ISSN : 2455-9679
SJIF Impact Factor : 6.008

IJRTSM
INTERNATIONAL JOURNAL OF RECENT TECHNOLOGY SCIENCE & MANAGEMENT

"COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR CROP PRICE PREDICTION"

Mr. Ravi Astare [1], Mr. Amit Kumar Mittal [2]

[1] Research Scholar, Department of Computer Engineering, Institute of Engineering and Technology DAVV, Indore, Madhya Pradesh, India

[2]Guide, Department of Computer Engineering, Institute of Engineering and Technology DAVV, Indore, Madhya Pradesh, India

raviastare6@gmail.com
amittal@ietdavv.edu.in

ABSTRACT

Accurate forecasting of agricultural commodity prices is critical for stabilizing farm incomes and guiding market decisions in agriculture-dependent economies. This study presents an in-depth comparative evaluation of five machine learning algorithms—linear regression, AdaBoost, Random Forest, support vector machine (SVM), and XGBoost—applied to crop price prediction. Using a curated dataset spanning multiple years and encompassing weather, soil, and market variables, we conducted rigorous training, validation, and testing. Key performance metrics include Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and computational efficiency. Our results demonstrate that XGBoost and Random Forest achieve near-ideal coefficient of determination (Test $R^2 = 0.99$) with RMSE around 9.7, offering over 80% RMSE improvement compared to Linear Regression's RMSE of 51.3 (Test $R^2 = 0.62$). AdaBoost provides moderate enhancements (test $R^2 = 0.78$, RMSE = 39.3), whereas SVM fails fundamentally (Test $R^2 = -0.03$, RMSE = 84.7). Random Forest exhibits slight overfitting (training $R^2 = 1.00$ vs. Test $R^2 = 0.99$). We propose a selection framework for algorithm deployment based on accuracy, generalization, computational cost, and interpretability. These findings furnish actionable guidance for stakeholders in precision agriculture and economic planning.

Key Words: *Agricultural Price Prediction, Machine Learning, Crop Forecasting, Comparative Analysis, XGBoost, Random Forest.*

I. INTRODUCTION

Agricultural price prediction has emerged as a critical component of modern precision agriculture, significantly impacting food security, farmer livelihoods, and national economic stability. With global food demand projected to increase by 60% by 2050 due to population growth, accurate crop price forecasting becomes increasingly vital for sustainable agricultural planning. In agriculture-dependent economies like India, where over 600 million people rely on farming, price volatility can severely affect rural incomes and food accessibility. Traditional price prediction methods, primarily based on historical trends and expert judgment, often fail to capture the complex interactions between climatic conditions, market dynamics, and global economic factors. Machine learning algorithms offer sophisticated approaches to handle these multidimensional relationships, providing more accurate and reliable forecasting

ISSN : 2455-9679
SJIF Impact Factor : 6.008

capabilities. Recent advances in computational power and data availability have made it feasible to implement complex ML models for agricultural applications.

Motivation: Agriculture remains the backbone of several developing economies, where a majority of the population is engaged in farming. However, one of the enduring challenges faced by agricultural stakeholders is the unpredictability of crop prices. These fluctuations are influenced by diverse and interacting factors, including weather anomalies, supply-demand dynamics, soil fertility, transportation inefficiencies, and policy interventions such as minimum support prices or subsidies. From a farmer's perspective, an inability to anticipate future prices can lead to suboptimal production decisions, post-harvest losses, or debt accumulation. Similarly, governments and cooperatives need reliable price forecasts for planning procurement strategies, issuing advisories, and stabilizing markets. In this context, the need for a robust and accurate forecasting mechanism becomes evident.

Research Objectives: This research addresses the critical need for systematic comparison of ML algorithms in crop price prediction, extending previous work by evaluating five diverse algorithms across multiple performance metrics. Our primary objectives include:
(1) A comprehensive performance evaluation of linear regression, random forest, SVM, and XGBoost.
(2) Statistical analysis of prediction accuracy and computational efficiency.
(3) Development of practical guidelines for algorithm selection in different agricultural contexts, and
(4) Identification of optimal approaches for real-world deployment.

