Crop Yield Prediction: A Comprehensive Survey

Mohd Rafi Lone
VIT Bhopal University, Bhopal, India
mohdrafilone@vitbhopal.ac.in

KV Monish
VIT Bhopal University, Bhopal, India
kv.monish2020@vitbhopal.ac.in

Aryan Kashyap
VIT Bhopal University, Bhopal, India
aryan.kashyap2020@vitbhopal.ac.in

Aryan Singh Rajput
VIT Bhopal University, Bhopal, India
aryan.singh2020@vitbhopal.ac.in

Shrey Khanduja
VIT Bhopal University, Bhopal, India
shrey.khanduja2020@vitbhopal.ac.in

Sagar Maheshwari
VIT Bhopal University, Bhopal, India
sagar.maheshwari2020@vitbhopal.ac.in

***Abstract* —** This survey explores crop yield prediction, tracing its evolution from traditional to data-driven methods crucial for modern agriculture. Highlighting its impact on food security and economic sustainability, it dissects traditional approaches and reveals limitations, propelling the shift to advanced techniques. Focusing on machine learning algorithms like Random Forest and Artificial Neural Networks, it evaluates their strengths and limitations in crop yield prediction. The paper delves into data mining, emphasising association rule mining for pattern extraction. Integration of remote sensing and IoT into prediction models for real-time data collection is discussed, along with challenges like data quality [1]. Case studies illustrate practical applications, and the paper addresses ongoing research to overcome challenges. Looking ahead, it envisions hybrid models, considerations for climate change, and the role of explainable AI in building farmer trust [2]. This comprehensive survey blends empirical evidence, critical analysis, and future perspectives for a holistic view of crop yield prediction in contemporary agriculture.

I. Introduction

India's primary and most prominent culture has long been thought to be agriculture. Because the ancient people cultivated their own food on their own land, their necessities were met. As a result, natural crops are grown and utilised by a variety of animals and birds, including humans. A healthy and happy life is made possible by the green products that the creature took from the soil. The sector of agriculture is gradually deteriorating since new, creative technology and approaches have been developed. As a result of these numerous inventions, individuals have focused on creating hybrid, artificial goods that might lead to an unhealthy lifestyle. The growing of crops at the proper time and location is not something that modern people are aware of. Food insecurity results from these cultivation practices because they also alter seasonal climate conditions that are detrimental to basic resources like soil, water, and air. After examining all of these concerns and issues, including the weather, temperature, and other variables, there is no appropriate technology and solutions to get us out of our current predicament. There are several strategies available in India to boost agricultural economic growth. There are several approaches to raise and enhance the crop yield as well as crop quality.

II. TRADITIONAL APPROACHES

A. Farmer Expertise

Traditionally, farmers have heavily relied on their expertise, acquired through years of hands-on experience and generational knowledge. This experiential wisdom involves an intimate understanding of local soil conditions, climate patterns, and the nuances of cultivating specific crops. Farmers observe subtle changes in nature, such as the behaviour of animals or the colour of the sky, to make predictions about upcoming weather conditions.

B. Historical Data Analysis

Analysis of historical data, including records of past crop yields and environmental conditions, has been a foundational approach. Farmers and agricultural experts often refer to historical patterns to anticipate potential challenges and optimise crop selection. This method involves assessing trends over several years, identifying cycles, and adjusting planting and harvesting schedules accordingly.

Limitations*:* The major drawback of historical data analysis is its assumption that the future will mirror the past. In a rapidly changing climate, this assumption becomes less reliable. Moreover, historical data might not encompass extreme events or new challenges, rendering traditional approaches less effective in adapting to unforeseen circumstances.

C. Statistical Models

Simple statistical models, such as linear regression, have been used to predict crop yields based on various factors like rainfall, temperature, and soil quality. These models aim to establish quantitative relationships between different variables and forecast yields for upcoming seasons.

Limitations*:* While statistical models provide a more structured approach, they often oversimplify the complex interdependencies within agricultural systems. Linear models may struggle to capture nonlinear relationships, and their predictive accuracy diminishes when faced with intricate, multifaceted factors influencing crop yields.

