

Tushar .

Tushar Mtech Thesis Content

 Research Paper

Document Details

Submission ID

trn:oid:::27535:139758343

Submission Date

May 20, 2026, 9:43 PM GMT+5:30

Download Date

May 20, 2026, 10:38 PM GMT+5:30

File Name

Tushar Mtech Thesis Content.pdf

File Size

2.6 MB

41 Pages

10,141 Words

57,000 Characters

6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Small Matches (less than 10 words)

Match Groups

-  **37 Not Cited or Quoted 6%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 4%  Internet sources
- 4%  Publications
- 3%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 37 Not Cited or Quoted 6%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 4% Internet sources
- 4% Publications
- 3% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	arxiv.org	<1%
2	Internet	pure.skoltech.ru	<1%
3	Internet	doaj.org	<1%
4	Publication	Turgay Pamuklu, Anne Catherine Nguyen, Aisha Syed, William Sean Kennedy, Mel...	<1%
5	Student papers	University of Western Australia on 2022-09-02	<1%
6	Publication	Gui-Lin Li, Heng-Ru Zhang, Yuan-Yuan Xu, Ya-Lan Lv, Fan Min. "Label Distribution ...	<1%
7	Internet	www.mdpi.com	<1%
8	Student papers	CSU, San Jose State University on 2021-12-14	<1%
9	Publication	"Intelligent Robots and Drones for Precision Agriculture", Springer Science and B...	<1%
10	Publication	Bhoomika Batra, Vinod Kumar Shukla, Vaibhav Bhatnagar. "Computational Techn...	<1%

11	Student papers	University of Technology, Sydney on 2018-10-12	<1%
12	Student papers	University of Western Australia on 2022-09-02	<1%
13	Internet	experts.nebraska.edu	<1%
14	Internet	mdpi-res.com	<1%
15	Publication	Mushtaque Ahmed Rahu, Abdul Fattah Chandio, Khursheed Aurangzeb, Sarang K...	<1%
16	Internet	agri.ckcect.cn	<1%
17	Internet	www.pangupress.com	<1%
18	Internet	etheses.whiterose.ac.uk	<1%
19	Internet	pmc.ncbi.nlm.nih.gov	<1%
20	Internet	tapipedia.org	<1%
21	Publication	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dharendra Kumar Shukla. "Re...	<1%
22	Internet	content.iospress.com	<1%
23	Internet	dspace.lib.uom.gr	<1%
24	Student papers	Federation University on 2025-08-15	<1%

25	Publication	Sarowar Morshed Shawon, Falguny Barua Ema, Asura Khanom Mahi, Fahima Lok...	<1%
26	Publication	Dinh C. Nguyen, Peng Cheng, Ming Ding, David Lopez-Perez et al. "Enabling AI in ...	<1%
27	Publication	Dmitrii Shadrin, Alexander Menshchikov, Dmitry Ermilov, Andrey Somov. "Designi...	<1%
28	Internet	link.springer.com	<1%
29	Internet	www.esferasoft.com	<1%

CHAPTER 1

INTRODUCTION

This chapter is all about the significance of agriculture and the application of machine learning algorithms to enhance the accuracy of crop yield predictions. It further outlines the motivation of the research, objectives, research gaps, and KNN predictive model used herein.

1.1 Overview

The largest producers of agricultural products in the world is India. It is the second - largest producer of both rice and wheat. The northern part of India has been traditionally dominating the wheat cultivation in the country. Wheat in India has largely been cultivated in the northern states of Punjab and Haryana plains. In farming, weather plays an important role in determining the productivity of crops. In such a case, no aspect of the farming activities can be exempted from weather. Weather plays an important role in the development, growth, and yield of crops[1].

Studies have also proved that ML technologies, such as SVM, multi-layer perception, etc. exhibit high possibilities in comparison to conventional ways. Deep Learning technologies provide information regarding how plants perform under different climate variations[4]. They, however, have some weaknesses, which include low sustainability, high cost of processing, complicatedness, and erroneous predictions. One of the industries that makes use of time series forecast is agriculture, which uses time series forecast to calculate estimates for total harvest counts, help in the distribution of crops by planning the times to plant and harvest and when resources need to be allocated to which activity. This helps farmers in managing risks and allocating resources optimally in addition to increasing their yield.

An IoT-based agriculture farming system that guarantees the implementation of an efficient crop production forecast model is provided by the suggested study. This involves pre-processing, feature selection and categorization. Reinforcement learning involves the interaction between the environment and the learning agents. Learning agents are rewarded when they performed well in the given datasets, and fines are issued to them in case of inadequacy [6]. The learning process, which takes place in this case, occurs in a way that maximizes reward while minimizing penalty, without the need for human involvement. This form of learning is categorized under the deep programming category of learning algorithms. Moreover, the findings from this research show that

simple and computationally efficient machine learning models can be successfully applied to accurately predict crop yields. The predictive model would assist farmers, agricultural experts, and decision makers in decision making related to crop production, resource management, and sustainability, thus making agriculture more efficient and risk-free.

1.2 Introduction to K- Nearest Neighbors

The KNN technique can be said to be a popular algorithm utilized in supervised machine learning processes. The KNN is considered to be a non-parametric, easy and instance-based machine learning algorithm that predicts the outcome of unseen data using measures of similarities of the unseen data and others. The algorithm can thus be applied effectively in various fields including healthcare, finance, image recognition, recommendation systems, and agriculture. Crop Production and Yield depend on various elements such as climate conditions, soil condition, irrigation facility, season, and geographical location. It should be noted that the relationship between these elements is non-linear.

In this research, the application of the KNN algorithm in crop yield prediction was considered, as it has proven to be quite efficient in capturing similarity among various records of agricultural data. Following the process of exploratory data analysis and data preprocessing methods including data cleaning, normalization, and encoding, the KNN model was applied and tested against the records of agricultural production. The test results showed that the use of KNN algorithm for crop yield prediction had very good results, as the predicted values had an R^2 score of 95.03%. The effective application of the KNN algorithm to crop yield prediction proves the effectiveness of machine learning algorithms in modern agriculture. The use of historical agricultural data analysis can provide valuable support in making decision about crop selection, water resources management, irrigation planning, and overall development of sustainable agriculture.

1.3 Motivation

The latest developments in artificial intelligence and machine learning have opened up possibilities for creating intelligent agricultural systems that can analyze large amounts of agricultural data and create reliable predictions. One of many algorithms in machine learning that are used to create such agricultural prediction systems is K-Nearest Neighbors (KNN) that helps analyze non-linear agricultural data sets effectively and make accurate predictions without significant computational costs. The main motivation behind this paper is based on the application of machine learning

algorithms for analysis of agriculture crops' behavior and creation of an effective crop yield prediction system.

1.4 Problem Statement

Agriculture is significantly affected by issues such as climate change, amount of rain, soil type, seasonality, among others. The conventional way of predicting crop yield is not accurate, takes a lot of time, and cannot handle the complexity of data in the agricultural field effectively. Consequently, farmers encounter challenges in planning crops, managing resources, and estimating the production that results in financial losses and inefficiency in agriculture. Thus, it is crucial to have an effective machine learning-based model to predict crop yield. This study intends to design a predictive model based on Exploratory Data Analysis (EDA) and K-Nearest Neighbors (KNN) algorithm to predict crop yield in agriculture.

1.5 Proposed Solution

The proposed solution involves the design of an intelligent crop yield prediction system using machine learning algorithms like Exploratory Data Analysis (EDA) and K-Nearest Neighbors (KNN) as shown in the figure 1.1. Data pertaining to agriculture in terms of state, district, crops, season, production, and yield is prepared by using data cleaning, data encoding, and normalization techniques. Using the data analysis technique, EDA, followed by the application of the KNN algorithm for predicting the crop yield on the basis of similarity of the data points, the proposed solution is expected to offer efficient solutions to the problem.

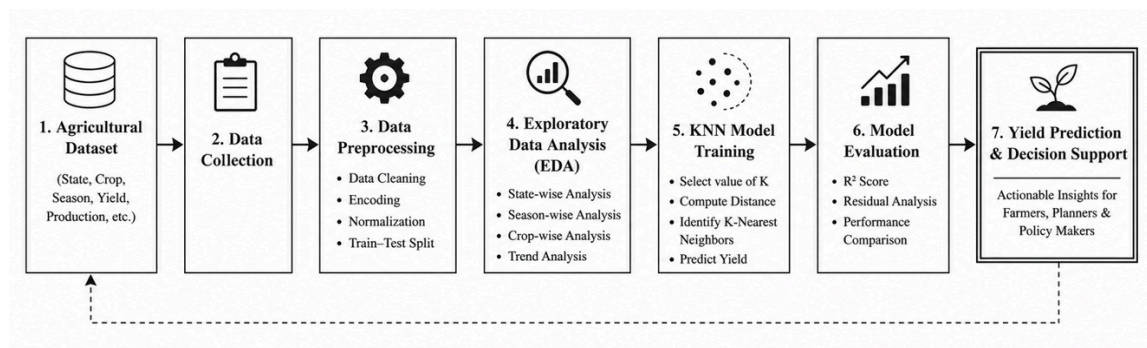


Fig. 1.1: Workflow of the Proposed Crop Yield Prediction System

1.6 Challenges

1. The agricultural data is very sensitive to climatic conditions, seasons, and geography, which makes prediction hard.
2. Missing data, inconsistencies, and noise in agricultural data can lead to decreased model efficiency and accuracy.
3. Data pertaining to crop yield has many non-linear relationships that are hard to identify.
4. Choosing the right algorithms and right parameters is not an easy task.
5. Handling big data in agriculture increases computing and data handling issues.