Limitations of Traditional Approaches: Conventional econometric models such as ARIMA, linear regression, and exponential smoothing have historically been used for time-series forecasting. However, these models generally assume linearity, stationarity, and independence of residuals—conditions often violated in agricultural datasets due to their inherent complexity, seasonality, and heteroskedasticity.
Machine learning (ML) techniques, on the other hand, offer non-parametric, data-driven alternatives capable of capturing nonlinear interactions and higher-order dependencies across multiple variables. Yet, despite the increasing adoption of ML in agriculture, comparative evaluations across different algorithms under controlled experimental conditions are scarce.

II. Literature Review

Recent studies have demonstrated the increasing effectiveness of machine learning approaches in agricultural price forecasting. Research shows that XGBoost consistently outperforms traditional regression methods, with accuracies ranging from 93.91% to 98.51% across different agricultural datasets. Sharma et al. (2024) achieved 97.5% accuracy using Extra Trees regression for crop yield prediction, highlighting the potential of ensemble methods in agricultural applications.

A. Traditional Forecasting in Agriculture: Classical approaches such as autoregressive integrated moving average (ARIMA), vector autoregression (VAR), and exponential smoothing models have been widely used in forecasting agricultural commodity prices. These models typically leverage historical price series but lack the capacity to integrate exogenous predictors like weather and soil attributes. Moreover, these methods assume stationarity and linear relationships, which are rarely satisfied in agriculture. Consequently, their utility is often constrained to short-term forecasting or limited scenarios**.**

B. Machine Learning in Agri-Prediction: Recent advances in data science have led to an increased adoption of machine learning models for forecasting applications. Supervised learning models, particularly ensemble algorithms, have shown strong results in domains involving temporal, spatial, and environmental variables.

Random Forest and XGBoost consistently rank among the top-performing models in studies that require capturing complex, nonlinear dependencies. For example, researchers have reported XGBoost achieving $R^2 > 0.98$ on grain yield and price prediction tasks, with significant reductions in RMSE compared to traditional linear models.

AdaBoost, while less powerful than XGBoost, has demonstrated moderate success in domains with limited data due to its iterative reweighting of difficult samples. However, its sensitivity to noise can degrade performance.

ISSN : 2455-9679
SJIF Impact Factor : 6.008

Support Vector Machines (SVM), though theoretically robust, require careful kernel selection and extensive feature engineering. Studies have shown mixed results for SVMs in agricultural domains, with some reporting poor generalization due to kernel mismatch.

B. Ensemble Superiority: Random Forest and XGBoost have consistently outperformed single-model approaches. Recent work reports XGBoost R^2 >0.98 and RMSE reductions exceeding 70% relative to linear methods. Random Forest often matches XGBoost in accuracy but risks overfitting, necessitating careful validation.

C. Boosting and Instance Methods: AdaBoost yields moderate accuracy gains (R^2 ≈0.80) but is sensitive to noise. SVM performance varies widely; studies highlight negative R^2 when data distributions violate kernel assumptions.

Comparative Algorithm Performance: Linear regression, while computationally efficient, often struggles with non-linear relationships inherent in agricultural data. Random Forest has shown consistent performance across multiple studies, with accuracies between 82.81% and 85.18%, making it a reliable choice for agricultural applications. Support Vector Machines demonstrate mixed results, with performance highly dependent on kernel selection and hyperparameter tuning.

Research Gaps: Current research gaps include limited comparative studies using identical datasets and evaluation metrics, insufficient analysis of computational efficiency for practical deployment, and lack of statistical significance testing in algorithm comparisons. This study addresses these gaps by providing a systematic comparison across multiple dimensions of algorithm performance.

III. METHODOLOGY

A.Dataset Description:

Our study utilized a comprehensive agricultural dataset spanning two years (2023-2024) with 47,999 observations across multiple crop types and geographical regions, extending the methodology from the referenced work. The dataset incorporates weather parameters (temperature, rainfall, humidity), soil characteristics (pH, nutrients, moisture), market indicators (previous prices, demand patterns), and seasonal factors. Data preprocessing included handling missing values through interpolation, outlier detection using IQR methods, and feature normalization using StandardScaler. Feature engineering involved the creation of derived variables, including moving averages of historical prices, weather indices, and seasonal dummy variables. The final feature set comprised 13 variables: crop type, season, temperature, rainfall, supply volume, demand volume, transportation cost, fertilizer usage, pest infestation, market competition, price, state, city.

