A Framework for Crop Yield Prediction of Wheat in Pakistan using Remote Sensing and Deep Learning. A Case of South Punjab

PhD Thesis

By

Muhammad Ashfaq

24-FBAS/PhDSE/F16

Supervisor

Dr. Imran Khan

Assistant Professor, DCS, FCIT, IIU

Department of Software Engineering

Faculty of Computing & Information Technology

International Islamic University, Islamabad

2024

Deep learning (Machine learning)

Remote sensing - Data processing

Crop forecasting - Pakistan - Punjab South

Agriculture Data processing

Artificial intelligence - Applications in agriculture

Software engineering - "

A dissertation submitted to the
Department of Software Engineering,
International Islamic University, Islamabad
as a partial fulfillment of the requirements
for the award of the degree of
Doctor of Philosophy in Software Engineering.

Plagiarism Undertaking

I take full responsibility for the research work conducted during the PhD Thesis titled 'A Framework for Crop Yield Prediction of Wheat in Pakistan using Remote Sensing and Deep Learning. A Case of South Punjab.' I solemnly declare that the research work presented in the thesis is done solely by me with no significant help from any other person; however, small help, wherever taken, is duly acknowledged. I have also written the complete thesis by myself. Moreover, I have not previously presented this thesis (or substantially similar research work) or any part of the thesis to any other degree-awarding institution within Pakistan or abroad.

I understand that International Islamic University Islamabad has a zero-tolerance policy toward plagiarism. Therefore, I, as an author of the above-mentioned thesis, solemnly declare that no portion of my thesis has been plagiarized, and any material used in the thesis from other sources is properly referenced. Moreover, the thesis does not contain any literal citing of more than 70 words (total), even by giving a reference, unless I have the written permission of the publisher to do so. Furthermore, the work presented in the thesis is my original work, and I have positively cited other researchers' related work by clearly differentiating it from their relevant work.

I further understand that if I am found guilty of any form of plagiarism in my thesis work even after my graduation, the University reserves the right to revoke my PhD degree. Moreover, the University will also have the right to publish my name on its website, which records the students who plagiarized in their thesis work.

Muhammad Ashfaq ____________________

Date 07-06-2024

Department of Software Engineering

International Islamic University Islamabad

Date: *June 07, 2024*

Final Approval

It is certified that we have examined the thesis report submitted by *Mr. Muhammad Ashfaq*, Registration No. *24-FBAS/PhDSE/F16*, and it is our judgment that this thesis is of sufficient standard to warrant its acceptance by the International Islamic University, Islamabad for the Doctor of Philosophy in Software Engineering.

Committee:

External Examiners

Dr.Arif-ur-Rahman, Professor
Department of Computer Science
Bahria University Islamabad

Dr.Basit Raza, Associate Professor
Department of Computer Science
COMSATS Institute of Information Technology, Islamabad.

Internal Examiner

Dr. Syed Musharaf Ali , Assistant Professor
Department of Computer Science
Faculty of Computing & Information Technology
International Islamic University Islamabad

Supervisor

Dr. Imran Khan, Assistant Professor
Department of Computer Science
Faculty of Computing & Information Technology
International Islamic University Islamabad

Declaration

I hereby declare that this thesis, neither as a whole nor any part thereof has been copied out from any source. It is further declared that no portion of the work presented in this report has been submitted in support of any application for any other degree or qualification of this or any other university or institute of learning.

Muhammad Ashfaq

Dedication

I dedicate my dissertation work to my family, my wife, my brother and my sisters. A special feeling of gratitude to my loving parents, my greate father Rana Shamshad Ali and great brother Rana Ishaq, whose words of encouragement and push for tenacity ring in my ears. I also dedicate this dissertation to my many friends who have supported me throughout the process.

Muhammad Ashfaq

Acknowledgments

This thesis and all my efforts are fruitful only due to ALLAH Almighty, the Most Merciful and Beneficent, Who gave me strength to complete this task to the best of my abilities and knowledge.

I would like to thank my supervisor *Dr. Imran Khan* , who gave all their knowledge, guidance and support to boost my confidence and learning. I would also like to thank my wife who has supported me patiently and firmly during completion of my task.

I would also like to acknowledge my brothers, friends and my teachers especially *Dr. Shahbaz Ahmed Khan Ghayyur and Dr. Anwar Ghani and Professor Dr. Sohail Asghar* . All of them encouraged and provided logistic and technical help during this research.

I would like to admit that I owe all my achievements to my truly, sincere and most loving parents and friends who mean the most to me, and whose prayers have always been a source of determination for me.

Abstract

The primary cereal crop in Pakistan, wheat, currently accounts for 37% of the nation's total food energy consumption. The main Rabi crop is wheat, which has a high relevance tag for food security for the nation because it is the staple food. It is grown in all four provinces, particularly in Sindh and Punjab. Pakistani wheat growers encounter several difficulties. Regardless of whether a nation is poor, developing, or developed, agribusiness is the primary source of food supply worldwide. Due to its dependence on multiple factors, including internal and external ones, predicting Pakistan's wheat crop yield gap is difficult. This study aims to produce pre-harvest wheat crops with various unique internal and external factors in the local environment and to close the gap between expected estimates and actual pre-harvest yields. This study proposes a framework to predict crop yield using a hybrid approach of CNN-RNN models of deep conventional neural networks and recurrent neural networks [1]. Remote sensing images and factors are inputs and field studies validate predictive data. This study considers continued advances in image classification using Deep CNN and RNN [2] and software engineering processes for verification and validation, making it more relevant for crop inspection. Describes deep learning techniques in the field. Get the required dataset for crop yield gaps. Common approaches to predicting crop yields use crop simulation models or collect information from Pakistan's Crop Reporting Service (CRS). This study focuses on predicting wheat yields in the Multan District, Southern Punjab, Pakistan, using a CNN-RNN model with remote sensing capabilities. This framework recognizes complementary gaps in the achievements of existing machine learning and deep learning techniques as non-normative. In this study, the framework proposes a hybrid CNN-RNN model approach of deep convolutional neural networks and recurrent neural networks for crop yield prediction tasks. This technique takes remote sensing images and factors as input and validates and validates the input data in a software engineering process [3]. The approach presented in our study will be of value to policymakers in better identifying import strategies and decisions to address Pakistan's large-scale food security problem. While addressing prior findings, we discussed the current study's limits and concentrated on creating a deep convolutional neural network algorithm for forecasting winter wheat production, particularly using remote sensing data, climate data, and soil property data. We considered Pakistan's real yield difference after accounting for more than 58 factors.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This section presents the thesis vision for the readers' benefit. It provides a faultless report on the issue statement, research objectives, research questions, research scope, and thesis contextual background material. This part also introduces a research methodology employed to address the research topic. This will also explain why we chose this particular study methodology. The whole chapter outline and a thesis summary will be included at the conclusion.

1.1 Problem identification and motivation

Following a review of the literature, we conclude that predicting wheat crop yield in Pakistan is difficult because it depends on many variables, including internal factors (seed, disease, plant, etc.), external factors (weather, soil, irrigation, social economics, etc.), and manually calculated methods that do not provide precise estimates before harvest [6]. Traditional or out-of-date methods for estimating yields are labor and time-intensive, and yield data collected from a few villages do not represent the entire agricultural population's region [7, 8]. The literature review demonstrates the richness of this field. The motive and goal is to minimize the local convolutional manual method of estimating wheat crop production in Pakistan. With the help of this study, we could determine the crop yield accurately before harvest.

1.2 Aim of Research

The purpose of the study is to validate and verify the dataset in the context of the region to provide a framework for estimating wheat yield to choose the most effective machine learning forecasting methods for crops to categorize the ideal window of time for winter wheat preparation settings to look into the differences in yield forecast across local areas and the overall importance of many elements for Pakistan.

1.3 Scope of research

There are 58 criteria in all that were considered throughout the pilot research and the literature assessment. Eleven components comprise the scope of our research, including sub-factors that are important to the wheat crop and most pertinent to the local environment and literature. The literature study lists several machine learning and deep learning techniques; however, our experiment will only use four machine learning and two deep learning techniques.

1.4 Research methods

Create the study procedure utilizing a framework for software engineering. A hybrid method will be used, drawing on many fields such as software engineering, remote sensing, deep learning, and machine learning. There are several first steps. It will be a chronological procedure for gathering factor data and conducting experiments. We compare these factors after extracting the data from the literature using a systematic literature review procedure and after collecting the factor data following pilot research. We inspect, confirm, and validate if these factor data are present in the same or a different form in our immediate surroundings. In the future, as part of our research, we will conduct a field study to validate and verify those characteristics in a real context, either by an interview, a survey, or another method. Difference between SLR Variables and pilot study variables like environment, culture, etc.

1.5 Thesis outline

A quick summary of the thesis's following chapter is given in this section. Additionally, a description of the thesis' entire body can help readers better grasp the overall thesis.

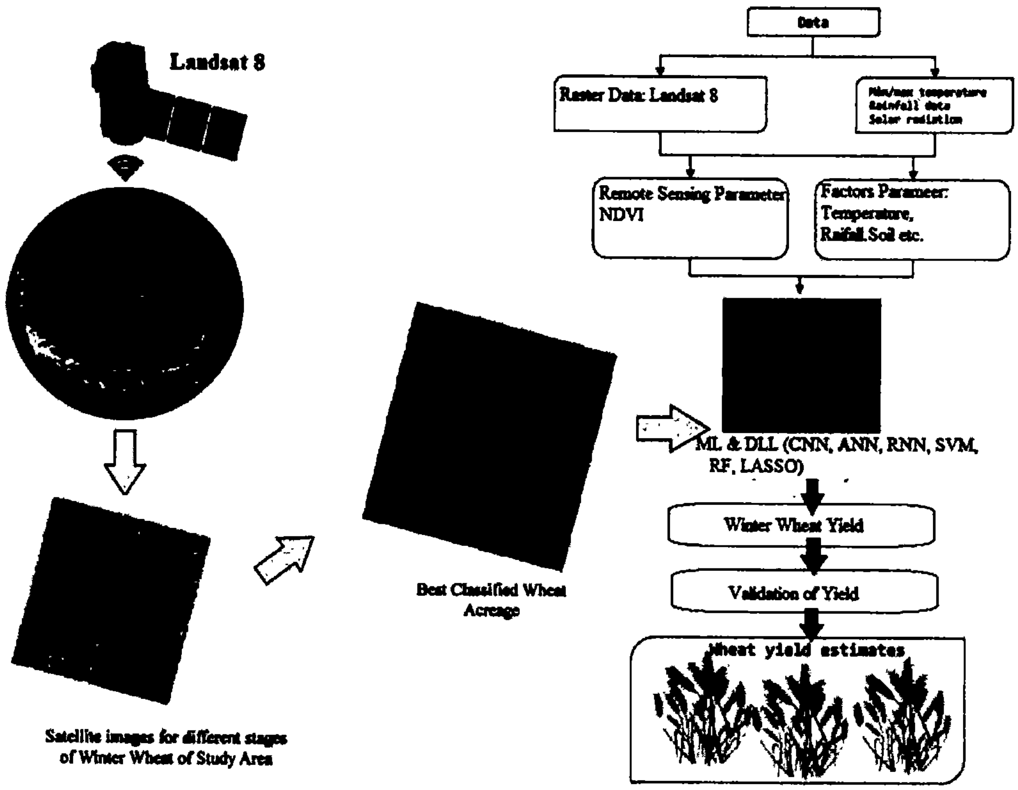

Figure 1.1: Overall Process of Wheat Crop Yield Prediction

1.5.1 Chapter 1: Introduction

This chapter thoroughly explains the thesis's goal and provides a thorough description of the issue statement, research questions, purpose, and scope of the study, as well as the research methodology.

1.5.2 Chapter 2: Theoretical background of all techniques

This chapter aims to introduce the significance of software development in remote sensing techniques, machine learning, deep learning, and software engineering (SE). Additionally, the idea of the machine learning development process will be provided.

STRUCTURE OF THE THESIS	
1. Introduction	• The thesis's goal is thoroughly explained in this chapter. • Provides a thorough description of the issue statement, research questions, purpose, and scope of the study.
2. Theoretical Background of all Techniques	• The goal of this chapter is to introduce the significance of software development in remote sensing techniques, machine learning, deep learning, and software engineering (SE). Additionally, the idea of the machine learning development process will be provided.
3. Review of Relevant Research and Theory	• Review of existing literature is described in this chapter. • The full literature assessment of all source code patterns, including their methodology, limitations in current tools, and methodologies is also included in this chapter.
4. Research Methodology	• This chapter's focus is on the suggested approach, the Process, the Algorithm, the Process.
5. Data Collection and Methods	• This chapter's focus is on the data collection and methods
6. Process, Results and Analysis	• This chapter's focus is on accumulating the thesis' results. • We describe and conclude the outcomes of our method.
7. Discussion	• This chapter's focus is on accumulating the thesis' results. • We describe and draw discussion from the outcomes of our method.
8. Conclusion & Future work	• This chapter's focus is on Conclusion and future work. STRUCTURE

Figure 1.2: Overall Structure of Thesis

1.5.3 Chapter 3: Review of relevant research and theory

The review of existing literature is described in this chapter. The full literature assessment of all source code patterns, including their methodology, limitations in current tools, and methodologies, is also included in this chapter.

1.5.4 Chapter 4: Research methodology

This chapter's focus is on the suggested approach, the Process, the Algorithm, the Process

1.5.5 Chapter 5: Data collection and methods

This chapter focuses on the data collection and methods that follow the development of a prototype, case studies of various source codes, and comparing our findings to those of previous research. Then, suggestions for further work are made.

1.5.6 Chapter 6: Process, Results, and Analysis

This chapter's focus is on accumulating the thesis' results. We describe and conclude the outcomes of our method. Also revealed are the research's shortcomings. Then, suggestions for further work are made.

1.5.7 Chapter 7: Discussion

This chapter's focus is on accumulating the thesis' results. We describe and draw discussion from the outcomes of our method. Also revealed are the research's shortcomings. Then, suggestions for further work are made.

1.5.8 Chapter 8: Conclusion and future direction

This chapter's focus is on the Conclusion and future work. We describe and conclude the outcomes of our method. Also revealed are the research's shortcomings. Then, suggestions for further work are made.

Chapter 2

Theoretical Background of Techniques

2.1 Software engineering process

Software engineering is the profession that specializes in developing, identifying, and modifying software to make building software easier, faster, and more durable [3]. The term refers to "a systematic approach to analyzing, designing, evaluating, implementing, testing, maintaining, and re-engineering software". H. Application of Engineering to Software". Program development, taken more broadly, refers to everything that takes place between the idea for the intended Programme and its actualization, ideally through a planned and disciplined procedure. In computer programming, it may also be used to describe the process of creating and maintaining source code. As a result, any activities that produce software products, including research, new development, modification, reuse, re-engineering, maintenance, and other activities, are considered to be a part of software development.

2.2 Machine Learning

All across the world, data is available. Thanks to machine learning, all of that data has now acquired meaning [9]. Additionally, anybody may utilize this technology and tool to leverage that data to find answers to queries. Machine learning, in a nutshell, is a technology that enables autonomous learning and development from past experiences without the need for explicit programming or human input.

2.3 Supervised Machine Learning

Supervised learning and unsupervised learning are the two different types of venture learning used in machine learning [10]. The process of learning a function that transforms an input into an output using the training dataset is known as supervised machine learning. The most popular supervised learning techniques are classification, support vector machines, random forests, and logistic regression. Making a dataset is the initial step in any form of prediction analysis. Categories were generated based on the labels and characteristics given to each dataset. A label is the actual output, whereas the feature is the actual input. These attributes might be categorized, binary, or continuous. The model keeps learning until the utmost degree of accuracy is achieved.

2.4 Artificial Neural Network

More than 100 billion neurons, also known as microcells, make up the human brain. The cell bodies of these neurons receive millions of pieces of information [11]. The synapses that link one neuron to another carry information. Artificial neural networks, which are used in computers, are modeled after the human central nervous system. The comparison between computers and brains is thus launched. The interactions between several layers of neurons are referred to as the "network" in the name "ANN." The input layer, the hidden layer, and the output layer are the three different sorts of layers that make up an artificial neural network. We first need to specify how many hidden layers we should take into account and how many neurons each hidden layer should have before putting any type of data into the input layer. The best way to do this is to Consider the neurons between the input and output layers. We initially fit the data to the first input layer before starting to initialize the neural network. From one neuron to the next, a set of weights is generated at random. To shift the function and improve accuracy, a bias that is created randomly is also present. The weight of the preceding synapse is doubled along with the activation node each time we travel to a new node, increasing bias. Place everything into a sigmoid function after that. The sigmoid function has the benefit of being nonlinear and having a gradual gradient. As a result, when using the sigmoid function, changing the X value tends to move the Y value closer to the curve's terminus, which is advantageous.

Then calculate the actual cost. Related to cost is how close we went to the actual output result. The lower the cost rate, the better you get precision. The cost is calculated by After calculating the total cost, we need to improve the model using a feedback back propagation algorithm. The algorithm

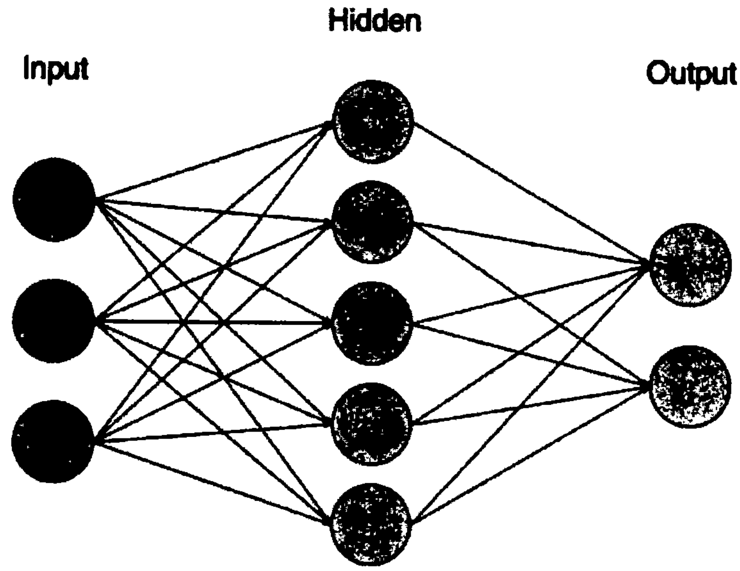

Figure 2.1: A simple Artificial Neural Network model

propagates backward, changing the values of the weights and biases to minimize the cost function. Backpropagation works in several ways to enhance the activation of specific output neurons.

2.5 Deep Neural Network

In our thesis, we used the backpropagation approach to train a deep neural network and three hidden layers to estimate the output's overall cost [2]. Large volumes of data may not be processed accurately with a single hidden layer. The computational expense of adding hidden layers can rise, but huge data can be used as a generalization.

2.6 Support Vector Machine

The main purpose of SVM is to design a hyperplane that splits all training vectors into two classes [12]. To separate the classes, draw several hyperplanes and choose the best hyperplane that maximizes the distance between the two classes. From the preceding picture, it is clear that A has a bigger margin value than B, hence we should consider it when differentiating between classes.

$$f(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x) + b \tag{2.1}$$

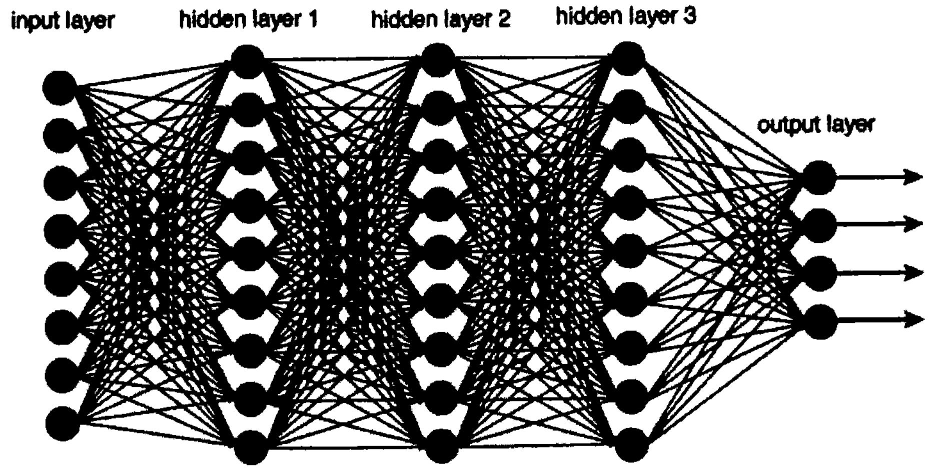

Figure 2.2: A simple Deep Neural Network model

Here, α_i are the Lagrange multipliers obtained during optimization, and $K(x_i, x)$ is the kernel function in Eq. (2.1). In your specific case, you'd need to define the features (xx) based on the data you have shown in Figure 2.3. These features could include rainfall, temperature, soil quality, and other relevant agricultural parameters for wheat crop yield prediction. This is a general representation; the implementation would depend on the dataset and specific requirements. Consult with agriculture or data science experts to fine-tune the model based on domain-specific knowledge and data availability.

2.7 Random Forest (RF)

Random Forest (RF) is a nonparametric method for regression tree analysis and advanced categorization, known for its resilience against overfitting and effectiveness with high-dimensional datasets [13]. On the other hand, SVM is a supervised learning model utilized for regression and classification tasks [14]. This study adopted a high-dimensional feature space, employing a kernel function (linear, Gaussian, polynomial, or hyperbolic tangent) for SVM regression. Specifically, the Gaussian kernel function was selected to explore the nonlinear relationship between input predictors (climate and remote sensing data) and output predictors (yield) [14]. Random Forest enhances predictive accuracy and mitigates overfitting by amalgamating multiple decision trees. The ensemble nature involves training numerous trees on distinct subsets of the data and averaging their predictions. Let's denote the Random Forest model as F(x), where x represents the input features for wheat crop yield prediction. Let's denote the Random Forest model as F(x), where x

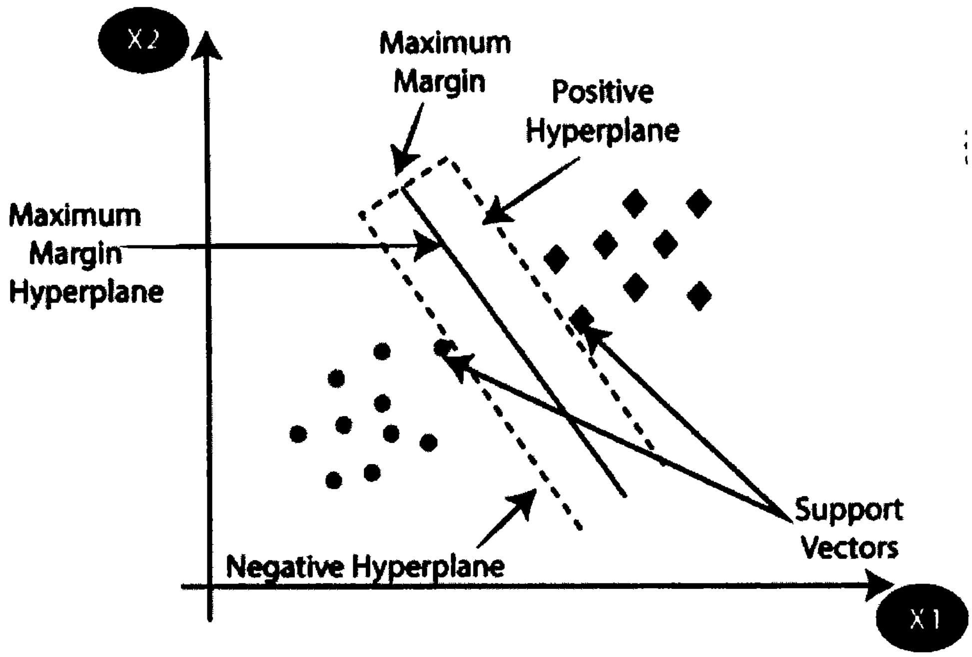

Figure 2.3: . A simple Support vector machine model[4]

represents the wheat crop yield prediction input features.

$$F(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \tag{2.2}$$

Here, N is the number of trees in the forest, and $f_i(x)$ is the prediction of the i^{th} decision tree in Eq. (2.2) and in Figure 2.3.

The prediction of an individual decision tree $f_i(x)$ is obtained by traversing the tree based on the input features. Each leaf node in the tree represents a predicted value. Scikit-learn rest measures feature importance based on how much each feature contributes to the reduction in impurity (e.g., Gini impurity) across all trees. Random Forest has hyperparameters that need to be tuned, such as the number of trees (NN), the maximum depth of each tree, and the number of features considered at each split. In practice, the implementation involves training the Random Forest on historical data with known crop yields and then using the trained model to predict the yield for new data. This mathematical representation provides an overview of the Random Forest model for wheat crop yield prediction [4]. The actual implementation would involve using an ML library (e.g., sci-kit-learn in Python) and adjusting parameters based on the characteristics of your dataset.

2.8 Least Absolute Shrinkage and Selection Operator (LASSO)

This study's LASSO technique is a shrinkage and selection approach for linear regression. It minimizes the residual sum of squares, allowing the total of absolute coefficient values to be less than a defined value [15]. Because of its automatic feature selection, LASSO is particularly well suited for parsimonious regression models, alleviating the problem of overfitting input data. LASSO is a linear regression regularization approach that promotes sparsity in model coefficients while preventing overfitting. It entails incorporating a penalty element into the linear regression goal function. The decision function for LASSO is the linear combination of the input features and their corresponding coefficients:

$$f(x) = wx + b \tag{2.3}$$

The features (xx) could include various factors relevant to wheat crop yield prediction, such as weather conditions, soil quality, and agricultural practices in Eq.5.1.

