Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 11, 2025

FROM DATA TO HARVEST: A PCA-BASED FEATURE SELECTION APPROACH WITH K-NEAREST NEIGHBOR AND SUPPORT VECTOR REGRESSION FOR CROP YIELD PREDICTION

Bashir Ahmad[1], Asad Ullah[2], Malak Roman[*3], Awrang Zaib[4], Toufeeq Ur Rehman[5]

[1,2,5]*BS-Computer Science, Department of Computer Science, University of Chitral, KP-Pakistan.*
[*3,4]*Lecturer, Department of Computer Science, University of Chitral, KP-Pakistan.*

[*3]malak_5116@uoch.edu.pk

DOI: https://doi.org/10.5281/zenodo.17976189

Keywords
Machine Learning, Crop Yield Prediction, K-Nearest Neighbor, Support Vector Regression, Principal Component Analysis, Agriculture

Article History
Received: 15 September 2025
Accepted: 12 November 2025
Published: 28 November 2025

Corresponding Author: *
Malak Roman

Abstract

Crop yield prediction is a key component of modern agricultural management, as it provides reliable estimates that support farmers and agricultural planners in making decisions related to resource allocation, crop scheduling, and food production planning. Machine Learning (ML) and modern computational techniques are increasingly used to address major challenges in agriculture, including low productivity, climate variability, disease outbreaks, and inefficient resource use. These methods allow agricultural systems to process large amounts of data from soil records to satellite images and convert them into practical recommendations for farmers and planners. The expansion of agricultural datasets provides an opportunity to use machine learning for more reliable forecasting. This research evaluates and contrasts two Machine Learning approaches, K-Nearest Neighbors (KNN) and Support Vector Regression (SVR), in predicting crop yield after applying Principal Component Analysis (PCA) for data reduction. A comprehensive preprocessing pipeline was implemented clean, encode, standardize, and transform the data using PCA to remove redundancy while retaining 95% variance. Model assessment employed multiple statistical metrics including the coefficient of determination (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Explained Variance. Results indicate that SVR is more effective in capturing yield variation when used with PCA preprocessing. The study concludes that dimensionality reduction and proper preprocessing significantly enhance model robustness and prediction accuracy, providing a practical framework for applying machine learning in agricultural yield forecasting.

INTRODUCTION

The estimation of crop yield before it is harvested is known as crop yield prediction. Agricultural productivity plays a vital role in ensuring food security, especially in regions where increasing population growth and climate change have made farming outcomes highly uncertain. Agricultural production has a critical role in warranting food safety, chiefly in the areas where incremental population growth and environmental variation have made agricultural outcomes extremely unreliable [1]. The increasing temperatures, lack of freshwater, and soil nutrient weakening continue to pressure overall food production. The United Nations agreed 2030 as the target time to improve food security and

reduce hunger [2]. Policymakers require exact crop estimates to plan decisions on sustenance transfers and importations [3]. Even more, the economic significance of agriculture in Pakistan underscores the need for accurate yield prediction. Agriculture contributes about 23 percent to the national GDP and delivers employment to approximately 37 percent of the employees in Pakistan [4]. Even though, crop yields in Pakistan persist to be erratic. Irregular rainfall, prolonged water shortages, and inadequate implementation of up-to-date agricultural technologies affect yield erraticism. These ongoing challenges show the significance of reliable crop yield prediction [5].

Machine Learning (ML) and advanced computational methods are now widely applied in agriculture to tackle key challenges such as declining productivity, climate variability, crop diseases, and inefficient use of resources. Early crop yield predictions used statistical models that considered factors like rainfall, temperature, and soil conditions [6]. The rising complexity of agricultural data, which involves many variables such as soil health, pest attacks, and climate change, requires more advanced modeling methods. In recent years, machine learning (ML) techniques have gained attention for yield prediction due to their ability to handle nonlinear relationships and large datasets more effectively than traditional statistical methods. Agriculture has become more fruitful, effectual, and viable with the dawn of the Machine Learning Technology (MLT), making farmer's work easier [7]. MLT uses Machine Learning Algorithms (MLA) for the prediction of situations. MLAs have two types [8]: supervised machine learning algorithms and unsupervised machine learning algorithms. Supervised machine learning algorithms depend on labelled data like linear regression, Decision Tree (DT), Random Forest (RF), Bayesian Network, Artificial Neural Network (ANN), and Support Vector Machines (SVM), etc., to more effectively represent the relationship between the predictor variables and target features [9]. On the other hand, unsupervised ML algorithms like Markov chain model, K-means clustering, EMA, DBSCAN, and Apriori algorithm, etc., work on unlabeled data [10].

Data mining, a branch of machine learning, refers to the process of examining large datasets to discover useful patterns, trends, and relationships that can support decision-making and problem-solving [11,12]. Recent progress in data mining, a branch of machine learning, has produced strong results for crop yield prediction by using large datasets from different sources [13]. Latest studies have highlighted inconsistencies in the predictive accuracy of widely used algorithms for instance [14] showed that KNN underperforms when applied to time-series yield data, while [15] demonstrated that SVR performs poorly when growth-stage-based data are used without proper preprocessing. These findings suggest that raw input features may reduce the predictive strength of these algorithms and that dimensionality reduction techniques such as Principal Component Analysis (PCA) could enhance their performance.