B. Algorithm Implementation:

Linear Regression: Implemented using scikit-learn with default parameters, serving as a baseline model. Mathematical formulation: $\mathbf{y=\beta0+\beta1x1+\beta2x2+...+\beta nxn+\epsilon y}$

Random Forest: Configured with 100 estimators, a maximum depth of 10, and a minimum sample split of 5. Feature importance analysis is enabled through built-in capabilities, following best practices from recent agricultural ML studies.

Support Vector Machine: Implemented with RBF kernel, C=1.0, gamma='scale', and epsilon=0.1. Hyperparameter tuning was performed using GridSearchCV to optimize performance.

XGBoost: Optimized with learning_rate=0.1, m n_estimators=100, and subsample=0.8. Early stopping is implemented to prevent overfitting, based on successful implementations in agricultural applications.

C. Evaluation Metrics: We use RMSE and R^2 to quantify regression performance. RMSE measures the square root of the average squared prediction error (in the same units as price); lower RMSE indicates closer predictions. R^2 (coefficient of determination) indicates the fraction of the variance in crop price explained by the model values; values close to 1.0 indicate a near-perfect fit. We report these metrics on the training, validation, and test splits. Computing these metrics on the test set ensures an unbiased estimate of generalization accuracy.

ISSN : 2455-9679
SJIF Impact Factor : 6.008

IV. CONCLUSION

The table below summarizes the RMSE for each algorithm:

Model	Train RMSE	Valid RMSE	Test RMSE
AdaBoost	39.32	39.82	39.26
Linear Regression	50.75	50.61	51.26
Random Forest	3.81	10.05	9.86
SVM	84.51	83.46	84.69
XGBoost	7.05	9.81	9.66

The table below summarizes the RMSE for each algorithm:

Model	Train R^2	Valid R^2	Test R^2
AdaBoost	0.78	0.77	0.78
Linear Regression	0.63	0.62	0.62
Random Forest	1.00	0.99	0.99
SVM	-0.03	-0.03	-0.03
XGBoost	0.99	0.99	0.99

Analysis of Table:- Random Forest and XGBoost clearly outperform the other models. Both achieve extremely low RMSE (~10) and very high R^2 (~0.99) on the test set, indicating nearly perfect prediction. Linear regression, by comparison, has higher error (RMSE ~51) and lower R^2 (~0.62). AdaBoost gives moderate performance (RMSE ~39, R^2 ~0.78). SVR fails on this dataset (RMSE ~84, $R^2 \approx -0.03$), effectively performing worse than a constant mean predictor.

