Yield Forecasting

The significance of optimization algorithms in machine learning lies in their ability to improve model performance by finding the best possible parameter values. In machine learning, the accuracy of a model depends not only on the algorithm itself but also on the proper selection of hyper parameters such as learning rate, regularization factors, kernel parameters, and error thresholds. Manual tuning of these parameters is time-consuming and may not guarantee the best solution. [7]Optimization algorithms automate this process by searching the solution space efficiently to identify the optimal parameter combinations that minimize error and maximize prediction accuracy. In agricultural, data is often nonlinear, noisy, and high-dimensional, and so here optimization play a crucial role. For example, when combined with Support Vector Regression, optimization techniques can tune parameters like **C**, **γ**, and **ε** to achieve better generalization and lower prediction errors. Algorithms such as Particle Swarm Optimization, Genetic Algorithm, and Crayfish Optimization Algorithm improve convergence speed, avoid local optima, and enhance computational efficiency. Therefore, optimization algorithms are significant because they strengthen machine learning models by improving accuracy, stability, and robustness, making them highly effective for complex predictive tasks such as agricultural crop yield estimation.

Problem Statement: Due to the nature of linearity, the Regression, and Classification model was failed to provide the précised prediction in a situation such as extreme value data and nonlinear data.

Objective: The primary objective of this research is to develop an effective Crop Yield Forecasting system using Optimized Support Vector Regression and Weighted Classification to enable farmers in getting informed decisions so as to improve agricultural productivity.

II. Literature Review

Thomas van Klompenburg *et al* [8] conducted to identify and summarize the algorithms and features commonly applied in crop yield prediction research. Based on predefined search criteria, a total of 567 studies were collected from six major electronic databases. The analysis Artificial Neural Network emerged as the most commonly adopted algorithm in machine learning-based crop yield prediction. From this analysis, Convolutional Neural Network (CNN) was identified as the most widely implemented deep learning technique, followed by Long Short-Term Memory (LSTM) and Deep Neural Networks (DNN).

Muthukumaran *et al* [9] proposed a Hybrid Machine Learning Model with Combined Wrapper Feature Selection Techniques **.** Five feature selection techniques were employed. This reconstructed dataset was then utilized to train machine learning models for paddy yield prediction. In addition, the extracted knowledge was represented through association rules. Performance evaluation results indicate that the SVM and integrated feature selection strategy significantly improves the identification.

Rajaswee Surana *et al* [10] examined the essential factors required for accurate crop yield prediction by applying various machine learning techniques. To improve prediction performance, algorithms such as Random Forest, AdaBoost, Gradient Boosting, and Support Vector Machine were utilized. The agricultural dataset used for experimentation consisted of 2201 instances, where 80% of the data was allocated for training and the remaining 20% for testing the predictive capability of the models. The comparative analysis of these algorithms showed that the Random Forest model achieved the highest prediction accuracy among all methods.

Vijay H. Kalmani *et al* [11] introduced a hybrid deep learning framework for improving crop yield prediction by combining CNN with LSTM, and an attention mechanism. The proposed model is specifically designed for two of the most significant crops cultivated in India. To enhance predictive performance, the hybrid CNN-LSTM architecture incorporates advanced modifications such as multi-head attention and multiplication-based skip connections. Experimental results demonstrate that the proposed model outperforms traditional regression techniques and achieved superior prediction accuracy with a high coefficient of determination (R^2) of 0.967.

Swati *et al* [12] offered an integrated and intelligent platform that connects conventional agricultural practices. In this framework, Random Forest is utilized as the core model for crop suitability analysis and yield forecasting. For plant disease detection, a CNN is employed to analyze crop images and identify diseases effectively. The system further enhances decision-making by integrating external data sources obtained through the OpenWeather API and contextual agricultural recommendations. By combining reliable machine learning models with live environmental data, the system provides a comprehensive smart agriculture solution that improves productivity.