This study addresses this gap by conducting a comparative analysis of PCA-enhanced KNN and SVR models for crop yield prediction. By systematically evaluating their performance, the research aims to determine whether preprocessing methods like PCA can overcome the limitations of these models and provide more reliable predictions across varying agricultural datasets.

LITERATURE REVIEW

In modern years, machine learning has developed to be increasingly important for agricultural yield prediction [16]. Machine learning has been discovered as a potent tool for gaining insight and patterns from information industries [17] and other fields including agricultural environment. Machine learning empowers computers to achieve innovative capacities without the necessity of direct encoding [3]. The implementation of modern implements such as Artificial Intelligence AI, Machine Learning ML, and Deep Learning DL, Internet of things IoTs, have converted the traditional farming enhancing optimal input utilization and efficient resource allocation improving crops harvests. Additionally, Machine Learning ML, Deep Learning DL, offer enormous prospects to analyze the growth, health and yield on time, which helps in tactical decision making to accomplish supportable nutrition safety [18].

Traditional models, often reliant on linear relationships, frequently flop to catch the complex, non-linear relations between the multitude of features influencing crop growth [13]. KNN, a non-

parametric, instance-based learning algorithm, has been extensively used in several agricultural studies due to its easiness and effectiveness in handling nonlinear data; however, its performance often declines when faced with high-dimensional or noisy datasets, making it highly dependent on preprocessing techniques [19]. K-Nearest Neighbors (KNN) is a simple and widely used supervised machine learning algorithm applied in both classification and regression tasks. It operates on the principle that similar data points lie close to each other in the feature space. For a new input, KNN calculates its distance from all training samples, selects the k^{th} nearest points, and predicts the output through majority voting or averaging. Its effectiveness has been demonstrated in several real-world domains, including text and medical data analysis. The performance of KNN largely depends on key factors, particularly the choice of *k*. A very small *k* makes the model highly sensitive to noise, while a very large *k* can blend different classes and reduce accuracy, especially with imbalanced datasets [20,21]. KNN is also computationally expensive because it stores the entire training set and computes distances for every new query, earning it the label of a "lazy learner." Additionally, proper feature scaling is essential; without it, distance calculations become dominated by features with larger numeric ranges, leading to distorted results [22]. Recent studies have highlighted inconsistencies in the predictive accuracy of widely used algorithms such as Support Vector Regression (SVR) and K-Nearest Neighbors (KNN). KNN underperforms when applied to time-series yield data [14].

Support Vector Regression (SVR) is a supervised machine learning algorithm that extends Support Vector Machine (SVM) from classification to regression, predicting continuous values. It fits a function to the data with maximum margin, focusing on key data points called support vectors, while penalizing errors outside an epsilon-insensitive tube [15]. SVR has also been adopted in crop yield modeling because of its ability to catch multifaceted nonlinear relationships, and studies have shown that SVR can outperform models such as Random Forest and Artificial Neural Networks in certain datasets, including county-level corn yield predictions [10,23]. Despite these strengths, SVR can underperform when redundant or highly correlated features are present, highlighting the need for dimensionality reduction or feature selection strategies [24]. SVR performs poorly when growth-stage-based wheat crop yield data are used without appropriate preprocessing [15]. SVR performed poorly for crop yield prediction giving a negative R2 score of 0.68, which indicates that it cannot accurately represent and struggle with non-linear and complex data patterns in a large dataset. SVR requires careful data scaling and parameter tuning to perform efficiently. Concurrently, a study compared several ML algorithms comprising Random Forest, XGBoost, LightGBM, and Support Vector Regression (SVR) for predicting crop yields in Maharashtra. Their results positioned ensemble methods at the forefront, with LightGBM (R^2 = 0.87) and XGBoost (R^2 = 0.86) outperforming SVR (R^2 = 0.64), which showed laziness to catch the complex patterns in the agricultural data [25].

PCA is a highly effective procedure for dimensionality reduction which can drastically decrease the amount of variables, summarizing the information from eight original variables to just two prime components which are uncorrelated undeviating blends of the original variables that are chosen to achieve maximum variance with each subsequent component accounting for the next highest possible variance under the constraint of orthogonality [26]. PCA has therefore been increasingly used in agricultural prediction tasks as a preprocessing method to reduce data redundancy and noise while preserving the most essential variance. For example, PCA has improved classification accuracy in crop variety identification and enhanced regression accuracy when combined with SVR in yield forecasting [27,28]. Recent reviews also emphasize that while KNN is highly sensitive to raw, high-dimensional input, and SVR's robustness depends heavily on effective feature representation, PCA consistently improves model efficiency and generalizability by compressing correlated variables into fewer uncorrelated components [29,30].