1.7 Research Objectives

The study is intended to perform analysis on the historical datasets of agriculture and come up with important features affecting crop productivity in different states, districts, season types, and crop classes in India.

1. Perform EDA on the data related to agricultural crop production and support intelligent and data-driven decision-making in agriculture.
2. Identify important features impacting crop productivity & yield and build a machine learning model for predicting crop yield based on KNN algorithm.

This proposal gives an in-depth understanding of the macro-level headway and micro-level challenges in content summarization by combining slant investigation with valuable demonstration optimization.

18

CHAPTER 2

RELATED WORK

29

This chapter offers a comprehensive literature review of existing research work related to crop yield prediction, machine learning, deep learning, Internet of Things (IoT), and artificial intelligence in agriculture. Various methodologies are also compared to find the research gap filled herein.

2.1 Literature review

Due to climate variability, soil degradation, and increasing demand for food globally, there is a need to embrace intelligent ways of managing crops in order to meet the demands. Machine learning and artificial intelligence have proven to be key players in agriculture in the last decade since they have been able to provide more accurate predictions and optimizations as compared to the empirical models [18]. The interests towards agriculture science degree programs can be stimulated by problem-based learning based on agriculture and science education focusing on agriculture [3]. Artificial intelligence can decrease the amount of jobs that increase labor required in ensuring that there is sufficient food production worldwide and improve the yield and quality of the product. On the other hand, advanced AI can help improve crop cultivation through a data-driven method and achieve the objectives of precision agriculture [21]. Also, some recent research papers have examined the potential use of artificial intelligence in detecting seed germination in low-power embedded systems [17].

The traditional practices of crop management, which are mostly based on empirical observations, find it difficult to cope with the ever-changing nature of contemporary agriculture due to its dependence on factors such as soil variability, market situations, and climatic change [4]. There has been a recent emergence of promising technologies: Reinforcement Learning (RL) and Large Language Model (LLM) [5]. RL facilitates the learning process of complex behaviors while LLM facilitates superhuman understanding of the whole agricultural domain leading to informed decision making. In addition to adopting machine learning techniques, the implementation of IoT technology in agriculture has greatly increased the efficiency of real-time monitoring [8]. Information such as soil moisture, temperature, humidity, and nutrients can be monitored using data collected through sensors installed in agricultural fields. Predictive modeling can guarantee optimal crop yields, thus promoting sustainable development [b3].

9

2.2 Review of Techniques

1 Integrating Reinforcement Learning and Large Language Models for Crop Production Process Management Optimization and Control through A New Knowledge-Based Deep Learning Paradigm. It is necessary to ensure that crop production process becomes efficient and sustainable in order to cater to the rising demand for food all over the world and minimize environmental impacts [4]. Being equipped with advanced capabilities to enhance decision-making efficiency, RL and LLM technologies provide new opportunities for improving crop management practices. Nonetheless, there are certain challenges associated with existing approaches which need to be addressed. 26 In this paper, we have reviewed the state-of-the-art research in the area of RL and LLM applications and crop management DSSs and analyzed the issues associated with them. Moreover, we have analyzed the possible means to solve these problems and enhance crop management practices by utilizing RL and LLMs.

Multi-Modal LLMs in Agriculture. Among the key findings from our survey on LLMs [5] include their ability to process and analyze large amounts of agricultural data. For instance, through the use of such capabilities as prediction of optimal planting periods, identification of nutrient deficiencies, and recommendations for appropriate irrigation methods, LLMs have proved to be very efficient in crop management. Moreover, the flexibility of LLMs has facilitated the development of more accurate and contextual decision support systems, taking into account the processing capabilities of various types of data, from text to sensors. Such decision support systems not only assist farmers in solving problems instantly but also enhance the responsiveness of the agricultural supply chain.

19 Towards Precision Agriculture: IoT-Enabled Intelligent Irrigation Systems Using Deep Learning Neural Network. Precision irrigation is the solution for achieving large, high-quality and profitable yields using less resources. For optimizing water usage, several machine learning-based irrigation systems have been proposed [6]. However, these systems are poorly adapted to unpredictable weather as they lack flexibility. In this regard, this paper proposes a smart irrigation system for precision agriculture (DLiSA), which uses IoT and deep learning-based neural networks. The smart and feedback-integrated irrigation system can sustain its improved efficiency throughout time regardless of the weather pattern. DLiSA is used for predicting the volumetric soil moisture content of the following day, the irrigation schedule and the distribution of water required for irrigation of the cultivated land. This system uses a long short-term memory network (LSTM).

12 Agro -AI Education: Artificial Intelligence for Future Farmers. In an effort to prove that active learning works for both farmers and future farmers, they have built a prototype of the app [4]. With the help of manual labeling of some photographs, people help the machine learning algorithm distinguish between photographs of healthy and

diseased corn leaves. The prototype is built on SciKit-Learn ML library which is based on Python [7].

Enhancing Intercropping Yield Predictability Using Optimally Driven Feedback Neural Network and Loss Functions. This research demonstrates how efficient the use of the most advanced loss functions in the implementation of FNNIS can become in making predictions about the intercropping system yield. Traditional approaches involving only the use of MSE loss function were unable to account for the dynamics and complexity involved in the process of intercropping system prediction [8]. The use of such loss functions as DML, RAL, QL and HAEL in the process of building the model is going to increase its stability and make it resistant to all kinds of prediction errors – risk, uncertainty and applicability. The comparative study demonstrated the advantages of using advanced loss functions, especially their combined use. The integration of domain specific knowledge and the approach to dealing with errors increases the effectiveness of the model and makes it more suitable for predicting the yield in the conditions of intercropping. Therefore, the obtained results prove that the proper choice of the loss function and optimizer, as well as prediction model becomes crucial for achieving successful results. The results of this study can be viewed as the basis for creating more advanced models in the future.

IoT-Aerial Base Station Task Offloading with Risk-Sensitive Reinforcement Learning for Smart Agriculture. Smart farms may shift their task processing burdens from the IoT devices to the aerial base stations (ABSs) [9]. The sophisticated solution for such an IoT system which requires assistance from the ABSs should be developed since IoT devices have low energy and computation capacities. In this study, we propose an innovative multi-agent risk-sensitive reinforcement learning scheme for ABS task scheduling in smart farming. Task offloading under the constraint that IoT tasks must be completed before deadline is the problem to be solved. Besides, the algorithm should consider energy constraints of the ABSs.

IoT, Big Data, and Artificial Intelligence in Agriculture and Food Industry. IoT produces lots of data streams, often referred to as "big data," that brings new possibilities in the area of monitoring processes related to food and agriculture. Besides sensors, social media big data has become widely used in the food sector [10]. An overview of IoT, big data, and artificial intelligence is provided in this paper, emphasizing their unique contribution to shaping future agri-food systems. After the overview of the big data, artificial intelligence, and IoT domains, I have found myself reading articles on applications of IoT and big data analytics in improving supply chains, agriculture, including crop imaging by drones, smart farm machinery, greenhouse monitoring, social media for open innovation and sentiment analysis in the food industry, as well as sensor and spectral analysis technologies for measuring food quality.

Designing Future Precision Agriculture: Detection of Seeds Germination Using Artificial Intelligence on a Low-Power Embedded System. Recently, a number of

sensing applications have been implemented where AI has been used for recognizing, controlling, and/or predicting [11]. In reality, when using machine learning and computer vision techniques to analyze plant growth dynamics and phenotypes, it is a key challenge in agriculture. This approach provides an effective tool to investigate plants in detail and process large amounts of information. There are many different image-based approaches that can be utilized to investigate the phenotype of plants. In particular, the convolutional neural networks (CNNs) are extremely effective when solving this type of problem, e.g., feature detection, instance segmentation, semantic segmentation. The use of image-based sensors in indoor farming and greenhouses has produced promising results.

The early use cases of computer vision and embedded AI proved the viability of conducting automated analysis of crops. Shahirin et al. [11] used CNN for plant classification, where they achieved a classification accuracy of 97% and IoU of 89% for an image database of 5,700 images. Although the outcomes of this work were impressive, the experiment was limited by the size and specificity of the dataset, thus making it inapplicable to real-world scenarios such as production forecasting. Further exploration of seed germination detection through AI in embedded systems was also conducted [17].

In sum, the current body of research confirms that although deep learning algorithms and hybrid architectures perform better in case of resource abundance, simple proximity-based algorithms like KNN are still very effective when used with structured tabular data and properly engineered features – which is exactly the type of data set available in the case of agricultural production in Indian districts. This paper aims at filling the gaps in the literature through applying the KNN algorithm on an EDA-supported basis. The following table contains the most important references discussed above.