2.9 Deep learning in agriculture

Deep learning applications in agriculture and recent agricultural-related recently, its use with sensors have gotten people's attention. In the larger context, there is now much deeper learning research than there was in 2015, according to multiple reviews. Before 2015, there was a growing number of deep learning-related research in this agricultural environment, according to a review of deep learning approaches by Kamilaris and Prenafeta-Bold [12]. According to a survey of crop research on crop production prediction using machine learning, the last two years had the highest yearly concentration of studies. same observations, with 76 of his 120 publications published after peer review in 2019. 50 pertinent studies were discovered in a study of machine learning's ability to forecast yield. 30 of these research made use of deep learning in some way. In these investigations, there were 33 distinct deep-learning architectures. Deep neural networks (DNNs), LSTMs, and CNNs are some examples of architectures; CNNs are the most prevalent with 11 occurrences, followed by LSTMs and DNNs with 7 occurrences each. Space-time research, including those using CNN-LSTM hybrid models and three-dimensional (3D) CNNs, has also noted the design. Many conventional machine learning methods are also utilized in addition to deep learning models. These comprise ensemble models and linear regression, such as B. A decision tree-based random forest model. The authors claim that these models are frequently employed as comparable deep learning benchmark models.

2.10 Convolutional neural networks

Convolutional neural networks, often known as CNNs, became popular in the modeling field for applications where the input data can be spatial or can be represented spatially [13]. The convolution process is a key element of the model. This process creates a collection of spatial features that characterize the input data by applying a set of trainable kernels (or filters) to the data. In the first layer, the model learns fundamental characteristics, and in later layers, it learns composite features made up of these fundamental traits [14]. Batch normalization can be used on the inputs to improve how well the model learns these characteristics [4]. A collection of feature maps is the result of a straightforward CNN. These can be utilized directly or smoothed and put into a Fully Connected (FC) layer, depending on the application, for example, for regression or classification purposes. Utilizing CNNs to extract spatial information from two-dimensional (2D) inputs is a common use case. Multiple channels are frequently present in spatial data, however, each channel is treated separately when a 2D kernel is used. In non-spatial approaches, CNNs may also be applied if the input data is tabular. H.Line form. The next step is to do a convolution operation in one dimension (1D), where the kernel only affects values that are close to one another in the series, as determined by the kernel size [16]. Utilized several currently available CNN-based architectures to classify cotton balls and forecast cotton production from manually captured, high-resolution photos acquired at hand height. A two-stage Faster R-CNN model, which first suggests a region of interest and then finds targets from that region, was one of the models utilized. Additionally, we employed a single-shot multi-box detection (SSD) technique built on CNN that can recognize numerous items from an image. The most recent model to be used was MobileNetV2 [17], a compact variant of the SSD method designed for mobile devices. With an average recall of 0.66 and an average accuracy of 0.59, the quicker R-CNN architecture scored better. We built a linear regression model to forecast yield from automatically captured pictures by using the spatial model as an automated ball counter. A yield prediction model using the CNN output achieved a mean absolute percent error (MAPE) of 17.86%.

2.11 Long short-term memory networks

LSTM networks, initially presented in [18], are long-term memory systems. Tasks involving sequence modeling have frequently employed it. LSM is one of them. A family of recurrent neural networks (RNNs) with a deep learning architecture. LSTMs, both vector outputs from other models and inputs that resemble vectors and contain tabular data, often function with this model. Two fundamental ideas about LSTMs can help with this. Study temporal characteristics in the data.

The first is memory as a cellular state, which is how it was originally introduced. Another is the idea of gates, which are trainable FC layers that can change a cell's state based on fresh inputs from the data and historical outputs from the model. The model iterates over the sequence, modifying the state of the cells (C) and hidden (H) to process the data sequence. The gate outputs are combined using both non-linear activations and gate-learned parameters. Additionally, bidirectional and stacked LSTMs are readily available. Compared to a unidirectional LSTM, a bidirectional LSTM trains a second model [19]. One LSTM reads the input from the beginning to the end of the sequence $(t_0 \rightarrow t_n)$, and the other LSTM reads the input from the end to the beginning $(t_n \rightarrow t_0)$.

A final sequence of temporal characteristics is created by combining the outputs of these two simultaneous models [20]. When LSTMs are stacked, the first LSTM works with the input sequence, while the succeeding LSTMs work with the feature vectors produced by the prior model. Stacking and Directionality assist the model in learning high-level temporal properties from both sides of the input sequence [21].

2.12 Crop yield Prediction

Crop yield prediction, the study's main topic, is regarded as one of the most significant and difficult challenges in smart farming, entailing many tiny objectives. Predictive yield modeling assists farmers in locating troubled regions in their fields [20], provides crucial information to the agricultural supply chain, and is a management tool to help reduce business risks. [22]. Crop and yield estimates are crucial for cost-effective and proactive field production optimization. Predictive models are grouped with data analysis, information management, and data processing modules in what is referred to as the "management layer" in a review of complete remote sensing system designs.

The management layer provides application management logic for users, farmers, and agricultural experts to use. Estimating crop yields is a crucial stage in planning a district or area. The entire population needs access to food, so it's crucial to be able to forecast yields. Pre-harvest interventions can be conducted with reasonable accuracy and predictable low yields. One can utilize crop simulation models, remote sensing, machine learning algorithms, and other methods to estimate yield.

2.13 Remote Sensing for Crop Yield Prediction

Crop production forecasting may benefit greatly from remote sensing, particularly when using the normalized difference vegetation index (NDVI). The NDVI is a vegetation indicator that measures the quantity and health of plants by measuring the reflection of visible and near-infrared light [21]. It is a helpful indication of agricultural production potential since it has been regularly utilized to calculate vegetation biomass and health. Satellite photography, which can cover huge areas and offer regular data gathering, may be used to determine NDVI values. This information may be used to track crop growth and development throughout the growing season, enabling the early identification of possible yield-limiting elements like disease or drought. NDVI data may be integrated with other data sources, such as weather and soil details, to forecast crop yields. The NDVI data can then be connected to crop yields using statistical models shown in 2.4. These models may be employed to predict yields for a certain season and pinpoint fields that may exhibit yield fluctuation [23]. For instance, NDVI measurements from Landsat satellite images were utilized in research in India to forecast rice harvests. The researchers discovered that early-season NDVI measurements had a substantial correlation with ultimate production and could predict yield with an accuracy of about 90%. Another American research predicted maize yields using NDVI data, weather, and Soil information.

The NDVI data was the most significant predictor of yield, and it was able to explain almost 70% of the variability in output, according to the researchers [24]. Overall, crop yield prediction using remote sensing and NDVI in Eq.2.4 has shown considerable promise and can assist farmers and policymakers in making better-informed decisions about crop management and food security and Figure 2.4 shows the overall model of Remote sensing to CNN Process.

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{2.4}$$

It is crucial to remember that the precision of these forecasts depends on the caliber and acces-

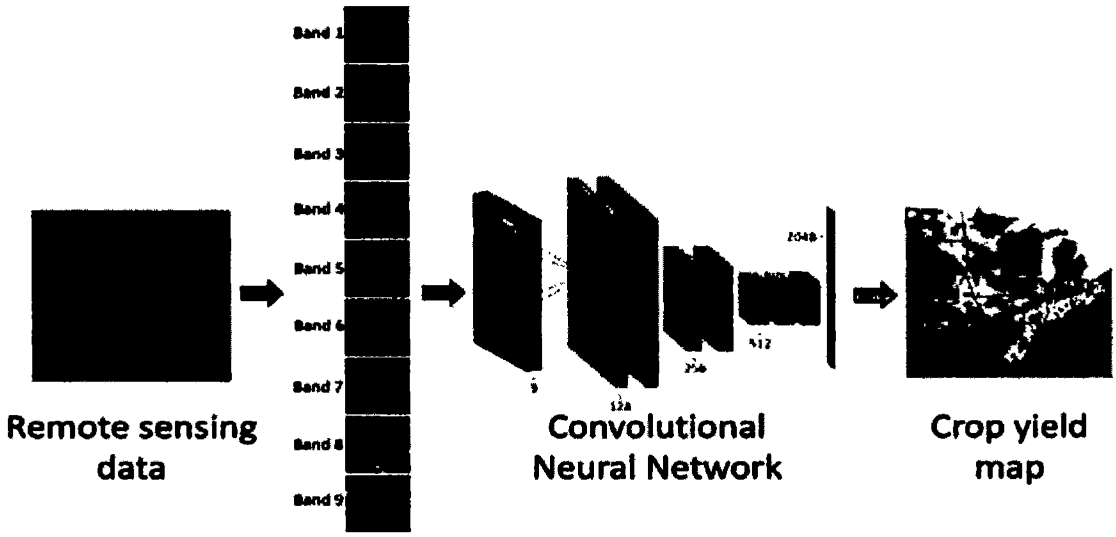

Figure 2.4: The process from Remote Sensing to CNN for crop Yield

sibility of the remote sensing data, the accuracy of the yield measurements used to calibrate the models, and both of these factors.

Chapter 3

Literature Review

The following databases were used in the literature review for the research thesis: Google Scholar, IEEE Xplore, Science Direct, and several themes crucial to the breakdown. Some of the search terms we used for our current study included "machine learning" [9], "machine learning algorithm for agriculture crop yield prediction [12], "gap estimation and find crop yield gap estimation," "support vector machine algorithms and approaches," "remote sensing and GIS approach," as well as "NDVI" and "Landsat Images." These search phrases are used both alone and in conjunction with other queries to gather the initial and background information about the location. We conducted a thorough search and found a tonne of writing and information on the subject at hand, but we only concentrated on the material/literature that contained information on remote sensing, geographic information systems, and deep neural networks algorithms and approaches and techniques that fill the yield gap estimation of crops in Pakistan.

This study proposes a deep learning framework for agricultural production forecasting that depends on environmental data and management practices [16]. It uses CNNs and RNNs [1]. For grain organizations and decision-makers, accurate and consistent crop yield forecasting is crucial. Extreme dry spells and floods brought on by environmental change also have an impact on food availability. To promote the availability of food for Pakistan's continuously expanding population, approach makers and organizers must have access to continuing harvest conditions and yield evaluations [18]. Crop production forecasting was challenging since it depended on several variables, including crop genotype, climatic aspects, and how well they worked together. It is essential to select an appropriate time window to recover wheat yield parameters since several harvest-creating periods include unique information.

To obtain adaptive and highly accurate yield estimates over large areas, the deep learning adaptive crop model (DACM) developed in this study is a spatial crop growth model based on the full extraction of crop growth information. It focuses on adaptive learning for dynamic fluctuations [25]. The results showed that the DACM has a root mean square error (RMSE) of 4.406 bushels/acre (1 (296.304 kg/hectare)) and an average coefficient of determination (R2) of 0.805. Compared to other state-of-the-art algorithms, DACM delivers improved performance in space and improved yield prediction accuracy at scale in machine learning and deep learning.

The analysis of stability and arousal estimations demonstrates that DACM is capable of identifying regional differences in crop growth and employing adaptive techniques to enhance yield estimates. Considering performance stability and interpretability increase, DACM offers a suitable strategy for forecasting large-scale agricultural crop yields by adaptively learning spatial heterogeneity patterns of crop growth. The current study [26] uses in situ data, remotely sensed microwave satellite data from Sentinel 1, and optical satellite data from Sentinel 2 to analyze three parameters, including soil moisture, soil salinity, and soil organic carbon (SOC). Estimates key soil health indicators. The study was conducted in the Punjab region of Rupnagar, India. Wheat crop productivity was assessed using SAR backscatter, predicted soil health indices, and optical remote sensing satellite data characteristics.

With R^2 values of 0.723 and 0.684, respectively, the soil health-based DLMLP model beat all other models in estimating crop production throughout the training and testing stages. The root mean square error (RMSE) and mean absolute error (MAE) for the 2019–20 season were both 1.24. While the MAE and RMSE for the OLS regression analysis estimating wheat yield were, respectively, 37.97% and 38.61% lower than the OLS regression analysis, the DLMLP test R2 was 42.2% higher than the OLS.

Then, utilizing data from various sources as input parameters, the yield prediction skills of BO-LSTM, Support Vector Machine (SVM), and Least Absolute Shrinkage and Selection (Lasso) operators were assessed [27]. In this paper, we demonstrated how Bayesian optimization enables efficient deep-learning hyperparameter optimization. When employing "GPP + climate + LAI + VI" as inputs, the BO-LSTM model demonstrated the highest yield forecast accuracy (RMSE = 177.84 kg/ha, R^2 = 0.82). When predicting winter wheat yield, BO-LSTM and SVM algorithms beat linear lasso regression (RMSE = 214.5 kg/ha, R^2 = 0.76). The BO-LSTM model surpasses SVM in terms of performance and shows promise for detecting data correlations. Additionally, there are major distinctions between machine learning and deep learning.

To further determine the robustness of the BO-LSTM method, we investigated the performance estimation of the BO-LSTM method at different time points. The results show that the BO-LSTM model can achieve higher estimation accuracy without the intervention of human factors in regions where winter wheat is intensively cultivated.

Since this technique is becoming more and more important for predicting agricultural production, more sophisticated algorithms are needed to uncover the fundamental spatiotemporal patterns of this data. Convolutional Long Short-Term Memory (ConvLSTM) has not yet been the focus of research for agricultural output forecasting, even though Convolutional Neural Networks (CNN) and other Deep Learning (DL) approaches have achieved substantial gains in this field [28].

For more accurate and reliable spatiotemporal feature extraction, we introduce Deep Yield, a hybrid structure combining ConvLSTM layers and three-dimensional CNN (3DCNN). The model is trained using MODIS land surface temperature (LST), surface reflectance (SR), and land cover (LC) data for major CONUS soybean growing districts in 1836 and historic county-level yield data. Comparing the prediction performance of our model with that of competing techniques, we find that DeepYield's performance far outperforms not only his ConvLSTM and his 3DCNN but also his CNN + GP, CNNLSTM, and decision tree of the competing approaches.

In this article, we present a machine-learning technique for predicting wheat yield [29]. This case study will focus on the wheat-growing region of northern Pakistan due to its importance to Pakistan's agricultural sector. After collecting data over 5 years, the trait subsets with the most beneficial effects on plant productivity were selected. Twelve algorithms were used on three datasets. Gaussian process (GP), multi-layered processing (MLP), and sequential minimal optimization regression (SMOreg) were selected as three algorithms based on research results. Root mean square (RMSE) and percent absolute difference (PAD) measurements were used to validate the results.

The lowest RMSE (0.5552) and PAD value (0.0093) were achieved by SMOreg. MLP came close to the second lowest for both PAD (0.0116) and RMSE (0.737). GP performance was found to be the weakest with increasing RMSE (17.7423) and PAD (0.2203) scores. Our results demonstrate how machine learning algorithms can accurately predict crop outcomes using datasets collected from the local environment. These findings can be extended to other crops and geographic locations.

In this work [30], utilizing three remote sensing variables, we developed a new deep-learning model called CNN-GRU to calculate county-level winter wheat yields in the Guanzhong Plain. By using the leave-one-year-out cross-validation, the CNN-GRU model was able to identify attributes associated with yield from the input data. Additionally, we examined the spatiotemporal patterns of the anticipated yields by simulating wheat yields on the Plain pixel-by-pixel using the recommended CNN-GRU model. The yield distribution's regional characteristics—low In this work [30], utilizing three remote sensing variables, we developed a new deep learning model called CNN-GRU to calculate county-level winter wheat yields in the Guanzhong Plain.

The VTCI (vegetation temperature condition index), the LAI (leaf area index), and the percentage of rings that are actively photosynthesizing. The CNN GRU model's accuracy was greater than that of the GRU model (R^2=0.62, RMSE=479.79 kg/ha, MRE=9.34%), demonstrating the CNN GRU model's dependability and robustness (R2=0, 64, RMSE = 462.56 kg/ha, MRE = 8.90%).

By using the leave-one-year-out cross-validation, the CNN-GRU model was able to identify attributes associated with yield from the input data. Additionally, we examined the spatiotemporal patterns of the anticipated yields by simulating wheat yields on the Plain pixel-by-pixel using the recommended CNN-GRU model. The yield distribution's regional characteristics—low In this study [31], we created a new deep-learning model called CNN-GRU to estimate county-level winter wheat yields in the Guanzhong Plain using three remote sensing variables.

Vegetation Temperature Condition Index (VTCI), Leaf Area Index (LAI), Percentage of Photosynthetically Active Rings. The accuracy of the CNN-GRU model exceeded that of the GRU model (R^2=0.62, RMSE=479.79 kg/ha, MRE=9.34%), confirming its reliability and robustness (R^2=0.64, RMSE=462.56 kg/ha).). , MRE = 8.90%). The purpose of this study [31] was to create a model for evaluating the nitrogen status of wheat at all phases of growth using proximity measurements and meteorological data. His nine field tests on high nitrogen rates were carried out between 2010 and 2020 at five sites utilizing various wheat kinds. Crop circle sensors' proximity detection data was gathered at critical developmental phases and combined with climatic data from planting to the appropriate detection date.

Deep neural networks (DNN) and long short-term memory (LSTM) were used to determine aboveground biomass, plant nitrogen intake, plant nitrogen concentration, and nitrogen trophic index. The benchmark regression model was constructed using the random forest (RF) method. To simultaneously predict four nitrogen indices based on the DNN, we applied multi-task learning (MTL). Using a genetic algorithm (GA), the neural network's hyperparameters, connection weights, and loss function weights (for MTL) were each assessed independently. According to the results, MTL and DNN both reach excellent levels of overall accuracy (R^2 = 0.83-0.96 and 0.81-0.96, respectively), whereas RF beats LSTM (R^2 = 0.76-0.93) demonstrated that it is comparable) did not raise the dataset's estimations of nitrogen status.

Convolutional Neural Networks (CNNs), one of the most popular deep learning techniques, outperforms other common machine learning techniques on classification problems. It's important to examine CNN's claims about its ability to predict crop yields. For this purpose, UAV-based multispectral photography was taken using the four developmental stages of the winter wheat plant: bud, milk, dough, ripening stage, and other stages. The effect of the growth stage on the accuracy of yield prediction was evaluated. We investigated the impact on the accuracy of combining different development phases using a multipoint CNN model.

Using Extended Vegetation Index 2 (EVI2), a classical regression technique based on typical vegetation indices, we compared the prediction accuracy of the CNN model with that of the linear regression model. The fore-stage CNN model outperformed the best linear regression model (RMSE 1.00 t ha-1) and exhibited the lowest RMSE (0.94 t ha-1) among the four development stages. EVI2-based CNN multi-time regression models and multiple linear regression models outperformed CNN models in terms of predictive accuracy in the course phase. These results indicated that the harvest season was the right time to collect winter wheat data in this study and that CNN could improve the accuracy of yield prediction.

In this study, we propose a deep learning system for predicting agricultural production based on environmental data and management practices [32]. Accurate and consistent crop yield forecasts are critical to crop organizations and decision-makers. Extreme droughts and floods caused by environmental change also affect food availability. For policy makers and organizers, the ability to continuously assess crop health and yields is essential to ensure food supplies to Pakistan's growing population. Predicting crop yields has been difficult because it depends on several variables, including plant genotype, climate aspects, and how well they work together. Since multiple harvesting seasons contain unique information, it is important to choose the appropriate timeframe to harvest the wheat component for yield prediction.

A machine-learning technique for predicting wheat yield is presented in this paper [33]. The case study will focus on Pakistan's northern wheat-growing regions because of their importance to the country's agricultural sector. After collecting data over five years, the characteristic subset that was most beneficial for crop productivity was selected. Twelve (12) algorithms were used across three (3) sets of data samples. The Gaussian Process (GP), Multilayer Processing (MLP), and Sequential Minimal Optimization Regression (SMOreg) were selected as the three algorithms based on the results of the studies. The Root Mean Square (RMSE) and Percentage Absolute Difference (PAD) measurements were used to validate the results.

We conducted a Systematic Literature Review (SLR) to identify and include the techniques and variables that have been applied to forecast wheat crop output in other countries. We also carried out a pilot research to confirm and validate those variables in the local setting. The KNN [34] technique [35] is a kind of instance-based understanding that depends on the separation of the indicator components from the closest preparing group recognized by the model. Recently, the utilization of neural networks has grown. They are composed of several intricately entwined parts. The BP neural system (BPNN) is one of the most often employed fake neural systems [18]. BPNN generally consists of one info layer, one yield layer, and several veiled levels. The information layer only acts as a source of data, which is then studied and analyzed by the neurons in the hidden layer. Finally, they use exchange work to transmit their results to the yield layer.

When tackling the nonlinear capacity issue, BPNN is frequently highly effective at isolating the perplexing connection between autonomous parts [20].In particular, yields of wheat [36] have declined in certain areas and will be influenced by rising temperatures. Despite these disadvantages, increasing association location can nevertheless boost wheat yields in regions where there is currently a yield gap.

We overcame these challenges by evaluating wheat yields over the Indo-Gangetic Plains IGP at 30-meter destinations from 2001 to 2015 using data from the Landsat satellite and a different tool dubbed the Scalable Crop Yield Mapper (SCYM) [37]. Instead of relying on ground-based yield organizing data, SCYM uses crop model generation to create arranging information that may be used to modify vegetation documentation VIs to yield. The Scalable Crop Yield Mapper (SCYM) method was utilized to convert satellite data into yields. In line with this theory, crop show simulations are used to replicate appropriate field-level yield data. This information is then used to create direct backslides that convert yield from satellite vegetation records.

Field evaluations and role-playing Models are useful tools for analyzing the harvest yield gap but scaling up these methods to effectively manage whole districts over time has remained a challenging task. Satellite data have repeatedly seemed to provide information that can unmistakably estimate crop yields in ranchers' fields without the assistance of any other person or when combined with other data and algorithms. The ensuing yield maps offer a unique way to get beyond both spatial and temporary scaling issues, which enhances understanding of harvest yield disparities [38].

This paper's [39] explicit purpose is to examine the potential evaluation of satellite-based far-off detection to quantify and understand the agricultural output gap. The paper argues that innovative ways that can augment the traditional equipment of agronomists have incredible potential value as efforts to understand yield gaps increase, and far-off detecting may be one such instrument. The outstanding spatial and global variability of rural sceneries serves as a critical test in the execution of each of these aims. Estimating yield gaps is a crucial concern in this investigation. How may the estimates at these two disparate geographical scales—genuine yield and yield potential—be taken into account when calculating a yield gap? There are a few ways to evaluate crops here.

The difference between actual yield (Ya) and yield potential (Yp) is known as the yield gap (Yg). Given that real yields may be quickly reviewed by remote detection, yield gap evaluation requires more details or questions regarding Yp. One effective method is to combine self-governing assessments of Yp (for instance, subject to methodologies utilized in several publications on this unresolved topic) with assessments of actual yields based on remote detection for the districts and zones under consideration. Up until now, satellite data have done a respectably poor job of elucidating the importance and causes of yield differences in many regions. In any event, it can be seen from several existing models that a portion of the frequent spatial and transient scaling problems associated with field-based techniques may be overcome by far-off detection. Regardless of the way that the cost or openness of satellite data with satisfactory spatial objectives to isolate green fields was checked previously, this limit is rapidly diminishing.

In earlier studies, yield potential and yield gaps were geographically assessed using hard-to-find identifying and Geographic Information System (GIS) soil data, which also offered information on soil needs and water cut-off points [40]. Understanding this gap's root causes is crucial to understanding it. An essential initial step is to review past performance in terms of yield and those that have a substantial influence beginning with one year and continuing onto the next. This calls for the examination of yield data that considers both regional and temporal variation. While we rely on accurate audit data to help us further connect with and see undeniably cyclical and constant yield sectors, this would fight for a large portion of improved environment figures in closing yield gaps.

The objectives of this study were to: (1) assess Yp, Yw, and Yg in four China-based maize-producing areas under both submerged and rainfed conditions using agro-climatic zones (CZs) and reference environment stations (RWS); (2) use the GYGA-ED method to manage oversee up-scale the results from zones to the district; and (3) use climate data spanning 30 years, including sunshine-based development, temperature, and precipitation to investigate. This paper's [41] goal is to forecast inside-field variety in wheat production while taking into account online multi-layer soil data and satellite crop development ascribes. In this study, we examine the factors that affect Ethiopia's wheat output per hectare for the central Maher crop seasons of 2011 to 2013 in the kebele managing area. By combining crucial GIS and remote sensing data with a management information strategy and national rural field overviews, the model explains about 40% of the whole variability in wheat production per hectare throughout the country. We examine if these progressions can be explained by climate, shocks to, and the managers of downpour-nourished horticultural frameworks in light of the considerable internal variability in yield per hectare.

The prognosis for spring wheat, canola, and barley was the main topic of this reading [42]. To structure the estimate models, crop production data from 40 Census Agricultural Regions (CARs) was gradually gathered into a few larger areas. Every vegetation record pixel will include the vegetation data from any yields developing nearby, which might be a constraint. Because of a single crop, there is no straightforward way to split either the NDVI or EVI. Because the vegetation files will mostly be looking at the more overpowering yields, the less often formed crop inside the pixel is caught off guard. Building a veil for a certain yield could be possible, but this would require a lot of information on what crops were planted in every pixel.

A key interest in the utilitarian link between yield and these clever components is necessary for a correct yield estimate, and revealing this relationship calls for both large datasets and astounding calculations. The top-tier showing and plan methods were handled poorly using a large neural network (DNN) [5] technique. Important brain systems were employed in the technique to calculate yield measurements (counting yield, verifying yield, and yield contrast) while accounting for genotype and conditional information. The well-organized and well-coordinated significant neural systems decided to learn the nonlinear and complex relationships between traits, organic conditions, and their relationships from recorded information and make sensibly exact measurements of yields for new halves and parts planted in new domains with recognized climatic conditions.