D. Local Knowledge Exchange

Farmers within a community often exchange knowledge about successful practices and challenges. This informal network serves as a valuable source of information, allowing farmers to adapt their strategies based on the experiences of their peers. This approach fosters a sense of community resilience.

Limitations: While local knowledge exchange is robust, it tends to be localised and may not account for broader regional or global trends. Additionally, it relies on anecdotal evidence, which might not always align with scientifically validated practices.

E. Challenges and Ongoing Relevance

Traditional approaches have sustained agriculture for centuries and continue to play a crucial role in many regions. However, their limitations are increasingly pronounced in the context of contemporary challenges such as climate change, globalised markets, and the demand for precision agriculture. As a result, there is a growing recognition of the need to complement traditional methods with advanced, data-driven approaches to enhance the resilience and sustainability of agricultural systems.

III. DATA DRIVEN APPROACHES

A. Machine Learning Algorithms

Data-driven approaches in crop yield prediction have witnessed a paradigm shift with the advent of machine learning algorithms. Techniques such as Random Forest, Decision Trees, Support Vector Machines (SVM), and Artificial Neural Networks (ANN) have gained prominence. These algorithms analyse large datasets, identifying patterns and relationships that may not be apparent through traditional methods.

- Random Forest: This ensemble learning method constructs a multitude of decision trees during training and outputs the mode of the classes for classification problems or the mean prediction for regression problems. Random Forest is robust, handles non- linearity well, and can manage high-dimensional datasets [3]. The initial phase of this prediction method involves gathering and refining the database, followed by its integration with modules dedicated to weather and temperature prediction. Subsequently, the model is trained using the Random Forest Algorithm. Ultimately, by inputting parameters such as District Name, Crop Name, Area, Soil Type, the system predicts the yield with accuracy.

- Decision Trees: Decision trees offer a valuable approach for predicting crop yield by analysing historical agricultural data. Initially, gather and preprocess comprehensive datasets, focusing on factors like weather conditions, soil quality, crop types, and farming practices. Through feature selection, pinpoint the most influential variables affecting yield. Train the decision tree algorithm using this data, constructing a hierarchical structure to make predictions based on input variables like upcoming weather forecasts, soil characteristics, and crop types. Evaluate the model's accuracy by comparing predicted yields against actual outcomes and refine it as needed for improved performance. Decision trees excel in capturing complex relationships between diverse agricultural factors, offering transparent and effective predictions for crop yield estimation[4].

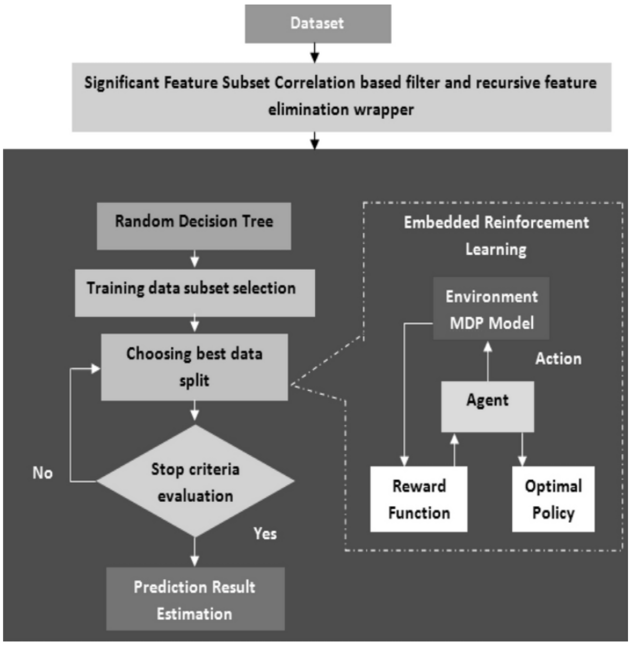
Dataset
Significant Feature Subset Correlation based filter and recursive feature elimination wrapper
Random Decision Tree
Training data subset selection
Choosing best data split
No
Stop criteria evaluation
Yes
Prediction Result Estimation
Embedded Reinforcement Learning
Environment MDP Model
Action
Agent
Reward Function
Optimal Policy