Principal Component Analysis (PCA) is useful technique in exploratory analysis, as eight major Indian crops, the first two principal components explained 93% of the total variability, effectively reducing data complexity while revealing key

correlations between crops [26]. PCA is more directly for yield forecasting of maize in Uttarakhand, India. They used principal components derived from numerous weather indices as regressors in a Multiple Linear Regression (MLR) framework, finding that their most comprehensive model, which incorporated PCs from all weighted and unweighted indices, achieved the best performance with an adjusted R^2 of 79.78%, the lowest Root Mean Square Error (RMSE) on test data. This underscores PCA's value in creating robust, parsimonious models by distilling the most critical information from a large set of correlated predictors [31]. Principal Component Analysis (PCA) has also been effectively combined with Genetic Algorithms (GA) for agricultural crop classification, where it improved dimensionality reduction and enhanced model accuracy [32].

Collectively, these studies suggest that the integration of PCA with KNN or SVR can address their inherent weaknesses, yet there is still limited comparative research analyzing how PCA specifically alters their performance in agricultural yield prediction, thus creating a clear research gap.

RESEARCH METHODOLOGY

In this study we use an experimental and comparative approach to analyze the effectiveness of K Nearest Neighbors (KNN) and Support Vector Regression (SVR) to predict crops yield. Both algorithms were trained and tested on large scale of synthetic agricultural dataset comprising numerous environmental and crops related attributes. Principal Component Analysis (PCA) were used to decrease dataset dimensionality and increase computational proficiency. The methodology was organized in progressive order, including data preparation, preprocessing, model training, evaluation, and comparative analysis as shown in figure 1.

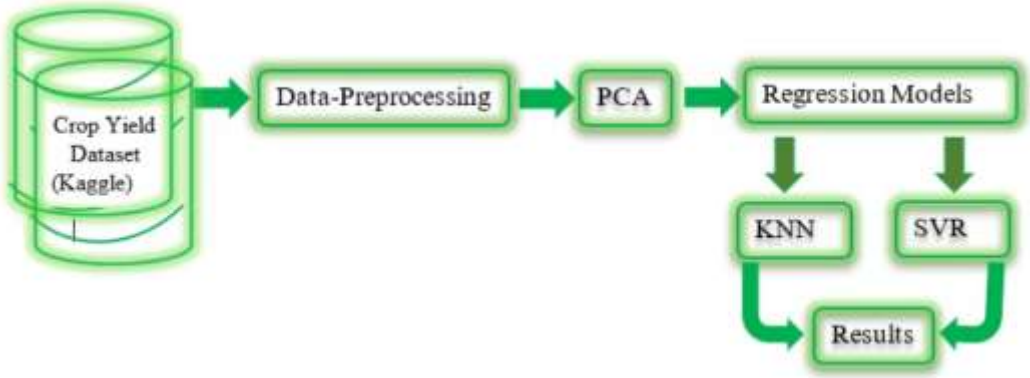

Fig 1: Research Methodology Flow Chart

Dataset Collection

The dataset was simulated from a widely used online repository known as Kaggle, consisting of 1,000,000 crop records and 10 initial features related to regional and environmental factors. The dataset contains data about different crops and related features from different Geographical information like North, East, West and South. Moreover, the dataset also includes information about the type of soil like Clays Oil, Loam Soil, Sandy Soil, Silt Soil, Peaty Soil and Chalky Soil. Additionally, the data set consists of information about the number of yields produced by each type of crop per hectare of the total area as shown in figure 2. Experimental modeling is performed in Python using pandas, numpy, and sickit-learn libraries.

