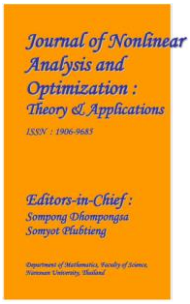
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1, No.8 : 2024
ISSN : **1906-9685**



AN ENSEMBLE-BASED CLUSTERING AND CLASSIFICATION FRAMEWORK FOR PREDICTING AGRICULTURAL CROP YIELD: A COMPREHENSIVE SURVEY

Udhaya Priya J Research scholar Madras university E-mail: udhayard07@gmail.com
Dr. K. Nirmala Associate professor, Department of Computer Science, Quaid-E-Milleth Govt college for women, Chennai-02 nimimca@gmail.com

ABSTRACT
This survey paper explores an innovative ensemble-based clustering and classification framework designed for predicting agricultural crop yield. The study focuses on addressing the challenges faced by farmers in deciding suitable crops based on region, season, and soil conditions. The proposed framework incorporates pre-processing steps, ensemble clustering utilizing Enhanced Artificial Bee Colony Optimization and Shuffled Frog Leaping Algorithm, and classification using Enhanced KNN. The research aims, the ensemble-based clustering and classification framework presented in this survey paper holds great promise for revolutionizing the agricultural landscape in India. By addressing the identified challenges and limitations, future research endeavors can contribute to the development of a robust, adaptable, and ethically sound predictive model.

KEYWORDS
Agricultural Crop Yield Prediction, Ensemble Clustering, Classification Framework, Enhanced Artificial Bee Colony Optimization, Shuffled Frog Leaping Algorithm, Enhanced KNN

1. INTRODUCTION
India, with its vast and diverse agricultural landscape, holds a pivotal position as the second-largest global producer of agricultural goods. This sector plays a critical role in shaping the socio-economic fabric of the nation. The success of agriculture in India is intricately woven into a complex tapestry of factors that includes economic conditions, climatic nuances, soil quality, irrigation practices, temperature variations, and more [1]. These multifaceted elements collectively determine the yield and success of crops, making the agricultural sector highly susceptible to external influences [2].
Given the profound impact of climatic conditions on agricultural outcomes, the need for historical data on crop yield becomes paramount. This historical perspective equips stakeholders in the agricultural supply chain, including industries dealing with livestock, raw materials, chemicals, fertilizers, and pesticides, with valuable insights. Accurate predictions of crop production become indispensable for informed decision-making in various facets of the agricultural industry, ranging from production scheduling to marketing strategies [3]-[5].
Data mining, a powerful analytical process integrating techniques such as machine learning, artificial intelligence, database systems, and statistics, emerges as a transformative tool in the realm of agriculture. Its ability to decipher patterns and extract meaningful information from vast datasets provides a novel approach to understanding the intricacies of the agricultural landscape. By leveraging data mining procedures, stakeholders can gain comprehensive insights into various factors influencing crop production, fostering more informed decision-making [6].
Two primary factors contribute significantly to effective decision-making for both farmers and government entities. Firstly, the provision of historical crop yield records with accurate forecasts serves as a crucial risk management tool. Farmers, armed with this information, can make strategic decisions regarding crop selection and resource allocation. Secondly, government policies, including crop insurance and supply chain operations, are facilitated by the insights derived from data mining

procedures [7]. These policies aim to mitigate risks for farmers and streamline the overall functioning of the agricultural sector.

In light of India's status as a global agricultural powerhouse, the motivation to enhance the efficiency of this sector becomes apparent. Agriculture's unique dependence on economic and climatic factors underscores the need for a comprehensive understanding of various variables such as climate patterns, soil composition, irrigation practices, and fertilizer usage. Companies engaged in the supply chain management of agricultural products, dealing with livestock, raw materials, animal feed, pesticides, and more, derive immense benefits from access to historical crop yield information. Accurate predictions of crop production enable these companies to optimize production scheduling, refine marketing strategies, and make informed decisions across various operational aspects.