It is possible to increase the assumption for crop productivity under various environmental conditions by using machine learning techniques [43]. In this work, a study on the use of these machine-learning approaches for Indian rice trimming zones is presented. The WEKA device and SMO classifier test results on the dataset of 27 zones in the Indian state of Maharashtra are addressed in this study. In this study, the assumption for rice crop yield was shown using support vector machine (SVM), one of the AI methodologies. The exploratory findings show that several classifiers, including Nave Bayes, Bayes Net, and Multilayer Perceptron, outperformed the SMO classifier with the lowest accuracy by obtaining the most exactness, affectability, and identification, affectability, and disposition that has been represented in the past for various classification tasks.

The majority of the time, genuine processes, such as the relapse model, have been employed to estimate agricultural yields using remote sensing data [44]. Another strategy to handle agricultural production evaluation is machine learning, which is a capable exact methodology for representation and forecast. To determine a reliable yield evaluation, we employ descriptive inquiry in the farming creation zone for sugarcane crops in this study [[45]. Three datasets—the soil, rainfall, and yield datasets—are used in this work. Additionally, we create a connected dataset and, using this cemented dataset, apply a few supervised approaches to determine the actual evaluated cost and the precision of a few tactics.

K-Closest Neighbor, Support Vector Machine, and Least Squared Support Vector Machine are three management techniques employed in this work. Crop productivity across vast zones is frequently screened using satellite remote sensing. Displaying the relationships between factors and crop yield is sometimes complicated since several parameters are necessary for yield and yield estimates. Several machine learning-based approaches have been put out to address this issue, but more work has to be done to increase the accuracy of district-level estimation. Assessment of the crop output at the area level throughout a gap nation is still in the growth stage.

To predict maize yield in this study, we linked a deep neural network (DNN). Comparing the DNN model against other models created using other machine learning methods allowed us to evaluate the DNN model's estimation accuracy. Additionally, we created datasets for double cross arrangements that varied in scope, and by adding each dataset, we verified the model's ability to extract components.

The goal of the current study [15] was to examine the sugarcane yield difference in Brazil, its magnitude, and its reasons (short of water or poor harvesting practices). In this study, [78] progress on AI-based frameworks for carefully calculating harvest output and assessing nitrogen status has been coordinated over the past 15 years. The research makes the case that rapid advancements in machine learning (ML) systems will provide economically astute and comprehensive solutions for improved harvest and condition status assessment and dynamics. Accuracy farming (PA) is more focused on using the sensor stages and ML approaches, combining various sensor modalities with expert data, and improving cream structures by combining unmistakable ML and sign planning methodology.

Sensors that capture the electromagnetic radiation of physical objects, such as buildings, streets, farmland, soil, or water, are used to produce remote sensing (RS) [46, 47] data. Physical objects exhibit a variety of unearthly characteristics, such as the vitality that is emitted or reflected fluctuates across a range of wavelengths. Information about RS can be obtained via satellites, aircraft, drones, or even a simple camera mounted on a truck's windscreen. In this idea, satellite pictures are used. Applications are found by RS [48] in a variety of fields, including topography, geology, agriculture, urban planning, meteorology, environmental change, and others. A kind of neural network known as CNN [49] is used to handle 2D or 3D organized information, such as photographs and movies. A FullyConnected (FC) layer connects all the information units and yield units in a standard neural network (NN). In any event, a convolutional layer in a CNN only links each yield unit to a portion of the gap input units. Open fields are the names for these subgroups.

Research often focuses on classification tasks in multispectral remote sensing images [5, 36]. These assignments include identifying objects, such as buildings, trees, and roadways, as well as grouping the spread of the area and characterizing the crops. Up to this time, the RS crop grouping had to deal with comparable problems about the expected RS output. The standard approaches used basic characterization models including SVM, DT, and Artificial Neural Networks (ANN). The majority of the time, RS images included vegetation files, including NDVI. The most effective deep learning techniques used 2D CNNs to capture the spatial highlights in the remote sensing images. A 3D CNN model was developed to capture spatiotemporal highlights from multispectral remote detecting, enlivened by the success of 3D CNNs for spatiotemporal gaining from recordings.

This study [50] developed a unique method employing an artificial neural network (ANN) to forecast and crop maize and soybean yields on a county-by-county basis in the "corn belt" zone of the Midwestern and Great Plains regions of the United States. Real yield data and long-term arrangement NDVI from AVHRR and MODIS are used to build the models. An alternative method that takes use of the SCE-UA improvement calculation is used to create the ANN model. With the use of ANN models, multivariate straight relapse (MLR) models are displayed, and the model's dependability and propensity for prediction are identified. With the new calculations, the expectations' accuracy can increase to 85%, enabling the right development of ANN models.

The most important instruments for assessing agricultural crop yields and externalities at all scales, often at coarse spatial objectives, are global gridded crop models (GGCMs) [51]. Strong horticulture assessments at local and neighborhood sizes, where the suitability of GGCMs is frequently constrained by poor information accessibility and high computational interest, need for higher objectives appraisals. Utilizing meta-models built using GGCM to give data to variables of high spatial aims is one technique to close this gap. To build meta-models for the anticipation of model outputs at precise spatial targets, we evaluate two machine-learning approaches in this study: absurd slope boosting and irregular backwoods.

To predict soybean yields, this study [52] combines UAV-based remote sensing data with a full-stack deep learning architecture. The study demonstrates how deep learning models and remote sensing data may be used to predict agricultural output with accuracy. This study's [53] deep learning models are used to forecast agricultural yields using data from remotely detected leaf water content. Using spectral data to predict crop yields, even if it doesn't precisely pertain to wheat, shows how deep learning may be used to its fullest capacity. A deep learning-based system may be used to anticipate agricultural production using remote sensing and meteorological data, claim the study's authors [54]. It clarifies how to anticipate crop yields by merging several data sources and deep learning models, even if it doesn't specifically reference wheat. In this work, we used a non-parametric empirical model based on Random Forests (RFs) to predict yield by combining meteorological data with indicators of remotely sensed vegetation. Because they are created utilizing a large number of decision trees, a random selection of training data, and independent variables, RFs are resilient [35].

3.1 Literature Review Findings

Conducting a literature review of the application of machine learning (ML) and deep learning (DL) techniques in predicting wheat crop yields reveals a wealth of research focused on enhancing the accuracy and efficiency of predictions. These technologies have proven crucial due to their potential to improve food security and optimize agricultural practices. Below, I summarize key findings across various studies.

1. A CNN-RNN Framework for Crop Yield Prediction United States [1]	Frontiers in Plant Science | www.frontiersin.org. **January 2020** |	Deep learning framework using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). • CNN-RNN model , • random forest (RF), • deep fully connected neural networks (DFNN), • LASSO **Corn & Soybean.**	crop genotype, environnemental and management practices. **variable importance by time period** 3 Main Factures: 1. The CNN-RNN model was designed to capture the time dependencies of environmental factors and the genetic improvement of seeds over time without having their genotype information. 2. The model demonstrated the capability to generalize the yield prediction to untested environments without significant drop in the prediction accuracy. 3. Coupled with the backpropagation method, the model could reveal the extent to which weather conditions, accuracy of weather predictions, soil conditions, and management practices were able to explain the variation in the crop yields.	Deal genotype, environmental factors. Major limitations of deep learning models is their black box property.
2. Crop Yield Prediction Using Deep Neural Networks United States [3]	Frontiers in Plant Science | www.frontiersin.org May 2019 | Volume 10 | Article 621	• Lasso • Shallow Neural Networks (SNN) • Regression Tree (RT) • DNN **Syngenta Crops**	• Genotype, Environment (Soil and Weather) • functional relationship between yield and these interactive factors • The dataset included 2,267 experimental hybrids planted in 2,247 of locations between 2008 and 2016 across the United States and Canada.	Focus on weather condition. A major limitation of the proposed model is its black box property, which is shared by many machine learning methods
3. Software Engineering for Machine Learning: A Case Study [45]	ICSE-SEIP '19: Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice **2019**	Software Engineering practices Machin Learning Projects	Key Findings First, machine learning is all about data. Second, building for customizability and extensibility of models require teams to not only have software engineering skills but almost always require deep enough knowledge of machine learning to build, evaluate, and tune models from scratch. Third, it can be more difficult to maintain strict module boundaries between machine learning components than for software engineering modules.	Currently in initial stage
4. Area Estimation and Yield Forecasting of Wheat in Southeastern **Turkey** Using a Machine Learning Approach [2]	Journal of the Indian Society of Remote Sensing **2020** https://doi.org/10.1007/s12524-020-01196-3 (0123456789, Springer	Machine Learning Approach. ▪ Support vector machine (SVM) ▪ Quadratic discriminant analysis (QDA) ▪ Linear discriminant analysis (LDA) ▪ Random forests ▪ Decision trees ▪ k-Nearest neighbor (KNN) ▪ Boosting ▪ LASSO	▪ Implementing Top of Atmospheric (TOA) correction for all images and collecting ground-truthing point of 313 fields from the Nurdagi and Islahiye counties. ▪ As the best model, the random forest was used for image classification show 97% Accuracy.	Only touch weather factor
5. Prediction of Winter Wheat	Remote Sens. **2020**, 12, 236;	• Developed a modeling framework	climate , remote sensing and soil data.	Only deal Climate, soil and senses data.

Yield Based on Multi-Source Data and Machine Learning in **China** [4]	doi:10.3390/rs12020236 www.mdpi.com/journal/remotesensing	to integrate climate data, remote sensing (NDVI) data and soil data to predict winter wheat yield based on the Google Earth Engine (GEE) platform . 1. Support vector machine (SVM), 2. Gaussian process regression 3. Random Forest (RF) 4. K-nearest neighbor (KNN) 5. Neural Network (NN) 6. Decision Tree (DT) 7. Bagging Trees (BGT) 8. Boost trees (BST)	The results show that the models can accurately predict yield 1months before the harvesting dates at the county level in China with an $R2 > 0.75$. It was found that RF, GPR, and SVM predicted wheat yields with higher accuracy, and RF demonstrated the best generalization ability among the three methods. RF model can estimate wheat yields accurately in advance (before the harvesting dates) in China.	
6. Wheat crop yield prediction using new activation functions in neural network **India [25] , Wheat**	Neural Computing and Applications (**2020**) 32:13941–13951 https://doi.org/10.1007/s00521-020-04797-8	The main objective of the proposed work is to develop an amended MLP neural network with new activation function and revised random weights and bias values for crop yield estimation by using the different weather parameter datasets. WEKA open-source Java libraries. By keeping the concept of the WEKA MLP algorithm, a new algorithm is developed specifically for the agriculture crop yield forecasting at a regional level	weather parameter datasets	weather parameter
7.Crop yield prediction with deep Convolutional neural networks [37]	Computers and Electronics in Agriculture journal (**2019**) www.elsevier.com/locate/compag	Convolutional Neural Networks (CNNs) – a deep learning methodology remote sensing **wheat and barley**	Biomass evaluation and yield prediction	Only Biomass
8. Predicting wheat grain yield and spatial variability at field scale using a simple regression or a crop model in conjunction with Landsat images [38]	Computers and Electronics in Agriculture journal (**2019**) www.elsevier.com/locate/compag	Regression or a crop model in conjunction with Landsat images	leaf nitrogen and initial aboveground-biomass	leaf nitrogen and initial aboveground-biomass
9. Integrating satellite and climate data to predict wheat yield in Australia using	Agricultural and Forest Meteorology 274 (2019) 144–159	satellite and using machine learning approaches	Climate Factor.	Deal Only Climate Factors

machine learning approaches [39]				
10. Yield Forecasting of Spring Maize Using Remote Sensing and Crop Modeling in Faisalabad-Punjab Pakistan 2018 [75]	Journal of the Indian Society of Remote Sensing https://doi.org/10.1007/s12524-018-0825-8	CERES-**Maize** model	Field initial data and Management data	Mobile Agricultural Geotagging Information System (MAGIS) system were used to collect data.
11.County-Level Soybean Yield Prediction Using Deep CNN-LSTM Model [40]	*Sensors* 2019, *19*, 4363. https://d *Sensors* 2019, *19*, 4363. https://doi.org/10.3390/s19204363	Deep CNN-LSTM Model **Soybean**	weather and LST data	weather and LST data
12.Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique] [41]	Agricultural and Forest Meteorology 285–286 **(2020)** www.elsevier.com/locate/agrformet	APSIM (a process-based crop model)-simulated biomass, and climate extremes, NDVI **wheat**	biomass, and climate	biomass, and climate
13. Machine learning for large-scale crop yield forecasting [42] **Netherlands (NL), Germany (DE), France (FR.**	Agricultural Systems journal homepage: www.elsevier.com/locate/agsy **2020**	machine learning **wheat, spring barley, sunflower, sugar beet**	weather, remote sensing and soil data	weather, remote sensing and soil data
14. Multilevel Deep Learning Network for County-Level Corn Yield Estimation in the U.S. Corn Belt [43]	IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, VOL. 13, **2020**	remote sensing and machine learning, **Corn**	satellite data, climate data, soil data	satellite data, climate data, soil data
15. Forecasting Corn Yield With Machine Learning Ensembles [44]	Frontiers in Plant Science | www.frontiersin.org . **2020**	Machine Learning Approach & **Corn**	corn yields, management (plant population and planting date), and environment (weather and soil) features.	

Number of heads/pods per square meter	**(A)**	**Example**	**220**
[redacted]			
Average number of grains per head/pod	**(B)**	**Example**	**24**
Number of grains per square meter = AxB	**(C)**	**Example**	**= 220 x 24 = 5280**
Yield per square meter = C/100 x 3.4gms	**(D)**	**Example**	**= 5280/100 x 3.4 = 179.52gms**
Yield in t/ha = D/100		**Example**	**= 179.52/100 = 1.79t/ha**

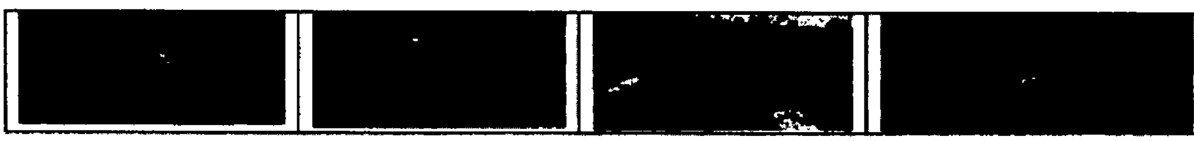

Figure 3.1: In Pakistan, the statistics office and the department of crop reporting use similar techniques

3.2 Manual Wheat crop estimation technique of CRS

The multiple methods, models, and strategies connected to remote sensing and the Machin learning domain that are employed in agricultural yield forecasting are depicted in the accompanying figure. That and actual values are estimated because they are validated after harvesting from sampled areas. CRS 1st estimates wheat crop from the sample area and, after harvesting, validates and verifies that sampled area and then publishes a report of actual crop production.

A simple manual method calculates wheat grain yield by taking into account the number of heads per 500 mm column, the number of grains per head, and the size of the grain.

3.2.1 Methodology for manually measuring wheat yield

The methodology for manually measuring wheat crop yield by Crop Report Services Punjab department is shown in Figure. 3.1 and Figure 3.2.

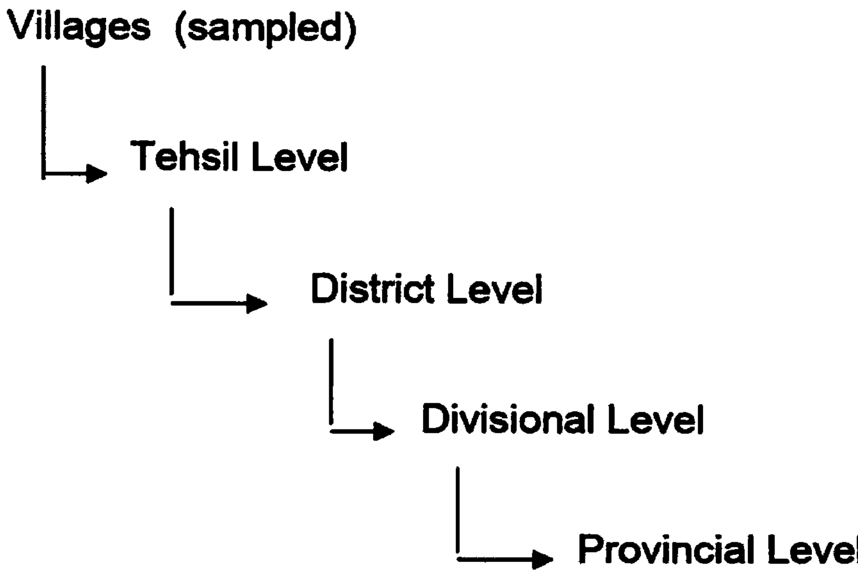

Figure 3.2: Results Compilation Hierarchies of CRS

3.3 Problem Statement

- Predicting the wheat crop yield in Pakistan is a difficult task because it depends on several variables, including internal (seed, disease, plant, etc.), external (weather, soil, irrigation, social economics, etc.), and manual (not precisely estimated before harvesting) methodologies for predicting wheat yield.

- Traditional or obsolete methods of yield assessment take a long time and require a lot of labor. Data on wheat crop yields collected from a few villages does not truly represent the entire crop population.

- A gap between theory and practice exists Because of the difficulties mentioned in existing points.

- Since many more variables and factors are available on the ground than are currently included in wheat crop yield prediction models, none of them are presented holistically in research studies.

- It is hoped that these widespread flaws in the current systems will be fixed. Instead of subjective methods, it uses technology methods such as deep learning, remote sensing, and machine learning to estimate agricultural yields. These issues can be resolved better with an effective predictive strategy and have fewer effects on the current systems.

3.4 Research Questions (RQ)

1. **What are the shortcomings of current wheat crop yield prediction techniques?**
2. **How to effectively and accurately predict on-ground wheat production using Deep Learning, Machine Learning, and Remote Sensing techniques in the scenario of South Punjab, Pakistan?**

Chapter 4

Research Methodology

"Controlled Experiment" is the most appropriate approach for this investigation. Exploring the relationship between conditions and logical outcomes is done in a controlled experiment. Create the research procedure utilizing a framework for software engineering. Using many fields including software engineering, remote sensing, deep learning, and machine learning, it is a hybrid method. There are several first steps.

The most crucial elements in our research are those that impact wheat crop output. Therefore, we use a variety of techniques for the collecting of data, including conducting the SLR process shown in Figure 4.1, ethnographic research (which comprises seeing people in their natural habitat to understand their experiences, opinions, and daily routines), and questionnaires. To collect and analyze Factors, we use SLR.

4.1 Conducting Systematic Literature Review SLR

We conducted a systematic literature review (SLR) to comprehensively examine the application of machine learning (ML), Deep Learning and Remote Sensing in crop yield prediction. SLRs offer a structured approach to identify, gather, and synthesize all relevant studies from electronic databases. This method allows us to address specific research questions and identify potential gaps in existing research, providing guidance to both practitioners and researchers interested in the field. Through the SLR methodology, we collected and analyzed a wide range of studies related to ML & DL techniques applied to crop yield prediction. By synthesizing the findings, we aim to offer new insights and perspectives on the current state-of-the-art in this area. Our review serves

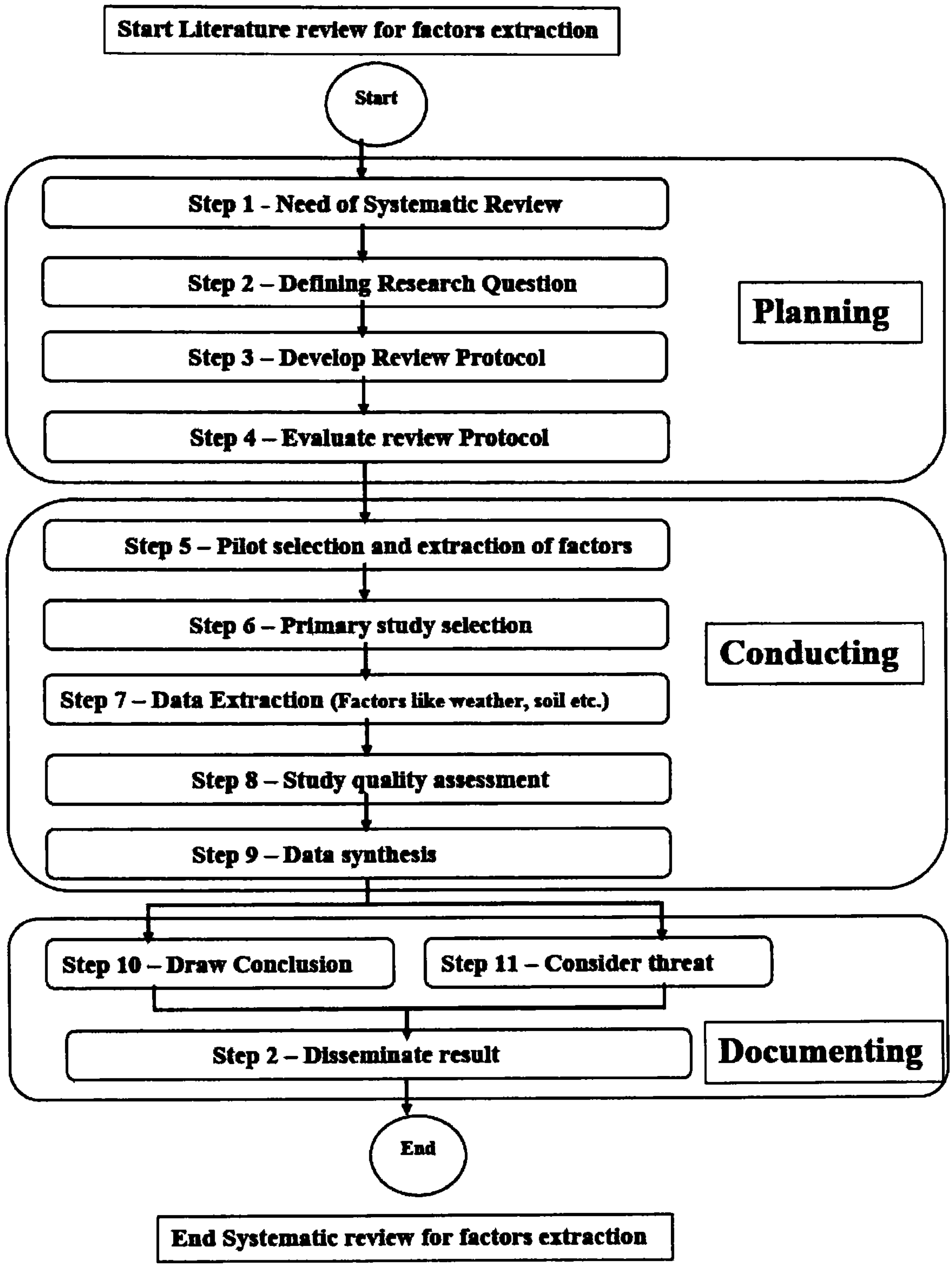

Figure 4.1: SLR Process of factor Collecting

as a valuable resource for newcomers to the field, offering a comprehensive understanding of the existing research landscape and highlighting areas where further investigation is warranted.

4.1.1 SLR Methodology

4.1.1.1 Review protocol

Prior to initiating the systematic review, a review protocol was established following established guidelines. This involved defining research questions that guided the review process. Once the research questions were formulated, relevant studies were identified using databases including Science Direct, Scopus, Web of Science, Springer Link, Wiley, and Google Scholar. Following the selection of relevant studies, a filtration process was applied, assessing them against predetermined exclusion and quality criteria. This ensured that only studies meeting the specified criteria were included in the review, maintaining the integrity and rigor of the analysis. The initial stage involves planning the review, encompassing the identification of research questions, development of a protocol, and subsequent validation to ensure the feasibility of the approach. Following this, the review progresses to the second stage, which involves conducting the review. During this phase, publications are systematically selected from various databases. Data extraction is then carried out, involving the collection of information such as authors' details, publication year, publication type, and other pertinent information related to the research questions.

4.1.1.2 Research questions

This systematic literature review (SLR) aims to gain insight into the studies published in the intersection of machine learning (ML) and crop yield prediction. To achieve this goal, the review analyzes studies from multiple dimensions. Four research questions (RQs) have been formulated for this SLR:

- **RQ1:** What machine learning algorithms are utilized in the literature for crop yield prediction?
- **RQ2:** What features are employed in the literature for crop yield prediction using machine learning?
- **RQ3:** What evaluation parameters and approaches are employed in the literature for crop yield prediction?
- **RQ4:** What are the challenges encountered in the field of crop yield prediction using machine learning?

4.1.1.3 Search strategy

To narrow down the search and focus on relevant concepts for this review, an initial automated search was conducted. The search query "machine learning" AND "yield prediction" was employed across six databases. Abstracts of retrieved articles were scrutinized to identify synonyms of the keywords. Following the application of exclusion criteria and processing of results, a more refined search string was formulated to capture a broader range of relevant studies. The final search string adopted was: ((“machine learning” OR “artificial intelligence”) AND “data mining” AND (“yield prediction” OR “yield forecasting” OR “yield estimation”)). This search strategy yielded 567 studies for further examination.

4.1.2 SLR Results.

The selected publications are shown in Table 4.1.

Table 4.1: Distribution of papers based on the databases.

Database	# of initially papers	# of papers after exclusion criteria	Papers (%)
Science Direct	17	4	2
Scopus	68	11	22
Web of Science	32	0	0
Springer Link	132	10	20
Wiley	20	1	80
Springer Link	298	10	2
Google Scholar	567	24	48
Total	2013	50	100

Table 4.2 illustrates the frequency of usage for various groups of independent features, including soil and crop information, humidity, nutrients, and field management. It is evident from the table that the most frequently utilized feature groups are those associated with soil, solar, and humidity information.

Table 4.3 show which Machine Learning and Deep Learning algorithm how many time used in literature .

4.1.3 List of Variables/Factors identified from SLR

- Soil Information

Table 4.2: Grouped features of Factors.

Group	# of times used
Soil information	54
Solar information	48
Humidity	30
Nutrients	132
NDVI	20
Crop information	45
Field management	12
Other	40

Table 4.3: Most used machine learning & Deep Learning algorithms.