Manjula *et al* [13] employed association rule mining to discover meaningful relationships among agricultural factors influencing yield. The collected data undergoes preprocessing to remove inconsistencies. The cleaned dataset is then grouped using the K-means clustering algorithm to organize similar data patterns into clusters. Following clustering, association rule mining is applied within each cluster to extract significant rules related to crop yield. The training process concludes with the generation of a set of association rules that form the basis for prediction. In the testing phase, these learned rules are used to estimate yield values for new input data. Experimental analysis identified 151 unique frequent patterns across all clusters, which further generated 193 distinct association rules using a minimum support threshold of 0.3 and a minimum confidence level of 0.7.

Jayanarayana Reddy *et al* [14] reviewed various machine learning techniques applied in crop yield prediction and provides a comparative analysis. Generally, the collected data undergoes preprocessing to eliminate noise, inconsistencies, and missing values, ensuring better data quality for analysis. After preprocessing, feature extraction is performed and these extracted features are then used as input for machine learning models to perform yield prediction. Existing crop yield prediction models have widely employed techniques such as Artificial Neural Network, Random Forest, and K-Nearest Neighbors regression for improving prediction performance. Additionally, advanced deep learning models such as CNN, Long Short-Term Memory LSTM, and Deep Neural Networks DNN have become increasingly common in crop yield prediction studies.

Mahmoud Abdel-salam *et al* [15] presented a novel framework for crop yield prediction designed that combines an advanced hybrid feature selection strategy with an optimized Support Vector Regression (SVR) model. The framework is organized into three major phases: preprocessing, hybrid feature selection, and prediction. In the preprocessing phase, the agricultural data is normalized to improve consistency, followed by the application of K-means clustering along with Correlation-based Feature Selection (CFS) to create a reduced and more relevant dataset. In the final prediction phase, an enhanced optimization technique called improved Crayfish Optimization Algorithm is proposed to fine-tune the hyper-parameters. This optimization process significantly improves the prediction capability of the model while reducing computational complexity.

Iii. Research Methodology

Proposed Methodology – Crop Yield Forecasting using Crayfish Optimized Support Vector Regression CYF-CSVR, Weighted KNN Classification CYF-WKNN