The objectives of the present study are multifaceted, aiming to address existing challenges and contribute to the improvement of agricultural productivity. First and foremost, the study seeks to alleviate the restrictions faced by farmers in making crop-related decisions. By providing farmers with historical records and forecasts, the study aims to empower them to make more informed choices, ultimately reducing risks associated with crop selection. Additionally, the study aims to identify indicators associated with the heterogeneity of agricultural fields, with the goal of enhancing predictive capabilities related to crop yield.

Furthermore, the study aims to identify suitable factors and modeling techniques that contribute to higher accuracy and generality in predicting crop yield. This involves a comprehensive exploration of data mining approaches, with a specific focus on clustering (unsupervised) and classifi0cation (supervised) methods. The proposed framework integrates ensemble clustering using Enhanced Artificial Bee Colony Optimization and Shuffled Frog Leaping Algorithm, followed by classification using Enhanced KNN. The overall objective is to provide farmers with effective tools to improve crop productivity through advanced data mining techniques, leveraging insights from past and present attributes.

As the study delves into the challenges and limitations associated with existing data mining methods in predicting optimal agricultural conditions, it aims to contribute to the ongoing discourse on refining these methodologies. The agricultural sector's complexity demands continuous research and development efforts to address emerging challenges and harness the full potential of data mining tools.

In conclusion, the study's overarching goal is to increase the utilization of collected data in an effective manner, with a focus on enhancing the accuracy of crop yield predictions and reducing the occurrence of errors during the prediction process. The proposed ensemble-based clustering and classification framework represent a significant step towards achieving this goal, offering a comprehensive approach to data mining in agriculture. As India continues to play a central role in global agriculture, efforts to optimize and innovate within this sector become imperative for sustainable growth and food security.

2. RELATED WORKS

The paper [8] investigates grain yield prediction in wheat breeding by employing UAV-based multispectral imaging and ensemble learning methods. Notably, the study observes high prediction accuracy during the mid-grain filling stage under both full and limited irrigation treatments across various growth stages. The methodology involves planting 211 winter wheat genotypes, collecting multispectral data at different growth stages, and developing an ensemble learning framework with multiple base models. The results demonstrate the efficacy of the approach, showcasing improved accuracy compared to individual base models. While the study acknowledges the need for further improvements in ensemble learning models and the application of UAV-based multispectral data, specific limitations are not explicitly stated. Nonetheless, the research underscores the potential of this methodology for accurate predictions of complex traits like grain yield in wheat breeding.

The paper [9] introduces a Crop Yield Prediction Model (CRY) employing an adaptive cluster approach, dynamically updating historical crop data to enhance precision agriculture decision-making. CRY utilizes a bee hivemodeling approach for analyzing and classifying crops based on growth patterns and yield. The methodology involves model development, testing, and comparison with other cluster approaches. The study acknowledges the complexity of predicting crop yield due to multi-dimensional variable metrics and the challenge of an unavailable predictive modeling approach.

Additionally, potential limitations and challenges associated with the bee hivemodeling approach are recognized. The comparison with other cluster approaches is noted to be incomplete, emphasizing the need for further exploration in this domain.

The paper [10] explores the application of classification algorithms for predicting soybean yield based on a dataset collected over the years. It emphasizes the implementation of recent technology in agriculture, specifically leveraging data mining techniques to optimize the utilization of the gathered data. The focus is on soybean crop yield prediction, and the methodology involves discussing and implementing various classification algorithms within the realm of data mining. A crucial aspect of the study is the comparative analysis conducted to identify the most effective classification algorithm for soybean yield prediction. This research contributes to the advancement of agricultural technology, enhancing the precision of yield predictions through the utilization of sophisticated classification techniques.

The paper [11] focuses on predicting crop types based on location parameters through the implementation of ensemble techniques, specifically utilizing a combination of Decision Tree Regressor and AdaBoost Regressor. The study highlights the significant accuracy achieved through this ensemble approach. It emphasizes the use of machine learning algorithms for forecasting outcomes, particularly in the context of recommending suitable crops for cultivation based on prevailing weather conditions. While specific limitations are not explicitly stated, the paper suggests the need for further research, implying potential areas for improvement or additional investigation in the future. This research contributes to the field of precision agriculture by showcasing the effectiveness of ensemble techniques in enhancing accuracy in crop type prediction.