Fig. 2. Architecture of Random Forest Approach

- Support Vector Machines (SVM) stand out as powerful tools in supervised learning, excelling in both classification and regression tasks. Operating by discerning the optimal hyperplane to segregate data points into distinct classes, SVM showcases its prowess in high-dimensional spaces. Its versatility extends to handling diverse data types, making it applicable to a wide array of real-world scenarios. The underlying principle of SVM involves maximising the margin between classes, enhancing its robustness and generalisation capabilities. In high-dimensional contexts, where relationships may be intricate, SVM's ability to identify complex decision boundaries contributes to its effectiveness, rendering it a valuable asset in the realm of data-driven approaches for crop yield prediction and beyond.[5].

- Artificial Neural Networks: ANNs consist of interconnected nodes that process information. They excel at capturing intricate patterns in data and are highly adaptable to different types of agricultural variables [6]. Using Artificial Neural Networks (ANNs) for crop yield prediction involves gathering historical data on crop yields, weather, soil, and farming practices. After preparing this data, the ANN learns patterns between past yields and these factors. Inputting upcoming weather, soil details, and crop types, the ANN predicts future yields. Regularly comparing these predictions to actual results helps refine and improve the ANN's accuracy in forecasting crop yields, offering a valuable tool for informed decision-making in agriculture.

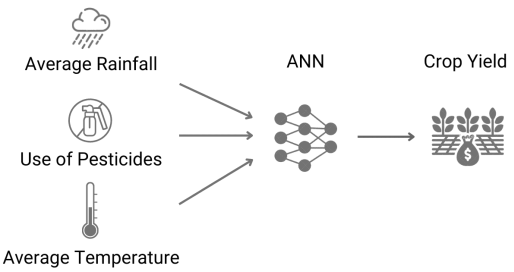

Fig. 2. Feed Forward Neural Network

B. *Strengths and Limitations*

Data-driven approaches offer several advantages. They can handle large and complex datasets, uncover non-linear relationships, and adapt to changing environmental conditions. However, they require substantial amounts of high-quality data for training, and their predictions may lack interpretability, posing challenges for user acceptance and trust in the model outcomes.

IV. INTEGRATION OF TECHNOLOGIES

The integration of cutting-edge technologies has marked a transformative phase in crop yield prediction, enhancing the precision and efficiency of forecasting models. Remote sensing and the Internet of Things (IoT) have emerged as pivotal components in this evolution. Remote sensing technologies, including satellites and drones, provide real- time data on various agronomic indicators such as soil moisture, crop health, and weather patterns. This wealth of information enables a more accurate assessment of crop conditions across large geographic areas. Simultaneously, IoT devices, deployed in the form of sensors and smart agricultural equipment, contribute to the continuous collection of on-field data. These devices monitor factors like temperature, humidity, and nutrient levels, creating a comprehensive dataset for analysis. The synergy of remote sensing and IoT facilitates a holistic understanding of the agricultural landscape, allowing farmers and researchers to make informed decisions about crop management practices. The integration of these technologies fosters a data-driven approach, minimising uncertainties and optimising resource allocation for sustainable and resilient agricultural systems. It underscores the role of real-time data in mitigating risks, enhancing productivity, and shaping the future trajectory of precision agriculture [1].

V. CHALLENGES AND SOLUTIONS

A. *Data Quality:* One of the primary challenges in crop yield prediction is ensuring the quality of the data used for training models. Agricultural datasets may contain missing values, outliers, or inaccuracies due to variations in data collection methods. Inconsistent data quality can lead to biassed models and inaccurate predictions [7].

Solutions: Implementing rigorous data cleaning and preprocessing techniques is crucial. This involves handling missing values, identifying and rectifying outliers, and ensuring consistency across diverse datasets. Collaboration between researchers, farmers, and data scientists is essential to improve data quality through standardised collection protocols.

B. *Interpretability of Models:* Many machine learning models, especially complex ones like neural networks, are often viewed as "black boxes" because understanding how they arrive at specific predictions can be challenging. In agricultural contexts, where interpretability is crucial for farmer trust and adoption, this lack of transparency poses a significant hurdle.