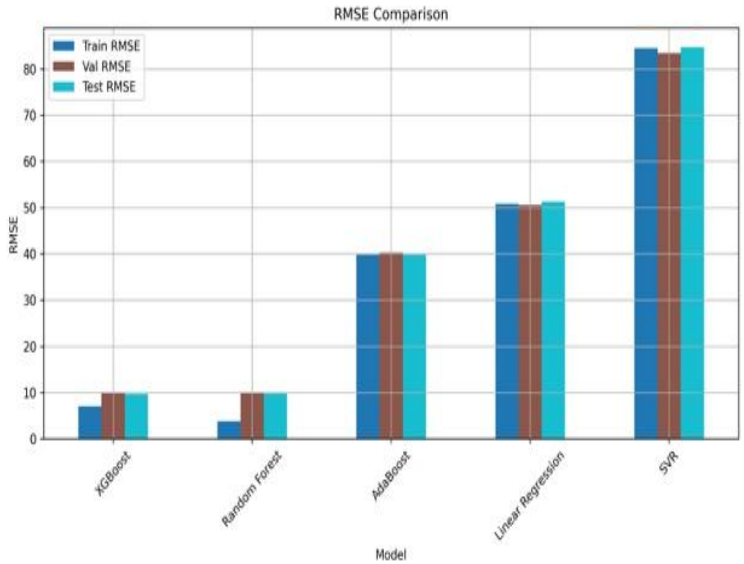

Fig 1: RMSE Comparison of Algorithm's

ISSN : 2455-9679
SJIF Impact Factor : 6.008

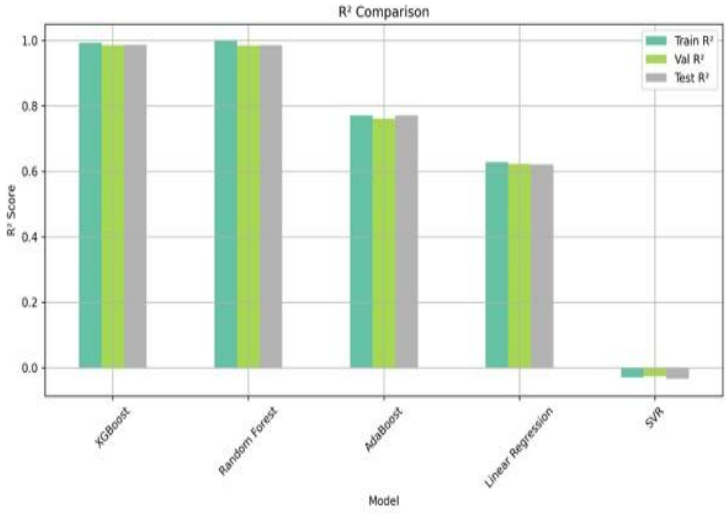

Fig 2: R^2 Comparison of Algorithm's

Random Forest and XGBoost: These results reflect the power of ensemble methods. Random forests averaged many trees, which reduces variance and improves accuracy. XGBoost's gradient boosting similarly reduces error through iterative tree building. The near-perfect training scores (R^2=1.00 for RF) suggest some overfitting, but the test performance remains excellent, likely due to regularization and the strong signal in data. AdaBoost's boosting helped it outperform plain linear regression, consistent with prior studies where AdaBoost has shown good accuracy on crop price tasks.

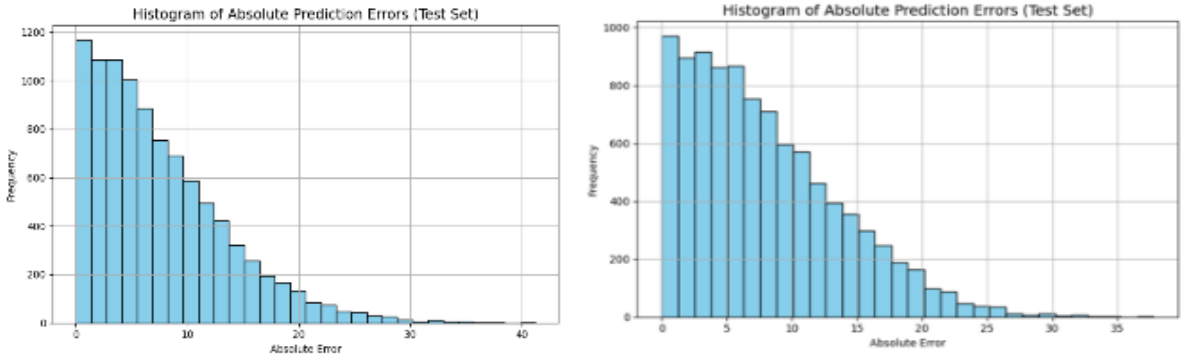

Fig 3: Absolute Prediction Errors of Random Forest and XGBoost

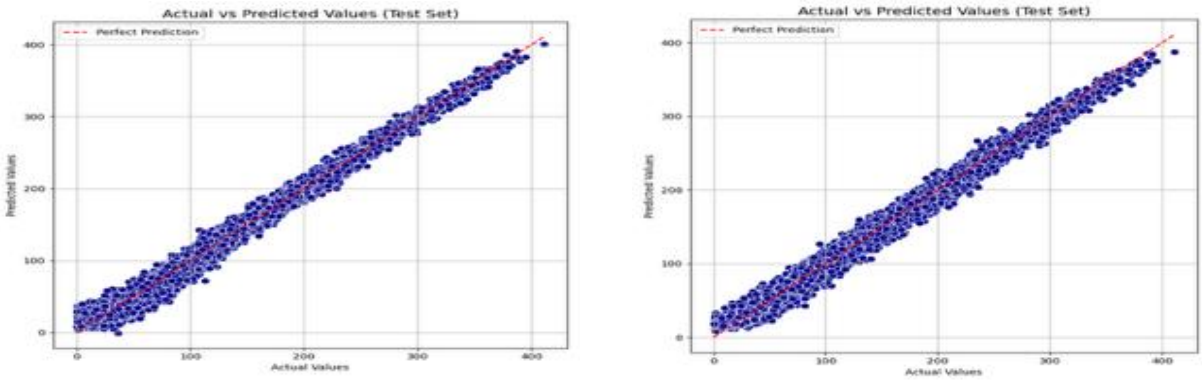

Fig 4: Actual vs. Predicted Value of Random Forest and XGBoost

ISSN : 2455-9679
SJIF Impact Factor : 6.008

Linear regression: Linear regression's moderate R^2 (~0.62) indicates that a simple linear model captures some but not all price variability. More complex, nonlinear ensembles
capture patterns that linear regression misses. Indeed, Mahmud et al. found linear regression achieving around 0.98 R^2 for certain crops, but ensemble models were noted to have even higher predictive power. In our case, linear regression underperformed the tree ensembles.