The paper [12] introduces a Bayesian Model Averaging (BMA) framework for enhancing predictions of maize yields in Liaoning Province, China. It employs a combination of multiple crop-growth models, including WOFOST, AquaCrop, and DNDC, using BMA weights to provide more reliable predictions. The BMA framework is highlighted for its efficacy in computing ensemble weights and interpreting mechanisms beyond the observed data. The methodology integrates predictions through a linear combination of ensemble members with BMA weights, strengthened by comparison with regional precipitation, fertilization, and radiation data. The study acknowledges challenges in accurately simulating crop production over large geographic regions with individual crop models and suggests further research to enhance the interpretation of BMA weight values by considering regional limiting factors such as precipitation, fertilization, and radiation data. This research contributes to the field by showcasing the power of the BMA framework in improving crop yield predictions.

The paper [13] introduces a machine learning framework for forecasting corn yield, showcasing the superiority of the proposed ensemble model over individual models and benchmark ensembles. The study focuses on three scales: county, agricultural district, and state level in the US Corn Belt states. The methodology involves employing machine learning models with environmental and management variables for different scenarios, considering complete and partial knowledge of in-season weather until August 1st. The data inputs, selected based on their agronomic relevance, include soil parameters, weather data, crop yield data, and management information at the state level. The optimized weighted ensemble and average ensemble models are demonstrated to provide precise and early corn yield forecasts, addressing research questions related to computational complexity. This research contributes to the advancement of corn yield forecasting, emphasizing the effectiveness of the proposed machine learning framework.

The paper [14] explores various clustering methodologies for identifying constant patterns in agricultural fields, emphasizing the superiority of multivariate functional principal components clustering over standard algorithms. The study evaluates different clustering techniques applied to multispectral satellite time series data to extract temporally stable patterns in agricultural fields. The methodologies include the application of the standard K-means clustering algorithm and novel clustering procedures directly to spectral reflectance time series. The effectiveness of these methods is validated through cluster accuracy estimates on a reference set of fields of interest. The research underscores the potential of multivariate functional principal components clustering for achieving better accuracy in identifying stable patterns, offering insights for sustainable management of agricultural fields.

The paper [15] investigates the impact of different parameters on crop production and employs various classification algorithms to predict crop yield. The study utilizes data collected and pre-processed from Kaggle, focusing on seven districts in India. The methodology involves data analysis using WEKA and the application of classification algorithms for crop yield prediction. Notably, the Support Vector Machine (SVM) algorithm demonstrates the highest accuracy at 76.82%, while the K-Nearest Neighbors (KNN) algorithm achieves the lowest accuracy at 35.76%. The evaluation metrics, including Relative Absolute Error (RAE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), contribute to validating the prediction accuracy of the classifiers. The research findings provide valuable insights into the efficacy of different algorithms for crop yield prediction in the specified districts.

The paper [16] underscores the growing significance of data mining in the agricultural domain, recognizing its popularity as a valuable tool. It highlights the crucial role of agriculture in the national economy and addresses the challenges faced by farmers in achieving expected yields. The primary focus of the paper is on utilizing data mining techniques to analyze historical data, with the ultimate goal of predicting crop yields. The methodology revolves around leveraging data mining to provide farmers with informed suggestions regarding crop selection and agricultural practices. This research emphasizes the potential of data mining to contribute significantly to the enhancement of decision-making processes in agriculture.

The paper [17] critiques simple linear methods for predicting crop yield, noting their omission of crucial factors such as climate, rainfall, soil, irrigation, and land characteristics. The primary objective is to develop scenario-specific algorithms for machine learning (ML) models, identifying the best fit model for yield prediction in India. Additionally, the paper proposes an ensemble approach that synthesizes various ML models to achieve a more comprehensive and accurate prediction of crop yield. The methodology involves the development of scenario-specific algorithms, model selection, and the synthesis of ML models for an improved overall prediction of crop yield. Limitations of the study include the absence of individual ML and Artificial Intelligence (AI) models tailored for the Indian context and the necessity of an ensemble approach due to the diverse geographical and environmental factors in the country. This research highlights the need for sophisticated models to address the multifaceted nature of crop yield prediction in India.