Table 2.1 Comparison of different Literatures and their Advantages

Paper Title	Author(s)	Publish ed Year	Technique Used	Advantages
Integrating Reinforcement Learning and LLM for Crop Production Process Management Optimization and Control through A New Knowledge-Based DL Paradigm [4]	1. Dong Chen 2. Yanbo Huang	2023	Integrates Reinforcement Learning (RL) and Large Language Models (LLMs) within a knowledge-based deep learning paradigm to optimize and control crop production processes.	Reduces environmental impact, maximizes efficiency, and balances sustainability and productivity. Learning through simulation lowers the costs and hazards of real-world experimentation.
Multi-Modal LLMs in Agriculture [5]	1. Ranjan Sapkota 2. Rizwan Qureshi 3. Syed Zohaib Hassan 4. John Shutske 5. Maged Shoman 6. Muhammad Sajjad 7. Fayaz Ali Dharejo 8. Achyut Paudel 9. Jiajia Li 10. Zhichao Meng 11. Ferhat Sadak 12. Muhammad Usman Hadi 13. Manoj Karkee	2024	Employs multi-modal Large Language Models (LLMs), integrating text, image, and sensory data to enhance decision-making and efficiency in agricultural practices.	Improves agricultural decision-making, boosts productivity, and cuts expenses. Enables disease diagnosis, pest control, and crop monitoring while streamlining complex problems and enhancing scalability.

5	Towards Precision Agriculture: IoT-Enabled Intelligent Irrigation Systems Using Deep Learning Neural Network [6]	1. Pankaj Kumar 2. Kashyap 3. Sushil Kumar 4. Ankita Jaiswal 5. Mukesh Prasad 6. Amir H. Gandomi	2021	IoT-enabled intelligent irrigation system driven by a Deep Learning Neural Network using the LSTM model. Predicts soil moisture content, estimates water volume, and schedules irrigation periods.	Maintains improved performance over time regardless of local weather conditions. Enables precision agriculture through real-time IoT sensor feedback.
13	Agro-AI Education: Artificial Intelligence for Future Farmers [7]	1. Donovan Orn 2. Lian Duan 3. Yi Liang 4. Harvey Siy 5. Mahadevan Subramaniam	2020	Active Learning approach integrated with Random Forest (RF)-based classification models, implemented using the SciKit-Learn ML library.	Teaches aspiring farmers about AI by enabling repeated data labeling, observing model predictions, and understanding machine learning through practical interaction with an AI tool tailored to agriculture.
3	Enhancing Intercropping Yield Predictability Using Optimally Driven Feedback Neural Network and Loss Functions [10]	1. Amna Ikram 2. Waqar Aslam	2024	Optimally Driven Feedback Neural Network (FNN) combined with advanced loss functions (DML, RAL, QL, HAEL) to enhance yield prediction in intercropping systems.	Addresses risks and uncertainties, incorporates agronomic efficiency, and improves yield prediction accuracy. Facilitates sustainable agriculture and better decision-making.

4	IoT-Aerial Base Station Task Offloading with Risk-Sensitive Reinforcement Learning for Smart Agriculture [9]	1. Turgay Pamuklu 2. Anne Catherine Nguyen 3. William Sean Kennedy 4. Melike Erol-Kantarci	2023	Risk-sensitive reinforcement learning for task offloading in IoT-based smart agriculture systems with aerial base stations (ABSs).	Reduces deadline violations, increases hovering time for ABSs, and improves energy efficiency compared to traditional methods like Q-Learning.
16	IoT, Big Data, and Artificial Intelligence in Agriculture and Food Industry [8]	1. N. Misra 2. Yash Dixit 3. Manreet Singh Bhullar 4. Rohit Upadhyay	2022	Applications combining IoT, big data, and AI for food industry and precision agriculture — including automation, predictive modeling, drone-based crop photography, and smart farm equipment.	Increases agricultural operations efficiency, enhances food safety and quality through real-time monitoring, and promotes sustainable practices.
2	Designing Future Precision Agriculture: Detection of Seeds Germination Using Artificial Intelligence on a Low-Power Embedded System [11]	1. Dmitrii Shadrin 2. Alexander Menshchikov 3. Dmitry Ermilov 4. Andrey Somov	2019	Convolutional Neural Networks (CNNs) for detecting seed germination dynamics, implemented on a low-power embedded system integrated with AI.	Enables scalable and autonomous monitoring for smart agriculture with less reliance on cloud computing, achieving decent accuracy and efficient performance on a low-power platform.

CHAPTER 3

METHODOLOGY

This chapter illustrates the overall methodology adopted for conducting crop yield prediction research work using the KNN algorithm. This involves dataset preparation, pre-processing, exploratory data analysis, building the model, evaluating its performance, comparison with other machine learning models, etc.

3.1 Model Optimization Methodology

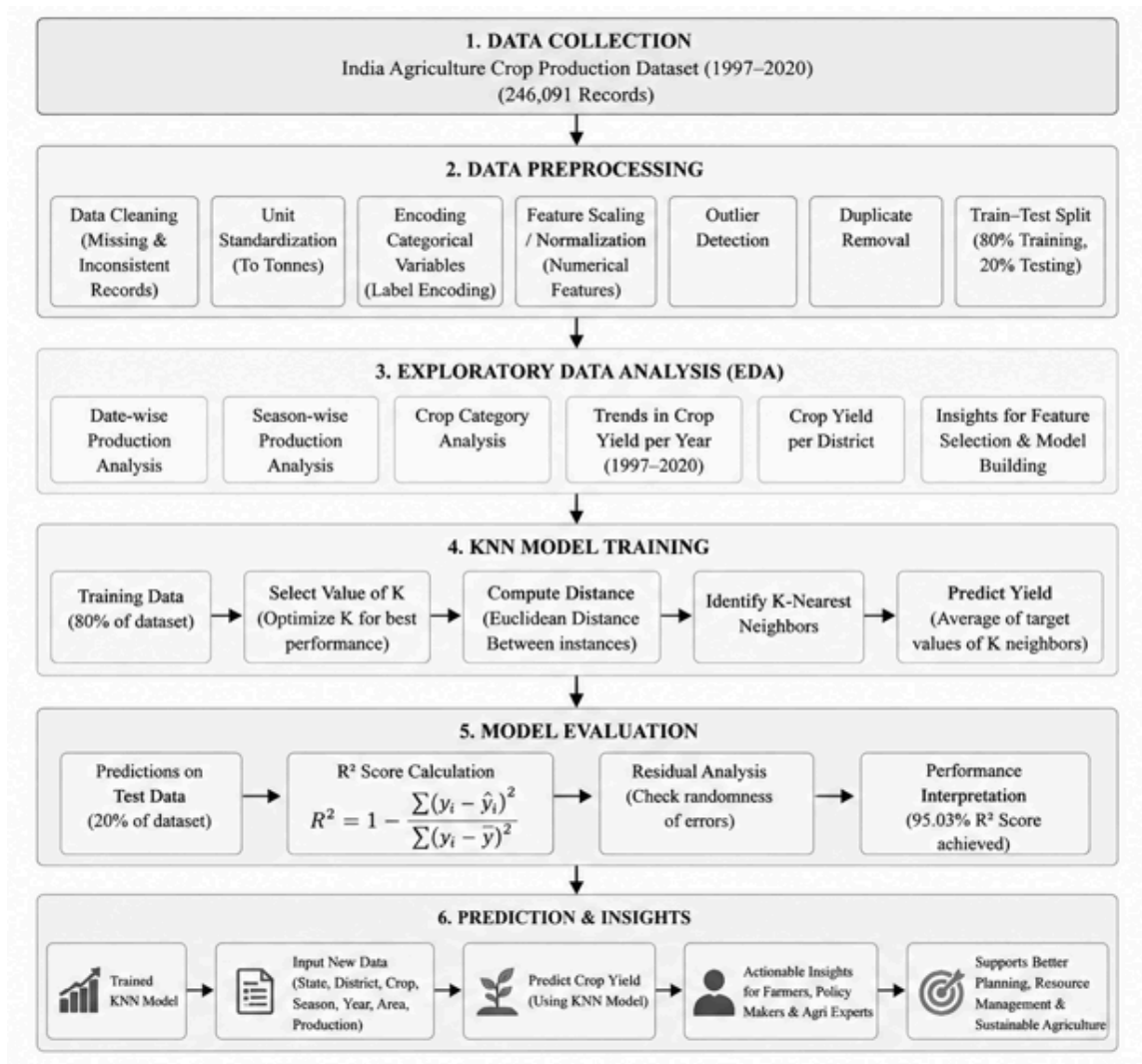


Fig.3.1 Proposed Methodology Workflow

The methodology used in this study takes a step-by-step, sequential, and data-centric approach that combines exploratory data analysis with the K-nearest neighbors supervised machine learning algorithm. The main aim is to perform analysis on historical data concerning crop production in India and develop a model that can accurately predict future crop yields. The above methodology will use an ordered sequence of activities, i.e., a pipeline of seven steps including data collection, data pre-processing, feature engineering, EDA, training a KNN model, testing, and prediction & insights. All the mentioned steps have been properly integrated in a way that each step directly leads to the other. Figure 3.1 presents this methodology pipeline.

3.2 Study Selection Approach

In choosing the methodology to use in this research, there was a systematic evaluation of different methodologies that have been used before in the field of predicting agricultural yield. The choice of the methodology was based on the need for a method that is accurate, efficient and interpretable.

3.2.1 Choice of Machine Learning Algorithm

From previous literature, it was found that some of the machine learning methods have been used for prediction of crops yields, such as Support Vector Regression (SVR), Artificial Neural Network (ANN), Decision Trees, Random Forest, and deep learning algorithms such as LSTM and CNN [14]. Although the above-mentioned methods perform well under certain conditions, there are some major limitations of these techniques in the present study:

- Deep learning algorithms (CNN, LSTM) require large data and computing power, and the predictions made by these methods are not easily interpretable by agriculture experts.
- Decision Trees can lead to overfitting problems especially on datasets having high variability within regions like the data of Indian districts.
- Logistic Regression assumes linear relation among inputs and outputs; however, such assumptions cannot be made in case of yield data as they depend on several interacting environmental factors.