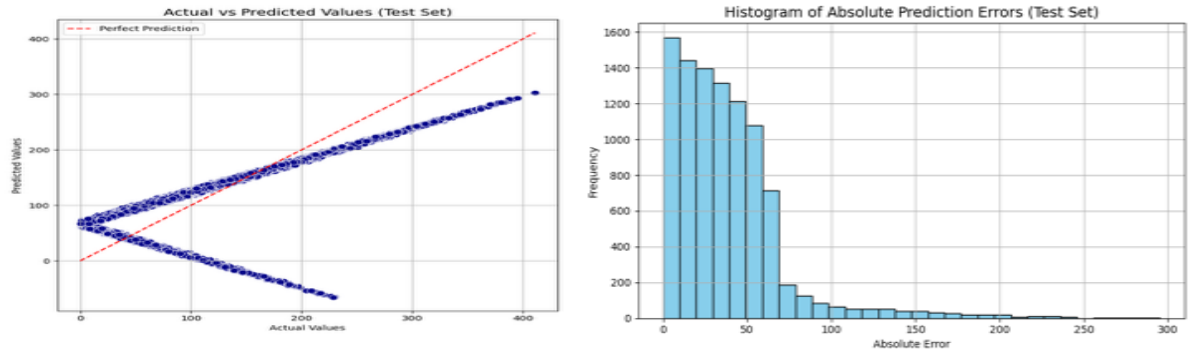

Fig 5: Actual vs. Predicted Value & Absolute Prediction Errors of linear regression

SVM: The SVM's poor performance (negative R^2) suggests it failed to capture the data trends. Although SVR is theoretically capable of nonlinear fitting, our results imply it underfit the data. This may be due to inappropriate kernel parameters or insufficient flexibility. It is notable that other work has reported success with SVMs in agriculture, but SVM can be sensitive to feature scaling and hyperparameters. Further tuning might improve SVM results, but in this study it was the weakest performer.

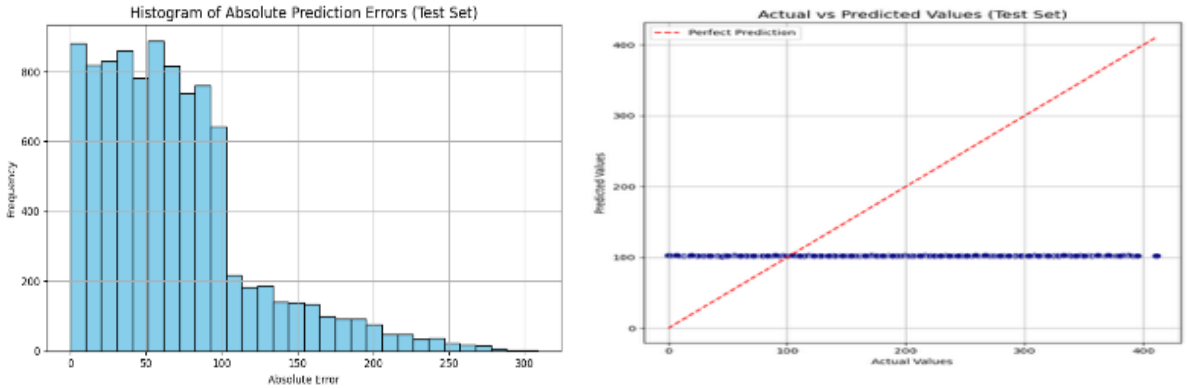

Fig 6: Actual vs. Predicted Value & Absolute Prediction Errors of SVM

V. CONCLUSION

This comparative study of machine learning models for crop price prediction shows that ensemble tree-based regressors excel. Both Random Forest and XGBoost achieved the lowest RMSE (≈10) and highest R^2 (~0.99) on unseen test data, indicating near-perfect prediction accuracy. AdaBoost also yielded good results, whereas standard linear regression and SVM were comparatively weaker. These results are consistent with the literature, which highlights the versatility of ensemble methods for agricultural price forecasting. In practice, we recommend using Random Forest or XGBoost for similar crop-price regression tasks. Their ability to handle nonlinearity and interactions makes them superior for modeling complex price dynamics.