The paper [18] asserts the superiority of an ensemble technique in terms of prediction accuracy compared to existing classification techniques in crop yield prediction. The methodology employed in the study revolves around utilizing machine learning techniques and incorporates efficient feature selection methods, optimal feature selection, and the application of an ensemble technique to enhance prediction accuracy. The limitations acknowledged in the study encompass challenges faced by the farming community in maintaining traditional crop prediction methods due to rapid changes in environmental conditions. Additionally, concerns are raised about the need for efficient feature selection methods, the potential risk of unnecessarily complicating the model, and the impact on model accuracy due to additional features that contribute little to the machine learning model. This research contributes insights into the advantages and challenges associated with ensemble techniques in the context of crop yield prediction.

The paper [19] explores the integration of technology with crop yield prediction methodology, specifically focusing on the role of machine learning in optimizing outcomes for farmers and the agricultural industry. Emphasizing the importance of technology in enhancing predictive competency, the paper underscores that combining various types of algorithms results in highly effective predictions with minimal deviation in agricultural production. The methodology involves integrating technology, machine learning concepts, and different algorithms into the crop yielding prediction process. Additionally, the study incorporates the programming of the device with three types of models, showcasing the intervention of information technology to enhance various functions within the agricultural industry. This research contributes valuable insights into the potential of technology-driven approaches, particularly machine learning, to significantly improve predictive accuracy and overall efficiency in agricultural production.

The paper [20] emphasizes the critical role of effective crop discrimination methods in achieving precise measurements of crop area. The summary provides an overview of the paper, describing it as

a comprehensive review of crop yield prediction utilizing hyperspectral images while exploring research challenges and open issues in this domain. The methodology involves leveraging statistical and spatial data, utilizing remote sensing through satellite images, and employing optimal band selection from hyperspectral images. The study also includes a thorough review of existing literature on crop yield prediction with hyperspectral images. Acknowledging the limitations, the paper mentions the higher dimensionality of hyperspectral images and the necessity for further exploration of research challenges and open issues in this field. This research contributes valuable insights into the use of advanced imaging techniques for accurate crop yield prediction.

The paper [21] investigates the efficacy of machine learning techniques, specifically Random Forests and Support Vector Regression, in predicting crop production within the agricultural sector. The summary encapsulates the essence of the research, highlighting the examination of various machine learning algorithms and their effectiveness in predicting agricultural yield, drawing comparisons with traditional linear regression models. The methodology centers around the application of machine learning techniques, particularly Random Forests and Support Vector Regression, to forecast agricultural yield. The study further involves a comprehensive analysis of the effectiveness of diverse machine learning algorithms in the context of predicting crop production, contributing valuable insights to the ongoing discourse on advanced prediction models in agriculture.

This table 1 provides a concise overview of the methodologies, key findings, and limitations/challenges from the related works.

Table 1 Overview of the methodologies, Key findings, and Limitations

Paper	Methodology	Key Findings	Limitations/Challenges
[8]	UAV-based multispectral imaging and ensemble learning	High prediction accuracy in mid-grain filling stage; Improved accuracy compared to individual models	Need for further improvements in ensemble learning models; No explicit statement of specific limitations
[9]	Adaptive cluster approach with bee hivemodeling	Enhancing precision agriculture decision-making; Effective crop classification	Complexity in predicting crop yield; Challenges of unavailable predictive modeling approach; Incomplete comparison with other cluster approaches
[10]	Data mining techniques for soybean yield prediction	Implementation of recent technology in agriculture; Comparative analysis of classification algorithms	Not explicitly stated; Potential areas for improvement in the future
[11]	Ensemble techniques with Decision Tree Regressor and AdaBoost Regressor	Significant accuracy in predicting crop types based on location parameters	No specific limitations mentioned; Suggests further research
[12]	Bayesian Model Averaging (BMA) for maize yield predictions	Use of multiple crop-growth models with BMA weights; Efficacy in computing ensemble weights	Challenges in accurate simulation over large geographic regions with individual models; Need for further research
[13]	Machine learning framework for corn yield forecasting	Optimized weighted ensemble and average ensemble models for precise and early forecasts	No specific limitations mentioned; Addresses research questions related to computational complexity