Anusandhanvallari
Vol 2026, No.1
July 2026
ISSN 2229-3388

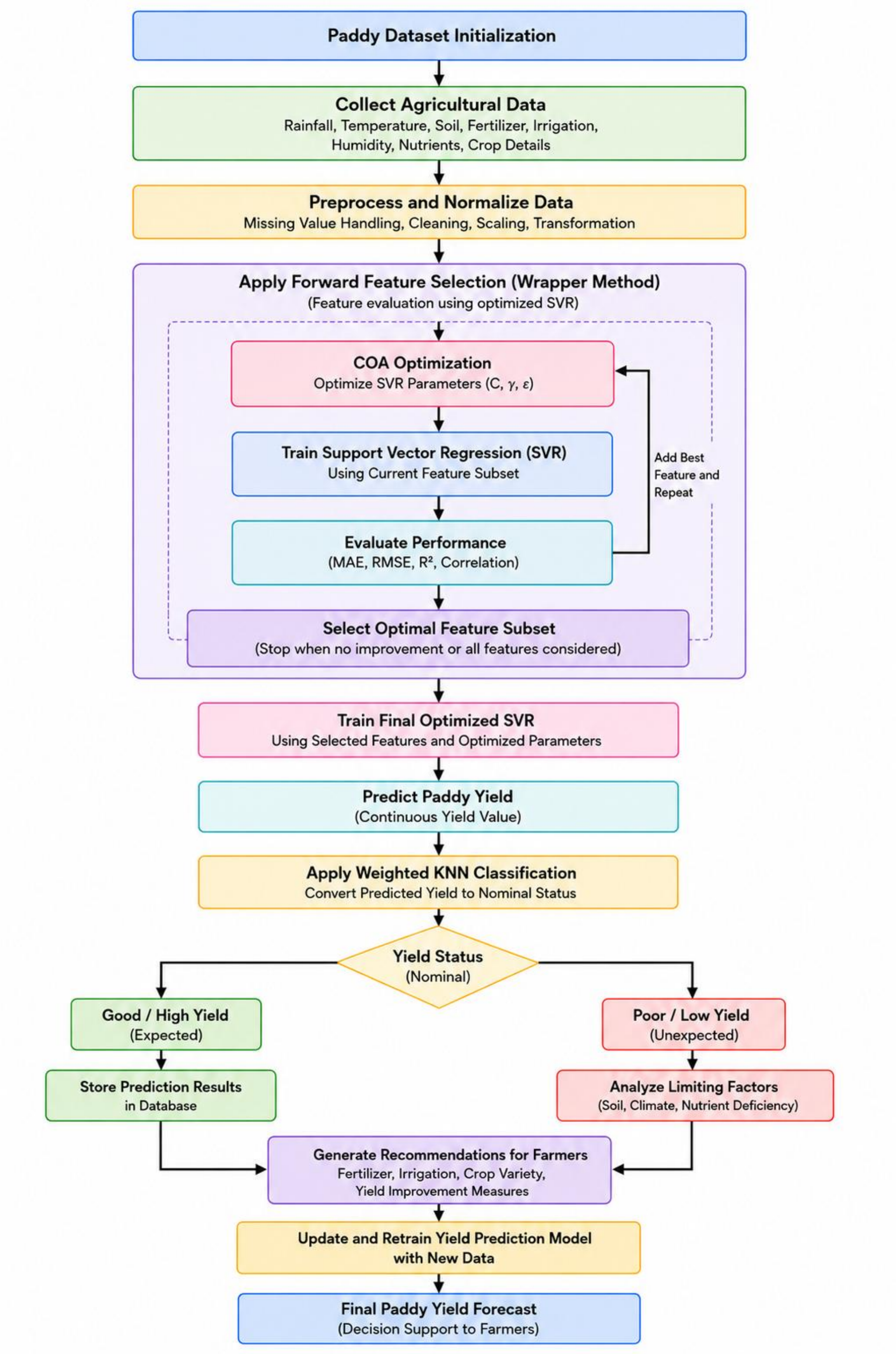

Figure 1. Flow of Work CYF-CSVR, CYF-WKNN

Preprocessing

The missing values replaced with the mean values, the normalization is done with Min Max normalization, and the feature selection is done using Wrapper Forward Feature selection method.

Min-Max Normalization-$X_{norm} = \frac{X - X_{min}}{X - X_{max}}$ $\quad - (1)$

Where, X = original value, X_{min} = minimum value of the attribute, X_{max} = maximum value of the attribute, X_{norm} = normalized value.

Feature Selection –Forward Feature Selection (FFS) is a wrapper-based feature selection technique used to identify the most relevant features for a machine learning model. When combined with Support Vector Regression (SVR), it helps improve prediction accuracy by selecting only the important agricultural attributes. Forward selection starts with an empty feature set and adds one feature at a time based on model performance.

Regression Analysis CYF-CSVR

The method expands as 'Crop Yield Forecasting-CrayFish with Support Vector Regression.

Crayfish Optimization Algorithm

Crayfish Optimization Algorithm is a recent nature-inspired metaheuristic optimization technique developed based on the food-searching, foraging, and survival behavior of crayfish. It is designed to solve complex optimization problems by balancing exploration (searching globally) and exploitation (refining locally). COA is particularly effective in machine learning for optimizing model parameters because it can efficiently search large solution spaces and avoid getting trapped in local optima. In machine learning applications such as crop yield prediction, COA is mainly used to optimize hyper parameters of models like Support Vector Regression, thereby improving prediction accuracy and reducing error. COA is a recent bio-inspired Meta-Heuristic optimization technique based on the food-searching and survival behavior of crayfish. It is used to optimize the hyper parameters of SVR by searching for the best combination of C, γ, and ε. In the proposed hybrid framework, COA is applied after preprocessing and feature selection but before model training. The workflow is:

Input Data → Preprocessing → Feature Selection → COA Optimization → SVR Training → Yield Prediction

The fitness function used by COA is generally the prediction error:

Fitness = RMSE (or) Fitness = MAE – (2)

COA iteratively updates candidate solutions to minimize the error:

Best(C,γ,ϵ) = Min(Error) –(3)

Working Principle – It has three stages namely Summer Resort Behavior, Competition Behavior, and Foraging Behavior.

Summer Resort – It represents global search and explores different solution. It increases diversity in search.

$$X_i^{t+1} = X_i^t + r_1(X_{best}^t - X_i^t) + r_2(X_j^t - X_k^t) \quad - (4)$$

Where, X_i^t = current solution, X_{best}^t = best solution, X_j^t, X_k^t = random solutions, r_1,r_2 = random coefficients

Competition Behavior - Focuses on improving nearby good solutions and fine-tunes the solution.

$$X_i^{t+1} = X_{best}^t + \beta(X_i^t - X_{best}^t) \quad - (5)$$

Foraging Behavior – Moves closer to optimum solution.

$$X_i^{t+1} = X_i^t + \alpha(X_{Optimum} - X_i^t) \quad - (6)$$

In paddy yield prediction, this hybrid COA-SVR framework improves model accuracy by automatically tuning the SVR parameters. COA determines the optimal parameter set for SVR. These results in lower prediction error and higher generalization compared to standard SVR.

Thus, the integration of SVR with COA provides an effective and optimized approach for agricultural crop yield forecasting, particularly in complex and nonlinear paddy datasets.

Support Vector Regression (SVR)

Support Vector Regression is a powerful supervised learning method used for predicting continuous values such as crop yield. It is an extension of Support Vector Machine (SVM) designed for regression tasks. The main objective of SVR is to identify an optimal regression function that predicts output values with minimum deviation while maintaining maximum generalization ability. Unlike traditional regression methods, SVR does not attempt to minimize all errors directly; instead, it introduces an **epsilon-insensitive boundary**, where small errors within the threshold are ignored. This makes SVR more robust to noise and outliers, which are common in agricultural datasets.

The regression function of SVR is represented as:

$$f(x) = w\varphi(x) + B - (1) \qquad -(7)$$

where: *x* represents the input features (rainfall, temperature, soil nutrients, fertilizer, etc.), w denotes the weight vector , *b* is the bias term, *φ(x)* is the kernel transformation function

The optimization objective of SVR is:

$$\text{Minimize } \frac{1}{2}\left|\, \middle| w|\phi|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)\right) \qquad -(8)$$

subject to:

$$y_i - (w\phi(x_i) + b) \leq \epsilon + \xi_i \qquad -(9)$$

Where,

C = penalty parameter, *ε* = error tolerance, *ξ, ξ** = slack variables.

The performance of SVR depends heavily on its hyperparameters such as C, γ (gamma), and **ε**. Improper selection of these parameters can reduce prediction accuracy and increase computational complexity. To address this challenge, the Crayfish Optimization Algorithm (COA) is integrated with SVR.

Prediction to Classification using CYF-WKNN

An attribute created in the name 'Status' .To convert the predicted yield (continuous value) into a nominal status like Good or Poor, you need an if–else condition based on a threshold value. The status attribute is classified to get the overall yield status using weighted KNN(K=2).

If (Predicted Yield ≥ Threshold) **Good Yield** else **Poor Yield** – (10)

Weighted KNN - Weighted KNN (WKNN)**,** closer neighbors are given higher importance than distant neighbors. This improves classification accuracy, especially when data points are unevenly distributed. WKNN is used after SVR prediction to convert the predicted continuous yield into nominal classes such as Good or Poor.