	A	B	C	D	E	F	G	H	I	J	K
1	Region	Soil_Type	Crop	Rainfall_n	Temperat	Fertilizer_	Irrigation_	Weather_	Days_to_H	Yield_tons_per_hectare	
2	West	Sandy	Cotton	897.0772	27.67697	FALSE	TRUE	Cloudy	122	6.555816258	
3	South	Clay	Rice	992.6733	18.02614	TRUE	TRUE	Rainy	140	8.527340906	
4	North	Loam	Barley	147.998	29.79404	FALSE	FALSE	Sunny	106	1.127443336	
5	North	Sandy	Soybean	986.8663	16.64419	FALSE	TRUE	Rainy	146	6.517572508	
6	South	Silt	Wheat	730.3792	31.62069	TRUE	TRUE	Cloudy	110	7.248251218	
7	South	Silt	Soybean	797.4712	37.70497	FALSE	TRUE	Rainy	74	5.898416312	
8	West	Clay	Wheat	357.9024	31.59343	FALSE	FALSE	Rainy	90	2.652391665	
9	South	Sandy	Rice	441.1312	30.88711	TRUE	TRUE	Sunny	61	5.829542349	
10	North	Silt	Wheat	181.5879	26.75273	TRUE	FALSE	Sunny	127	2.943716457	
11	West	Sandy	Wheat	395.049	17.6462	FALSE	TRUE	Rainy	140	3.707293127	
12	North	Peaty	Wheat	385.1353	21.65619	FALSE	FALSE	Sunny	73	2.56444246	
13	East	Sandy	Cotton	145.3007	19.75553	TRUE	TRUE	Cloudy	141	4.367612094	
14	South	Peaty	Cotton	607.1503	15.56216	TRUE	TRUE	Sunny	136	6.52518616	
15	East	Clay	Barley	929.1237	29.6773	FALSE	TRUE	Rainy	134	6.493030752	
16	North	Peaty	Barley	621.7784	26.84317	TRUE	FALSE	Rainy	77	4.573219248	
17	East	Chalky	Rice	874.4567	27.25687	TRUE	FALSE	Sunny	115	5.839291311	
18	East	Peaty	Wheat	787.0843	25.67292	FALSE	FALSE	Cloudy	68	4.36688061	
19	North	Clay	Cotton	416.8986	23.19081	TRUE	TRUE	Sunny	95	4.85892438	
20	North	Sandy	Barley	977.2591	17.6041	FALSE	TRUE	Sunny	93	5.778099045	
21	South	Clay	Maize	888.2076	39.94551	TRUE	FALSE	Rainy	76	7.173036677	
22	East	Clay	Cotton	990.2674	24.07205	FALSE	TRUE	Sunny	110	6.187396154	
23	West	Loam	Barley	183.9397	34.22792	TRUE	FALSE	Cloudy	99	3.017923632	

crop_yield

Fig 2: Kaggle Dataset Description

Feature Selection

Feature selection, a key step in data preprocessing, has demonstrated its effectiveness in preparing data particularly high-dimensional datasets for diverse data mining and machine-learning tasks [33]. Its primary goals are to create simpler, more interpretable models, enhance data-mining performance, and deliver clean, well-structured data for analysis [34]. Prior to feature selection, several preprocessing steps were carried out using Principal Component Analysis PCA.

Principal Component Analysis PCA

The principal component analysis is mainly aimed at reducing the dimensionality level of a dataset with several interrelated variables, retaining as much of the difference existing in the dataset. This is achieved by transformation to a new variable set-the PCs, which are uncorrelated and which are organized in such a way that the first few retain most of the variation present in all the original variables [35]. This process helps in the removal of redundant information, reducing noise while keeping useful information and speeding up the training of the models. It improves the efficiency of the model and reduces overfitting by reducing feature complexity.

Its basic objectives are the development of simpler and more interpretable models, enhancement in the performance of data mining, and providing spotless and well-structured data for analysis [12]. The principal component analysis is mainly aimed at reducing the dimensionality level of a dataset with several interrelated variables, retaining as much of the difference existing in the dataset. One technique used in statistics and biometrics is Principal Component Analysis (PCA). Equation 1 illustrates how it uses a unique transformation to transform a set of related data points into a new set of unrelated data points.

$$PC_p = a_{1p}X_1 + a_{2p}X_2 + \ldots + a_{np}X_n$$

equ.1

Feature	Variety	Description
Region	Categorical	It represents the area of cultivated crops.
Soil-Type	Categorical	It describes the type of soil in which the crop is cultivated.
Crop	Categorical	It tells us which crop is cultivated in a specific area.
Rainfall-mm	Numeric	Tells the average rainfall in the described region in millimeter.
Temperature	Numeric	Represents the average temperature of the region in centigrade.
Fertilizer-Used	Numeric	It represents whether fertilizers are used or not and the ratio.
Yield-tons/ hectare	Numeric	Describes the number of crops that yield in tons per hectare

Table 1: Attributes Picked by PCA

In this study, Principal Component Analysis (PCA) is applied to reduce the dimensionality of the dataset and highlight the most relevant variables that preserved 95% of the total variance, reducing the redundancy between correlated features while preserving important information for model training as shown in figure 3 and filtered attributes are shown in table 1.