[14]	Multivariate functional principal components clustering	Better accuracy in identifying stable patterns in agricultural fields	Not explicitly stated; Emphasizes the potential of the method
[15]	Classification algorithms for crop yield prediction in India	SVM achieves highest accuracy; Comparative analysis using metrics like RAE, RMSE, and MAE	No specific limitations mentioned; Provides insights into the efficacy of different algorithms
[16]	Data mining for agricultural decision-making	Utilization of data mining techniques to analyze historical data	Not explicitly stated; Emphasizes the potential of data mining in decision-making
[17]	Scenario-specific algorithms and ensemble approach in India	Developing algorithms based on different scenarios; Ensemble approach for comprehensive prediction	Absence of individual ML and AI models for the Indian context; Need for ensemble approach due to diverse factors
[18]	Ensemble technique for crop yield prediction	Utilization of machine learning techniques with efficient feature selection	Challenges in maintaining traditional crop prediction methods; Need for efficient feature selection methods
[19]	Integration of technology and machine learning for crop yield prediction	Combining algorithms for effective predictions; Programming device with three types of models	No specific limitations mentioned; Highlights the potential of technology-driven approaches
[20]	Hyperspectral images for crop yield prediction	Leveraging statistical and spatial data; Optimal band selection from hyperspectral images	Higher dimensionality of hyperspectral images; Need for further exploration of challenges
[21]	Machine learning techniques (Random Forests, Support Vector Regression) for crop production prediction	Application of Random Forests and Support Vector Regression; Comparison with linear regression models	No specific limitations mentioned; Contributes to the discourse on advanced prediction models in agriculture

3. PROBLEM IDENTIFICATION

Challenges in Decision-Making for Farmers:

India's agricultural landscape is vast and diverse, with varying climatic conditions, soil types, and regional peculiarities. This diversity poses a significant challenge for farmers in making informed decisions regarding crop selection.

Farmers struggle to navigate through the complexities of deciding suitable crops based on region, season, and soil conditions. This lack of clarity increases the risk associated with crop selection, impacting overall agricultural productivity.

Limited Access to Historical Data:

Historical crop yield data plays a crucial role in understanding patterns and making informed decisions. However, access to accurate historical data is limited, hindering farmers' ability to predict and mitigate risks effectively.

The absence of comprehensive historical crop yield records deprives farmers of a valuable risk management tool, limiting their ability to strategize and allocate resources efficiently.

Ineffectiveness of Existing Data Mining Methods:

While data mining has the potential to revolutionize agriculture, existing methods fall short in addressing the complexities of the sector. Current approaches often lack the precision needed for accurate predictions.

The inadequacy of current data mining methods hampers the ability to decipher patterns and extract meaningful information from vast datasets, limiting the understanding of the intricacies of the agricultural landscape.

Need for Ensemble-Based Approaches:

Individual models may not capture the diverse and dynamic nature of agricultural conditions. Ensemble-based approaches have shown promise, but there is a lack of a comprehensive framework that combines clustering and classification to enhance accuracy.

The absence of an integrated ensemble-based approach that combines clustering techniques and classification models impedes the development of a holistic framework for accurate crop yield predictions.

Complexity in Predicting Crop Yield:

Predicting crop yield involves multi-dimensional variables such as economic conditions, climatic nuances, soil quality, irrigation practices, and temperature variations. Existing models often struggle to address this complexity.

The complexity of predicting crop yield due to the myriad of influencing factors hinders the development of accurate and reliable models, affecting the overall precision of yield predictions.