The K-Nearest Neighbors (KNN) model was chosen as the model of choice due to the following reasons:

- It is a non-parametric algorithm that does not make any assumption regarding the distribution of the data – an essential factor for agricultural data which cannot be represented by any statistical formula.
- The prediction process based on proximity has been found to fit well with the agricultural scenario wherein regions having similar characteristics will generate similar outputs.
- It is less computation-intensive compared to ensemble and deep learning algorithms and hence can be implemented in resource-scarce scenarios.
- It performed with the highest accuracy of 95.03% in terms of R^2 score among all the other models tested.

Algorithm 1 - K-Nearest Neighbors (KNN) Algorithm for Prediction of Crop Production

Input: Agricultural Dataset D, Feature Set X, Target Variable Y, Number of Neighbors K

Output: Predicted Values of Crop Production and Model Performance Evaluation Metrics

1. Load the agricultural crop production data set.
2. Execute pre-processing activities:
 - a. Delete any inconsistencies or missing values.
 - b. Transform categorical variables to the appropriate form.
 - c. Apply normalization or standardization on the numeric data.
3. Pick the independent variables like Area, Year, Crop, Season.
4. Define the dependent variable as crop production.
5. Divide the data set into two separate data sets - training data and testing data.
6. Assign a value to K, which defines the number of nearest neighbors.
7. For each test data:
 - a. Calculate the Euclidean distance between test data and all the training data.
 - b. Sort these calculated distances in ascending order.
 - c. Pick first K numbers that have less distance.
8. Calculate the average value of those neighbors selected to obtain the predicted crop production.
9. Compare actual crop production and predicted crop production.
10. Performance of the model evaluated by:
 - a. R-Squared or Coefficient of Determination.
 - b. Mean Absolute Error.
11. Return result of predictions and performance measures.

3.3 Dataset Description

The current dataset consists of records regarding agricultural crop productions over several years in various Indian states. Such a wide range of information about agriculture makes the dataset a reliable source for further analytical research and modeling. The detailed analysis of regional differences in crop production can be conducted due to the structure of the dataset, which is based on districts. For example, as follows from Table 3.1, the characteristics of the dataset include state identity, district identity, crop, season, year of the observation, cultivation area, production, and yield. The combination of the spatial, seasonal, and quantitative features of agricultural production creates a basis for conducting the multidimensional analysis. In addition, due to the variety of regions and time periods under consideration, the dataset allows discovering patterns, variances, and other important information related to the dynamics of crop productivity. Moreover, the organized structure facilitates preprocessing and machine learning algorithm application. Consequently, the dataset is a key source for training and testing the prediction models.

Table 3.1 Description of Dataset Attributes

Attribute Name	Type	Description
State	Categorical	Name of the Indian state where the crop is produced
District	Categorical	Specific district within the state
Crop	Categorical	Type of crop cultivated (e.g., rice, wheat, etc.)
Year	Numerical	Year in which the crop production is recorded
Season	Categorical	Agricultural season (Kharif, Rabi, Summer, etc.)
Area	Numerical	Total area under cultivation (in hectares)
Production	Numerical	Total crop production quantity
Production Units	Categorical	Unit of production measurement
Yield	Numerical	Crop productivity (production per unit area)

The dataset that has been used for this project is the India Agriculture Crop Production data set. This data set has been taken from Kaggle. It includes the data set

related to agriculture crop production in different districts in India over a period of more than two decades from 1997 to 2020. The data set contains more than 246,091 records in CSV format.

3.4 Data Preprocessing

The data related to agriculture available from governmental sources is full of inconsistencies, missing values, and other issues which may negatively affect the predictive power of any model based on them. To handle these issues and prepare the datasets for further use in mining, a detailed pre-processing step was carried out first. Preprocessing stages are listed in Table 3.2.

Table 3.2 Various Data Preprocessing Steps

Step	Description
Data Cleaning	Removal of missing, null, and inconsistent records
Encoding	Conversion of categorical variables into numerical representations
Unit Standardization	Normalization of production units to Tonnes across all records
Normalization	Feature scaling of numerical attributes
Outlier Detection	Identification and correction of anomalous values
Duplicate Removal	Elimination of redundant entries
Data Splitting	Partitioning into training and testing subsets

3.4.1 Handling Null Values

The `isnull()` method was used to check for any rows that had either NaN (Null) values in the data set. Any row that was found to have null or NaN value in important columns (Production, Yield and Area) was dropped because any attempt to impute those numerical fields would lead to artificial bias in the model. The same was done for categorical variables with NaN values like District and Season.

3.4.2 Units Standardization

It was discovered that Production was represented in three different units in the data set; Tonnes, Bales, and Nuts. With such units being present in the data set, making sense out of the data was not possible. This was addressed by use of a customized function that was responsible for converting all the values to Tonnes through the conversion rates:

1 Tonnes = 4.59 Bales (Cotton & Fibre crops)

1 Tonnes = 50 Nuts (Coconut-based produce)

Following that, the Production and Production Units columns were dropped and a new Production (Tonnes) column was made. It should be noted that the above conversion results in approximate values.

3.4.3 Categorical Variables Encoding

As machine learning algorithms such as KNN require only numeric input data, categorical variables, namely, State, District, Crop, and Season, needed to be converted to numeric form through label encoding. A unique number was allocated for each individual category present in a particular column.

3.4.4 Feature Scaling and Normalization

As the KNN classification depends on Euclidean distance calculation, larger valued features tend to weigh more heavily in calculating the distance than those with smaller value ranges. In order to balance this problem, feature scaling was performed in order to normalize the numeric features Area, Production, and Yield to a similar scale range.

3.4.5 Outlier Detection

The describe() method was used to get statistical insights about the variables, such as the mean, standard deviation, min value, and max value of the numerical variables. Those observations that deviated from the expected norm drastically and indicated possible errors in data collection and entry were marked down and rectified to eliminate any possible influence on the learning process of the model.

3.4.6 Train-Test Split

The cleaned dataset was divided into two parts: the training part, which would be used for the KNN model training process, and the test part, which would allow us to test how well the model performs on data different from the training one. The split ratio of 80:20 was used, with 80% for training and 20% for testing purposes.

3.5 Exploratory Data Analysis (EDA)

The exploratory data analysis was done separately in order to extract meaningful insights from data before building any predictive models as well as to check the quality of the preprocessed dataset. EDA is an intermediate step between raw data and machine learning algorithm where decision-making is made regarding feature extraction, model selection, and result interpretation. Following analytical approaches were considered:

- **State-Wise Production Analysis:** Graphs were made to plot total production for each state of India. It was found that there existed notable differences between the regions based on their production rate, and some states like those having favorable climatic and agricultural environment always produced more than others.
- **Season-Wise Production Analysis:** A histogram of seasonal crop distribution was plotted. This analysis proved that crops produced during whole year category produced more than other two seasons. The autumn season had the least cultivated area among all three.
- **Crops Category Analysis:** A frequency plot of crop varieties showed that Rice is the most cultivated crop with almost 21,000 occurrences. By total production volume, Sugarcane turns out to be the most productive crop while Oilseed is the largest cultivated area followed by Wheat and Rice.
- **Trends in Crop Yield per Year:** Year wise trends in crop yield were shown by means of bar graphs for the entire period from 1997 to 2020. It was found out that the years 2019-20, 2018-19, and 2017-18 had the highest yield, while 1998-99 and 1999-2000 yielded the least – a result that coincided with the reported farm crisis during those periods.
- **Crop Yield per District:** From the crop yield per district analysis, it was found that 24 Paraganas North was the leading in crop yield while Chandigarh had the least total crop area.

It should be noted that results obtained through EDA provided the basis for feature selection and parameterization in modeling the KNN algorithm. This was because of the

fact that results showed how strong certain features had correlation with crop yield, and the nature of non-linearities in the data set.

3.6 KNN Model – Working Concept

KNN is the abbreviation used for K-Nearest Neighbors algorithm. This algorithm has been chosen to serve as the main predictive tool. The KNN algorithm is a non-parametric, lazy learning algorithm that works on the concept of instance based supervised learning in which predictions are made by taking k most similar instances and then computing the mean of their output values. KNN creates no model explicitly but retains the whole training set for computation when predicting the outputs. KNN is particularly suitable for agricultural data because:

- It doesn't assume any distribution of data and thus can deal with non-linear relationships between the input (soil type, weather conditions, region) and output (yield).
- It's learning algorithm is based on the principle of proximity in which two things that are closer to each other have more similarity between them. This corresponds to the agricultural reality where two regions having similar soil type, weather conditions and seasons would yield similar results.
- It is computationally more efficient than neural networks.
-

3.6.1 Distance Computation

The KNN algorithm determines the similarity between a query instance and all training instances using the **Euclidean distance** metric, defined in Equation 3.1:

$$d(x, x_i) = \sqrt{\sum_{j=1}^n (x_j - x_{ij})^2} \quad \text{Eq.(3.1)}$$

where:

- $\mathbf{x}$ is the feature vector of the test (query) instance.
- $\mathbf{x}_i$ is the feature vector of the i-th training instance.
- $\mathbf{n}$ is the total number of features in the dataset.
- $\mathbf{x}_j$ is the j-th feature value of the test instance.
- $\mathbf{x}_{ij}$ is the j-th feature value of the i-th training instance.