Crop Yield Forecasting using Weighted K-Nearest Neighbor (CYF-WKNN)

Anusandhanvallari
Vol 2026, No.1
July 2026
ISSN 2229-3388

Class = Weighted Vote (K), mi - min value, ma - max value.

Normalized Manhattan Distance (with Normalized Features)– Range between 0-1 $D_{norm}(a,b) = \frac{1}{n}\sum_{i=1}^{n}\frac{|a_i - b_i|}{ma_i - mi_i}$ $-(11)$

Algorithm CYF-CSVR, CYF-WKNN

Input: Paddy dataset *D* with agricultural attributes (Rainfall, Temperature, Soil Nutrients, Fertilizer, Irrigation, Humidity, etc.)

Output: Predicted yield, Yield status (Good/Poor)

Step 1: Initialize paddy dataset D

Step 2: Collect agricultural parameters from dataset

Step 3: Preprocess the dataset

Step 4: Regression Analysis CYF-CSVR

Step 5: Apply Wrapper Feature Selection

5.1 For each unselected Feature

5.2 Temporarily add the feature to the subset

5.3 Initialize COA population

5.4 Evaluate fitness of each solution

5.5 Update crayfish positions

5.6 Train SVR using optimized parameters

5.7 Select the feature that gives best performance

5.8 Add the selected feature to the permanent subset

5.9 Repeat step 5.1 – 5.8 until no further improvements occurs

Step 6: Classification Analysis CYF-WKNN

Step 7: Status of the paddy yield is calculated by using (10) to create 2 binomial class: Yield Good/Poor.

Step 8: Apply Weighted KNN.

Step 9: Evaluate model performance

Flow sequence:

Crop Parameters → Preprocessing → Feature Selection → COA Optimization → SVR Training → Performance Evaluation → Yield Prediction → Yield Classification

Advantages:

Improved Data Quality Reduced Dimensionality, Optimized Parameter Tuning, Higher Prediction Accuracy, Better Generalization and Support System.

IV. Reults And Discussion

Dataset Description – The Agriculture related Paddy dataset (https:// www. kaggle. com/ datasets/ mdnaimislam165436/paddy-dataset) is taken from Kaggle repository. The dataset has 45 attributes and 2789 instances which has the details about paddy yield. Regression and classification analysis is done on the dataset to get the prediction of paddy yield.

Feature Selection - Among 45 Features, 12 are selected by the Wrapper Forward Feature Selection method.

Regression Analysis Performance Measures - Four evaluation measures are used for Regression analysis namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Error (RE), Correlation (r).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \qquad -(12)$$

$$\mathrm{RMSE} = \mathrm{SQRT}(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad -(13)$$

$$\mathrm{RE} = \frac{\mathrm{Actual} - \mathrm{Predicted}}{\mathrm{Actual}} \qquad -(14)$$

$$r = \frac{(x_i - \overline{x}) * (y_i - \overline{y})}{\mathrm{Sqrt}(x_i - \overline{x})^2 * (x_i - \overline{x})^2} \qquad -(15)$$

Table 2, shows the performance of Regression analysis on the methods namely SVR, and CYF-CSVR. The proposed method outperforms the existing method by having less error values and high correlation values.

Table II. Regression Analysis Performance of CYF-CSVR

Model	MAE	RMSE	Correlation	Relative Error
SVR	0.61	0.84	0.72	0.68
CYF-CSVR	0.34	0.51	0.91	0.41