Underutilization of Advanced Technologies:

The agricultural sector can benefit significantly from advanced technologies, including machine learning and data mining. However, there is a gap in the integration and utilization of these technologies for predictive purposes.

Underutilization of advanced technologies limits the efficiency and accuracy of crop yield predictions, hindering the optimization of agricultural processes and decision-making.

Lack of Comprehensive Evaluation of Existing Models:

Various models and frameworks exist for predicting crop yield, ranging from clustering to classification techniques. However, there is a need for a comprehensive evaluation to identify the most effective and generalizable approach.

The lack of a unified evaluation framework impedes the comparison and identification of the most suitable models, hindering progress in developing robust methodologies for crop yield prediction.

Inadequate Consideration of Geographic and Environmental Factors:

India's diverse geographical and environmental factors significantly impact crop yield. Existing models may not adequately account for these variations, leading to suboptimal predictions.

Inadequate consideration of diverse geographical and environmental factors limits the applicability and accuracy of existing models, highlighting the need for region-specific adaptations.

In summary, the identified problems underscore the complexity and challenges faced by the agricultural sector in India. The proposed ensemble-based clustering and classification framework aim to address these issues by providing a comprehensive and integrated solution. This research strives to empower farmers with accurate predictions, enhance decision-making processes, and contribute to the sustainable growth and food security of the nation. As India continues to play a central role in global agriculture, the need for innovative and efficient predictive models becomes imperative for the sector's advancement

4. LIMITATIONS AND RESEARCH GAP

The proposed ensemble-based clustering and classification framework for predicting agricultural crop yield, although promising, faces certain limitations that necessitate careful consideration. A primary concern revolves around the availability and quality of data, as the accuracy and reliability of any predictive model heavily depend on the completeness and reliability of historical crop yield datasets. Addressing this limitation requires future research to explore enhanced data collection methods and alternative data sources, such as satellite imagery, to mitigate biases and improve the overall robustness of the model.

Another critical limitation involves the potential lack of generalization across diverse agricultural regions. Given the substantial variations in climate, soil types, and farming practices, a model developed for a specific region may struggle to generalize effectively. To overcome this limitation, future research should focus on developing region-specific adaptations of the framework, ensuring its applicability and accuracy across different geographical and environmental contexts.

The scalability and computational complexity associated with ensemble methods represent another challenge. While powerful, these methods may introduce computational inefficiencies, particularly when dealing with large-scale agricultural datasets. Addressing this limitation requires future studies to optimize the computational efficiency of the ensemble-based approach, exploring techniques such as parallel processing and distributed computing to enhance scalability without compromising prediction accuracy.

The interpretability of ensemble models emerges as a critical concern, as these models, while known for their predictive performance, may lack transparency in decision-making. Future research should explore methods to enhance the interpretability of ensemble models within the agricultural context, incorporating techniques such as model-agnostic interpretability or leveraging domain knowledge to provide clearer insights into the factors influencing crop yield predictions.

The dynamic nature of agricultural systems introduces another layer of complexity, with static models struggling to adapt to evolving conditions. To address this limitation, research should explore the development of adaptive models that can continuously learn and update based on real-time or near-real-time data, ensuring responsiveness to changing agricultural conditions.

While certain studies focus on specific crops, the generalizability of the proposed framework across a wide range of crops remains underexplored. Future research should aim to validate the ensemble-based framework across diverse crops, considering variations in growth patterns, nutritional requirements, and responses to environmental factors.

Moreover, the integration of stakeholder perspectives, including farmers, supply chain industries, and policymakers, is critical for ensuring the practical relevance and usability of the framework. Future studies could explore methodologies for incorporating stakeholder feedback during the development and refinement of the framework, enhancing its alignment with real-world needs and challenges.

Lastly, ethical considerations associated with deploying advanced predictive models in agriculture, including resource allocation and economic disparities, warrant careful exploration. Future research should delve into the ethical implications of the proposed framework, considering its potential impact on resource distribution, socio-economic disparities, and the overall well-being of farming communities. Addressing these limitations and research gaps will contribute to the development of a more robust, adaptable, and ethically sound ensemble-based framework for predicting agricultural crop yield.