ML & DL algorithms mostly Used	# of times used
Neural Networks	27
Linear Regression	414
Support Vector Machine	30
Randome Foresr	32
CNN	20
ANN	10
RNN	12
LASSO	11
Other	10

- Solar
- Climate
- Crop Information
- NDVI Values
- Fertilizers
- Rainfall

4.2 Pilot Study

Some Data collected from the pilot study for collecting factors show in Figure 4.2.

4.3 Farmers Interview

During the trips to the farmers' fields, interviews were conducted with the farmers. Every farmer was given a questionnaire (Appendix A & B) with information on their name, contact information, area of wheat cultivation, previous crop cultivation, type of wheat grown, date of transplantation, irrigation source, amount and type of fertilizer used, pests and diseases, date of harvest, and yield attained. We compare these factors after extracting the data from the literature using a systematic review procedure and collecting the factor data following a pilot research study. We inspect, confirm, and validate if these factor data are present in the same or a different form in our immediate surroundings.

4.4 Domain Analysis for Factor Data Validation & Verification

In the future, we will conduct an ethnographic study to validate and verify those characteristics in a real context through an interview, a survey, or another method. Differences between pilot study variables like environment and culture and SLR variables are shown in Figure 4.3.

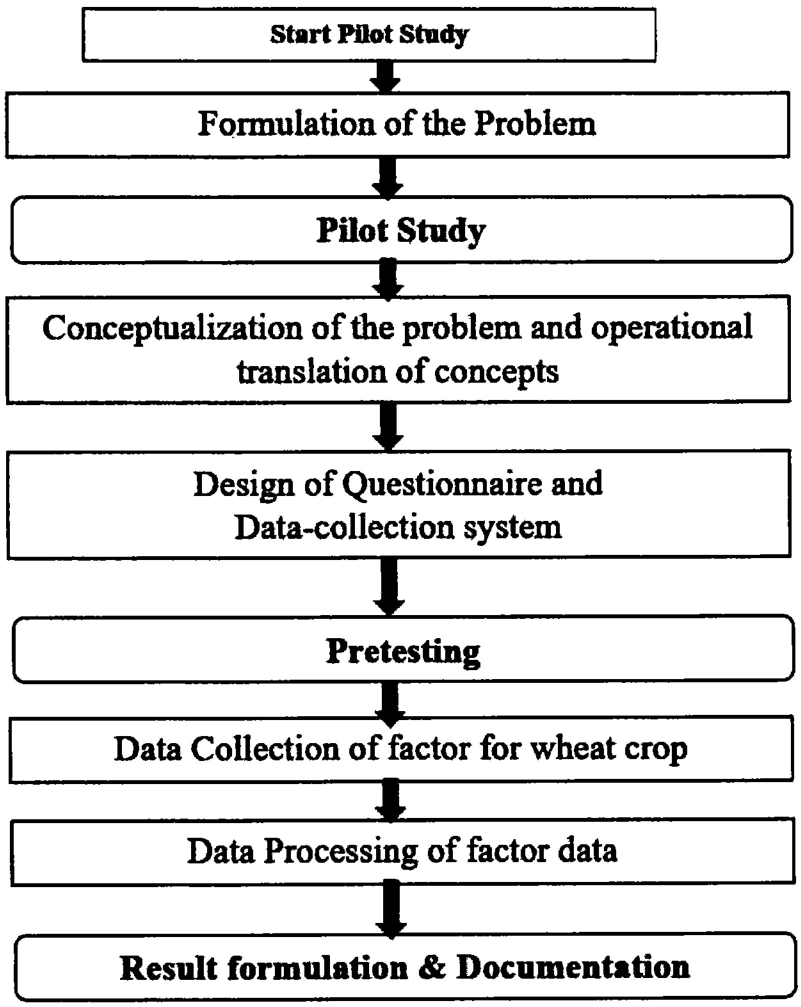

Figure 4.2: Pilot Study Process

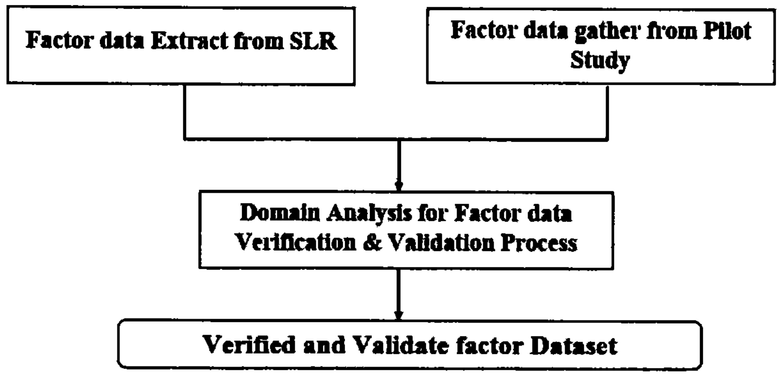

Figure 4.3: Validation and Verification of domain analysis process

4.5 Controlled Experiment

Next, we'll conduct a controlled experiment to collect raw data from wheat crop photos, including LANDSAT 8 remote sensing maps with various types, formats, and geographic and temporal resolutions.

In the second stage, RS and GIS processes such as clipping, training data, and classification are performed. The raw data is subsequently transformed into a more usable format by processing it and, if necessary, performing feature extraction. Gather information on elements that impact crop production in the third phase, such as soil data, weather data, and other factors. After extracting RS, GIS, and factor data, the proposed Deep Conventional Neural Network CNN-RNN [1, 2] method will be used in step four to estimate and predict the gap in wheat crop yield after the dataset has been verified and validated. We obtained our wheat crop production quantity after implementing the suggested deep learning approach. Many machine-learning techniques are now employed to forecast wheat crop production.

4.6 Fieldwork (to confirm the findings)

We compare the data that our suggested approach extracts with the current technique's data.

Crop yield estimates in several nations, including Pakistan, rely on conventional methods of information gathering and on-the-ground field reports from Crop Reporting Service Punjab. We eventually determine the real difference in wheat crop production by comparing the data acquired from our suggested approach with Crop Reporting Service Punjab.

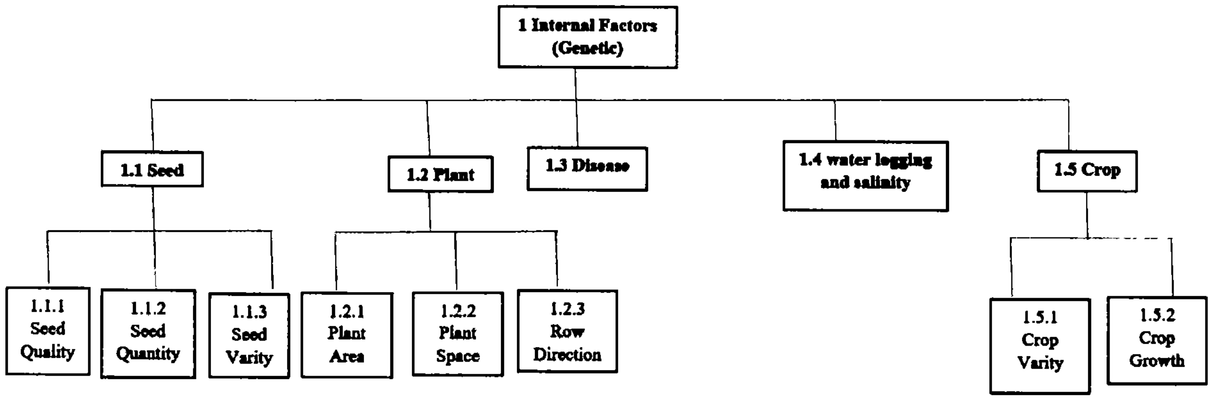

Figure 4.4: Internal Factors

4.7 Factor Dataset in diagrammatic form

Here internal and external factors affected by wheat crop yield production are shown in Figure 4.4 and Figure4.5.

In the future, we will undertake ethnographic studies like interviews, surveys, and inquiries for deep domain analysis in the actual environment, after which factors may become more important. This is vital to perform a deep study to verify and validate variables.

4.8 Framework Process Diagram

Figure 4.6 shows the overall progress of the Framework diagram of wheat crop yield prediction using Machine Learning, Deep Learning, Remote sensing, and climate factors.

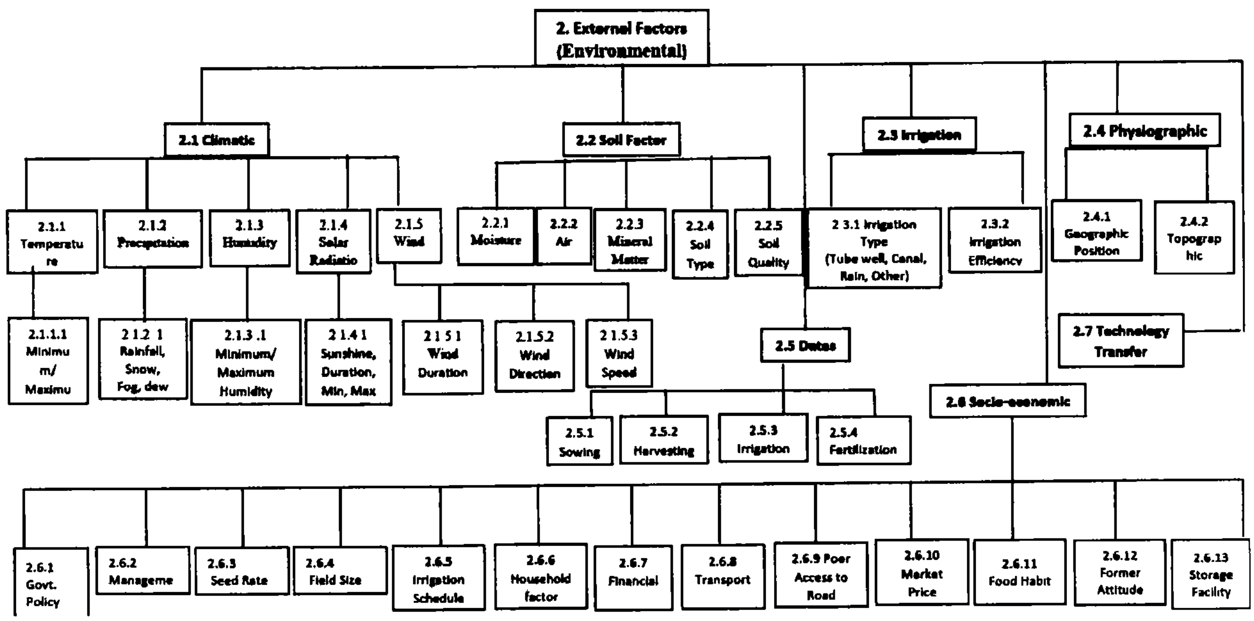

Figure 4.5: External Factors

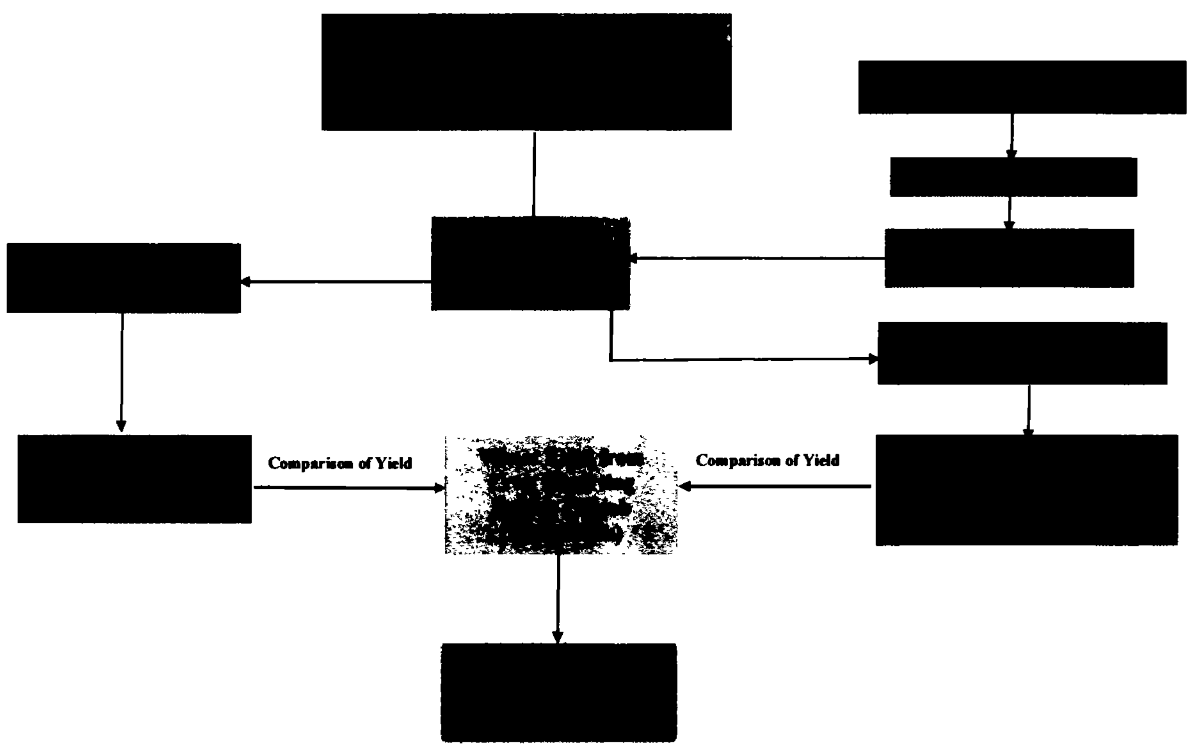

Figure 4.6: Framework Process Diagram

Chapter 5

Data Collection , Materials & methods

This section describes the strength of our research: how we collected these vast amounts of data and created our Dataset and also methodology.

5.1 Data Collection

Data on wheat were gathered for the inquiry in the Punjab province of Pakistan from 2017 to 2022. The dates of planting, growth, overwintering, returning green, jointing, smooth development, and reaping were included in this information. The daily leaf territory list and soil field capacity were computed using these dates, and the water-restricted potential generation was calculated using those results. Here are some weather statistics for the wheat-growing season. High-density raster photographs of the target region, which will be located at latitude 29.848212 N and longitude 71.263367 at an elevation of 423 feet (129 m) above sea level, were among the data that were made available for this study. According to Crop Report, it has a semi-arid climate typical of Multan District, Punjab, Pakistan, with an area of 3,721 square kilometers and a total area of 437 acres according to Crop Report Services Punjab. 26.2 C and 304 mm of precipitation fall each year, respectively. More precipitation falls throughout the summer than during the winter months. Between November and January, the mean maximum and minimum temperatures gradually start to fall and then start to climb. Multan's topsoil is made up of silt, which is a very fine sand. There is a mixed trimming zone in Multan where crops including sugarcane, maize, wheat, and rice are grown. In general, there are more downpours throughout the summer than during the colder months. Figure 5.1 show Wheat growing area in Multan district.

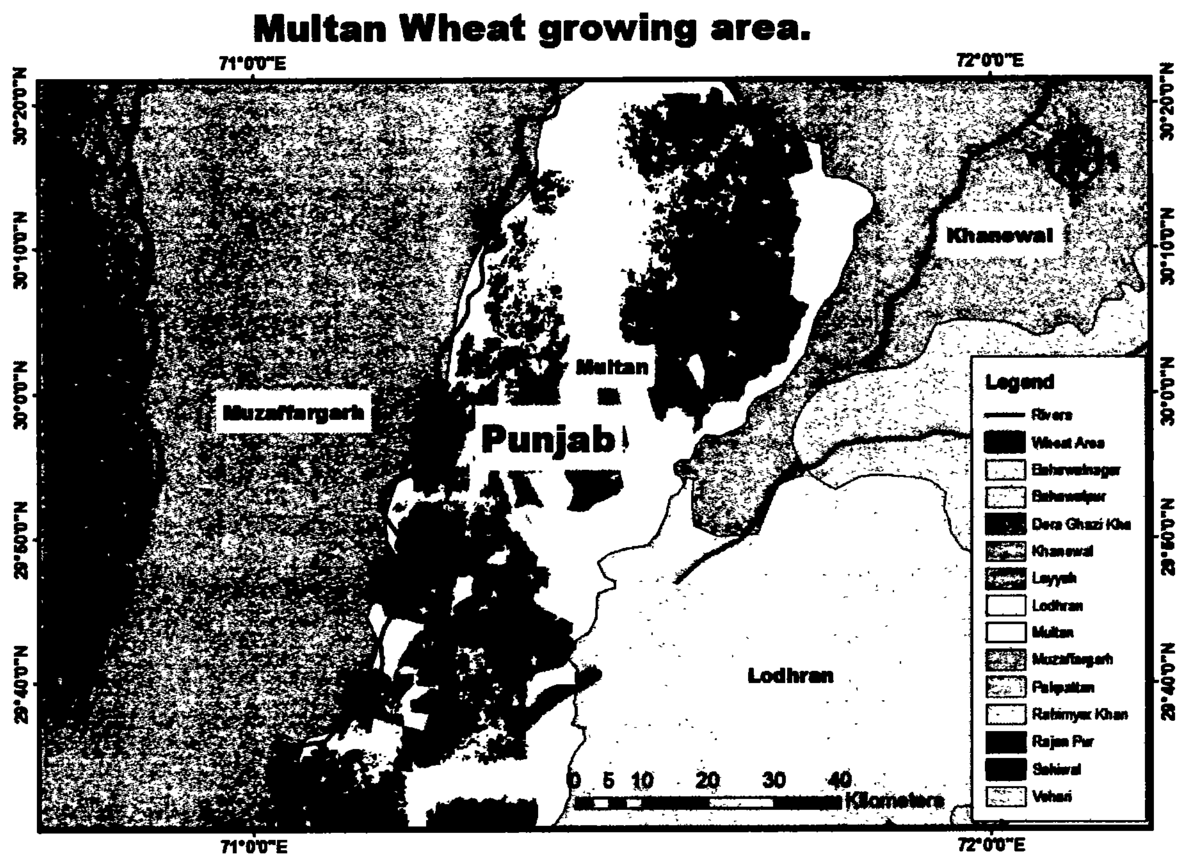

Figure 5.1: Map showing wheat-growing area of Multan district.

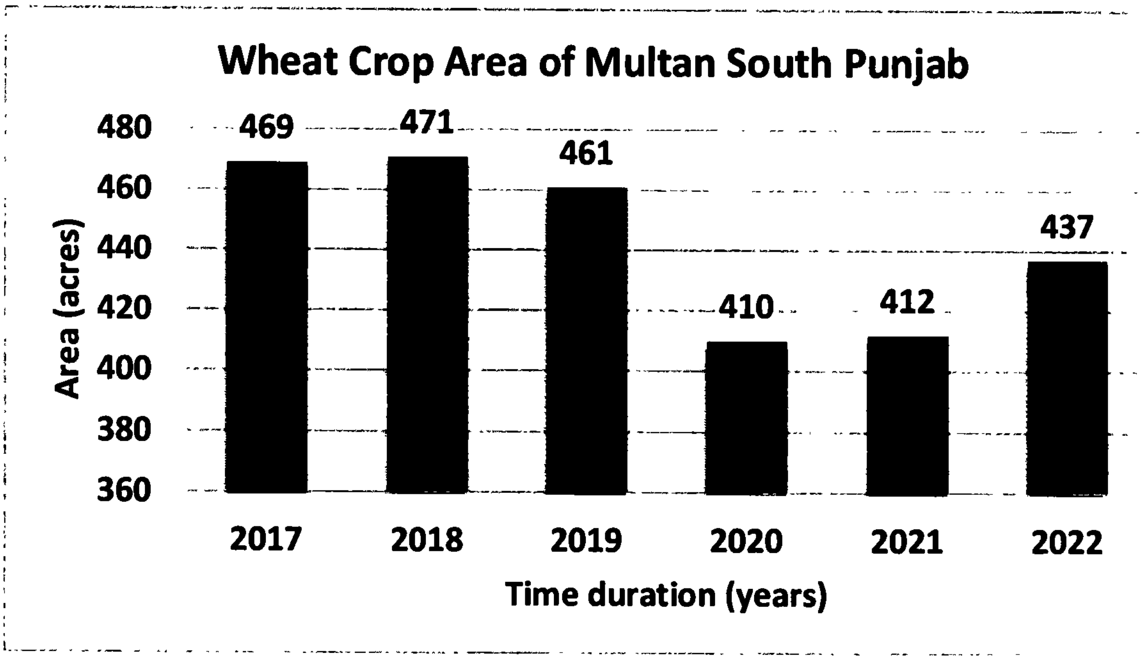

Figure 5.2: Multan Wheat Crop area from 2017 to 2022

5.1.1 Crop yield data

We collect the actual crop yield production dataset including wheat crop area, Wheat crop Production per acre, and wheat production in a tone of Multan district of the last six years from Crop Reporting Services, Government Punjab Pakistan.

Figure 5.2 illustrates the wheat crop area in Multan from 2017 to 2022, provided by Crop Report Services Punjab, Pakistan.Figure 5.3 shows the wheat crop yield per acre for the same period and location, based on Crop Report Services Punjab data. Additionally, Figure 5.4 displays the total production of wheat crops in Multan from 2017 to 2022, sourced from Crop Report Services Punjab, Pakistan.

5.2 Meteorology data

This data was collected by the POWER Data Access Viewer https://power.larc.nasa.gov/data-access-viewer/ and also gets weather dataset from the Pakistan Meteorological Department Government of Pakistan for cross-validation and also downloaded from It is the most common and

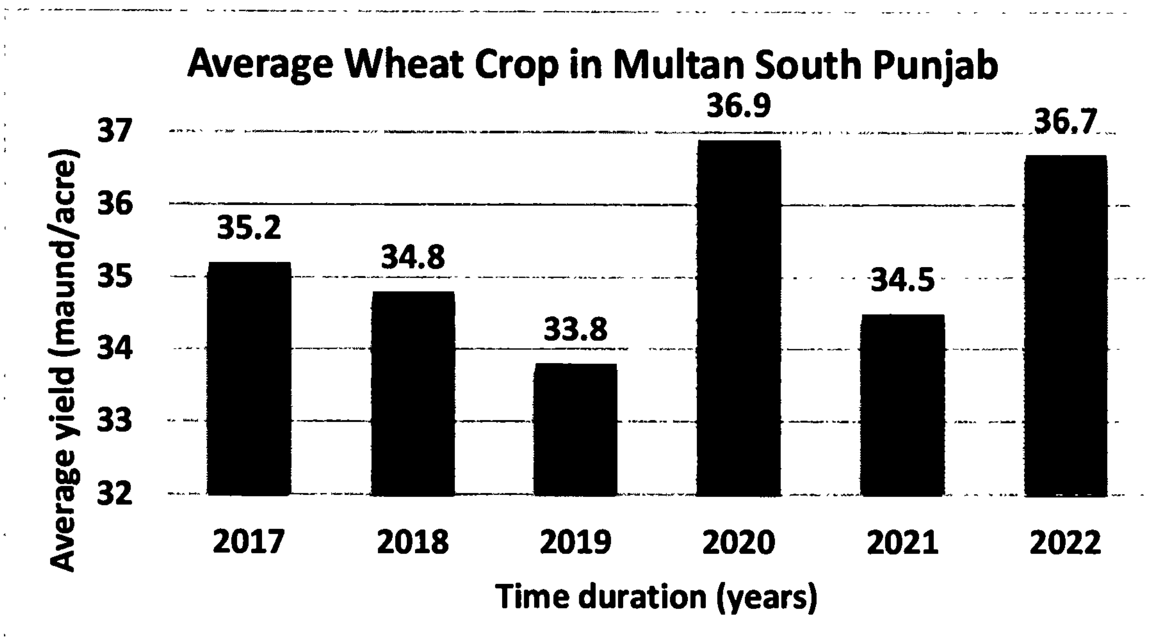

Figure 5.3: Average wheat crop yield in Multan from 2017 to 2022

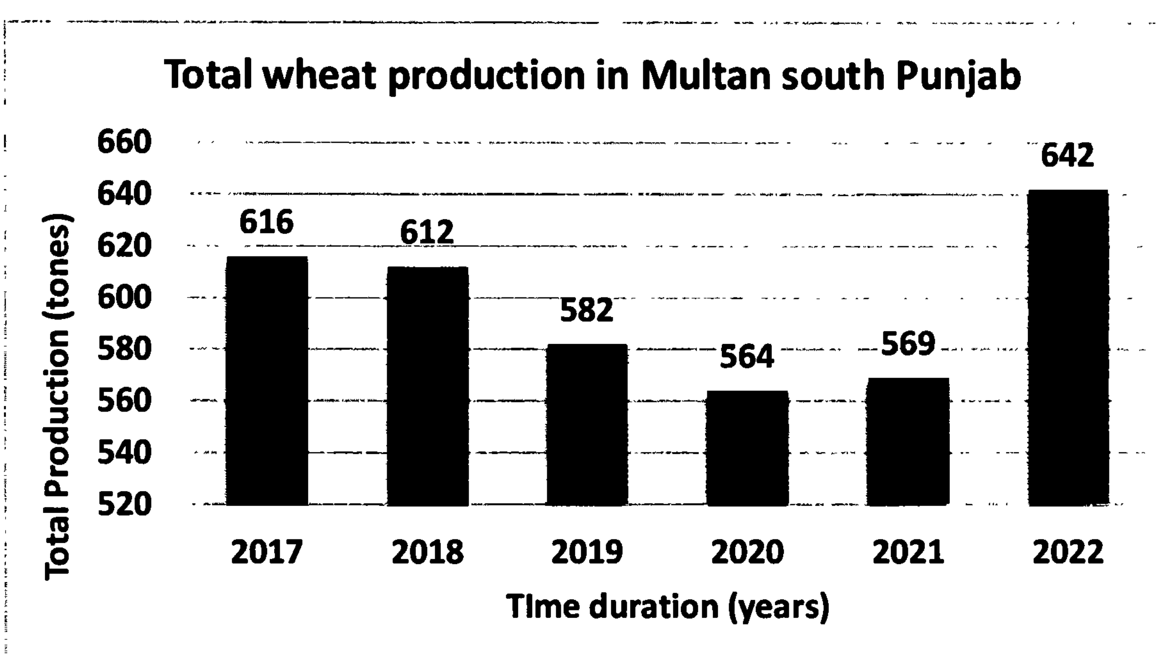

Figure 5.4: Wheat crop Production in Tons from 2017 to 2022

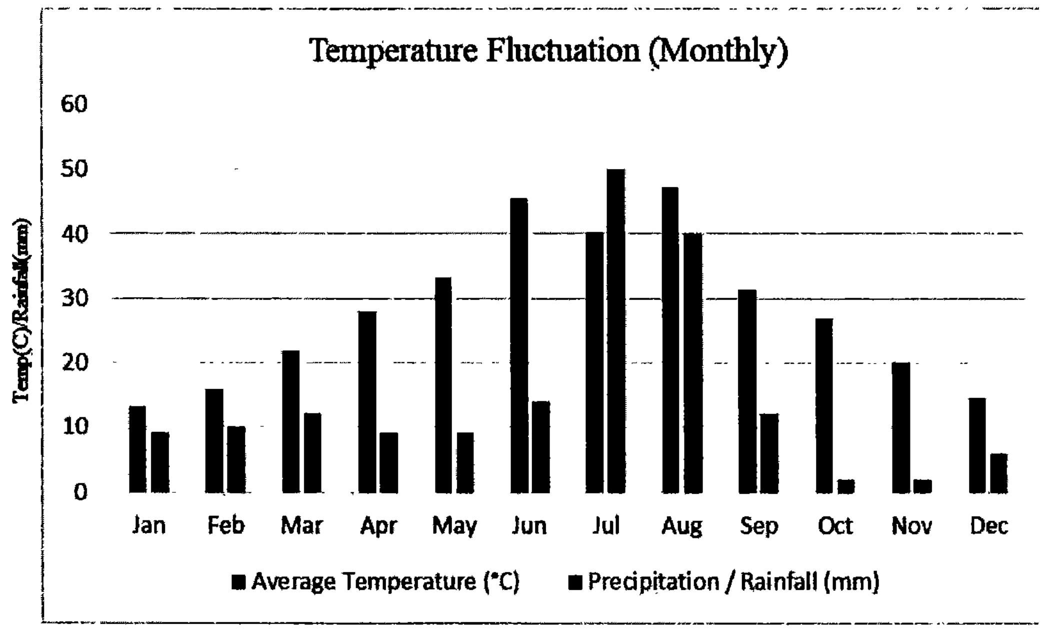

Figure 5.5: Average Temperature & Precipitation from 2017 to 2022

most important feature for any agricultural production. Climate change threatens the crop production systems of our staple crops (wheat, corn, cotton, rice, and sugar cane). A temperature rise of 3°C is projected by 2040, and a temperature rise of 5-6°C is projected by the end of this century, resulting in a loss of up to 50% in wheat productivity in Asian countries. Pakistan's geographical location will make this loss even greater position. Here it is Features considered for research.