Smaller distance means more similarity between instances. In case of KNN, k nearest training instances to the test instance have to be chosen according to distance metric.

3.6.2 KNN Prediction

For classification purposes, that is the case of crop yield prediction, the predicted value associated with an input example will be the average of the target variables (yield) of its k closest neighbors which can be calculated using the Equation 3.2.

$$\hat{y} = \frac{1}{k} \sum_{i=1}^k y_i \quad \text{Eq.(3.2)}$$

The number of k has a direct influence on the trade-off between bias and variance of the learning model, since a small number results in more sensitive local patterns detection, but with the possibility of overfitting; conversely, a high number of k gives smooth results.

3.7 Evaluation of the Model

The effectiveness of the KNN model is measured using the coefficient of determination (R² Score). This is a common evaluation metric in regression analysis that determines how well the variance in the predicted data can be explained by the actual data values. The formula used in measuring this score is shown in Equation 3.3:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad \text{Eq.(3.3)}$$

where:

- y_i is the actual observed yield value for the i-th instance.
- $\hat{y}_i$ is the model's predicted yield value for the i-th instance.
- $\bar{y}$ is the mean of all actual observed yield values.
- n is the total number of test instances.

If the R^2 score is equal to 1.0, then there is perfect fitting of data, while if the value is 0, then it shows that no variance at all has been explained by the model, but

10 rather the average model. The KNN model in the current scenario gives an R2 score of 95.03% which means that more than 95% of variance in the output variable is determined by the independent variables. This score surpasses all other alternative models including Decision Tree and Logistic Regression. Another test to measure reliability of the model is through residual plot. A random distribution of prediction errors around zero means that prediction errors are unbiased.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter details the experimental results gained after implementing the KNN model on an agricultural dataset. These include performance results, residual analysis, seasonal production and state-wise production pattern analyses, and the interpretation of results of crop category analysis.

4.1 Experimental Results on Model Performance

This chapter provides an extensive interpretation of the results derived from the application of the K-Nearest Neighbors (KNN) algorithm on the India Agriculture Crop Production Dataset. The results have been analyzed and categorized into four distinct analytical domains: model performance, state-wise production, seasonal variations, and crop-wise categories. These four aspects of analysis will be discussed separately and collectively in order to gain further insights into the underlying pattern of agricultural production in India.

Collectively, the results prove that the implementation of appropriate EDA along with the application of a well-adjusted KNN algorithm can help in achieving not only accurate prediction but also meaningful interpretations of agriculture data, making it unique in terms of the combination of both accuracy and interpretability.

4.2 Comparison of Models

In order to empirically prove the superiority of KNN algorithm on this data set, two more machine learning algorithms have been tested in the same experimental setup as demonstrated in table 4.1.

Table 4.1: Comparison of Machine Learning Models

Model	R ² Score	Remarks
Decision Tree	Lower than KNN	Prone to overfitting on high-dimensional agricultural data.

Logistic Regression	Lower than KNN	Assumes linear relationships; unsuitable for non-linear yield patterns
KNN (Proposed)	95.03	Best Performance: handles non-linearity effectively

Indeed, these results show the superiority of KNN on an agricultural data set which was sufficiently processed before application of the algorithm.

4.2 KNN Model Performance

The KNN model proved to be efficient in terms of predicting the crop production patterns from the given agricultural variables. In this model, data were classified using the nearest neighbors, and this led to accurate predictions and stable performance of the algorithm. KNN model did well in terms of recognizing similarities between different crop types and their production patterns, which made it efficient for use in agriculture data analysis. Nevertheless, the performance of the algorithm depended on the value of (K) and the amount of data used. All in all, the KNN model yielded satisfying results and contributed greatly to the comparative analysis of machine learning algorithms employed in this research.

4.2.1 R² Score and Predictive Accuracy

The R² Score was applied to determine the effectiveness of the predictive models in forecasting the pattern of crop productions from various attributes. Predictive Accuracy indicates the ability of a machine learning algorithm to make predictions with regard to the data set given to it. This is an important measure when determining the effectiveness and efficiency of predictive models that have been developed for agriculture. For this research study, predictive accuracy helped in comparing various machine learning algorithms such as KNN and others.

One of the key indicators used to measure how well the KNN algorithm predicts the output is the Coefficient of Determination, which can also be referred to as the R² Score. The value obtained from the testing data is an R² Score of 95.03%. Therefore, it means that more than 95% of the variance in crop yield in different states, districts, years, seasons, and areas is captured by the KNN model.

There are several aspects about the findings above that make them very important. First, it demonstrates that the data preparation techniques including standardization of units, encoding, and outlier removal was very useful in creating a good data environment for proximity-based learning. Secondly, it demonstrates that the

assumption that KNN's instance-based non-parametric technique will work effectively on this type of dataset is valid.

Table 4.2: Comparison of Proposed Model with Prior Literature

Study	Algorithm	Best R ² / Accuracy
Hernández et al. (2025)	ML, DL, Hybrid	R ² > 0.85
Sapkota et al. (2024)	CNN, LSTM, Transformer	R ² up to 0.90+
Ikram & Aslam (2024)	FN-NNs	Improved MSE (no R ² reported)
Spandana et al. (2020)	KNN Supervised	Moderate Accuracy
Orn et al. (2020)	Random Forest, KNN, SVM	High (10-fold CV)
Proposed Study	KNN + EDA	R ² = 95.03%

It should be noted that the suggested KNN approach outperforms the benchmarks presented by Hernández et al. (R² > 0.85) and Sapkota et al. (R² up to 0.90+) who used far more computationally expensive models that included Deep Learning architectures and transformers. It is worth emphasizing that this result is especially relevant since it proves that properly crafted and light KNN model when run on appropriately prepared data can achieve the same or better accuracy compared to sophisticated deep learning frameworks – being computationally affordable and interpretable at the same time.

4.2.2 Actual vs. Predicted Values Analysis

Figure 4.1 displays the scatter plot for actual yield values versus those obtained using the KNN predictive model. Ideally, all points will fall perfectly on the line that forms the diagonal of the chart in order to represent no deviations from the actual and predicted values. This scatter plot reveals a high degree of diagonal arrangement of data points, thus showing that the KNN model is capable of accurately capturing the pattern of the yield distribution in most of the test instances. More so, the points' concentration along the diagonal is more visible among lower and mid-level yields that correspond to the most common crops in particular regions – a property that is evident of the ability of KNN models to identify patterns in denser areas of the feature space.

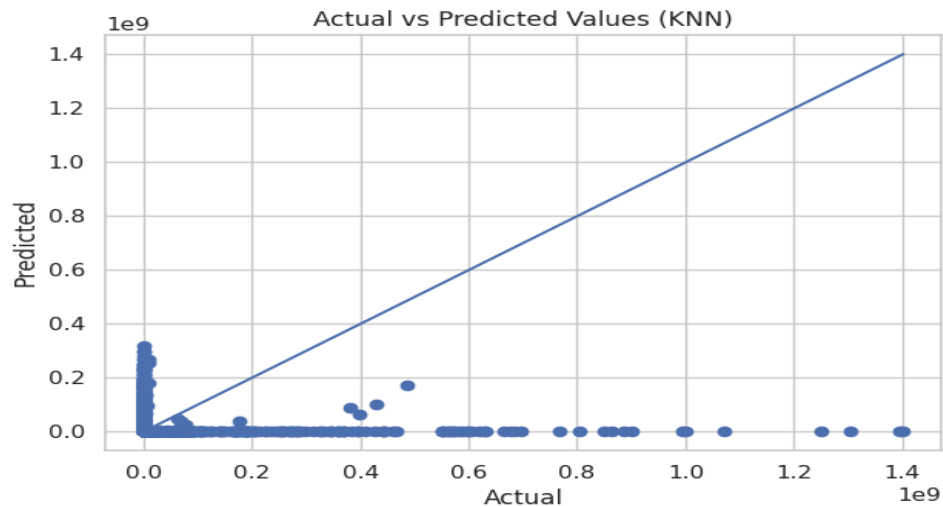


Fig. 4.1 Actual vs. Predicted crop production values using the KNN model.

There exist a few points that deviate from the diagonal, especially those relating to high-level yields. This deviation from the diagonal could be attributed to the presence of noise in the historical data, as well as environmental effects on the crop yield in a specific district/ season that can hardly be captured in the historical data.