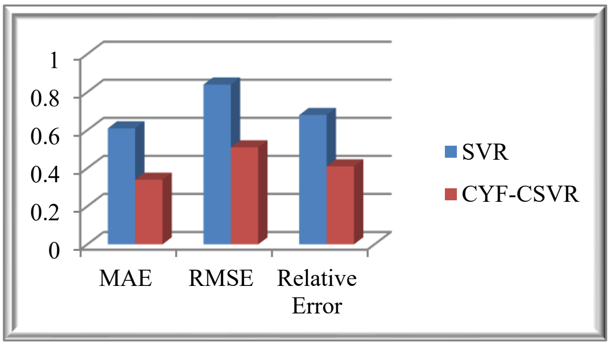

Figure 2. MAE, RMSE, RE for Regression Analysis Performance of CYF-CSVR

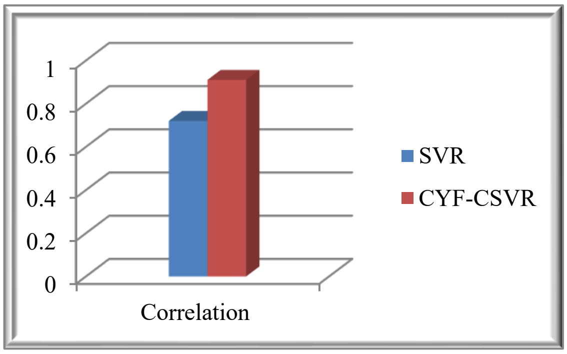

Figure 3. Correlation for Regression Analysis Performance of CYF-CSVR

Anusandhanvallari
Vol 2026, No.1
July 2026
ISSN 2229-3388

Figure 2, shows the error values for the method SVR, and CYF-CSVR. Among them, MAE decreased from 0.061 → 0.034, showing lower average prediction error. RMSE reduced from 0.084 → 0.051, indicating fewer large deviations. Correlation improved from 0.912 → 0.964, showing stronger agreement between actual and predicted yield. Figure 3, shows the Relative Error dropped from 0.078 → 0.041, proving better overall model efficiency.

Classification Analysis Performance Measures - Two most significant performance measures Accuracy, and Kappa Statistics is calculated and compared for proposed and other benchmark methods. Table 3, shows the performance analysis of CYF-WKNN with the benchmark data mining algorithms. Among them, the proposed outperforms the existing methods by having high values for Accuracy and Kappa Statistics.

Table III. Classification Analysis Performance of CYF-WKNN

enchmark Methods	Accuracy (in %)	Kappa Statistics (in range 0-1)
SVM	84	0.78
RF	91	0.85
CNN	90	0.84
FFNN	93	0.87
LSTM	95	0.89
SVR	87	0.81
CYF-WKNN	98	0.96

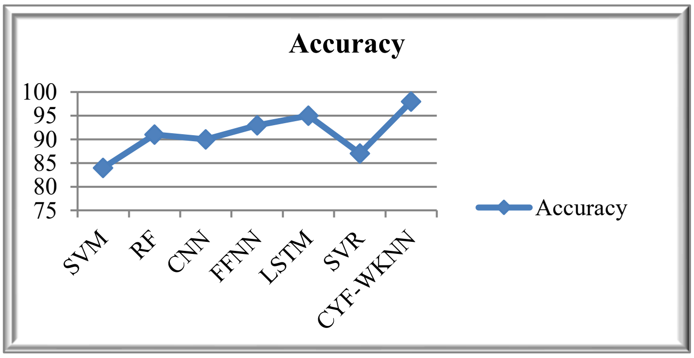

Figure 4. Accuracy for Classification Analysis Performance CYF-WKNN

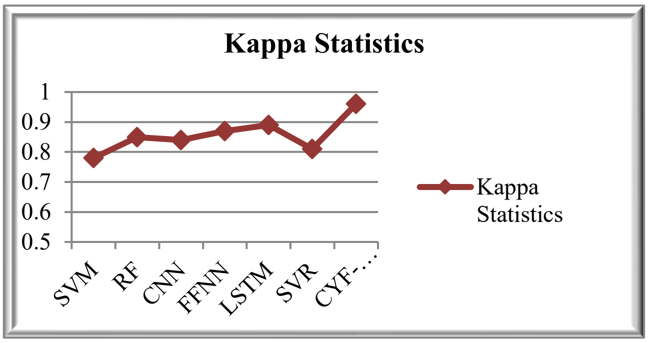

Figure 5. Kappa Statistics for Classification Analysis Performance CYF-WKNN

Figure 4, and Figure 5 shows the Accuracy and Kappa Statistics for the proposed method CYF-WKNN that is used for the classification of Yield status. The proposed method gives high Accuracy, and Kappa Statistics.