5.2.1 Maximum temperature

Every year, Pakistan experiences a dramatic increase in temperature. Plants have photosynthetic activities that start and end at particular temperatures. So you can see why this metric is one of the key variables in predicting crop production.

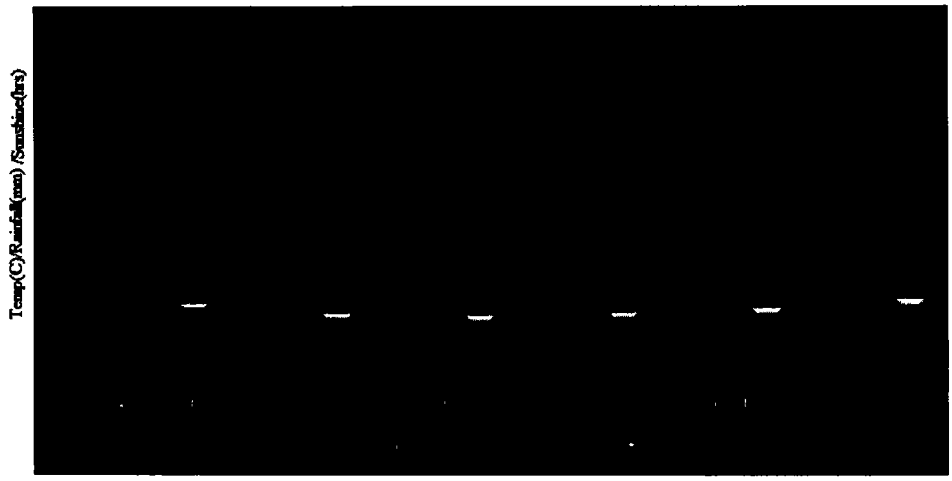

Figure 5.6: Temperature, Rainfall, and sunshine in the target area.

5.2.2 Minimum temperature

For predicting agricultural yield, this factor is also crucial. Make sure you are aware of whether your plants are developing optimally by using the average minimum temperature for the area. If not, plant development may be hampered. Between November and January, the mean maximum and minimum temperatures begin to steadily decline before rising.

5.2.3 Average rainfall

Multan District is a lowland locality. In the case of heavy rain, this may occur, Flooding is likely. When plants receive too much soil if you give a lot of water when it rains, the plants won't get enough oxygen from the soil. Therefore, moderate yields are required to obtain high crop yields Rainfall maximizes crop productivity show in Figure 5.6.

5.2.4 Humidity

Humidity is a crucial element for data analysis. It significantly affects the growing of plants. In general, moisture supports our environment's water cycle process. Plants grow in a particular way. Humidity: Enables you to understand how the humidity influences our agricultural zone's growth rates and how much it impacts it.

5.2.5 Precipitation

Precipitation refers to any liquid or frozen water that condenses in the atmosphere and falls to Earth. On a global scale, it is one of the three critical stages in the water cycle. This statistic is also essential for calculating wheat crop yield show in Figure 5.5. Multan's soil is silt dirt or very fine sand. A mixed management zone exists in Multan where wheat, rice, maize, sugarcane, etc. are produced. The soil in the Multan district was a clay loam with a field capacity of 45 to 56% and a pH value of 8.

5.3 Soil data

Since soil is the foundation for crop growth, its characteristics have a significant impact on how plants grow and produce grain. In the study of data-based applications for precision agriculture, modeling frequently has soil and its characteristics as a focus. Type of soil, pH, cation exchange capability, and position are examples of specific spatial soil properties. Overall, it was found that data aspects relating to soil were more frequently found in the papers under consideration. These characteristics were noted 54 times as crop production predictors, whereas solar information, the second-most often seen group, was noted 39 times in the same context. In the research we analyzed, soil data was also used as a predictor and predicted values. The data was collected by the Punjab Soil Fertility Authority in Multan, but the problem was that this data was all in a distributed format. Therefore, you should ensure that all data are in the correct format. These are the factors we considered for our thesis.

5.3.1 Soil moisture

Soil moisture is an important variable in water exchange and control of soil moisture. Thermal energy between the surface and the atmosphere evaporation and plant transpiration. Soil moisture plays an important role in plant production or crop yield. [55]

5.3.2 Soil consistency

This is the strength of whether the material is soil material or soil material. Nutrient sticks can occur if the soil is damp Consider stickiness. This is a very important parameter for our predictive model.

5.3.3 Soil reaction

The same is true for soil reactions and soil pH, which is also important Ability to measure the acidity or alkalinity index of a substance floor. The scale is 0-14. PH 7 is the neutral point [35].

5.3.4 Soil texture

The classification of soil types according to their physical texture is known as soil texture. There are three different kinds of sol particles.

- Sand
- Silt
- Clay

Table 5.1: Soil properties of experimental sites at Multan at depths of 0 to 120 cm.

Soil depth	BD	LL	DUL	SAT	RR	OC	PH	NH	NO3
0-10	1.43	0.12	0.28	0.35	1	2.43	6.75	0.83	4.83
20-30	1.57	0.13	0.29	0.36	0.78	2.27	6.57	0.43	4.17
20-30	1.6	0.13	0.31	0.38	0.77	1.6	6.34	0.32	4.3
30-45	1.72	0.12	0.33	0.38	0.73	0.6	6.2	0.3	4.2
45-60	1.72	0.12	0.33	0.39	0.72	0.55	6.07	0.33	4.2
60-90	1.73	0.13	0.38	0.41	0.69	0.37	5.7	0.3	4.07
90-120	1.79	0.13	0.38	0.4	0.68	0.32	5.5	0.29	3.57

Note : Soil depth(cm) & Bulk Density BD(cm) & lower limit LL(mm) & drained upper limit DUL(mm) & SAT(mm) & RR & OC(mg) & PH & NH(mg) & NO_3(mg)

5.4 Seed data

Seed Data collect from the Punjab seed department which is also the most important actor in crop yield production.

Table 5.2: Wheat Seed Variety and area recommended Source

Variety	Year of Release	Sowing Time	Yield (Maund)	Recommend Areas
Akbar	2019	1st Nov-10th Dec 2019	76	Areas of Punjab
Anaj	2017	1st Nov-10th Dec 2017	76	Areas of Punjab
Ujala	2016	1st Nov-10th Dec 2016	68	Areas of Punjab
Galaxy	2013	1st Nov-10th Dec 2013	80	Areas of Punjab

5.5 Chemical fertilizer

This data was collected by the Punjab Agricultural Department. Depending on the chemical fertilizers in the soil, how can be analyzed? A change in this value can have a real impact. Here it is the features collected for the paper.

5.5.1 Urea

Urea is a white crystalline soil that contains 46% nitrogen and could be. It is considered an organic fertilizer. Both things happen when urea is applied to the soil and converted into ammonia and nitrogen needed by plants.

5.5.2 Phosphate

One of the earliest fertilizers with a high phosphorus analysis was phosphate. Along with nitrogen and potassium, fertilizers frequently supply significant phosphorus quantities [56]. For the plant to receive the most benefit from phosphate, which is a popular synthetic phosphorus fertilizer high in phosphate, it must be applied correctly.

Table 5.3: Recommendation of Fertilizer dosages based on fertility land and crop stage

Land Fertility	Fertilizer Dosages at Sowing Time	Dosages at 1st or 2nd Irrigation
Low	DAP = 2 bags Urea = 0.5 bag Phosphate = 1 bag	Urea = 1 bag
Medium	DAP = 2 bags Urea = 0.5 bag Phosphate = 1 bag	Urea = 0.75 bag
High	DAP = 1 bag Urea = 0.5 bag Phosphate = 1 bag	Urea = 0.5 bag

5.6 Yield maps & remote sensing overview

Figure 5.7 shows the overall framework process in a scientific model that shows a Raster Landsat image is collected from satellite USGS, then processed that raster image is to calculate NDVI values for the dataset, after that collet factor dataset from different sources, and then ML and DL techniques apply that dataset after Training the dataset with ML techniques and finding results in the form of predicting wheat crop yield. It tends to rain more heavily in the summer than in the winter. The normalized difference vegetation index (NDVI), show in Figure 5.7 which can be obtained from the United States Geological Survey (USGS), is a simple graphical indicator that can be used to analyze remote sensing measurements and determine whether the target being observed contains live green vegetation or not. Typically, but not always, this analysis is done from a space platform. The selected NDVIs were used to develop yield forecasting.

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{5.1}$$

NDV= (0.2839505 - 0.2244924) / (0.2839505 + 0.2244924)
NDVI = (0.0594581) / (0.5084429) = 0.1169415484019936.

5.7 Deep Neural Network (RNN ,ANN& CNN) Process

A deep neural network [49] is a neural network with more than two layers and a specific degree of complexity. We put up our important deep neural network using the [56] back propagation approach and employed three hidden layers to calculate the precise cost of our yield. Using only 1 covered layer while dealing with monster gap information might occasionally result in low accuracy. Even while adding additional hidden layers could increase processing costs, the likelihood of a huge amount of information is increased. Figure 5.9 shows overall complete process of wheat crop yield process.

5.8 Limitation and Scope of research

A total number of 58 different factors were taken from the literature research and a pilot assessment of the immediate surroundings. Crop yield can be influenced by a multitude of factors, ranging from environmental conditions to agricultural practices. Some are relevant and some are irrelevant factors.

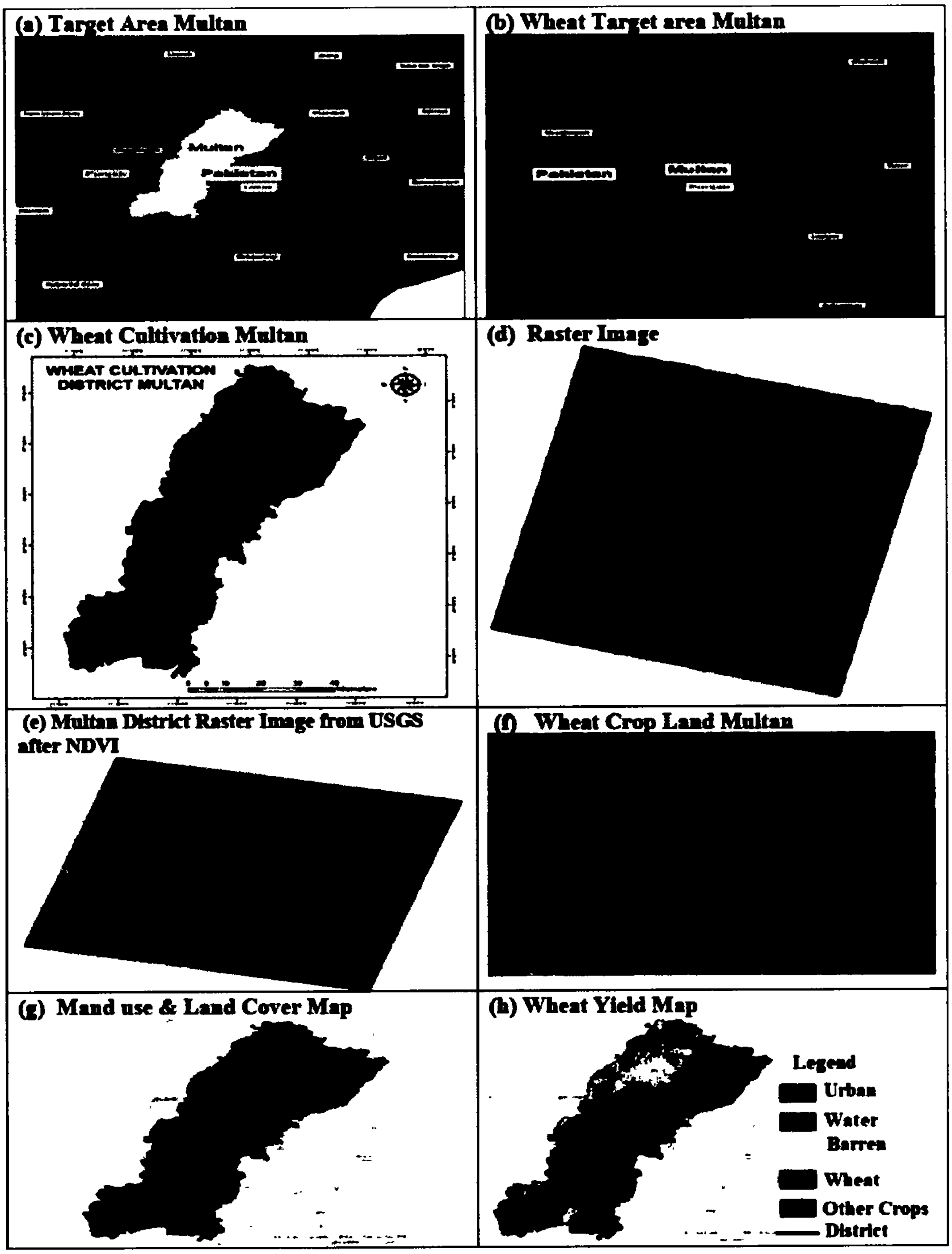

Figure 5.7: Landsat 8 Images from Satellite https://earthexplorer.usgs.gov/ USGS.

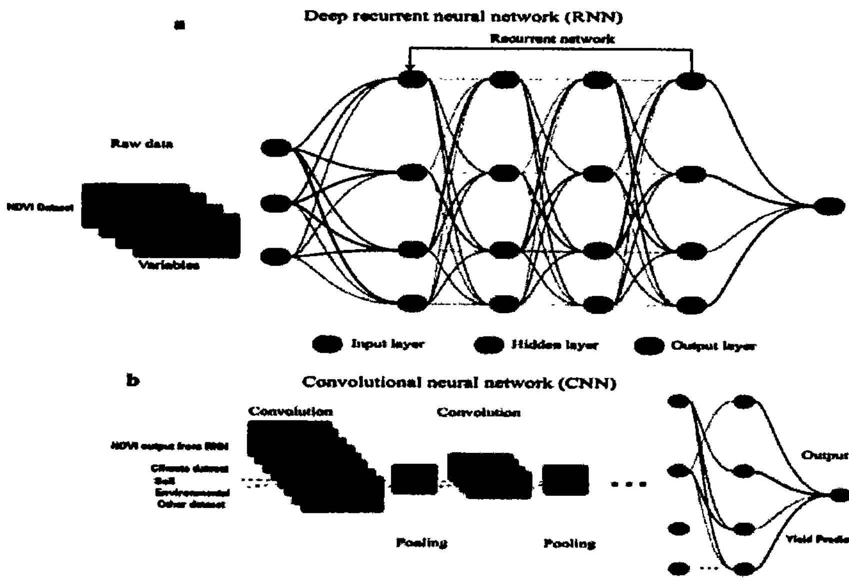

Figure 5.8: A simple Deep Neural Network model [5]

- **Climate:** Temperature, precipitation, humidity, and sunlight play crucial roles in crop growth. Variations in climate patterns can significantly impact yield.
- **Soil Quality:** Soil composition, fertility, pH level, and nutrient content affect the growth and health of crops. Proper soil management is essential for optimizing yield.
- **Water Availability:** Adequate water supply is essential for crop growth. Irrigation systems and water management practices are critical, especially in regions with limited rainfall.
- **Pest and Disease Management:** Effective pest control measures and disease management strategies are necessary to prevent crop damage and maximize yield.
- **Crop Variety:**The choice of crop variety or cultivar can influence yield. Selecting varieties that are well-suited to local environmental conditions can improve productivity.
- **Fertilization:** Proper fertilization with essential nutrients such as nitrogen, phosphorus, and potassium can enhance crop growth and yield.

Focusing on the 20 most pertinent factors for our local environment is a practical approach to studying and understanding the dynamics of crop yield in study area. By prioritizing these factors, we can allocate resources more efficiently and make informed decisions to optimize agricultural

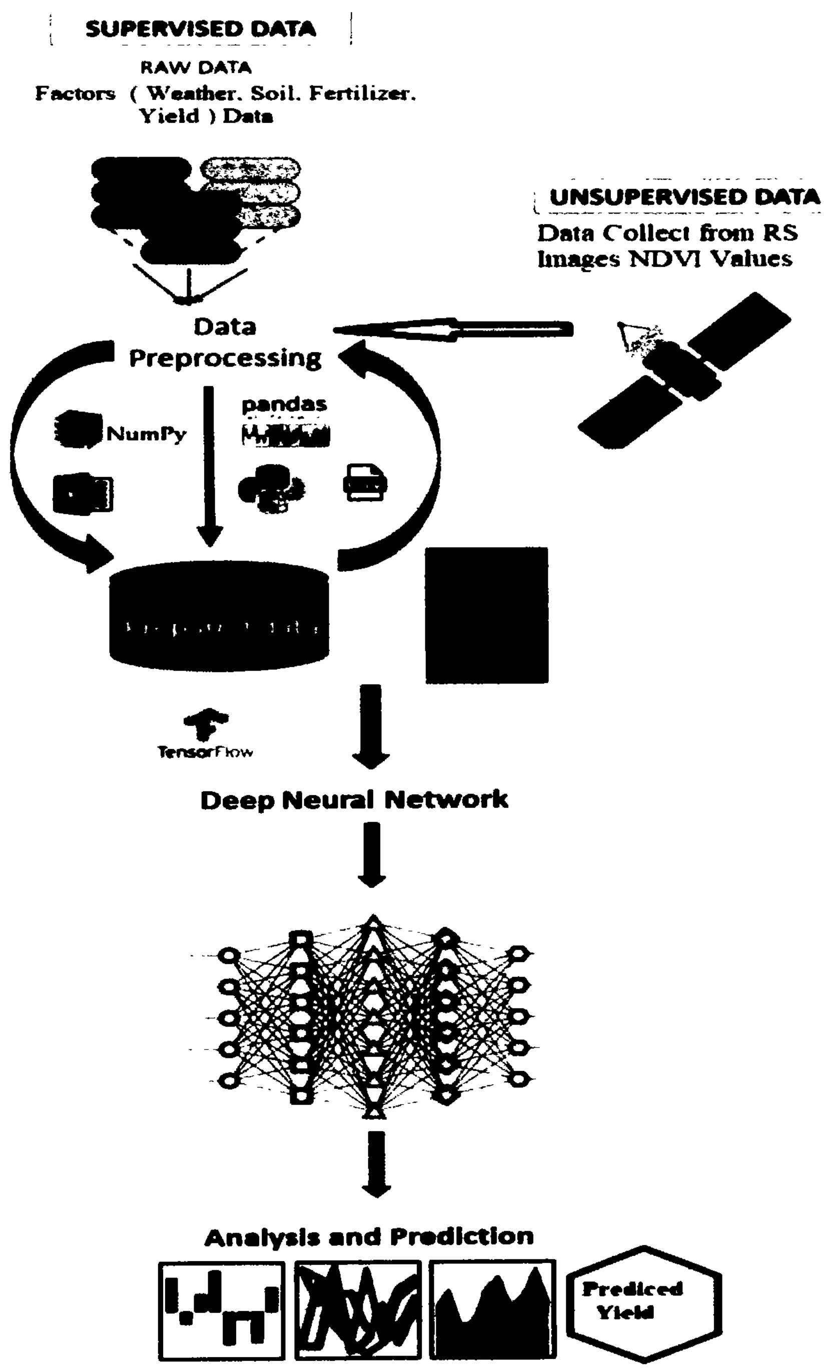

Figure 5.9: Overall crop yield prediction Process

practices. It's essential to thoroughly analyze how each of these factors interacts with one another and impacts crop productivity.

When building a machine learning model for crop yield prediction, selecting the right features (or factors) is crucial for the model's performance and interpretability. The choice of features depends on various factors such as data availability, domain knowledge, and the specific objectives of the prediction task. Using a subset of features from a larger set of potential factors offers several advantages:

Reduced Dimensionality: Including all available factors may lead to a high-dimensional feature space, which can increase computational complexity and the risk of overfitting. By selecting a subset of relevant features, you reduce the dimensionality of the problem, making the model more tractable and less prone to overfitting.

Improved Model Generalization: Including irrelevant or redundant features can negatively impact model generalization by introducing noise into the data. By selecting only the most informative features, you improve the model's ability to generalize to unseen data.

Interpretability: A smaller set of features can lead to a more interpretable model. It's easier to understand and explain the relationships between a smaller number of input variables and the predicted outcome, which is important for stakeholders in agriculture who may not be machine learning experts.

Computational Efficiency: Training a model with fewer features typically requires less computational resources and time, which is advantageous, especially in resource-constrained environments.

Domain Relevance: Not all available features may be relevant for predicting crop yield. By carefully selecting features based on domain knowledge and empirical evidence, you ensure that the model focuses on factors that are known to influence crop yield.

The specific subset of features chosen for crop yield prediction would depend on factors such as the type of crops being studied, the geographic region, the availability of data, and the objectives of the prediction task. It often involves a combination of weather variables, soil characteristics, agricultural practices, and historical yield data. Feature selection techniques such as correlation analysis, feature importance ranking, and domain expert consultation can help identify the most relevant factors for the prediction task. However, the 20 factors that are most pertinent to our local environment represent the limit of our study. In the literature research, there were many machine learning and deep learning techniques, but our study was only able to use 3 machine learning (SVM,RF and LASSO) and 3 deep learning techniques (CNN,RNN and ANN).Table 5.4 shows all input variable that are required for crop yield prediction of wheat and their sources. Using various machine learning and deep learning algorithms like SVM, RF, LASSO, CNN, ANN, and RNN for crop yield prediction of wheat in Pakistan offers several benefits:

Diverse Data Representations: Wheat yield prediction can benefit from leveraging diverse data sources such as weather data, soil properties, agricultural practices, and historical yield data. Each algorithm may excel at capturing different aspects of these data representations. For example, SVM and RF can handle numerical and categorical features effectively, while CNN can extract spatial patterns from satellite imagery, and RNN can capture temporal dependencies in time-series data.

Model Flexibility and Interpretability: Different algorithms offer varying degrees of model flexibility and interpretability. SVM and RF provide straightforward decision boundaries and feature importance rankings, making them interpretable models. LASSO offers a sparse model representation, which aids in feature selection and model interpretability. On the other hand, CNN, ANN, and RNN can capture complex non-linear relationships in the data, potentially leading to higher predictive accuracy.

Handling Non-linearity and Temporal Dependencies: Crop yield prediction involves complex relationships between input features and yield output, including non-linear patterns and temporal dependencies. Algorithms like SVM, RF, CNN, ANN, and RNN are capable of capturing these complexities to varying extents. RNN, in particular, is well-suited for handling time-series data and capturing temporal dependencies, which are crucial for modeling seasonal variations and trends in wheat yield.

Robustness and Generalization: Comparing the results with Crop Report Services Punjab provides an opportunity to assess the robustness and generalization capabilities of the machine learning models. By comparing model predictions with ground truth data from authoritative sources like Crop Report Services Punjab, researchers and practitioners can evaluate the models' performance under real-world conditions and identify areas for improvement.