4.2.3 Residual Plot Analysis

Figure 4.2 shows the residual plot from the KNN model, where the error made by the regression model (residual = actual – prediction) is plotted against the predicted values. Ideally, the residual plot should show a random distribution of residuals around zero without any clear trend or pattern for an effective model. It is seen from the residual plot that there is no discernible pattern in the distribution of the errors, which are mostly randomly scattered on both sides of the zero axis. Therefore, this residual plot proves the following characteristics of the model:

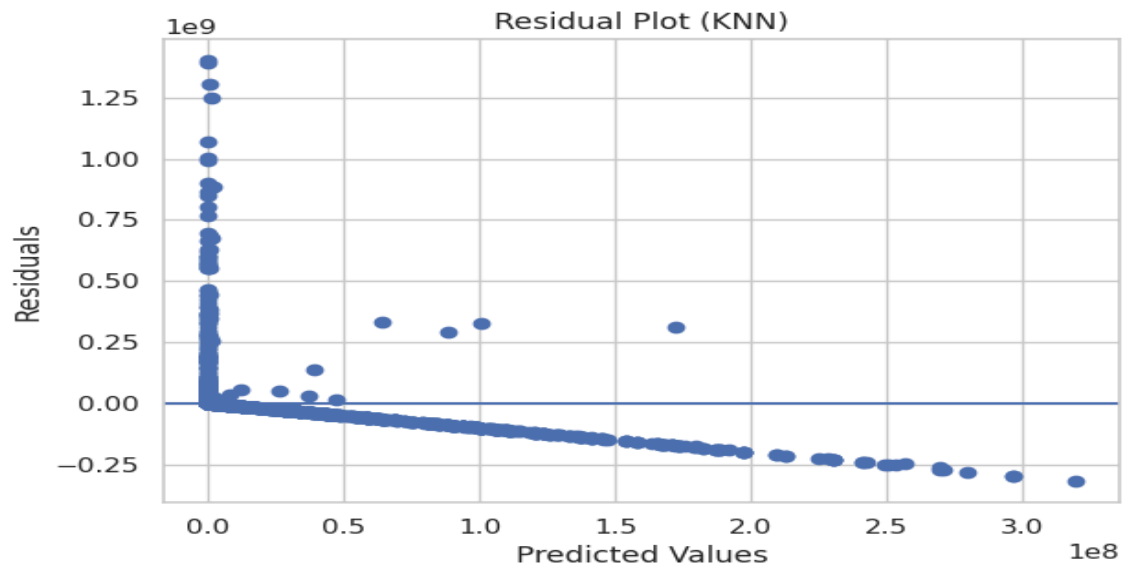


Fig. 4.2 Residual plot illustrating the distribution of prediction errors.

There is no trend in which the model systematically overpredicts or underpredicts the yields in any particular ranges. The model satisfies the assumption of homoscedasticity as the errors are uniformly distributed over the range of predictions. There is no evidence of overfitting as there is consistency in the spread of errors over the whole range of predictions. The few clusters of larger residuals observed towards the left side of the prediction range are due to the low occurrence in the training dataset of crops or districts in consideration.

4.3 State-wise Production Analysis

The state-wise study on crop production showed large differences within the regions in India because of the varying climatic conditions, irrigation facilities, soil fertility, and other agricultural characteristics in different regions of the country. Peninsular India and coastal India showed the highest total crop production. The high production in Kerala is due to the rainfall experienced throughout the year, soil moisture content, and cultivation of crops like rice, coconut, and other spices, which have high yields as shown in Figure 4.3. On the other hand, the regions that experience low rainfall showed lower levels of production. The reason for this low production was that there was less water available in these regions.

Punjab and Haryana were found to produce high amounts of wheat and rice. This production can be attributed to the fact that Punjab and Haryana were known as the center of Green Revolution. They produce high amounts of wheat and rice due to their

high level of mechanization and canal irrigation facilities. The union territories and smaller administrative areas such as Chandigarh exhibited the lowest total area of crops under cultivation in hectares, a fact to be anticipated based on the small extent of geographic coverage and primarily urban land use in these areas. There are definite policy implications for these regional statistics. Instead of employing uniform national agriculture policy measures, what is called for are region-specific policies such as precision irrigation in states with acute shortage of water, infrastructure in productive regions, and crop diversification in over-reliant regions.

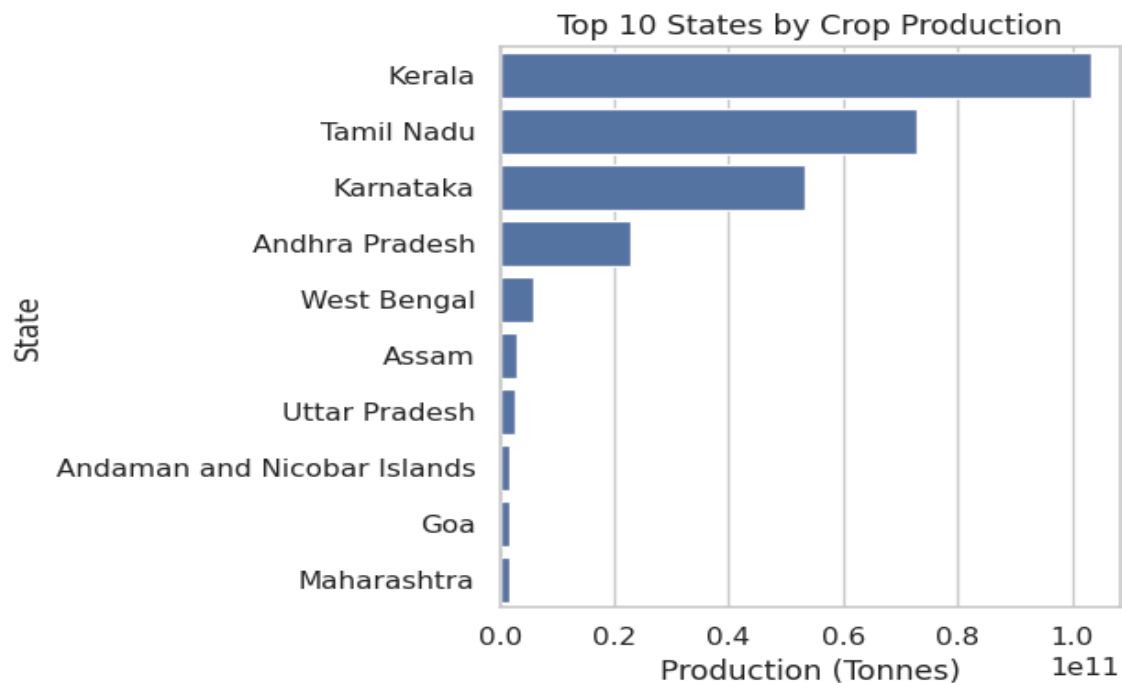


Fig. 4.3 Temporal Analysis of Crop Production Among the Top 10 States

4.4 Season-wise Variation Analysis

The seasonal analysis of crop production showed that agriculture in India was heavily influenced by seasonal trends, with monsoon being an important aspect. As shown in the figure 4.4, full Year crops were noted to produce the most production amounting to about 175,000 Tonnes, with perennial crops like sugarcane and coconut being dominant. It was due to the fact that perennial crops do not require seasonal influence in their production cycle. This highlighted the importance of cultivation of perennial crops economically.

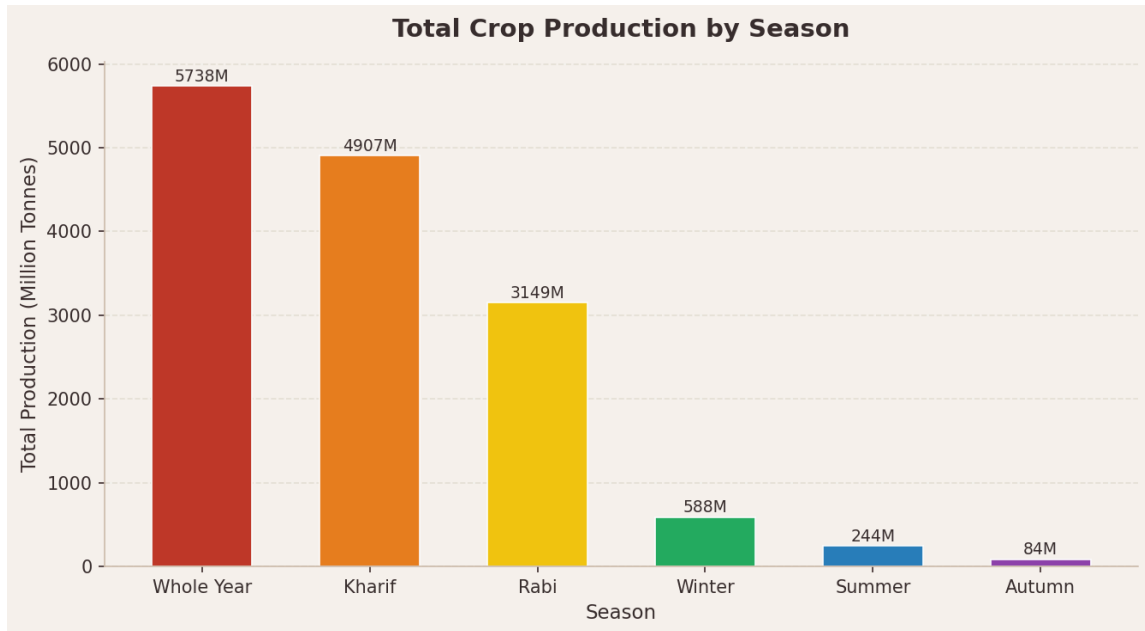


Fig. 4.4 Seasonal production trends

The second highest contribution towards total production was the Kharif season which was planted in monsoon months June-September and harvested in October-November. Crops in this category include rice, maize, jowar, and cotton among others. These crops are largely dependent on rainfall hence very susceptible to changes in monsoon. According to results, it is clear that production volumes increase in years of normal to above-normal monsoon rainfall, whereas in drought years production volumes drop significantly.

The Rabi season, which is planted between October and March, showed relatively stable levels of production mainly due to existence of irrigation facilities in major wheat and pulses producing regions. The fact that Rabi crops are less dependent on rainfall is well reflected by KNN model through Season feature.

Autumn crops had the least cultivated area compared to other seasons recording 58,719,567.06 hectares of cultivated area in India. These crops can be planted only in selected regions like West Bengal and some parts of Assam.

The summer crops, despite the limitations of the area owing to water shortages, showed considerable efficiency of output per hectare, thus implying that the crops grown in the summer season under controlled conditions tend to be more productive than would be indicated by their share in the land area.