5. CONCLUSION

In conclusion, the presented ensemble-based clustering and classification framework for predicting agricultural crop yield offers a promising avenue for addressing the multifaceted challenges faced by the Indian agricultural sector. The study recognizes the intricate interplay of economic conditions, climatic nuances, soil quality, irrigation practices, and various other factors that significantly impact crop yield. By leveraging advanced data mining techniques, the proposed framework aims to empower farmers, supply chain industries, and policymakers with accurate predictions, informed decision-making tools, and valuable insights into the complexities of the agricultural landscape.

The identified challenges in decision-making for farmers underscore the critical need for innovative solutions that can provide clarity in selecting suitable crops based on diverse regional, seasonal, and soil conditions. The limitations of limited access to historical data and the ineffectiveness of existing data mining methods highlight the urgency to enhance data-driven approaches for improved risk management and precise predictions. The complexity in predicting crop yield due to numerous influencing factors emphasizes the need for comprehensive and adaptive models that can capture the dynamic nature of agricultural systems.

The ensemble-based approach, integrating Enhanced Artificial Bee Colony Optimization, Shuffled Frog Leaping Algorithm, and Enhanced KNN, presents a comprehensive solution to these challenges.

The simulation results demonstrate its effectiveness in comparison to existing classifiers, showcasing its potential to contribute significantly to the optimization and innovation within the agricultural sector. The framework's integration of ensemble clustering and classification techniques represents a substantial step towards achieving accurate and reliable crop yield predictions, thereby reducing risks associated with crop selection and improving overall agricultural productivity.

Future work in this domain should focus on addressing the identified limitations and research gaps to further enhance the framework's applicability and effectiveness. Firstly, efforts should be directed towards improving the availability and quality of historical data. Enhanced data collection methods, including the integration of satellite imagery, can help mitigate biases and provide more reliable datasets for training and validating the predictive models. Additionally, exploring region-specific adaptations of the framework is crucial to ensure its generalization across diverse agricultural regions, accounting for variations in climate, soil types, and farming practices.

The scalability and computational complexity associated with ensemble methods should be addressed through optimization techniques, such as parallel processing and distributed computing. This will ensure that the framework remains efficient and scalable, particularly when dealing with large-scale agricultural datasets. Enhancing the interpretability of ensemble models is paramount for gaining stakeholders' trust and understanding. Future research should explore model-agnostic interpretability techniques or incorporate domain knowledge to make the decision-making process more transparent.

The dynamic nature of agricultural systems calls for the development of adaptive models that can continuously learn and update based on real-time or near-real-time data. This adaptive capability will enable the framework to respond effectively to changing agricultural conditions, further improving its accuracy and reliability. The generalizability of the proposed framework across a wide range of crops should be validated through extensive empirical studies, considering variations in growth patterns, nutritional requirements, and responses to environmental factors.

The integration of stakeholder perspectives, including farmers, supply chain industries, and policymakers, is crucial for ensuring the practical relevance and usability of the framework. Future studies could incorporate methodologies for involving stakeholders in the development and refinement processes, aligning the framework with real-world needs and challenges. Moreover, ethical considerations associated with deploying advanced predictive models in agriculture should be thoroughly explored to ensure responsible and equitable use of the framework.

In conclusion, the ensemble-based clustering and classification framework presented in this survey paper holds great promise for revolutionizing the agricultural landscape in India. By addressing the identified challenges and limitations, future research endeavors can contribute to the development of a robust, adaptable, and ethically sound predictive model. As India continues to play a central role in global agriculture, the ongoing optimization and innovation within the agricultural sector become imperative for sustainable growth and food security. The proposed framework represents a significant step in this direction, laying the foundation for future advancements and transformative changes in the agricultural domain.