Solutions: Research efforts are directed towards developing explainable AI techniques. Model- agnostic interpretability methods, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), aim to provide insights into model predictions. Striking a balance between model complexity and interpretability is an ongoing challenge.[8]

C. *Scalability:* As the scale of agricultural operations varies widely, models designed for one region or farm may not be directly applicable to others. Scalability issues arise when attempting to generalise predictive models across diverse agricultural landscapes [9].

Solutions: Customization and adaptation of models to specific regions or agro ecological zones are essential. Transfer learning, where models trained on one dataset are fine-tuned for another, offers a potential solution. Collaborative efforts between researchers and local agricultural experts can enhance scalability.

D. *Standardised Datasets:* The absence of standardised datasets hampers the comparability of different models and impedes advancements in the field. Datasets from various sources might use different units, scales, or formats, making it challenging to create universally applicable models [7].

Solutions: Initiatives to establish standardised datasets that encompass diverse geographic regions, crop types, and environmental conditions are underway. Open data collaborations and repositories facilitate the sharing of datasets, fostering a more collaborative and cohesive research environment.

VI. CASE STUDIES

A. *Geographical Variability:* Case studies illustrate the adaptability of crop yield prediction models to diverse geographical regions. For instance, a model trained on Indian agricultural data might be applied to African contexts with appropriate adjustments. These case studies demonstrate the versatility and transferability of predictive models [2].

B. *Crop Specific Insights:* Examining case studies for specific crops provides insights into the unique challenges and opportunities associated with different agricultural products. A model developed for rice cultivation might highlight the importance of specific climatic factors, irrigation practices, or soil conditions, offering valuable crop-specific recommendations [10].

C. *Impact on Decision Making:* Effective crop yield prediction models have a tangible impact on farmers' decision-making processes. Case studies showcase instances where accurate predictions have led to optimised resource allocation, improved harvest planning, and enhanced overall agricultural productivity.

D. *Real-Time Monitoring:* Case studies often emphasise the real-time monitoring capabilities of advanced prediction models. The integration of technologies like remote sensing and IoT in these studies showcases their practical applications, demonstrating how timely data can empower farmers to make informed decisions throughout the crop growth cycle.

E. *Societal and Economic Impacts:* Beyond the agricultural domain, case studies shed light on the broader societal and economic impacts of accurate crop yield predictions. Improved predictions contribute to food security, sustainable resource management, and the economic well-being of farming communities.

VII. FUTURE DIRECTIONS

The future of crop yield prediction is poised for transformative advancements across various dimensions. One notable avenue is the exploration of hybrid models that amalgamate traditional statistical methods with machine learning algorithms, offering a synergistic approach to enhance predictive capabilities. Climate change considerations are increasingly vital, necessitating the integration of evolving environmental factors into prediction models to ensure accuracy amidst changing conditions. Explainable Artificial Intelligence (XAI) takes centre stage in addressing transparency concerns, providing farmers with comprehensible insights into model outputs. Blockchain technology and smart contracts emerge as potential game-changers, promising secure and transparent data management in agriculture. Real-time processing through edge computing holds promise for on-farm data analysis, reducing dependence on centralised servers. Citizen science initiatives and participatory approaches highlight the importance of involving farmers actively in data collection, enriching datasets with localised knowledge. Open data initiatives, collaboration, and ethical considerations underscore the collective commitment to responsible AI practices, ensuring fairness, transparency, and accountability in the future landscape of crop yield prediction. These forward-looking directions aim to foster resilience, sustainability, and inclusivity in agricultural data science.

VIII. CONCLUSION

After conducting an extensive and comprehensive review of the existing body of research dedicated to forecasting crop yields, it becomes increasingly evident that the traditional methodologies employed for such predictions are marred by inherent vagueness and a lack of accuracy. This deficiency poses a significant challenge in agricultural domains where precise yield estimations are pivotal for effective planning and resource allocation.