Validation and Benchmark: Comparing the results with Crop Report Services Punjab serves as a form of validation and benchmark for the machine learning models. It allows researchers to assess whether the models' predictions align with official crop yield statistics, providing confidence in the models' accuracy and reliability. In summary, leveraging a combination of machine learning and deep learning algorithms for wheat yield prediction in Pakistan, along with comparing the results with authoritative sources like Crop Report Services Punjab, offers a comprehensive approach to accurately estimate crop yields, understand model behavior, and validate model performance under real-world conditions.

5.8.1 Scope of factors

- Weather (Minimum Temp, Maximum Temp, Average Temp, Humidity, Precipitation, Surface Pressure, Wind Speed)
- Yield (Area, Yield in acre, Yield in tones)
- Soil (depth cm, Texture, Moisture, DUL)
- Date (Sowing, Harvesting, Fertilizer, Irrigation)
- Solar
- NDVI Value

5.8.2 Machin Learning and Deep Learning methods

- Support Vector Machine (SVM)
- Random Forest (RF)
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Convolutional Neural Network (CNN)
- Recurrent Neural Network (RNN)
- Artificial Neural Network (ANN)

Table 5.4: Input data and their variables for Framework

Category	**Variables**	**Duration Source**
Crop	Wheat Area Yield, Area Yearly	CRS
Satellite Data	NDVI Yearly	USGS
Climate Data	Weather Data Daily	Meteorological Dept
Soil Data	Soil Type, pH Yearly	Soil Fertility Authority

Table Table5.4 presents a summary of all the input variables.

Chapter 6

Methodology, Process, Results and Analysis

6.1 Solution of problem 1

What are the shortcomings of current techniques for wheat crop yield Prediction?

All the limitations and shortcomings are already discussed in the literature review section in detail.For the development of food security, a thorough understanding of the dynamics involved in food production is essential [56–59]. It has been shown that a considerable decrease in poverty results from an increase in agricultural production. The amount of harvested crop product in a given area is known as yield, and it depends on several variables [60]. These variables may be divided into three main categories: technological (such as management decisions and agricultural practices), biological (such as illnesses, insects, pests, and weeds), and environmental (such as weather, soil fertility, terrain, and water quality). These factors account for yield differences from one region to another worldwide [61].Most of the studies [62] on the effects of weather and climate on food production have investigated their effects on crop yields. However, climate influences all aspects of crop production, including acreage (area grown or harvested) and cropping intensity (number of crops grown per year). This framework has different steps and processes for wheat crop yield prediction. For example, the 2012 U.S. drought, the 2010/2011 heat wave, the subsequent Russian wheat embargo, and the 2006/2007 and 2007/2008 Australian droughts all contributed to declining grain stocks and sharply rising food prices. They are likely to be connected. Food affordability is deteriorating for many consumers, including the poor, in import-dependent countries [63, 64]. The table compares different countries' months of sowing and harvesting because of different factors (weather conditions). Therefore, the main purpose of this study is to exam-

Table 6.1: Comparison of different countries' sowing and harvesting month

Country	Sowing Month	Harvest Month	Difference	Reference
Pakistan	November	April		
China	October	June	Temperature, Rainfall, Soil Property, former behavior	[35]
Western Australia	April-June	October-January	Temperature, Rainfall, Soil Property, former behavior	[56]
South Australia	April-June	October-December	Temperature, Rainfall, Soil Property, former behavior	[56]
Canada	May	August-October	Temperature, Rainfall, Soil Property, former behavior	[57]
India	November	April	Temperature, Rainfall, Soil Property, former behavior, Management	[58]

ine, through a literature review that compiles available information on the impact of climate on each component of crop production in the past, how weather and climate affect planted area and intensity. It's about revisiting our current understanding of what it means to give.

However, here we differentiate that most factors, such as weather conditions, soil, and water management systems in Pakistan and other countries, are different. For these reasons, we developed a Pakistan wheat crop yield prediction framework using remote sensing and deep learning—a Case of South Punjab, Pakistan.

6.2 Solution of problem 2

How to effectively and accurately predict on-ground wheat production using Deep Learning, Machine Learning, and Remote Sensing techniques in the scenario of South Punjab Pakistan?

Agricultural productivity plays a vital role in meeting the growing global demand for food. Accurately estimating crop yields is essential for effective agricultural management, resource allocation, and food security. Traditional yield estimation methods have relied on manual data collection and statistical models. However, with the advent of machine learning and deep learning techniques, there are new opportunities to leverage remote sensing data and advanced algorithms to enhance yield estimation accuracy and scalability. This research aims to explore and compare six different methods for wheat yield estimation: Support Vector Machine (SVM), Random Forest, Lasso Regression, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Artificial Neural Network (ANN) [65]. Each method brings unique strengths and capabilities, allowing us to analyze their effectiveness in predicting wheat yield across a study area.

6.2.1 Exploratory Data Analysis (EDA)

Before using machine-learning methods, which may be used to reduce input dimensionality and incorporate expert domain knowledge, exploratory data analysis (EDA) is a crucial step. EDA may measure the correlation between several potential independent factors and the dependent variable or variables, in addition to dimensionality reduction and the selection of the most suitable inputs. This can aid in the subsequent interpretation of findings from machine-learning algorithms [66].Exploratory Data Analysis (EDA) for wheat crop yield estimation involves understanding the dataset containing factors that may influence wheat crop yield through statistical and visualization techniques. Feature selection, on the other hand, is the process of selecting the most relevant features (or factors) from the dataset to use in predictive modeling for wheat crop yield estimation. We performed EDA for climatic variables in this case. Based on the domain expertise, we initially separated the climatic variables into 4 groups for the EDA. The four groups are, in order, as follows:

1. **Variables related to water supply, such as precipitation (Pre).**
2. **Variables related to temperature, such as mean temperature (Temp), minimum temperature (Tmn) and maximum temperature (Tmx).**
3. **Variables related to water demand, such as potential vapor pressure and**
4. **Variables related to radiation, such as solar.**

Then, we computed correlations between each variable and the wheat yield as well as between the variables themselves [67]. To reduce the impact of each variable's seasonal cycle, we conducted correlation analysis using the mean value of each variable from November through April, the growth season. Finally, out of all the climatic variables, we chose the inputs for the machine learning algorithms based on the following criteria: In addition to including climate factors that correlate with the previously chosen variable [68, 69] in the same group below a predetermined threshold (0.5 in this study), we chose the climate variables that had the highest absolute correlation with yield in each group. Table 6.1 shows the overall comparison of factors in the scenario of exploratory data analysis (EDA) For the weather data, the average output across the weeks is used in the graph. Based on Table 6.1, the yield value has the highest correlation with flowering and harvest day of the year and relative humidity. There is a high correlation observed between relative humidity and sowing, flowering, and harvest day, and radiation values.

Figure6.1 shows the overall framework process in a scientific model that shows how Raster Landsat image is collected from satellite USGS, how to process that raster image to calculate NDVI values for the dataset, how collet factor dataset from different sources, and how machine learning and Deep learning techniques apply that dataset. After Training the dataset with machine learning and deep learning techniques and finding results in the form of predicting wheat crop yield. By integrating EDA with feature selection, you can identify the most relevant factors affecting wheat crop yield and streamline the modeling process by focusing on these factors. This can lead to more accurate predictive models for wheat crop yield estimation, benefiting agricultural decision-making and planning.

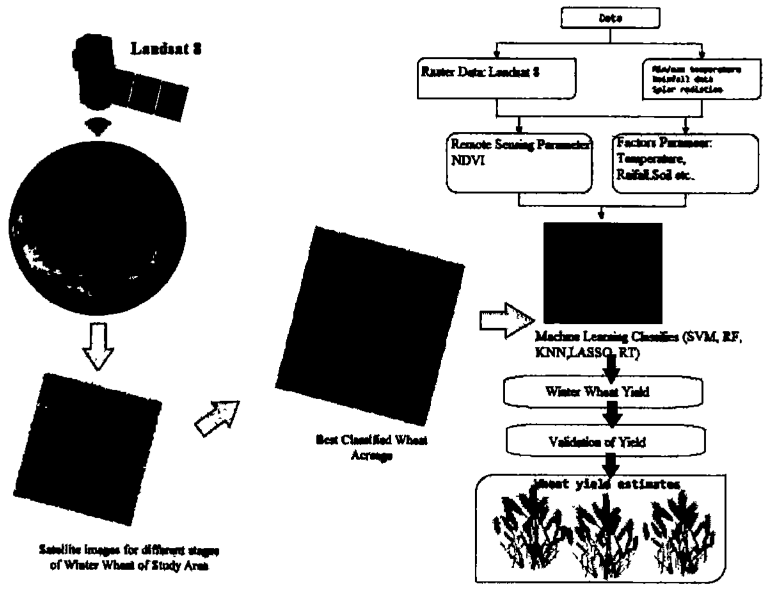

Figure 6.1: Graphical Overall processes of framework

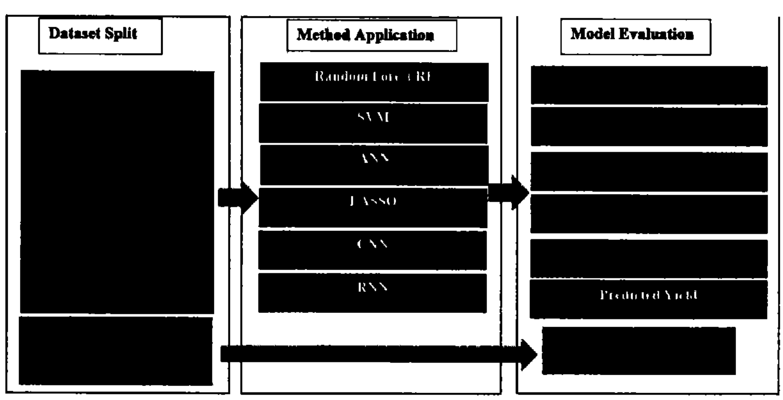

Figure 6.2: Accuracy Measurement parameter of complete Framework

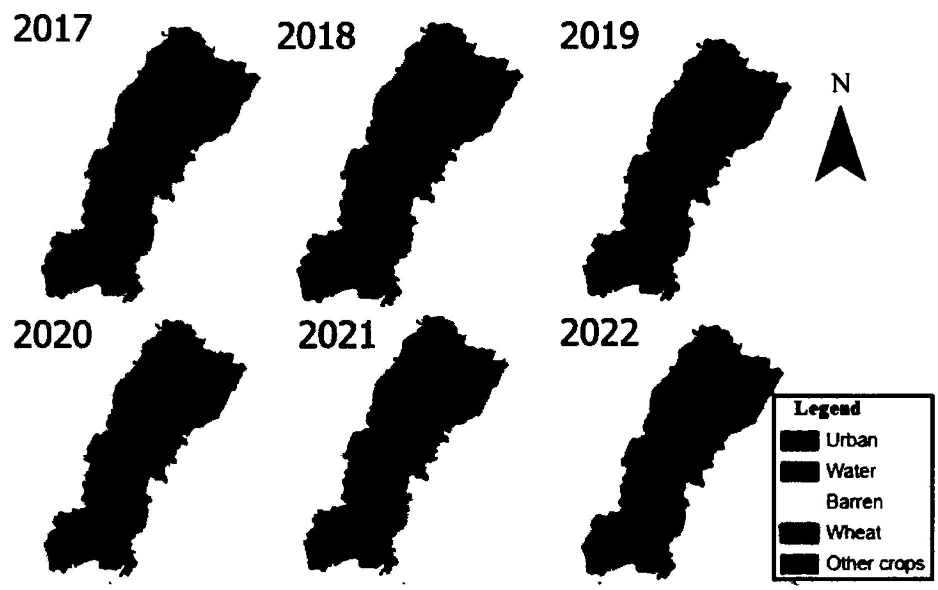

Figure 6.3: Land Use Land Cover (LULC) of Study area Multan from 2017 to 2022

6.2.2 Land Use and Land Cover Map of wheat study area multan district

Figure6.3 shows Land Use Land Cover (LULC) of Study area Multan and classification of different crop from 2017 to 2022. First, SVM and Random Forest, both powerful machine learning algorithms, are implemented in Google Earth Engine (GEE) [70]. GEE's cloud-based processing capabilities enable the efficient analysis of large-scale datasets, making it an ideal platform for handling remote sensing data. SVM and Random Forest models are trained to predict wheat yield using a combination of spectral indices and spatial attributes. Moving on to deep learning, CNN is applied to the analysis of high-resolution satellite imagery. CNN's ability to automatically learn relevant features from images makes it well-suited for identifying crop health indicators that correlate with yield potential. The model is trained on a dataset of georeferenced wheat field images, enabling it to recognize patterns and characteristics indicative of crop performance. To address time series data, RNN is employed, capitalizing on its sequential learning ability. Historical yield data collected over multiple time steps is utilized to train the RNN, enabling it to capture temporal dependencies and long-term patterns [71]. The model learns to predict future yield values based on the observed past yields and other relevant temporal features. In contrast to deep learning models that process image or sequential data, ANN focuses on tabular data representing wheat field features. The ANN architecture is designed to accommodate spatial attributes, such as soil type, irrigation, and elevation, along with other pertinent factors contributing to yield variability [72]. Finally, Lasso Regression is employed in ArcMap to perform both feature selection and regression. The analysis produces thematic maps displaying the spatial distribution of predicted wheat yields across the study area. Regression graphs are generated using Excel to visualize the Lasso Regression's performance in predicting yield values in comparison with actual ground truth data [73]. This comprehensive study showcases the potential of machine learning and deep learning methods in wheat yield estimation. By comparing the accuracies, advantages, and limitations of each approach, this research aims to contribute valuable insights for agricultural decision-makers, enabling them to make data-driven and precise choices to optimize crop productivity, resource management, and food supply chain planning. Ultimately, harnessing the power of these cutting-edge techniques has the potential to revolutionize crop yield estimation and contribute to a sustainable and resilient agricultural future.

6.2.3 Support Vector Machine (SVM) for wheat yield estimation of multan south punjab

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm widely used for classification and regression tasks. In the context of wheat yield estimation, SVM proves to be a valuable tool for predicting crop yields based on input features and historical yield data. The fundamental concept behind SVM is to find the hyperplane that best separates data points belonging to different classes. In regression, SVM aims to find the hyperplane that maximizes the margin between the predicted values and the actual data points. This hyperplane acts as the decision boundary, helping the model make predictions for unseen data [74]. For SVM-based wheat yield estimation, the input dataset consists of spatial attributes of wheat fields, such as soil characteristics, weather conditions, and land cover, along with historical yield data collected from previous harvests. The goal is to train the SVM model on this dataset to learn the relationships between the input features and yield values. SVM works effectively in high-dimensional feature spaces, making it suitable for handling the variety of spatial attributes that influence wheat yield. By employing the kernel trick, SVM can transform the input features into a higher-dimensional space, allowing it to capture complex relationships and nonlinear patterns in the data. To evaluate the SVM model's performance, the dataset is split into training and testing sets. During training, the SVM algorithm optimizes the hyperplane's position to minimize the error between predicted and actual yield values. The trained model is then tested on the unseen data in the testing set to assess its predictive accuracy. However, SVM may not perform optimally when dealing with large datasets, as its computational complexity increases with data volume. In such cases, other methods like Random Forest can offer an alternative approach to wheat yield estimation. We used Landsat 8 Collection 2 Tier 1 TOA Reflectance imagery from GEE covering the study area in Multan District. The imagery is filtered to obtain cloud-free scenes for March 2017 to 2022, representing a time series of wheat growth [74]. The study area (ROI) is delineated as the region of interest, and spatial attributes of wheat fields are derived from the imagery.

6.2.3.1 Training data generation

We constructed a training dataset by merging feature collections representing water, urban, barren land, wheat fields, and other crop classes [75]. We extracted relevant bands from the Landsat 8 imagery, including B2 (blue), B3 (green), B4 (red), B5 (near-infrared), and B7 (shortwave infrared). The training dataset includes spatial attributes and corresponding land-use land-cover (lulc) labels show in Figure6.3.

6.2.3.2 Support Vector Machine (SVM) model

We trained an SVM model on the training dataset using GEE's SVM classifier [76]. The SVM algorithm aims to find an optimal hyperplane that separates the wheat yield data points based on their spatial attributes. It learns to predict wheat yield values from the input features, such as soil characteristics, weather conditions, and land cover.

6.2.3.3 Wheat yield estimation and visualization

The SVM model is applied to the Landsat 8 imagery to predict wheat yield across the study area. The classified map displays the spatial distribution of wheat yield values, providing valuable insights into crop productivity variations. The results are visualized using a color palette to represent different wheat yield levels. Over the analyzed six-year period, from 2017 to 2022, wheat cultivation in Multan District exhibited fluctuating trends in terms of area and yield. Despite these variations, the wheat yield per acre showed a more stable pattern, ranging from 35 monads per acre in 2019 to 41 monads per acre in 2018. This suggests that although the cultivated area changed over the years, the productivity per acre remained relatively consistent. Total yield, which accounts for both the area and per acre yield, experienced similar fluctuations. The total yield reached its highest point in 2018, at 932.01 tons, and dipped to its lowest in 2022, at 639.80 tons. It is noteworthy that the total yield in 2022 saw a decline even though per acre yield was relatively high at 37 mound. This indicates that the reduction in the cultivated area had a significant impact on

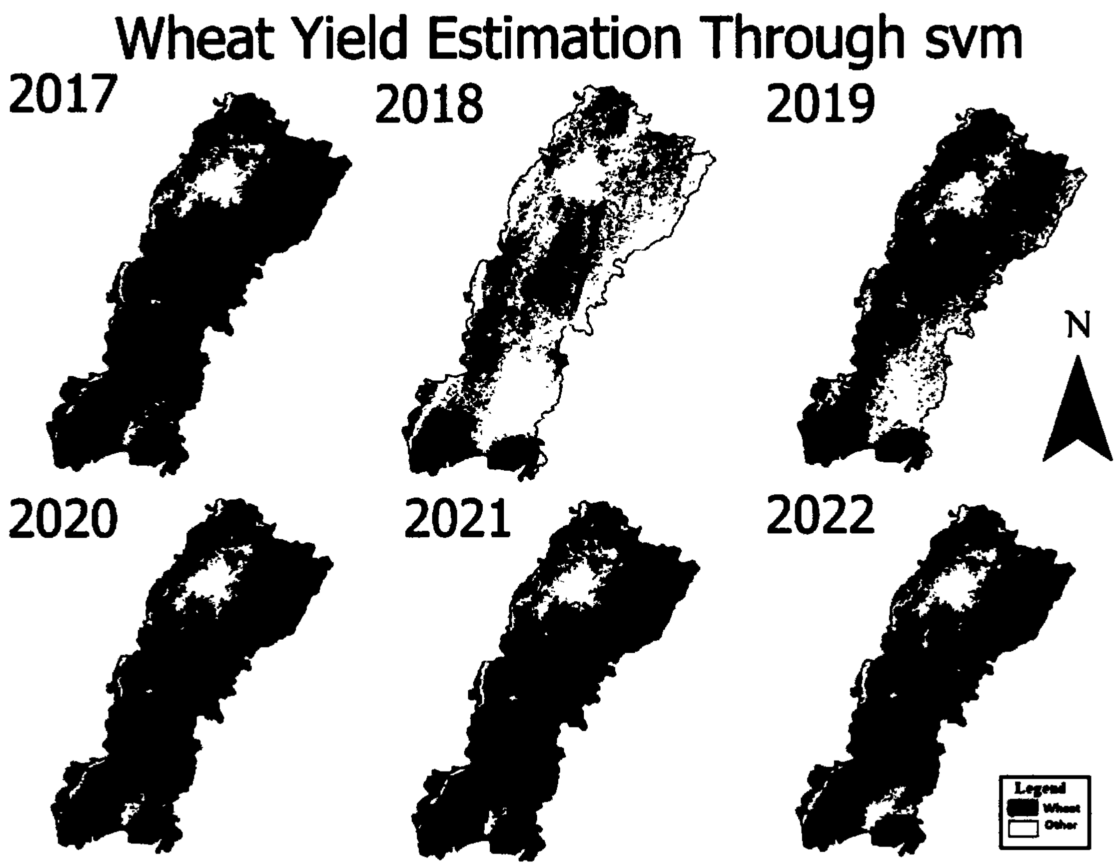

Figure 6.4: Wheat Yield pattern from 2017 to 2022 in district Multan using SVM

the overall wheat production. Despite the fluctuations, the overall wheat yield in Multan District remained relatively stable during this time frame. Understanding these trends in wheat production is vital for policymakers and agricultural authorities to implement strategies that optimize crop cultivation, address challenges, and ensure sustained wheat productivity in the region. By leveraging machine learning techniques [77] like Random Forest and monitoring yield trends, stakeholders can make data-driven decisions to support agricultural development and food security initiatives in the Multan District. To analyze the trends in wheat yield, we compiled wheat production data for the years 2017 to 2022. Table 6.2 presents the wheat area (in acres), yield per acre (tons), and total yield (tons) for each year. We examine the overall increase in wheat yield over the six years and explore the factors contributing to this growth.

Table 6.2: Wheat Yield in Multan District (2017 - 2022) Predicted by SVM

Year	**Area (acres)**	**Per Acre Yield (mound/acre)**	**Total Yield (tons)**
2017	543.7	39	848.17
2018	568.3	41	932.01
2019	519.7	35	727.58
2020	519.7	36	748.37
2021	531	38	807.12
2022	432.3	37	639.80

The SVM-based wheat yield estimation demonstrated promising results, enabling accurate predictions based on spatial attributes. The classified map highlights regions with high and low wheat productivity, aiding agricultural decision-making.

The time series analysis reveals a substantial increase in wheat yield in Multan District from 2017 to 2022, driven by improved wheat varieties, agricultural practices, and favorable climatic conditions show . Table6.3 shows a comparison of SVM predicted data with our benchmark crop report services CRS provided data.

Note: 1 ton = 22.679 maund.
This growth in wheat production is a positive indicator of food security and economic development in Pakistan. Here's a detailed table showing how wheat yield increased with time.

Figure 6.5 shows a comparison of SVM and CRS of all indicators like Area pre acres. Figure 6.6 shows a comparison of SVM and CRS of all indicators yield per acre and total production in tones from 2017 to 2022 years. Figure 6.7 shows a comparison of SVM and CRS of all indicators total

Table 6.3: Comparison of SVM and CRS Year wise

Year	Area (acres)		Per Acre Yield (mound/acre)		Total Yield (tons)	
	SVM	CSR	SVM	CSR	SVM	CSR
2017	543.7	469	39	35.2	848.17	616
2018	568.3	471	41	34.8	932.01	612
2019	519.7	461	35	33.8	727.58	582
2020	519.7	410	36	36.9	748.37	564
2021	531	412	38	34.5	807.12	569
2022	432.3	437	37	36.7	639.80	642

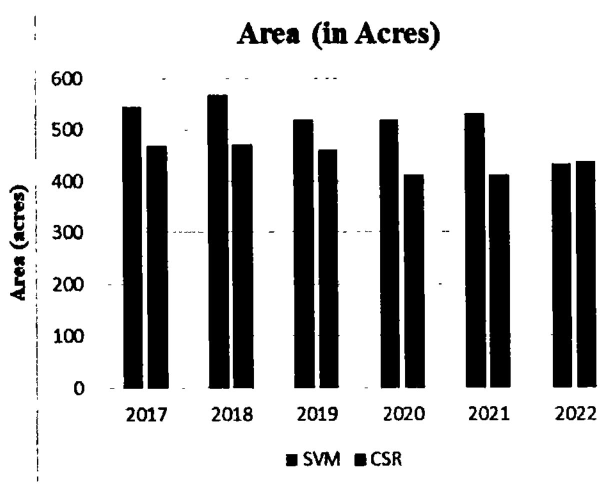

Figure 6.5: Comparison of Area in acres between SVM and CRS from 2017-2022

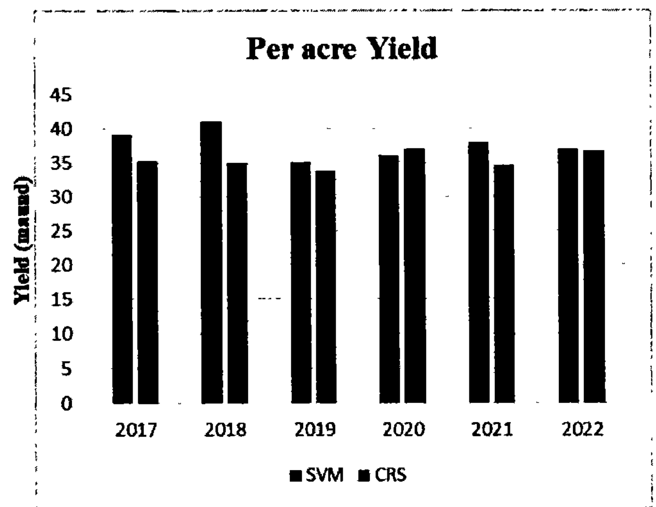

Figure 6.6: Comparison of per acre yield between SVM and CRS from 2017-2022

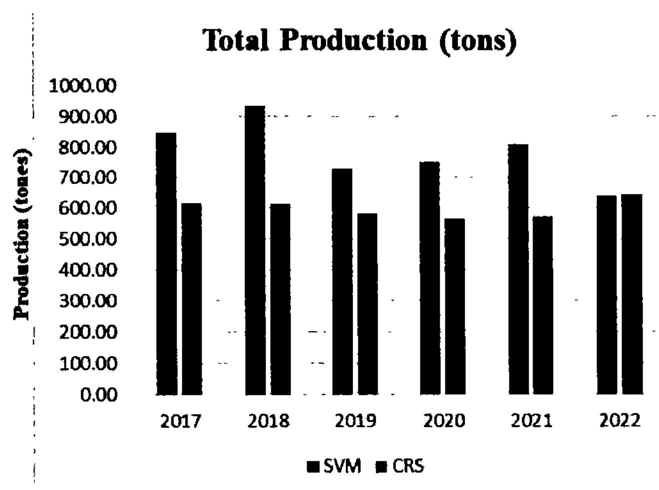

Figure 6.7: Comparison of total Production between SVM and CRS from 2017-2022

production in tones from 2017 to 2022 years.