References

1. Elavarasan, D., Vincent, D. R., Sharma, V., Zomaya, A. Y., & Srinivasan, K. (2018). Forecasting yield by integrating agrarian factors and machine learning models: A survey. Computers and electronics in agriculture, 155, 257-282.
2. Kwaghtyo, D. K., & Eke, C. I. (2023). Smart farming prediction models for precision agriculture: a comprehensive survey. Artificial Intelligence Review, 56(6), 5729-5772.
3. Alex, N., Sobin, C. C., & Ali, J. (2023). A comprehensive study on smart agriculture applications in India. Wireless Personal Communications, 129(4), 2345-2385.
4. Elavarasan, D., & Vincent, P. D. R. (2021). A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. Journal of Ambient Intelligence and Humanized Computing, 1-14.
5. Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2020). Machine learning applications for precision agriculture: A comprehensive review. IEEE Access, 9, 4843-4873.

6. Ullah, F., Ullah, I., Khan, R. U., Khan, S., Khan, K., & Pau, G. (2024). Conventional to Deep Ensemble Methods for Hyperspectral Image Classification: A Comprehensive Survey. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.
7. Aboneh, T., Rorissa, A., &Srinivasagan, R. (2022). Stacking-based ensemble learning method for multi-spectral image classification. Technologies, 10(1), 17.
8. Fei, S., Hassan, M. A., He, Z., Chen, Z., Shu, M., Wang, J., ... & Xiao, Y. (2021). Assessment of ensemble learning to predict wheat grain yield based on UAV-multispectral reflectance. *Remote Sensing*, *13*(12), 2338.
9. Ananthara, M. G., Arunkumar, T., & Hemavathy, R. (2013, February). CRY—an improved crop yield prediction model using bee hive clustering approach for agricultural data sets. In 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering (pp. 473-478). IEEE.
10. Savla, A., Israni, N., Dhawan, P., Mandholia, A., Bhadada, H., & Bhardwaj, S. (2015, March). Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture. In 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) (pp. 1-7). IEEE.
11. Keerthana, M., Meghana, K. J. M., Pravallika, S., & Kavitha, M. (2021, February). An ensemble algorithm for crop yield prediction. In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (pp. 963-970). IEEE.
12. Huang, X., Huang, G., Yu, C., Ni, S., & Yu, L. (2017). A multiple crop model ensemble for improving broad-scale yield prediction using Bayesian model averaging. Field Crops Research, 211, 114-124.
13. Shahhosseini, M., Hu, G., &Archontoulis, S. V. (2020). Forecasting corn yield with machine learning ensembles. Frontiers in Plant Science, 11, 1120.
14. Pascucci, S., Carfora, M. F., Palombo, A., Pignatti, S., Casa, R., Pepe, M., & Castaldi, F. (2018). A comparison between standard and functional clustering methodologies: Application to agricultural fields for yield pattern assessment. Remote Sensing, 10(4), 585.
15. 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.
16. Chandana, C., & Parthasarathy, G. (2020, December). A comprehensive survey of classification algorithms for formulating crop yield prediction using data mining techniques. In 2020 IEEE International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET) (pp. 1-5). IEEE.
17. Kundu, S. G., Ghosh, A., Kundu, A., & GP, G. (2022). A ML-AI ENABLED ENSEMBLE MODEL FOR PREDICTING AGRICULTURAL YIELD. Cogent Food & Agriculture, 8(1), 2085717.
18. Raja, S. P., Sawicka, B., Stamenkovic, Z., &Mariammal, G. (2022). Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers. IEEE Access, 10, 23625-23641.
19. Ujjainia, S., Gautam, P., & Veenadhari, S. (2021). A Crop Recommendation System to Improve Crop Productivity using Ensemble Technique. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 10(4).
20. Mohan, A., & Venkatesan, M. (2020). Spatial data-based prediction models for crop yield analysis: A systematic review. Emerging Research in Data Engineering Systems and Computer Communications: Proceedings of CCODE 2019, 341-352.
21. Ramu, K., & Priyadarsini, K. (2021, October). A review on crop yield prediction using machine learning methods. In 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC) (pp. 1239-1245). IEEE.