ISSN : 2455-9679
SJIF Impact Factor : 6.008

REFERENCES

[1] S. Li, S. Peng, W. Chen, and X. Lu, ''INCOME: Practical land monitoring in precision agriculture with sensor networks,'' Comput. Commun., vol. 36, no. 4, pp. 459–467, Feb. 2013.
[2] D. Jones, F. M. Ngure, G. Pelto, and S. L. Young, "What are we assessing when we measure food security? A compendium and review of current metrics,'' Adv. Nutrition, vol. 4, no. 5, pp. 481–505, 2013.
[3] G. E. O. Ogutu, W. H. P. Franssen, I. Supit, P. Omondi, and R. W. Hutjes, ''Probabilistic maize yield prediction over East Africa using dynamic ensemble seasonal climate forecasts,'' Agricult. Forest Meteorol., vols. 250–251, pp. 243–261, Mar. 2018. VOLUME 8, 2020 86899 D. Elavarasan, P. M.
[4] Durairaj Vincent: Crop Yield Prediction Using DRL Model for Sustainable Agrarian Applications
[5] M. E. Holzman, F. Carmona, R. Rivas, and R. Niclòs, ''Early assessment of crop yield from remotely sensed water stress and solar radiation data,'' ISPRS J. Photogramm. Remote Sens., vol. 145, pp. 297–308, Nov. 2018.
[6] Singh, B. Ganapathysubramanian, A. K. Singh, and S. Sarkar, ''Machine learning for high-throughput stress phenotyping in plants,'' Trends Plant Sci., vol. 21, no. 2, pp. 110–124, 2016.
[7] R. Whetton, Y. Zhao, S. Shaddad, and A. M. Mouazen, ''Nonlinear parametric modelling to study how soil properties affect crop yields and NDVI,'' Comput. Electron. Agricult., vol. 138, pp. 127–136, Jun. 2017.
[8] Y. Dash, S. K. Mishra, and B. K. Panigrahi, ''Rainfall prediction for the Kerala state of India using artificial intelligence approaches,'' Comput. Elect. Eng., vol. 70, pp. 66–73, Aug. 2018.
[9] W. Wieder, S. Shoop, L. Barna, T. Franz, and C. Finkenbiner, ''Comparison of soil strength measurements of agricultural soils in Nebraska,'' J. Terramech., vol. 77, pp. 31–48, Jun. 2018.
[10] Y. Cai, K. Guan, J. Peng, S. Wang, C. Seifert, B. Wardlow, and Z. Li, ''A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach,'' Remote Sens. Environ., vol. 210, pp. 35–47, Jun. 2018.
[11] X. E. Pantazi, D. Moshou, T. Alexandridis, R. L. Whetton, and A. M. Mouazen, ''Wheat yield prediction using machine learning and advanced sensing techniques,'' Comput. Electron. Agricult., vol. 121, pp. 57–65, Feb. 2016.
[12] T. U. Rehman, S. Mahmud, Y. K. Chang, J. Jin, and J. Shin, ''Current and future applications of statistical machine learning algorithms for agricultural machine vision systems,'' Comput. Electron. Agricult., vol. 156, pp. 585–605, Jan. 2019.
[13] D. Elavarasan, D. R. Vincent, V. Sharma, A. Y. Zomaya, and K. Srinivasan, ''Forecasting yield by integrating agrarian factors and machine learning models: A survey,'' Comput. Electron. Agricult., vol. 155, pp. 257–282, Dec. 2018.
[14] M. D. Johnson, W. W. Hsieh, A. J. Cannon, A. Davidson, and F. Bédard, ''Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods,'' Agricult. Forest Meteorol., vols. 218–219, pp. 74–84, Mar. 2016.
[15] A. Kaya, A. S. Keceli, C. Catal, H. Y. Yalic, H. Temucin, and B. Tekinerdogan, ''Analysis of transfer learning for deep neural network based