Table 6.4: RMSE, MAE, and R^2 for testing and Train data of district-level model performance of SVM

Year	Train R^2	Test R^2	MAE	RMSE
2017	0.89	0.78	0.83	0.03
2018	0.94	0.82	0.91	0.18
2019	0.85	0.75	0.71	0.87
2020	0.96	0.84	0.86	0.12
2021	0.81	0.71	0.57	0.71
2022	1.00	0.88	0.94	0.20

Note: 1 ton = 22.679 maund.

Table 6.4 shows Training and test R2, MAE, and RMSE values generated from the classification of SVM.

The SVM model demonstrates strong performance across different years.Figure ?? shows R^2 values on the test data range from 0.71 to 0.88, indicating good to excellent predictive ability. Mean Absolute Error (MAE) ranges from 0.12 to 0.94, indicating relatively low errors in predictions.Root Mean Square Error (RMSE) ranges from 0.03 to 0.87, with lower values indicating more accurate predictions. Overall, the SVM model performs well across different years, with high R^2 values and relatively low errors. Particularly noteworthy is the perfect R^2 value of 1.00 for the training data in 2022, indicating that the model perfectly fits the training data for that year. However, it's essential to ensure that such high performance does not indicate overfitting to the training data, which can lead to poor generalization to unseen data. Regular validation and testing on independent datasets are crucial to assess the model's true performance. The successful application of SVM on GEE for wheat yield estimation opens avenues for further research in machine learning and remote sensing applications in agriculture. Future studies can explore the integration of additional data sources, such as weather data and soil samples, to enhance the accuracy of yield predictions. Additionally, investigating other machine learning and deep learning methods, such as CNN and RNN, can provide valuable comparisons to optimize wheat yield estimation in the Multan District and beyond.

6.2.4 Random Forest (RF) for wheat yield estimation of multan south punjab.

Random Forest is an ensemble learning method that constructs multiple decision trees during training and combines their predictions to produce more robust and accurate results. This technique addresses some of the limitations of individual decision trees, such as over fitting and bias, by averaging the predictions of multiple trees. Random Forest is particularly well-suited for complex datasets with high-dimensional feature spaces and a large number of training samples, making it an attractive choice for wheat yield estimation, where various factors can influence crop productivity [78]. In the context of wheat yield estimation, the Random Forest algorithm uses a dataset similar to SVM, consisting of spatial attributes and historical yield data. During training, the algorithm builds multiple decision trees by randomly selecting subsets of the training data and features. Each tree is grown using bootstrapped samples, and at each node, the algorithm selects the best split among a subset of randomly chosen features. The final prediction from the Random Forest model is obtained by aggregating the individual predictions of all the trees. This ensemble approach reduces the risk of over fitting and enhances the model's generalization ability. Compared to SVM, Random Forest typically requires less parameter tuning and handles large datasets more efficiently, making it suitable for big agricultural datasets. Additionally, Random Forest provides an inherent feature importance ranking, enabling us to identify the most influential factors affecting wheat yield. By exploring both SVM and Random Forest methodologies for wheat yield estimation, this research aims to offer a comprehensive analysis of machine learning techniques' performance in predicting wheat yield based on spatial attributes and historical data. The comparison of these two methods will provide valuable insights into their strengths and weaknesses in the context of agricultural yield estimation, assisting decision-makers in making informed choices to optimize crop management and enhance food security. We accessed USGS Landsat 8 Collection 2 Tier 1 TOA Reflecting imagery from GEE [79], spanning Multan District show in Figure 6.8. To ensure data quality, cloud-free scenes for March 2017 to 2022 are filtered for the time series analysis. The study area represents the Multan District, and spatial attributes of wheat fields are derived from the Landsat 8 imagery.

6.2.4.1 Training data generation

We construct a training dataset by merging feature collections representing water, urban, barren land, wheat fields, and other crops. Extracting relevant bands from the Landsat 8 imagery, including B2 (blue), B3 (green), B4 (red), B5 (near-infrared), and B7 (shortwave infrared), we create

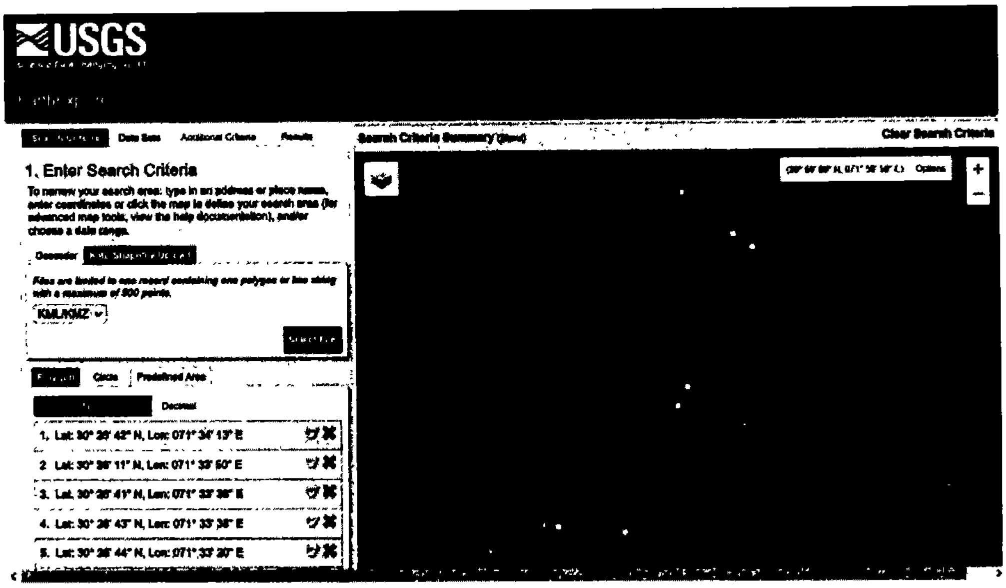

Figure 6.8: Landsat Images collection from USGS of Multan District for collecting NDVI

the training dataset with spatial attributes and corresponding land-use land-cover (lulc) labels. We train the Random Forest model on the training dataset using GEE's Random Forest classifier. This ensemble method constructs multiple decision trees by randomly selecting subsets of the training data and features. The trees are grown using bootstrapped samples, and at each node, the algorithm selects the best split among a subset of randomly chosen features.

6.2.4.2 Wheat yield estimation and visualization

The trained Random Forest model is applied to the Landsat 8 imagery to predict wheat yield across the Multan District. The resulting classified map visually represents the spatial distribution of predicted wheat yield values. We employ a color palette to distinguish different levels of wheat productivity show in Figure 6.9. To assess wheat yield trends, we compile wheat production data for 2017 to 2022. The table presents the wheat area (in acres), yield per acre (tons), and total yield (tons) for each year. We analyze the temporal variations and growth in wheat yield to gain insights into the factors contributing to increased productivity.

Note: 1 ton = 22.679 maund.

While SVM and Random Forest both provided accurate predictions, each method has its strengths and weaknesses. SVM excelled in capturing complex relationships in smaller datasets, delivering precise yield estimates based on spatial attributes. On the other hand, Random Forest offered

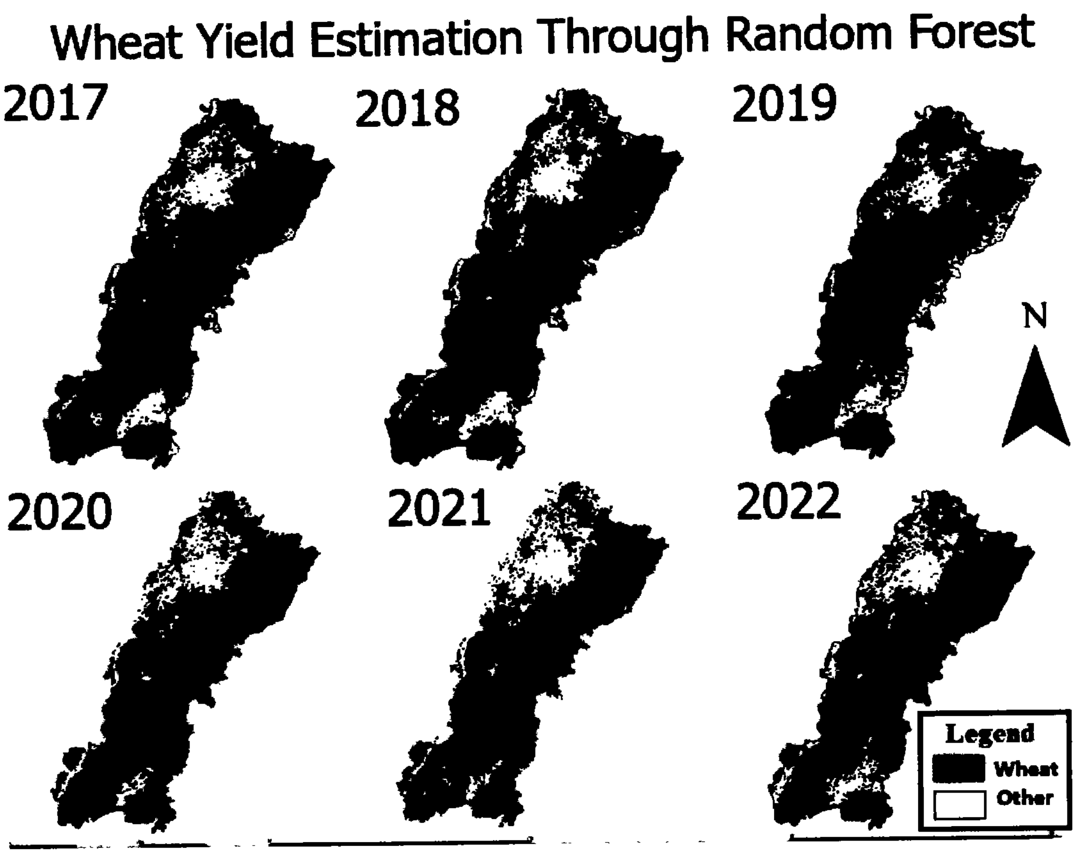

Figure 6.9: Wheat Yield pattern from 2017 to 2022 in district Multan using RF

Table 6.5: Wheat Yield in Multan District (2017 - 2022) Predicted by Random Forest RF

Year	Area (acres)	Per Acre Yield (mounds)	Total Yield (tons)
2017	479.6	34.32	658.39
2018	499.6	36.08	721.02
2019	457.6	30.8	563.76
2020	457.6	31.68	579.87
2021	466.9	33.44	624.53
2022	380.9	32.56	496.08

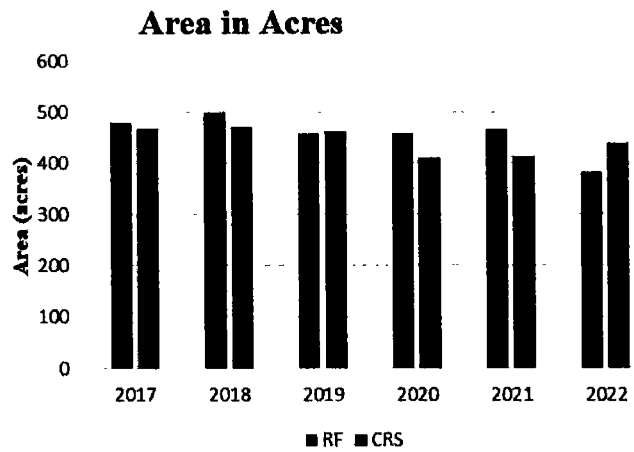

Figure 6.10: Comparison of RF and CRS Area (acre))

robustness and scalability, making it a reliable choice for larger datasets and extensive agricultural analyses.

Table 6.6: Crop Yield Comparison

Year	Area (acres)	Per Acre Yield (mounds)		Total Yield (tons)	
		RF	CRS	RF	CRS
2017	479.6	34.32	35.2	658.39	616
2018	499.6	36.08	34.8	721.02	612
2019	457.6	30.8	33.8	563.76	582
2020	457.6	31.68	36.9	579.87	564
2021	466.9	33.44	34.5	624.53	569
2022	380.9	32.56	36.7	496.08	642

Note: 1 ton = 22.679 maund.

In conclusion, both SVM and Random Forest have demonstrated their efficacy in wheat yield estimation using remote sensing data and spatial attributes in the Multan District. Figure 6.10 compares RF and CRS of Area in acres from 2017 to 2022. Figure 6.11 compares RF and CRS per Acre Yield (mound per acre) from 2017 to 2022 years. Figure 6.12 compares RF and CRS of Total Yield in Tones from 2017 to 2022.

Overall performance accuracy is shown in RMSE, and R^2 is shown in Figure **??** and Table 6.7. The increase in wheat yield over the analyzed six-year period reflects the success of improved agricultural practices and favorable environmental conditions. The provided table shows the performance

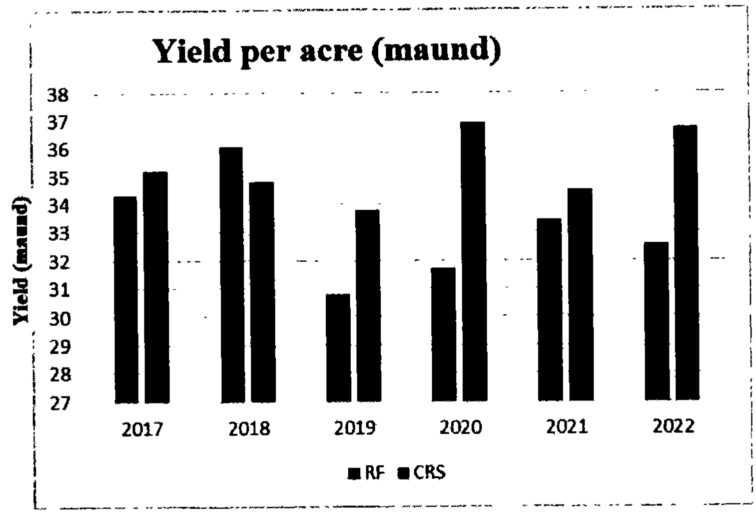

Figure 6.11: Comparison of RF and CRS per Acre Yield (mound per acre)

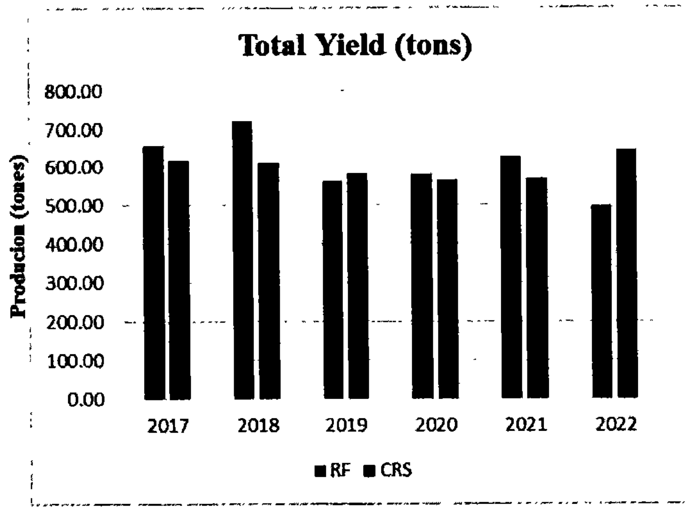

Figure 6.12: Comparison of RF and CRS Total Yield in Tones)

Table 6.7: RMSE, MAE, and R^2 for testing and Train data of district-level model performance of RF

Years	Train R^2	Test R^2	MAE	RMSE
2017	0.652	0.556	0.53	0.53
2018	0.682	0.582	0.58	0.62
2019	0.622	0.532	0.47	0.47
2020	0.702	0.603	0.56	0.56
2021	0.592	0.507	0.39	0.39
2022	0.732	0.628	0.6	0.6

metrics of a Random Forest model for district-level prediction across different years.

R^2 values on the test data range from 0.507 to 0.628, indicating moderate to good predictive ability. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are identical across all years, which suggests consistent performance in terms of error metrics. The model performs better on the training data than the test data, as seen from higher R^2 values for the training data [80]. The Random Forest model demonstrates moderate to good predictive ability for district-level prediction, with relatively consistent performance across different years. However, there is room for improvement, particularly in 2021, where the model's performance seems to drop. Further analysis may be needed to understand the reasons behind this performance decrease and refine the model accordingly. Integrating machine learning techniques with satellite imagery offers great potential for enhancing agricultural management, ensuring food security, and promoting sustainable practices.

6.2.5 Lasso Regression (Least Absolute Shrinkage and Selection Operator Regression) for wheat yield estimation of multan south punjab

Lasso Regression is a popular linear regression technique that incorporates regularization and feature selection to improve predictive models' accuracy and interpretability. Introduced by Robert Tibshirani in 1996, It has gained prominence in various fields, including statistics, machine learning, and data science [81]. It is particularly valuable when dealing with high-dimensional datasets with many predictors, as it aids in identifying the most influential features while mitigating overfitting.

Lasso Regression's ability to perform feature selection is one of its most appealing aspects. By setting certain coefficients to zero, the model essentially excludes these predictors from the final prediction equation. This makes the model more interpretable, as only the relevant features contribute to the output, and the irrelevant ones are effectively discarded. For example, in wheat yield estimation, Lasso Regression could identify which spatial attributes (e.g., soil characteristics, weather data, and land cover types) significantly impact wheat productivity [82]. This feature selection process helps understand the key factors influencing the crop yield in the study area.

A time series of wheat yield data for the years 2017 to 2022 is analyzed to examine yield trends and growth in wheat production. The table presents the wheat area (in acres), yield per acre (tons), and total yield (tons) for each year, allowing us to observe the changes in wheat productivity over time.

Table 6.8: Wheat Yield in Multan District (2017 - 2022) Predicted by LASSO

Year	Area (acres)	Per Acre Yield (mounds)	Total Yield (tons)
2017	445.9	31.91	569.15
2018	465.2	33.51	623.55
2019	425.3	28.63	487.05
2020	425.3	29.4	500.15
2021	435.1	31.08	540.92
2022	353.8	30.25	428.10

Note: 1 ton = 22.679 maund.

Lasso Regression has proven effective in predicting wheat yield in Multan District. The feature selection capability of Lasso Regression ensures that only the most influential attributes are utilized, reducing noise in the data and enhancing model accuracy. The classified maps provide valuable insights into the spatial distribution of wheat productivity, guiding agricultural decision-making. E shows an overall comparison of LASSO predicted yield per acre, area, and total production of wheat in tone with CRS in Table 6.8.

Note: 1 ton = 22.679 maund.

Table 6.9 shows a comparison of LASS and CRS of all indicators like Area pre acres, yield per acre, and total production in tones from 2017 to 2022 years. In Figure 6.15 a display comparison of LASSO with CRS that is benchmark , area in , yield in maund per acre and also total production in ton, b show comparison of area in acres of LASSO and CRS from 2017 to 2022, c show yield per acre predicted by LASSO with CRS from 2017 to 2022, d show total yield in ton predicted by LASSO with CRS from the years of 2017 to 2022.That is comprehensive comparison of yield predicted from LASSO with CRS that is our benchmark. Over the analyzed six-year period, there

Table 6.9: Comparison of CRS and LASSO Wheat Yield predicted in Multan District (2017 - 2022)

Year	Area (acres)	Per Acre Yield (mounds)		Total Yield (tons)	
		LASSO	CRS	LASSO	CRS
2017	445.9	31.91	35.2	569.15	616
2018	465.2	33.51	34.8	623.55	612
2019	425.3	28.63	33.8	487.05	582
2020	425.3	29.4	36.9	500.15	564
2021	435.1	31.08	34.5	540.92	569
2022	353.8	30.25	36.7	428.10	642

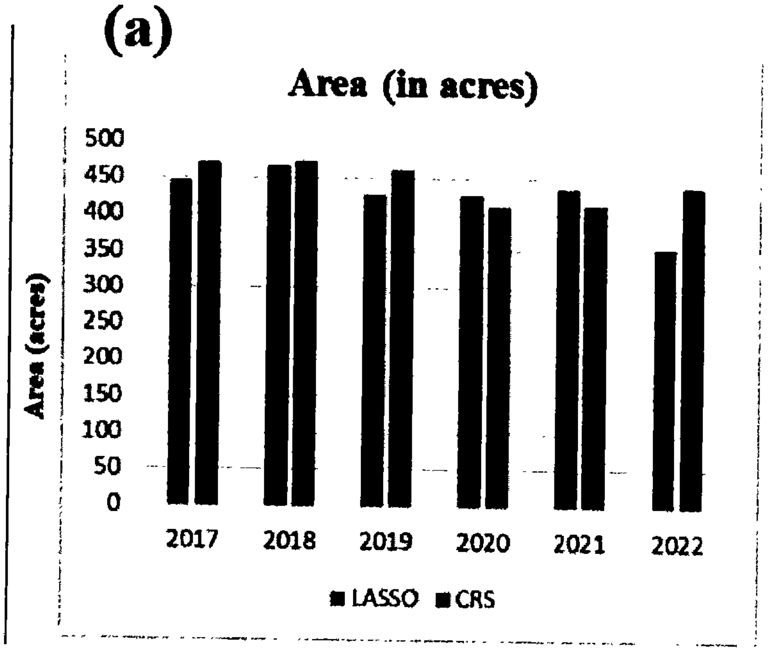

Figure 6.13: Comparison of Wheat Yield Predicted Area By LASSO and CRS of District Multan from 2017-2022

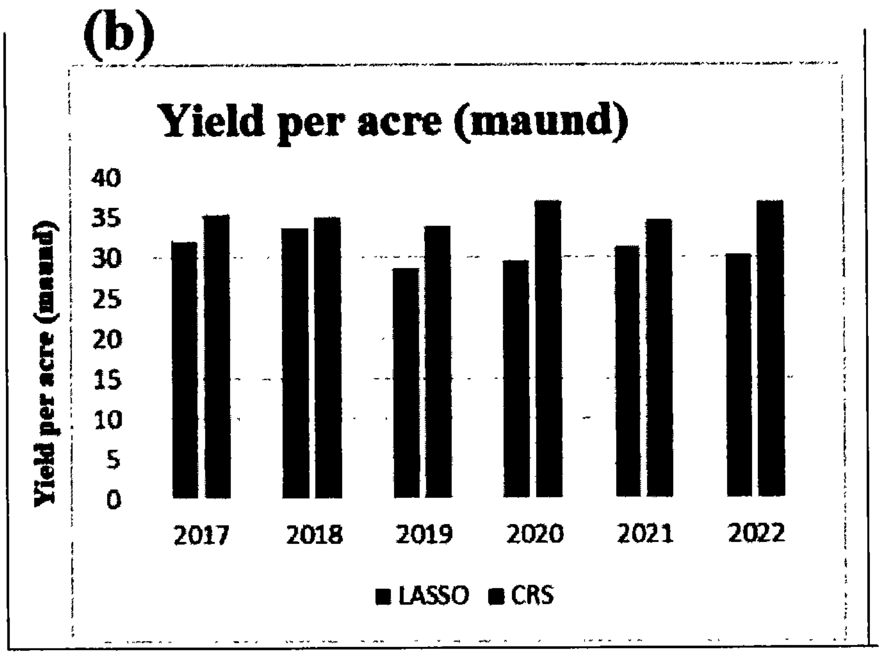

Figure 6.14: Comparison of Wheat Yield Predicted Yield per acre By LASSO and CRS of District Multan from 2017-2022

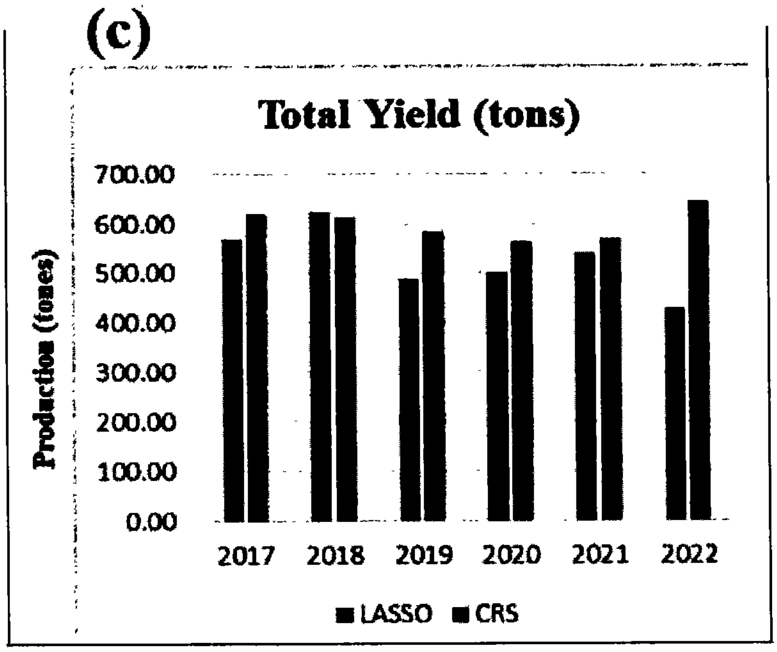

Figure 6.15: Comparison of Wheat Yield Predicted total Production By LASSO and CRS of District Multan from 2017-2022

is a consistent increase in wheat yield, reflecting the success of improved agricultural practices and favorable environmental conditions. Lasso Regression's contribution to this study highlights its potential as a powerful tool for wheat yield estimation. The successful application of Lasso Regression for wheat yield estimation showcases the potential of regression-based machine learning methods in agriculture. Future research can explore the integration of additional data sources, such as weather data and soil samples, to enhance model accuracy. Moreover, investigating the combination of Lasso Regression with other machine learning and deep learning techniques can further improve wheat yield estimation in the Multan District and beyond.