V. Conclusion

This research paper did an analysis on Crop Yield Forecasting (CYF) for the prediction of Paddy. The Regression analysis method CYF-CSVR used significant preprocessing, Support Vector Regression (SVR) with Optimization method CrayFish (C). This improved the existing SVR in terms of MAE, RMSE, RE, and Correlation. Further, the work extended by having classification on the dataset with the proposed attribute 'Yield Status' using the method CYF-WKNN. This improves the existing bench methods in terms of Accuracy, and Kappa Statistics. Thus the proposed two methods confirm the efficiency.

In Future, the work can be extended by having the same forecasting analysis on variant Agriculture datasets. The analysis can also be extended on a new proposed attribute.

REFERENCES

[1] N. Nandhini, and J. G. Shankar, "Prediction of crop growth using machine learning based on seed," Ictact journal on soft computing, vol. 11, no. 01, 2020

[2] Pritesh Patil, Pranav Athavale, Manas Bothara, Siddhi Tambolkar, and Aditya More. "Crop Selection and Yield Prediction using Machine Learning Approach." ISSN 2347-4688, 31 October 2023.

[3] R. Ghadge, J. Kulkarni, P. More, S. Nene, and R. L. Priya, "Prediction of crop yield using machine learning," Int. Res. J. Eng, Technolgy, vol. 5, 2018.

[4] Suresh, N. Manjunathan, P. Rajesh, and E. Thangadurai, "Crop Yield Prediction Using Linear Support Vector Machine," European Journal of Molecular & Clinical Medicine, vol. 7, no. 6, pp. 2189- 2195, 2020.

[5] M. Alagurajan, and C. Vijayakumaran, "ML Methods for Crop Yield Prediction and Estimation: An Exploration," International Journal of

[6] Engineering and Advanced Technology, vol. 9 no. 3, 2020

[7] P. Kumari, S. Rathore, A. Kalamkar, and T. Kambale, "Predicition ofCrop Yeild Using SVM Approach with the Facility of E-MART System" Easychair 2020

[8] Thomas van Klompenburg, et al."Crop yield prediction using machine learning: A systematic literature review", Computers and Electronics in Agriculture, Elsevier, 2020.

[9] Muthukumaran, et al. "A Hybrid Machine Learning Model with Combined Wrapper Feature Selection Techniques to Improve the Yield of Paddy", International Journal of Electronics and Communication Engineering Volume 10 Issue 12, December 2023, pp. 45-61.

[10] Rajaswee Surana et al. "Crop Yield Prediction Using Machine Learning: A Pragmatic Approach", Research Square, 2024.

[11] Vijay H. Kalmani, “Crop Yield Prediction using Deep Learning Algorithm based on CNN-LSTM with Attention Layer and Skip Connection”, Indian Journal of Agricultural Research, Volume 59, Issue 8, 2025.

[12] Swati P. Gade et al. “Crop Yield Prediction Driven by ML & DL”, International Journal Of Innovative Research in Technology, Volume 12, Issue 6, 2025, pp.6489-6492.

[13] Manjula et al. “A Model for Prediction of Crop Yield”, International Journal of Computational Intelligence and Informatics, Vol. 6: No. 4, March 2017.

[14] Jayanarayana Reddy et al. “Crop Yield Prediction using Machine Learning Algorithm”, IEEE Xplore, 2021.

[15] Mahmoud Abdel-salam, “A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning”, Neural Computing and Applications, 2024.