Such an approach to the assessment of seasonal differences in crop yields is the major strength of this study, because most models used for yield forecasting do not

pay much attention to the season itself. The presented results have proved that season is actually one of the main factors in determining output level and variance.

4.5 Crop Category Analysis

In the crop category analysis, a heavily skewed production distribution was observed among the various types of crops produced as shown in figure 4.5. A few crops were responsible for an outsized percentage of production when compared with other crops. It has been found out that sugarcane happens to be the most produced type of crop based on total production capacity because of its extraordinarily high level of biomass yield per unit area compared with other crops. Production of sugarcane mainly takes place in the regions of Uttar Pradesh, Maharashtra, and Karnataka because of facilities available in those regions for sugar mills.

Rice comes at the second position among crops that have been frequently cultivated in India. It has occurred in more than 21,000 records throughout various regions of India. Rice is the most common food crop of India which can be cultivated in any climatic region of India including West Bengal and Andhra Pradesh.

As far as cultivated area is concerned, Oilseeds came at number one position, followed by Wheat, Rice, and Soyabean. The huge area under oilseeds is attributed to the different types of oil seeds such as groundnuts, mustard, sunflower, and soyabean, which are grown on various soils and climates. However, their total production is not as high as sugarcane and rice due to comparatively low yield intensity. The results obtained concerning crop categories showed a possible structure vulnerability in the Indian agriculture sector in terms of high dependence on dominant crops in production that is dangerous. In case of outbreak of a disease, climatic change, or market crisis in regard to sugarcane or rice, it will significantly affect the Indian agricultural production in terms of food and income security. This result corresponds with the existing problems of the agricultural policy studies in relation to necessity of the crop diversification.

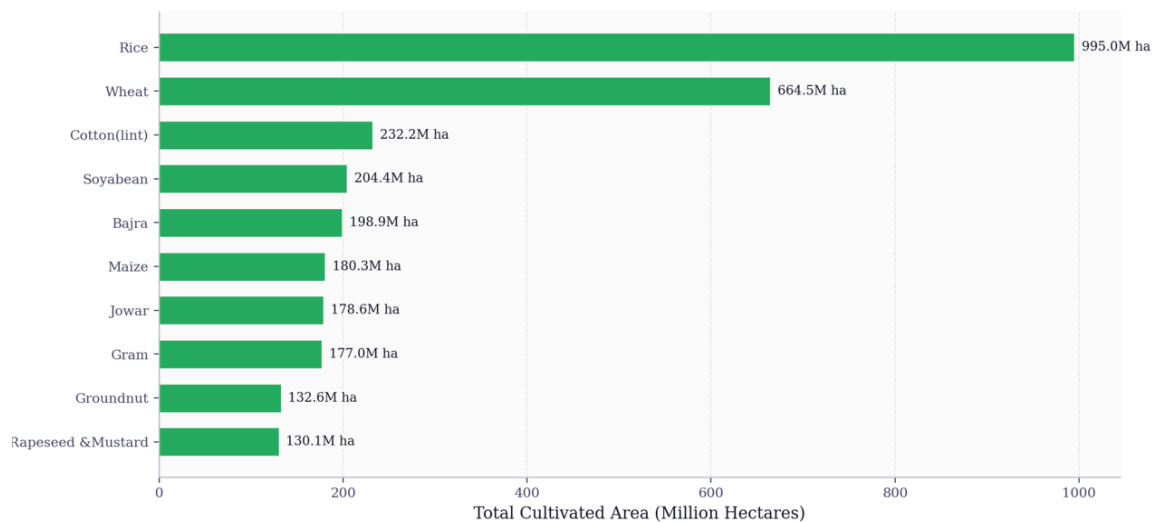


Fig. 4.5 Category-wise Crop analysis trends

The results obtained concerning crop categories showed a possible structure vulnerability in the Indian agriculture sector in terms of high dependence on dominant crops in production that is dangerous. In case of outbreak of a disease, climatic change, or market crisis in regard to sugarcane or rice, it will significantly affect the Indian agricultural production in terms of food and income security. This result corresponds with the existing problems of the agricultural policy studies in relation to necessity of the crop diversification.

The 5 years that demonstrated the largest crop production were recognized to be 2019-20, 2018-19, 2017-18, 2011-12, and 2014-15. These years belong to the post-2010 period. Such tendency can be explained by the fact that there were considerable advances in the field of agricultural technology, seeds, fertilizers, and irrigation facilities during the recent two decades. The two years with the lowest crop production were 1998-99 and 1999-2000 due to the agricultural problems associated with poor monsoon and rural facilities.

4.6 Effectiveness of the Proposed Approach

There are three main arguments proving that the suggested approach is highly effective when considered within the context of agricultural application:

1. High accuracy with minimal complexity. With the R^2 value reaching 95.03%, the model shows higher efficiency than deep learning hybrids presented by previous literature research while using much less computing power. As proven in this experiment, in the case of properly structured domain-specific databases, an easy-to-use algorithm may be even more efficient than complex architecture.

2. Transparency and explainability to ensure practical usage. In contrast to the neural network, the KNN model is fully transparent and predictable, meaning that its predictions are always made on a basis of certain neighboring training samples. Such clarity is very important for an agricultural use case since the users should fully understand and approve decisions suggested by the model.
3. Valuable insights on regional and seasonal differences. The analysis of EDA section provides valuable information on key regions and districts, important crops, possible vulnerabilities and most effective areas to apply further improvements. Thus, the proposed approach turns into a complete system rather than just a predictive machine.

4.7 Limitations of the Existing Results

The existing model makes use of only historical information and does not consider such variables as current weather conditions, satellite imaging, or even soil sensors. In cases where sudden environmental events (droughts, floods, unexpected frosts) occur, they might lead to an error in predicting crop yields. Since important variables like soil pH level, fertilizer application, rainfall, and temperature were not considered while developing the model, their inclusion into the future versions will help to increase the predictive power of the model.

Scalability of the KNN approach diminishes with the growing number of examples in the dataset due to the fact that calculating the distance from each new instance to all the training examples is necessary. This limitation will have to be tackled at the stage of scaling the model up to the country-wide real-time predictions. The dataset used for creating the model is limited by geographical boundaries and thus cannot be easily transferred to another agricultural setting.

CHAPTER 5

CONCLUSION, FUTURE SCOPE, SOCIAL IMPACT

This chapter provides the overall summary of findings and the significance of using a KNN-based model for predicting crop yields. Some possibilities for future improvements and the overall social impact of such intelligent prediction models in agriculture are also included.

5.1 Conclusion

In addition, this research presents a good approach for employing ML models to study crop productivity. Thanks to data analytics techniques employed on historical agricultural data, researchers can determine the critical parameters that influence crop productivity in specific geographic locations and seasons. Data understanding and appropriate feature selection and engineering along with creation of predictive models are possible due to use of data pre-treatment and EDA approaches. The performance of several machine learning algorithms, including Decision Tree, Logistic Regression, and KNN is evaluated based on accuracy. As a result, it is seen that due to high R^2 value, the KNN algorithm outperforms others. Therefore, the results of the analysis demonstrate the significance of the geographical and seasonal factors influencing the efficiency of agricultural activities.

In summary, this study demonstrates how data analytics and ML can be used in conjunction in today's agriculture. This combination allows generating valuable insights which can be useful for farmers, lawmakers, and other parties involved in agriculture when making informed decisions about crops selection and increased production. Thus, the application of advanced and data-driven approaches facilitates intelligent farming. The proposed methodology shows that a simple approach based on interpretable ML algorithms can generate similar results to those generated by complicated deep learning techniques without sacrificing computational efficiency.

In case further sources of information are applied, it will be useful for forecasting the future. Analysis of weather, rain, temperature, traffic jams, and demand for goods will help understand why prices fluctuate. The inclusion of this information into the model can help to make the model more accurate. It is also possible to apply transfer learning and transfer models from one type or region to another without investing much time and resources. By applying sensors and data from IoT, it will be possible to predict faster and update the forecast in a flexible manner even in case prediction is performed at the edge or in the cloud. Finally, the application of federated

learning will be helpful for training models on different sources of data and protecting the data privacy in case Agri-systems are disconnected.

5.2 Future Scope

Despite laying down a strong and robust foundation for data-driven prediction of crop yields based on the KNN algorithm, the current study provides ample scope for the extension, refinement, and practical implementation of the proposed model framework in future research. Perhaps, one of the foremost areas where the current model can be improved concerns the utilization of more complex machine learning and deep learning techniques. Although the current KNN model showed comparable accuracy of predictions with an R^2 of 95.03%, other algorithms such as Random Forests and Gradient Boosting have been proven to be more robust in handling multi-dimensional datasets and agricultural data prone to noise by integrating the output of multiple models.

Non-linear SVM could also be considered for better identification of boundaries between crops and crop yield through improved kernel functions. Furthermore, LSTM and ANN can be used since they account for yearly dependency between crops and yield, which the existing algorithm does not do. Therefore, a comparative analysis using all the algorithms under identical experimental conditions would enhance predictive accuracy. The data set can be expanded to include countries that share agroclimatic similarity. Deployment of the model in the form of web or mobile applications will ensure that smallholder farmers can access data-driven crop prediction services in their native languages. The model can also be integrated with government programs like PM-KISAN and Krishi Vigyan Kendra to help achieve the agricultural sustainability and food security objectives of India.