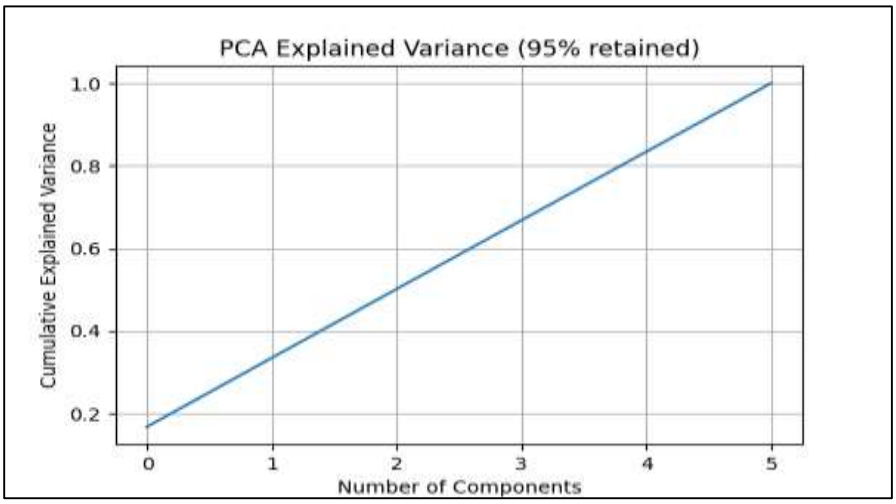

Fig 3. Demonstrating PCA Explained Variance

Classification Models Implementation

K-Nearest Neighbor Classifier

K-Nearest Neighbors (KNN) is widely used in agricultural data analysis because of its ability to capture the relationship between environmental factors and crop productivity. In crop yield prediction, KNN estimates yield for a target field by comparing it with historical cases that have similar conditions [8]. These conditions typically include soil characteristics, rainfall, temperature, fertilizer application, and crop management practices. The algorithm determines the "nearest" historical records using a distance metric [11], this similarity is evaluated through a distance metric commonly Euclidean distance shown in equation 2, which helps the algorithm locate the “nearest” students whose academic outcomes are already known [36]. Another advantage of KNN is that it does not require an

explicit training phase. Since agricultural datasets are updated frequently, KNN can easily adapt to new information without retraining a complex model [37].

For two points $x = (x1, x2, \dots, xn) and y = (y1, y2, \dots, yn)$

The Euclidean distance is:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots +}$$

equ.2

One of the strengths of KNN in agricultural modeling is its flexibility in handling nonlinear interactions among climatic and soil variables. Cropping systems often show complex patterns where yield is influenced by multiple biological and environmental factors. KNN does not assume any specific functional form, enabling it to model these nonlinear relationships effectively [38]. As a result, the method performs well when environmental variability is high or when the dataset includes diverse growing conditions. Another advantage is that KNN naturally adapts to updated datasets. As new yield records become available each season, they can be added directly without retraining a complex model. This is useful in agricultural forecasting systems where climatic conditions and management decisions constantly change [39]. Overall, KNN provides a practical and effective approach for crop yield prediction, allowing researchers and agricultural planners to analyze historical patterns and make timely yield estimates.

Support Vector Regression

Support Vector Regression (SVR) is machine learning method developed from the Support Vector Machine framework but adopted for predicting continuous values [40]. Instead of trying to pass exactly through all data points, SVR aims to find a balanced function that stays within a small acceptable margin around the actual values. Points that fall outside this margin are treated as errors and influence the final model, while points within the margin do not affect its shape [41]. Support Vector Regression (SVR) has become an important method in crop yield prediction because it can model complex, nonlinear relationships between environmental variables and crop productivity. Crop yield is influenced by multiple interacting factors such as temperature, rainfall, soil nutrients, fertilizer use, and crop management practices. These relationships are rarely linear, and SVR is well suited for such conditions because of its kernel-based framework, which allows it to capture nonlinear patterns in agricultural data [41].

In crop yield estimation, SVR attempts to fit a function that approximates the yield values within a specified error margin. This is achieved by focusing on support vectors data points that lie closest to the regression boundary thereby reducing the influence of noise and irregularities in climatic or soil measurements [42]. This property is valuable in agriculture, where variability caused by weather fluctuations or measurement inconsistencies can affect prediction accuracy.

SVR attempts to find a function $f(x)$that deviates from the actual target values by no more than ε, while keeping the model as flat as possible

$$f(x) = w^T x + b \quad equ.3$$

where:

w= weight vector
b= bias term
x= input feature vector

Another advantage of SVR is its robustness when working with small or medium-sized datasets. In agricultural studies, long-term datasets are not always available, especially for specific regions or crop varieties. SVR performs well under such constraints because it relies on a limited number of support vectors rather than the entire dataset [43]. Overall, SVR provides a reliable and flexible tool for crop yield prediction, enabling researchers and agricultural planners to analyze environmental data and produce accurate forecasts essential for farm management and policy planning.

RESULT AND DISCUSSION

For visualizing the performance of the models Python library named as Matplotlib is used to generate scatter plot, histogram and residual plots enabling a clearer understanding of the prediction performance.

Performance Indicators: To assess the predictive efficiency and reliability of the models, the following four statistical metrics are used [17].

R^2 Score: A greater R^2 rate points to that the model captures a greater portion of the variance in actual crop yields, reflecting better predictive power. Mathematically R^2 score can be represented as:

$$R^2 = 1 - \frac{\sum_{j=1}^{N}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{N}(y_j - \bar{y})^2} \quad equ.4$$

$$MAE = \frac{1}{N}\sum_{J=1}^{N}|y_j - \hat{y}_j| \quad equ.5$$

Mean Squared Error (MSE): MSE measures the average of squared differences among forecasted and real scores. Mathematically MSE can be represented as:

Mean Absolute Error (MAE): It provides a direct interpretation of average prediction error, with lower values indicating more precise predictions. Mathematically MAE can be represented as:

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2 \quad equ.6$$

Root Mean Squared Error (RMSE): RMSE is the radical expression of MSE, articulated in the similar divisions as of target variable (tons per hectare). Mathematically RMSE can be represented as:

$$RMSE = \sqrt{\frac{\sum_{j}^{N}(y_j - \hat{y}_j)^2}{N}} \quad equ.7$$

Algorithms	R^2 Score	Mae	Mse	Rmse	Explained Variance
KNN	0.7644	0.6833	0.6792	0.8241	0.7644
SVR	0.7878	0.6568	0.6117	0.7821	0.7878

Table 2: Demonstrates the results of PCA- enhanced KNN and SVR

The KNN model is tested with multiple values of k (3, 5, 7, and 9), The performance improved steadily with higher k values. The best results were obtained at k=9 with an R^2 score of 0.7644, Mean Absolute Error (MAE): 0.6833, Mean Squared Error (MSE): 0.6792, Root Mean Squared Error (RMSE): 0.8241, with an explained variance of 0.7644. Similarly, SVR model achieved R^2 Score: 0.7878, Mean Absolute Error (MAE): 0.6568, Mean Squared Error (MSE): 0.6117, Root Mean Squared Error (RMSE): 0.7821, as visualized in table 2 and figure 4.

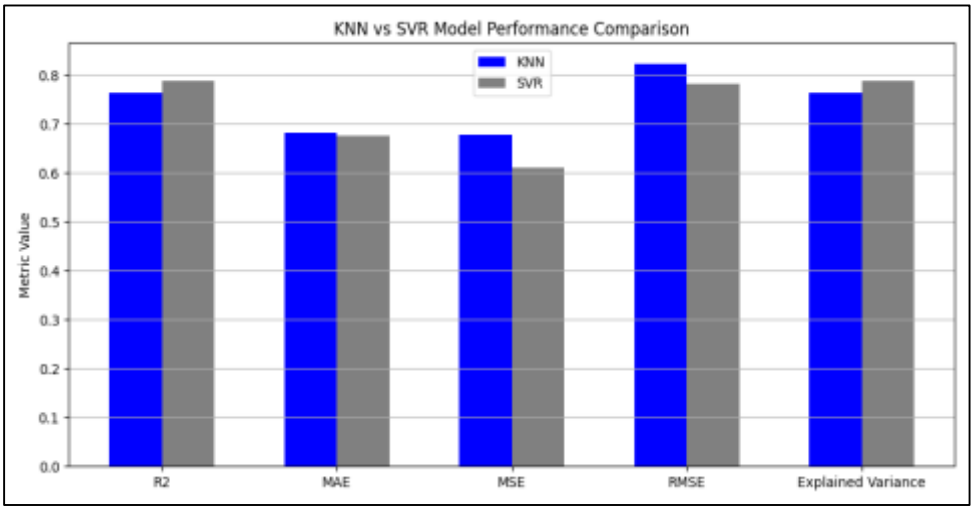

Fig 4: Graphically Representation of PCA-enhanced KNN and SVR

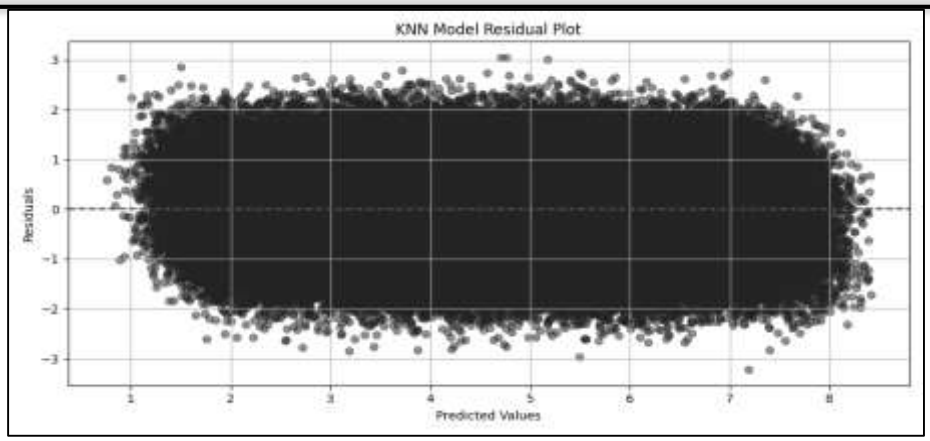

Fig 5: KNN Model Residual Plot

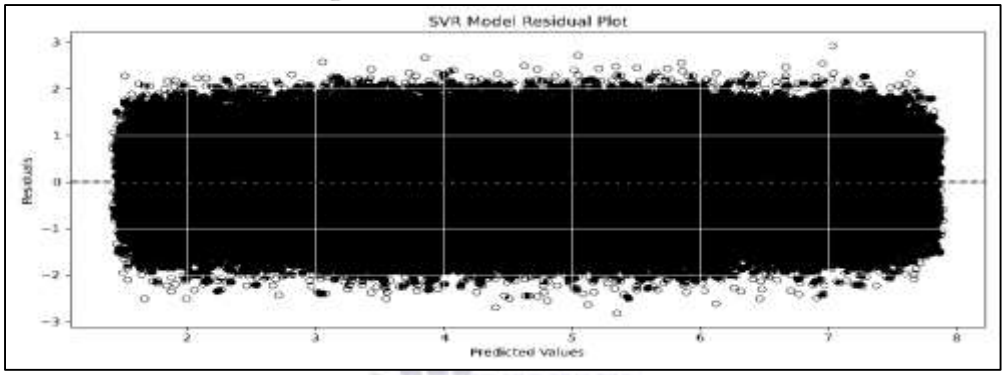

Fig 5: SVR Model Residual Plot

Discussions: The better performance of both KNN and SVR algorithms demonstrate the higher competence of PCA in reducing the dimensions and noise of complex data ultimately resulting in the improved performance of each one. The PCA-enhanced SVR model demonstrated slightly superior performance compared to KNN. The higher R^2 and lower error metrics confirm that SVR provided more accurate predictions than KNN. PCA reduced redundancy and improved feature representation, helping SVR capture complex nonlinear relationships more efficiently. The findings confirm that PCA-enhanced preprocessing significantly improves model performance and stability across both algorithms. Overall, this research high spots the power of data-driven approaches in supporting agronomic productivity forecasting, resource management, and sustainable farming decisions.

REFERENCES

Meena, R. S., Kumar, S., Datta, R., Lal, R., Vijayakumar, V., Brtnicky, M., ... & Marfo, T. D. (2020). Impact of agrochemicals on soil microbiota and management: A review. *Land*, *9*(2), 34.

Kolloju, N., Junuguru, S., Kumar, K. P., & Naveen, S. (2025). Rising Food Insecurity and the UNSDG 2030 Agenda: Challenges and Innovations in Achieving Zero Hunger. In *Climate Change, Food Security, and Land Management: Strategies for a Sustainable Future* (pp. 1-17). Cham: Springer Nature Switzerland.