However, the integration of historical data archives and the application of machine learning algorithms mark a transformative leap in the realm of crop yield prediction. Techniques such as random forests, decision trees, and support vector machines represent a new frontier in leveraging data-driven approaches to anticipate yields. Their adeptness at processing intricate patterns within historical data not only surpasses but also redefines the limitations of conventional methods. This shift toward advanced machine learning methodologies significantly enhances the predictive accuracy and reliability of crop yield estimations. The ability to discern intricate relationships between various influencing factors, coupled with the adaptability to evolving agricultural scenarios, positions these sophisticated techniques as game-changers in the pursuit of more precise and informed agricultural yield forecasts.

IX. REFERENCES

1. Gigi Annee Mathew, Varsha Jotwani, A. K. Singh, 2023, Integration of Iot and Data Analytics for crop Yield Prediction and Resource Management, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 12, Issue 10 (October 2023), DOI : 10.17577/IJERTV12IS100109

2. Qi Zhang, Kaiyi Wang, Yanyun Han, Zhongqiang Liu, Feng Yang, Shufeng Wang, Xiangyu Zhao, Chunjiang Zhao, A crop variety yield prediction system based on variety yield data compensation, Computers and Electronics in Agriculture, Volume 203, 2022, 107460, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2022.107460.

3. N. Suresh et al., "Crop Yield Prediction Using Random Forest Algorithm," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021, pp. 279-282, doi: 10.1109/ICACCS51430.2021.9441871.

4. Aditya Shastry, K., Sanjay, H.A., Sajini, M.C. (2022). Decision Tree Based Crop Yield Prediction Using Agro-climatic Parameters. In: Shetty, N.R., Patnaik, L.M., Nagaraj, H.C., Hamsavath, P.N., Nalini, N. (eds) Emerging Research in Computing, Information, Communication and Applications. Lecture Notes in Electrical Engineering, vol 789. Springer, Singapore. https://doi.org/10.1007/978-981-16-1338-8_8

5. K. Priyadharshini, R. Prabavathi, V. B. Devi, P. Subha, S. M. Saranya and K. Kiruthika, "An Enhanced Approach for Crop Yield Prediction System Using Linear Support Vector Machine Model," 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), Chennai, India, 2022, pp. 1-5, doi: 10.1109/IC3IOT53935.2022.9767994.

6. Md Samiul Basir, Milon Chowdhury, Md Nafiul Islam, Muhammad Ashik-E-Rabbani, Artificial neural network model in predicting yield of mechanically transplanted rice from transplanting parameters in Bangladesh, Journal of Agriculture and Food Research, Volume 5, 2021, 100186, ISSN 2666-1543, https://doi.org/10.1016/j.jafr.2021.100186.

7. Talaat, F.M. Crop yield prediction algorithm (CYPA) in precision agriculture based on IoT techniques and climate changes. *Neural Comput & Applic* 35, 17281–17292 (2023). https://doi.org/10.1007/s00521-023-08619-5

8. Hari Sankar Nayak, João Vasco Silva, Chiter Mal Parihar, Timothy J. Krupnik, Dipaka Ranjan Sena, Suresh K. Kakraliya, Hanuman Sahay Jat, Harminder Singh Sidhu, Parbodh C. Sharma, Mangi Lal Jat, Tek B. Sapkota, Interpretable machine learning methods to explain on-farm yield variability of high productivity wheat in Northwest India, Field Crops Research, Volume 287, 2022, 108640, ISSN 0378-4290, https://doi.org/10.1016/j.fcr.2022.108640.

9. Lontsi Saadio Cedric, Wilfried Yves Hamilton Adoni, Rubby Aworka, Jérémie Thouakesseh Zoueu, Franck Kalala Mutombo, Moez Krichen, Charles Lebon Mberi Kimpolo, Crops yield prediction based on machine learning models: Case of West African countries, Smart Agricultural Technology, Volume 2, 2022, 100049, ISSN 2772-3755, https://doi.org/10.1016/j.atech.2022.100049.

10. Thomas van Klompenburg, Ayalew Kassahun, Cagatay Catal, Crop yield prediction using machine learning: A systematic literature review, Computers and Electronics in Agriculture, Volume 177, 2020, 105709, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2020.105709.