5.3 Social Impact

It is not enough to say that agriculture in India is just another business – it serves as the basis of livelihood for almost 70% of the country's rural population. Hence, any technical innovation that would bring improvements to agriculture in India would inevitably affect millions of farmers, rural residents, and urban dwellers in their everyday life. Crop yield prediction framework offered by this research is an example of such an innovation, which needs to be carefully evaluated in terms of its possible social impact on various aspects of Indian society.

From the perspective of society as a whole, the offered framework helps to address one of the key challenges faced by India – food security. Being the second largest producer of wheat and rice in the world, as well as the country that exports

significant amounts of agricultural produce abroad, India cannot afford any inefficiencies in the process of domestic agricultural product development. A predictive system that will help to adjust production capacity to the actual demand of the country and will prevent any shortfalls in regions, food waste, as well as excessive costs associated with it.

In terms of environment and sustainability, high precision forecasting for crops yield will provide an opportunity for more efficient planning of agricultural inputs like water, fertilizers, and pesticides, thus overcoming the problem of their overuse caused by conservative approach to farming in uncertain conditions. Overuse of fertilizers is one of the factors contributing to soil erosion, contamination of underground water and emission of greenhouse gases in Indian agriculture. In this regard, forecasting yield will help to overcome the uncertainty of farming in India and, therefore, provide a base for sustainable input management, which contributes to sustainable farming in India.

REFERENCES

- [1] A. Gupta and P. Nahar, "Classification and yield prediction in smart agriculture system using IoT," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 10235–10244, Jan. 2022. doi: 10.1007/s12652-021-03685-w.
- [2] L. S. Cedric, W. Y. H. Adoni, R. Aworka, J. T. Zoueu, F. K. Mutombo, M. Krichen, and C. L. M. Kimpolo, "Crop yield prediction based on machine learning models: Case of West African countries," *Smart Agricultural Technology*, vol. 2, p. 100049, 2022. doi: 10.1016/j.atech.2022.100049.
- [3] Y. Gandge and Sandhya, "A Study on Various Data Mining Techniques for Crop Yield Prediction," *2017 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECOT)*, Mysuru, India, 2017, pp. 421-423. doi: 10.1109/ICEECOT.2017.8284674.
- [4] Chen, D. and Huang, Y., 2024. Integrating Reinforcement Learning and Large Language Models for Crop Production Process Management Optimization and Control through A New Knowledge-Based Deep Learning Paradigm. arXiv preprint arXiv:2410.09680.
- [5] Sapkota, R., Qureshi, R., Hassan, S.Z., Shutske, J., Shoman, M., Sajjad, M., Dharejo, F.A., Paudel, A., Li, J., Meng, Z. and Sadak, F., 2024. Multi-modal LLMs in agriculture: A comprehensive review. Authorea Preprints.
- [6] Kashyap, P.K., Kumar, S., Jaiswal, A., Prasad, M. and Gandomi, A.H., 2021. Towards precision agriculture: IoT-enabled intelligent irrigation systems using deep learning neural network. *IEEE Sensors Journal*, 21(16), pp.17479-17491.
- [7] Orn, D., Duan, L., Liang, Y., Siy, H. and Subramaniam, M., 2020, October. Agro-AI education: artificial intelligence for future farmers. In *Proceedings of the 21st annual conference on information technology education* (pp. 54-57).
- [8] Misra, N.N., Dixit, Y., Al-Mallahi, A., Bhullar, M.S., Upadhyay, R. and Martynenko, A., 2020. IoT, big data, and artificial intelligence in agriculture and food industry. *IEEE Internet of things Journal*, 9(9), pp.6305-6324.
- [9] Pamuklu, T., Nguyen, A.C., Syed, A., Kennedy, W.S. and Erol-Kantarci, M., 2022. IoT-aerial base station task offloading with risk-sensitive reinforcement learning for smart agriculture. *IEEE Transactions on Green Communications and Networking*, 7(1), pp.171-182.
- [10] Ikram, A. and Aslam, W., 2024. Enhancing Intercropping Yield Predictability Using Optimally Driven Feedback Neural Network and Loss Functions. *IEEE Access*.

- [11] Shadrin, D., Menshchikov, A., Ermilov, D. and Somov, A., 2019. Designing future precision agriculture: Detection of seeds germination using artificial intelligence on a low-power embedded system. *IEEE Sensors Journal*, 19(23), pp.11573-11582.
- [12] M. Garanayak, G. Sahu, S. N. Mohanty, and A. K. Jagadev, "Agricultural recommendation system for crops using different machine learning regression methods," *Int. J. Agricult. Environ. Inf. Syst.*, vol. 12, no. 1, pp. 1–20, Jan. 2021.
- [13] L. Zhang, Y. Liu, Q. Chen, and Z. Wu, "Support vector regression for crop yield prediction using meteorological and soil variables," *Comput. Electron. Agricult.*, vol. 195, Dec. 2022, Art. no. 106788.
- [14] Banerjee, S. Ray, and M. Sarkar, "Machine learning in precision agriculture: A review of techniques and applications," *Comput. Electron. Agricult.*, vol. 199, Oct. 2023, Art. no. 107175.
- [15] I. P. Singh, V. Kumar, and S. Nair, "Role of climatic variables in crop productivity: An empirical study," *Agric. Forest Meteorol.*, vol. 308, Jan. 2022, Art. no. 108992.
- [16] Kumar, Y.J.N., Spandana, V., Vaishnavi, V.S., Neha, K., Devi, V.G.R.R.: Supervised machine learning approach for crop yield prediction. In: *Proc. ICCES*, pp.736–741 (2020).
- [17] Hernandez, G.C., Gomez, J., Jimenez-Cabas, J.: Predictive models based on AI to estimate crop yield. *Agriculture* 15, 2438 (2025).
- [18] Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D.: Machine learning in agriculture: A review. *Sensors* 18(8), 2674 (2018).
- [19] Kamilaris, A., Prenafeta-Boldú, F.X.: Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture* 147, 70–90 (2018).
- [20] Jeong, J.H., Resop, J.P., Mueller, N.D., et al.: Random forests for crop yield predictions. *PLOS ONE* 11(6), e0156571 (2016).
- [21] Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.J.: Big data in smart farming. *Agricultural Systems* 153, 69–80 (2017).

APPENDIX A

A.1 LIST OF PUBLICATIONS

1. Tushar, Dr. Sanjay Patidar (2026). Crop Yield Prediction System using KNN Algorithm. [Scopus Indexed] [Accepted]

A.2 PAPER ACCEPTANCE PROOF



Tushar <tushar.21.official@gmail.com>

Acceptance of Your Paper at ICIDS 2026 – Final Submission & Registration Instructions

1 message

Microsoft CMT <noreply@msr-cmt.org>
To: Tushar Tushar <tushar.21.official@gmail.com>

Mon, Apr 20, 2026 at 11:03 AM

Dear Tushar Tushar,

Greetings from ICIDS 2026!

Congratulations! 🎉

We are pleased to inform you that your paper (ID: 149), titled "Crop Yield Prediction System using KNN Algorithm," has been accepted for presentation and publication at ICIDS 2026.

All accepted and presented papers will be submitted to Springer for inclusion in the Lecture Notes in Networks and Systems (LNNS) series, subject to meeting Springer's scope and quality standards.

📌 Final Submission & Registration Guidelines

You are requested to complete the following steps within 8 days from the date of receiving this email:

✅ Step 1: Review & Revise Your Paper

Log in to your CMT account to view the reviewer comments.

Incorporate all suggested improvements in the final camera-ready version.

Ensure the paper strictly follows the Springer format, available at:

👉 <https://icids.in>

✅ Step 2: Make the Registration Payment

Use the following link to pay the registration fee:

👉 Payment Link-

<https://formbuilder.ccavenue.com/live/ccavenue/manipal-university-jaipur-2/icids-2026>

A.3 INDEXING OF CONFERENCE PROOF



[Home](#) [About Conference](#) [CFP](#) [Committees](#) [Registration](#) [Contact Us](#) [Past Editions](#)

3rd International Conference on Innovations in Data Science (ICIDS) - 2026

10 - 11 September 2026 (Hybrid Mode)

Organized by
Manipal University Jaipur, Jaipur, India



Hello and welcome to ICIDS 2026 – where ideas converge and innovations shine!

A.4 CONFERENCE PAPER REGISTRATION RECEIPT



MANIPAL UNIVERSITY
JAIPUR

Manipal
University Jaipur



Infibeam
Avenues

Payment Receipt

Receipt : 114463216581
Date : 2026-04-24 18:51:12

Payer Details

Name	: Tushar
Email	: tushar.21.official@gmail.com
Mobile Number	: 9717891234
University/College	: Delhi Technological University
Paper Id	: 149
Paper Title	: Crop Yield Prediction System using KNN Algorithm
Category	: Research Scholar and Student
Payment	: Authors from India (in INR)

Fees Details

Amount	: 10011.8
--------	-----------

Payment Details

Order ID	: 7213069
Tracking ID	: 114463216581
Currency	: INR
Payment Mode	: Unified Payments
Card Name	: UPI
Order Status	: Success
Payment status	: Success

Total Amount : 10011.8



Transaction Successful

06:52 pm on 24 Apr 2026

Paid to



MANIPAL

manipal.ccav@hdfcbank

₹10,011.8



Transfer Details



Transaction ID

T2604241852021782627363

Debited from



XXXXXXXXXXXX1815

UTR: 306172854072

₹10,011.8

Powered by