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.

Khan, K. (2024). A Review on Fiscal and Debt Policies in Pakistan.

Saha, S., Kucher, O. D., Utkina, A. O., & Rebouh, N. Y. (2025). Precision agriculture for improving crop yield predictions: A literature review. *Frontiers in Agronomy*, *7*, 1566201.

Shi, W., Tao, F., & Zhang, Z. (2013). A review on statistical models for identifying climate contributions to crop yields. *Journal of geographical sciences*, *23*(3), 567-576.

Ismael, H. R., Abdulazeez, A. M., & Hasan, D. A. (2021). Comparative study for classification algorithms performance in crop yields prediction systems. *Qubahan Academic Journal*, *1*(2), 119-124.

Roman, M., Naz, I., Luqman, M. A., Ali, J., Jan, M. S., & Nawab, H. U. (2024). Stroke Disease Prediction Using K-Nearest Neighbor and Decision Tree Algorithms with Machine Learning Pre-Processing Techniques. *Migration Letters*, *21*(S4), 2015-2027.

Shah, F. A., Meraj, A. B. E., Roman, M., Anwar, M., Zaib, A., Shams, S. S., & Hussain, F. (2025). Enhancing Weather Forecasting Accuracy: A Machine Learning Approach Using Genetic Algorithm and Random Forest. *Global Research Journal of Natural Science and Technology*.

Khan, S. N., Khan, A. N., Tariq, A., Lu, L., Malik, N. A., Umair, M., & Zawaideh, F. H. (2023). County-level corn yield prediction using supervised machine learning. *European Journal of Remote Sensing*, *56*(1), 2253985.

Roman, M., Nawab, H. U., Ahmad, S., & Khan, I. A. (2022). K-Nearest Neighbor and Fuzzy K-Nearest Neighbor Algorithm Performance Analysis for Heart Disease Classification. *Webology(ISSN: 1735-188X)*, *19*(1).

Ullah, M. A., Ullah, A., Roman, M., Anwar, M., Siddiqi, M. U. M., Jaffar, M. H., & Ali, J. (2025). A Comparative Study of Machine Learning Algorithms for Cardiovascular Risk Prediction: Support Vector Machine, Gradient Boosting, And Rotation Forest. *Spectrum of Engineering Sciences*, 481-491.

Padia, N., & Sarvaiya, M. (2022). Crop Yield Prediction Using Data Mining Techniques. *Available at SSRN 5105365*.

Yan, Y., Wang, Y., Li, J., Zhang, J., & Mo, X. (2025). Crop yield time-series data prediction based on multiple hybrid machine learning models. *arXiv preprint arXiv:2502.10405*.

Lou, Z., Lu, X., & Li, S. (2024). Yield prediction of winter wheat at different growth stages based on machine learning. *Agronomy*, *14*(8), 1834.

Bali, N., & Singla, A. (2022). Emerging trends in machine learning to predict crop yield and study its influential factors: A survey. *Archives of computational methods in engineering*, *29*(1), 95-112.

Roman, M., Ullah, A., Ullah, M. A., Hussain, F., Shams, S. S., Bint-e-Meraj, A., & Ali, S. (2025). Predicting Academic Success: A Machine Learning Approach Using Decision Tables and Random Forests Algorithms. *Spectrum of Engineering Sciences*, *3*(5), 205-213.

Ajith, S., Vijayakumar, S., & Elakkiya, N. (2025). Yield prediction, pest and disease diagnosis, soil fertility mapping, precision irrigation scheduling, and food quality assessment using machine learning and deep learning algorithms. *Discover Food*, *5*(1), 1-23.

Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., & Khan, N. (2021). A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. *IEEE access*, *9*, 63406-63439.

Jingwen, S., Du, W., & Shi, N. (2018). A Survey of kNN Algorithm. *Information Engineering and Applied Computing*.

Suyal, M., & Goyal, P. (2022). A review on analysis of k-nearest neighbor classification machine learning algorithms based on supervised learning. *International Journal of Engineering Trends and Technology*, *70*(7), 43-48.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 986-996). Berlin, Heidelberg: Springer Berlin Heidelberg.

Patil, Y., Ramachandran, H., Sundararajan, S., & Srideviponmalar, P. (2025). Comparative analysis of machine learning models for crop yield prediction across multiple crop types. *SN Computer Science*, *6*(1), 64.

Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature

review. *Computers and electronics in agriculture*, *177*, 105709.

Sharma, S., Jain, A., Sharma, S., & Whig, P. (2025). Enhancing crop yield prediction through machine learning regression analysis. *International Journal of Sustainable Agricultural Management and Informatics*, *11*(1), 29-47.

Singh, V., & Singh, A. (2020). Analysis of agriculture data using principal component analysis. *International Journal of Multidisciplinary Research and Development*, *7*(1), 34-37.

Ahmad, I., Saeed, U., Fahad, M., Ullah, A., Habib ur Rahman, M., Ahmad, A., & Judge, J. (2018). Yield forecasting of spring maize using remote sensing and crop modeling in Faisalabad-Punjab Pakistan. *Journal of the Indian Society of Remote Sensing*, *46*(10), 1701-1711.

Mehta, A., Sengupta, P., Garg, D., Singh, H., & Diamand, Y. S. (2023). Benchmarking the effectiveness of classification algorithms and SVM kernels for dry beans. *arXiv preprint arXiv:2307.07863*.

Issac, A., Yadav, H., Rains, G., & Velni, J. M. (2022). Dimensionality reduction of high-throughput phenotyping data in cotton fields. *IFAC-PapersOnLine*, *55*(32), 153-158.

Zhang, Q., Wang, K., Han, Y., Liu, Z., Yang, F., Wang, S., & Zhao, C. (2022). A crop variety yield prediction system based on variety yield data compensation. *Computers and Electronics in Agriculture*, *203*, 107460.

Yadav, A., & Shukla, A. K. (2024). Prediction of maize crop yield using principal component analysis of weather parameters. *International Journal of Environment and Climate Change*, *14*(10), 189-195.

Cruz, G. B. D., Gerardo, B. D., & Tanguilig III, B. T. (2014). Agricultural crops classification models based on PCA-GA implementation in data mining. *International Journal of Modeling and Optimization*, *4*(5), 375.

Hussain, F., Shams, S. S., Roman, M., Anwar, M., Shah, F. A., Meraj, A. B. E., ... & Uddin, M. A. (2025). MACHINE LEARNING IN HEALTHCARE: PREDICTING CHRONIC KIDNEY DISEASE THROUGH FEATURE-DRIVEN HEURISTIC MODELS. *Frontier in Medical and Health Research*, *3*(7), 318-328.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys (CSUR)*, *50*(6), 1-45. Mining, M. D. Data Mining: Concepts and Techniques (2nd.edition).

Mishra, S., Singh, B. R., Naqvi, A. H., & Singh, H. B. (2017). Potential of biosynthesized silver nanoparticles using Stenotrophomonas sp. BHU-S7 (MTCC 5978) for management of soil-borne and foliar phytopathogens. *Scientific reports*, *7*(1), 45154

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

Ali, Z. M., Hassoon, N. H., Ahmed, W. S., & Abed, H. N. (2020). The application of data mining for predicting academic performance using k-means clustering and naïve bayes classification. *International Journal of Psychosocial Rehabilitation*, *24*(03), 2143-2151.

Jabed, M. A., & Murad, M. A. A. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon*, *10*(24).

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674.

Zhang, F., & O'Donnell, L. J. (2020). Support vector regression. In *Machine learning* (pp. 123-140). Academic Press.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, *14*(3), 199-222.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Hussain, S., Cheema, M. J. M., Saleem, S. R., Elbeltagi, A., & Aqib, M. (2025). Implementation of Artificial Intelligence in Agriculture: An Editorial Note. *AgriEngineering*, *7*(12), 401.