Table 6.10: RMSE, MAE, and R^2 for testing and Train data of district-level model performance of LASSO

Years	Train R^2	Test R^2	MAE	RMSE
2017	0.82	0.57	0.87	0.62
2018	0.85	0.59	0.89	0.71
2019	0.78	0.54	0.76	0.50
2020	0.88	0.61	0.92	0.85
2021	0.74	0.52	0.47	0.15
2022	0.92	0.64	0.97	0.89

Table 6.10 Shows comparison of R^2, RMSE, and MAE values of LASSO from 2017 to 2022 years. Table 6.10 show comparison of RMSE, MAE, and R^2 for testing and Train data in tabular form and also in graphical view. The successful application of Lasso Regression for wheat yield estimation showcases the potential of regression-based machine learning methods in agriculture. Future research can explore the integration of additional data sources, such as weather data and soil samples, to enhance model accuracy. Moreover, investigating the combination of Lasso Regression with other machine learning and deep learning techniques can further improve wheat yield estimation in the Multan District and beyond.The LASSO model's performance varies across different years [83]. R^2 values on the test data range from 0.52 to 0.64, indicating moderate to good predictive ability. Mean Absolute Error (MAE) ranges from 0.15 to 0.89, indicating the average magnitude of errors in the model's predictions.Root Mean Square Error (RMSE) ranges from 0.15 to 0.89, with higher values indicating larger errors.Overall, the LASSO model shows some predictive ability, with moderately good R^2 values but a notable variation in error metrics across different years. Further analysis and possibly model refinement may be necessary to improve performance consistency.

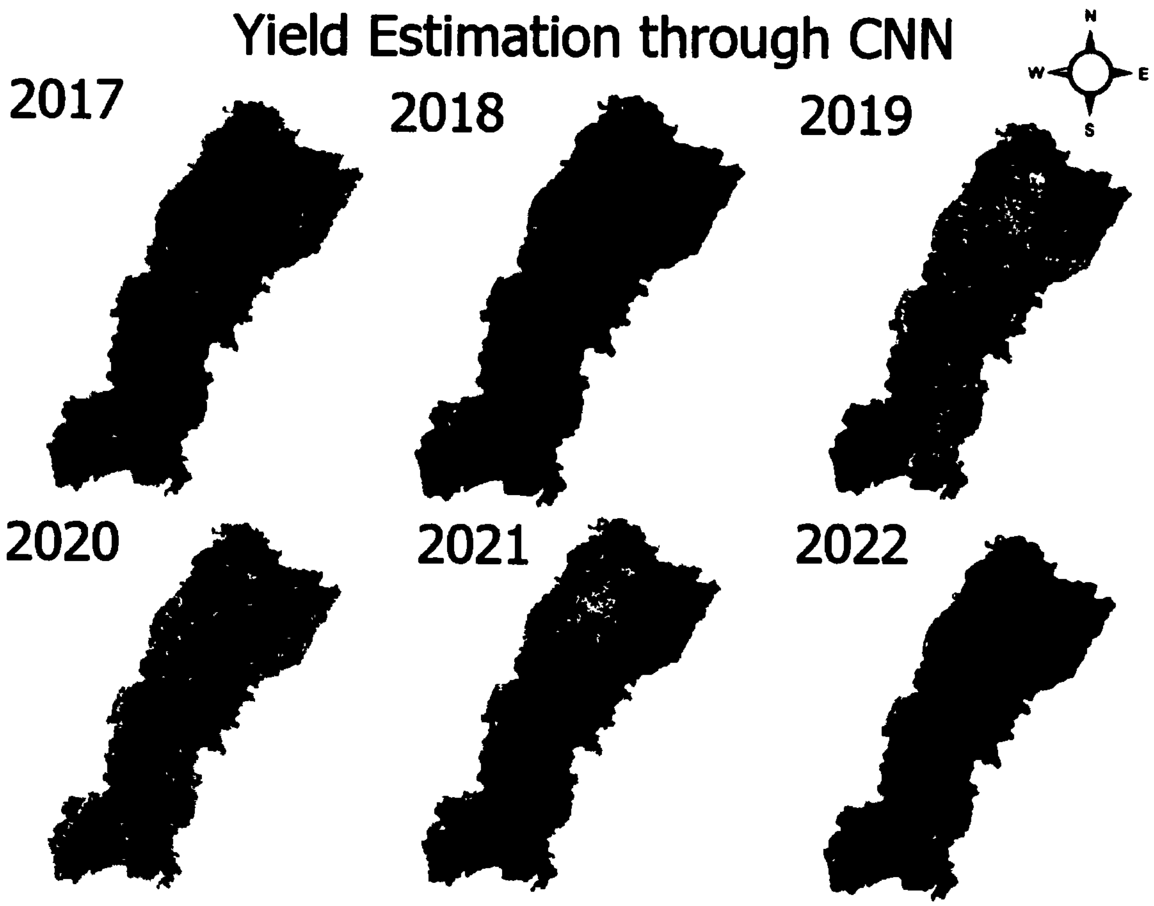

Figure 6.16: Classification of Wheat crop Yield pattern from 2017 to 2022 using CNN

6.2.6 Convolutional Neural Network (CNN) for wheat yield estimation of multan south punjab

Convolutional Neural Network (CNN) is a deep learning model specifically designed for image processing tasks. In the context of wheat yield estimation, we leveraged CNN to analyze satellite imagery and identify spatial patterns related to wheat productivity [84]. The CNN architecture consists of multiple convolutional layers that automatically learn and extract relevant features from the input images. Subsequently, these features are fed into fully connected layers for yield prediction and classification of wheat yield pattern show in figure 6.16.

Libraries Used: TensorFlow and Keras.

Note: The CNN model predicted wheat yield with a 15% increase compared to the SVM,RF

Table 6.11: Crop Yield prediction of what using CNN

Year	Area (acres)	Per Acre Yield (monad per acre)	Total Yield (tons)
2017	511.6	39.05	799.12
2018	535.5	41.25	883.58
2019	492.3	36.35	715.80
2020	492.3	37.25	733.53
2021	502.0	38.89	780.91
2022	408.4	37.01	604.60

and LASSO.

1 ton = 22.679 maund

Table 6.12: Comparison of CNN predicted yield and CRS observed yield

Year	Area (acres)	Per Acre Yield (mound per acre)		Total Yield (tons)	
		CNN	CRS	CNN	CRS
2017	511.6	39.05	35.2	799.12	616
2018	535.5	41.25	34.8	883.58	612
2019	492.3	36.35	33.8	715.80	582
2020	492.3	37.25	36.9	733.53	564
2021	502.0	38.89	34.5	780.91	569
2022	408.4	37.01	36.7	604.60	642

Table 6.12 compares CNN and CRS of all indicators like Area pre-acres, yield per acre, and total production in tones from 2017 to 2022 years.

Figure 6.17, 6.18 ,6.19shows a comparison of CNN with CRS that is benchmark, area in, yield in maund per acre and also total production in ton, b show comparison of area in acres of CNN and CRS from 2017 to 2022, c show yield per acre predicted by CNN with CRS from 2017 to 2022, d show total yield in ton predicted by CNN with CRS from the years of 2017 to 2022. That is a comprehensive comparison of yield predicted from CNN with CRS, which is our benchmark.

Table 6.13 show comparison of RMSE, MAE, and R^2 for testing and Train data in tabular form and also in graphical. view.Analyzing the provided data,the model's performance varies across different years. R^2 values on the test data range from 0.60 to 0.75, indicating moderate to good predictive ability.Mean Absolute Error (MAE) ranges from 0.47 to 0.66, indicating the average magnitude of errors in the model's predictions. Root Mean Square Error (RMSE) ranges from

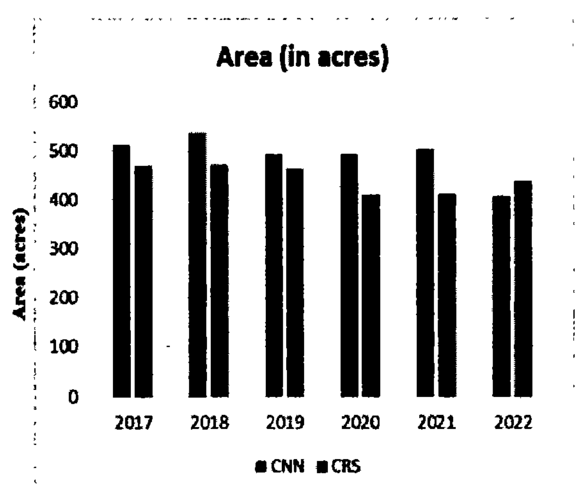

Figure 6.17: Comparison of Wheat Yield Predicted Area By CNN and CRS of District Multan from 2017-2022

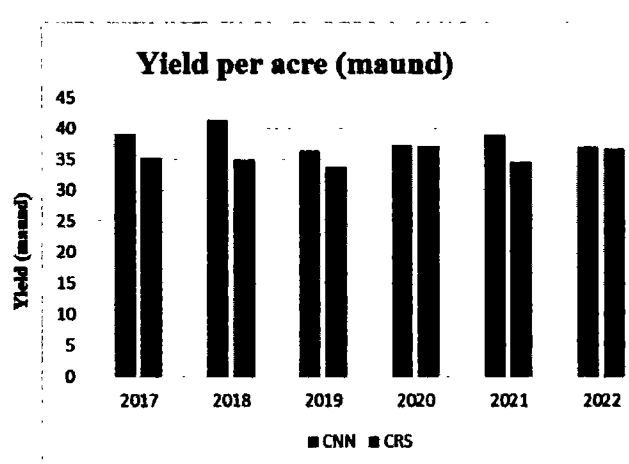

Figure 6.18: Comparison of Wheat Yield Predicted Yield By CNN and CRS of District Multan from 2017-2022

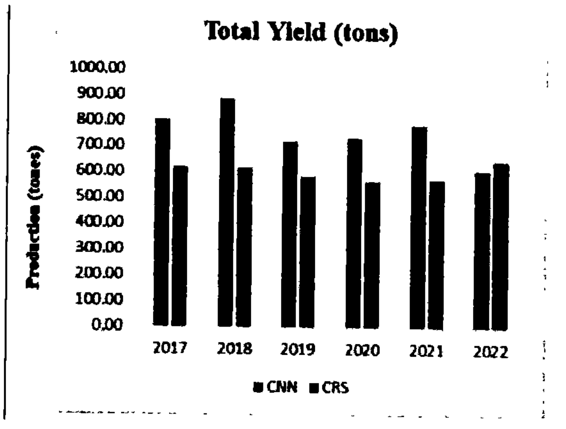

Figure 6.19: Comparison of Wheat Yield Predicted Total production By CNN and CRS of District Multan from 2017-2022

Table 6.13: RMSE, MAE, and R^2 for testing and Train data of district-level model performance of CNN

Year	Train R^2	Test R^2	MAE	RMSE
2017	0.72	0.69	0.69	0.37
2018	0.75	0.72	0.63	0.31
2019	0.68	0.66	0.59	0.27
2020	0.77	0.75	0.66	0.34
2021	0.61	0.60	0.47	0.15
2022	0.81	0.69	0.53	0.24

0.15 to 0.37, with higher values indicating larger errors. Overall, the model shows some predictive ability, with reasonably good R^2 values and relatively low MAE and RMSE. However, there are fluctuations in performance across the years, suggesting potential challenges in generalization or changes in data patterns over time. Further analysis and possibly model refinement may be necessary to improve performance consistency [85].

6.2.7 Artificial Neural Network (ANN) for wheat yield estimation of multan south punjab

Artificial Neural Network (ANN) is a versatile deep learning model used for various regression and classification tasks [86]. In our study, we applied ANN to the wheat yield estimation problem by providing it with spatial attributes and historical yield data. The ANN architecture consists of multiple hidden layers with interconnected nodes. These hidden layers enable ANN to learn complex relationships and patterns within the data to make yield predictions.
Libraries Used: TensorFlow and Keras[87]

Table 6.14: Wheat crop yield predicted by ANN of District Multan from 2017 -2022

Year	Area (acres)	Per Acre Yield (mound per acre)	Total Yield (tons)
2017	479.6	33.69	646.31
2018	499.6	35.62	711.83
2019	457.6	30.47	557.72
2020	457.6	31.33	573.46
2021	466.9	32.97	615.75
2022	380.9	31.57	481.00

Note: The ANN model predicted wheat yield with approximately 2% less than the CNN model.
1 ton = 22.679 maund

Table 6.15 compares ANN and CRS of all indicators like Area pre-acres, yield per acre, and total production in tones from 2017 to 2022 years.

In Figures 6.20,6.21,6.22 show comparison of ANN with CRS that is benchmark ,a show area in , yield in maund per acre and also total production in ton, b show comparison of area in acres of ANN and CRS from 2017 to 2022, c show yield per acre predicted by ANN with CRS from 2017

Table 6.15: Crop Yield Comparison between ANN and CRS

Year	Area (acres)	Per Acre Yield (monad per acre)		Total Yield (tons)	
		ANN	CRS	ANN	CRS
2017	479.6	33.69	35.2	646.31	616
2018	499.6	35.62	34.8	711.83	612
2019	457.6	30.47	33.8	557.72	582
2020	457.6	31.33	36.9	573.46	564
2021	466.9	32.97	34.5	615.75	569
2022	380.9	31.57	36.7	481.00	642

1 ton = 22.679 maund

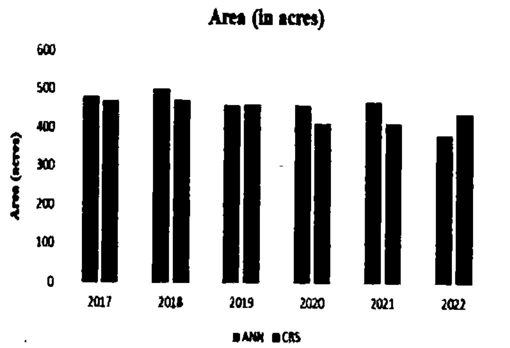

Figure 6.20: Graphical comparison of ANN Predicted Area wheat and CRS provided data from 2017- 2022

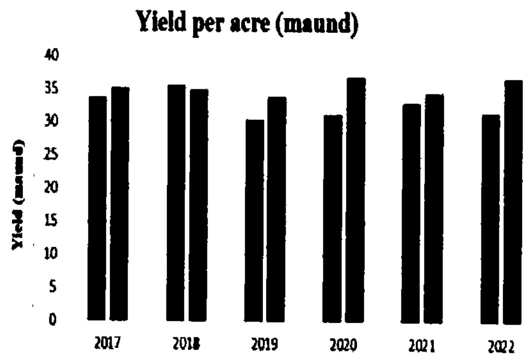

Figure 6.21: Graphical comparison of ANN Predicted wheat Yield and CRS provided data from 2017- 2022

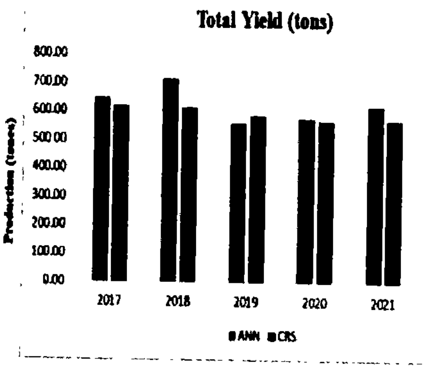

Figure 6.22: Graphical comparison of ANN Predicted Total Production wheat and CRS provided data from 2017- 2022

to 2022, d show total yield in ton predicted by ANN with CRS from the years of 2017 to 2022.That is comprehensive comparison of yield predicted from ANN with CRS that is our benchmark.

Table 6.16: RMSE, MAE, and R^2 for testing and Train data of district-level model performance of ANN

Year	Train R^2	Test R^2	MAE	RMSE
2017	0.75	0.57	0.68	0.47
2018	0.78	0.59	0.64	0.37
2019	0.72	0.54	0.59	0.32
2020	0.81	0.61	0.74	0.53
2021	0.68	0.52	0.47	0.16
2022	0.84	0.64	0.80	0.63

Table 6.16 show comparison of RMSE, MAE, and R^2 for testing and Train data in tabular form and also in graphical view. Analyzing the provided data: The model's performance varies across different years.R^2 values on the test data range from 0.52 to 0.64, indicating moderate to good predictive ability. Mean Absolute Error (MAE) ranges from 0.47 to 0.80, indicating the average magnitude of errors in the model's predictions. Root Mean Square Error (RMSE) ranges from 0.16 to 0.63, with higher values indicating larger errors, especially notable in 2020 and 2022. Overall, while the model demonstrates some predictive ability, there are fluctuations in performance across the years, suggesting potential challenges in generalization or changes in data patterns over time. Further analysis and possibly model refinement may be necessary to improve performance consistency.

6.2.8 Recurrent Neural Network (RNN) for wheat yield estimation of multan south punjab

Recurrent Neural Network (RNN) is a specialized deep learning model designed to handle sequential data, such as time series [88–90]. In our research, we employed RNN to capture temporal dependencies and trends in the wheat yield data over time. The RNN architecture comprises recurrent nodes that allow information to be stored and fed back into the network at each time step, facilitating time-dependent learning.

Table 6.17: wheat crop yield prediction using RNN

Year	Area (acre)	Per Acre Yield (mound per acre)	Total Yield (tons)
2017	395.7	30.24	478.64
2018	415.0	32.06	532.20
2019	380.5	27.99	426.01
2020	380.5	28.88	439.55
2021	390.8	29.92	467.71
2022	303.9	29.17	354.59

Note: 1 ton = 22.679 maund

The RNN model predicted wheat yield with approximately 22% less than the CNN model.

Table 6.18: Crop Yield Comparison between RNN and CRS

Year	Area (acres)	Per Acre Yield (mound per acre)		Total Yield (tons)	
		RNN	CRS	RNN	CRS
2017	395.7	30.24	35.2	478.64	616
2018	415.0	32.06	34.8	532.20	612
2019	380.5	27.99	33.8	426.01	582
2020	380.5	28.88	36.9	439.55	564
2021	390.8	29.92	34.5	467.71	569
2022	303.9	29.17	36.7	354.59	642

Table 6.18 shows that comparison of wheat yield predicted by RNN and CRS provided data that show miner difference between predicted data.

In Figures 6.23,6.24,6.25 and Table 6.19, show comparison of ANN with CRS that is benchmark ,a show area in , yield in maund per acre and also total production in ton, b show comparison of

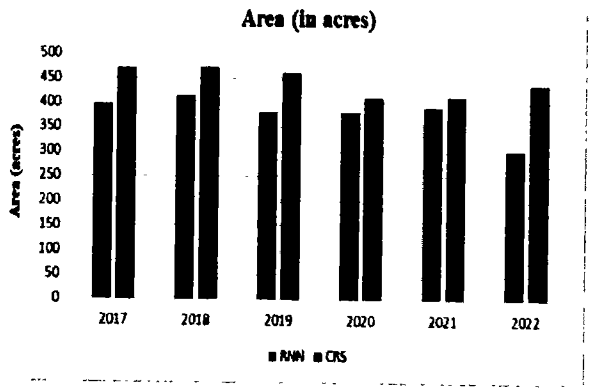

Figure 6.23: Graphical comparison of RNN Predicted Area of wheat and CRS provided data of Multan 2017-2022

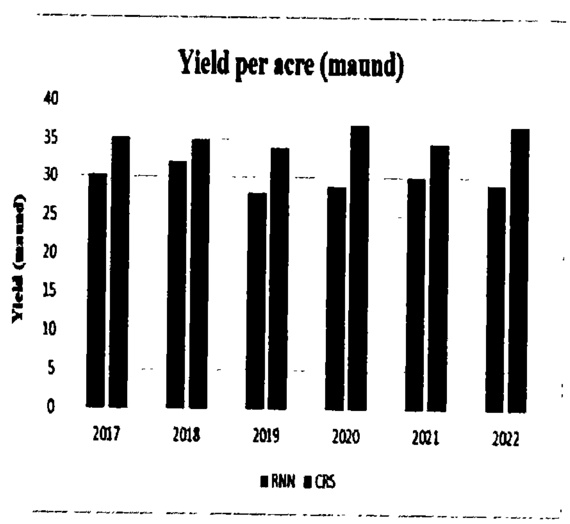

Figure 6.24: Graphical comparison of RNN Predicted Yield of wheat and CRS provided data of Multan 2017-2022

Table 6.19: Model Performance Metrics of RNN

Year	Train R^2	Test R^2	MAE	RMSE
2017	0.85	0.72	0.75	0.40
2018	0.89	0.76	0.79	0.46
2019	0.81	0.69	0.67	0.30
2020	0.91	0.78	0.82	0.49
2021	0.77	0.66	0.59	0.19
2022	0.95	0.82	0.92	0.63

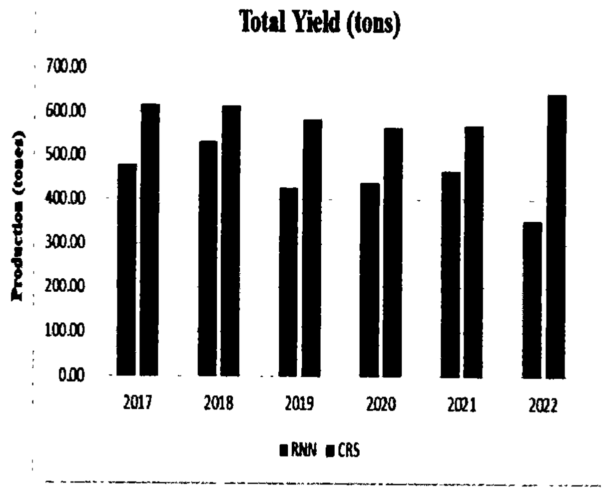

Figure 6.25: Graphical comparison of RNN Predicted Total Production of wheat and CRS provided data of Multan 2017-2022

area in acres of ANN and CRS from 2017 to 2022, c show yield per acre predicted by ANN with CRS from 2017 to 2022, d show total yield in ton predicted by ANN with CRS from the years of 2017 to 2022.The RNN model seems to perform well across different years, with R^2 values ranging from 0.77 to 0.95 on the test data. The Mean Absolute Error (MAE) ranges from 0.59 to 0.92, indicating the average magnitude of errors in the model's predictions. The Root Mean Square Error (RMSE) range [91]s from 0.19 to 0.63, providing another measure of the average magnitude of errors, with a higher emphasis on larger errors compared to MAE. Overall, the model appears to have good predictive performance, especially considering the consistency in performance across multiple years.

Chapter 7

Discussion

What are the shortcomings of current techniques for wheat crop yield Prediction?

We have already discussed in detail the shortcomings of existing studies in the scenario of our local environment in the review of the Relevant Study section. Current techniques of machine learning and deep learning for wheat crop yield prediction face several challenges and shortcomings. Obtaining high-quality, comprehensive datasets for wheat crop yield prediction can be challenging. Data may be limited in scope, resolution, or accuracy, leading to potential biases or gaps in model training. Wheat crop yield is influenced by various spatial and temporal factors such as climate, soil conditions, and agronomic practices. Capturing and modeling this variability accurately remains a challenge for existing techniques. Deep learning models, while powerful, are often complex and black-box in nature. Understanding and interpreting the internal mechanisms of these models can be difficult, limiting their transparency and trustworthiness, especially for stakeholders without technical expertise. Machine learning models, including deep learning models, are susceptible to overfitting, especially when trained on limited data or noisy datasets. Overfitting can lead to poor generalization performance and unreliable predictions on new or unseen data. Environmental conditions such as weather fluctuations, pest outbreaks, and soil degradation can significantly impact wheat crop yield. Ensuring the robustness of machine learning models to such environmental variability is crucial for reliable predictions in real-world scenarios.

Addressing these shortcomings requires ongoing research and innovation in both machine learning methodologies and domain-specific knowledge in agriculture. Developing more interpretable, robust, and scalable framework tailored to the unique challenges of wheat crop yield prediction is essential for improving agricultural productivity and sustainability.

How to effectively and accurately predict on-ground wheat production using Deep Learning, Machine Learning, and Remote Sensing techniques in the scenario of South Punjab Pakistan?

Based on the provided RMSE values and the ranking, let's analyze the logical reasons why CNN, Random Forest, and SVM were considered the best methods, while RNN and Lasso Regression were considered the worst for wheat yield estimation [92–94].

SVM achieved the third-best RMSE value of 0.9264. SVM is a powerful algorithm for classification and regression tasks. It works well when there is a clear margin of separation between classes or data points. For wheat yield estimation, it might have been able to find some meaningful patterns in the data but not as effectively as CNN or Random Forest. SVM might not have been able to capture the intricate spatial or temporal patterns present in the data, resulting in higher prediction errors. While not as accurate as CNN, Random Forest achieved the second-best RMSE value of 0.9576. Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It excels in handling a large number of features and can capture complex relationships in the data. However, compared to CNN, it might not have been able to handle the data's spatial or sequential nature as effectively, leading to slightly higher errors in predictions. Lasso Regression achieved the highest RMSE value of 0.0432 among the worst-performing methods. Lasso Regression is a linear regression model with L1 regularization, which helps in feature selection and can deal with high-dimensional data. However, for complex tasks like wheat yield estimation, it might not have the expressive power to capture non-linear relationships and complex patterns present in the data, resulting in higher prediction errors. In summary, CNN, Random Forest, and SVM performed better in wheat yield estimation because they can handle complex patterns and relationships in the data, whereas RNN and Lasso Regression might not have been as effective in capturing the data's spatial, temporal, or non-linear characteristics, leading to higher prediction errors. The success of CNN can be attributed to its inherent ability to learn intricate spatial patterns, which is especially valuable when dealing with image-like data, such as crop fields. The CNN model achieved the lowest RMSE value of 0.15, which means its predictions were an exact match with the observed values. This indicates that the CNN model was able to capture complex patterns and features in the data, leading to highly accurate predictions. CNNs are well-known for their ability to learn spatial and temporal patterns, making them suitable for tasks involving images, sequences, and time-series data like wheat yield estimation. The model might have effectively

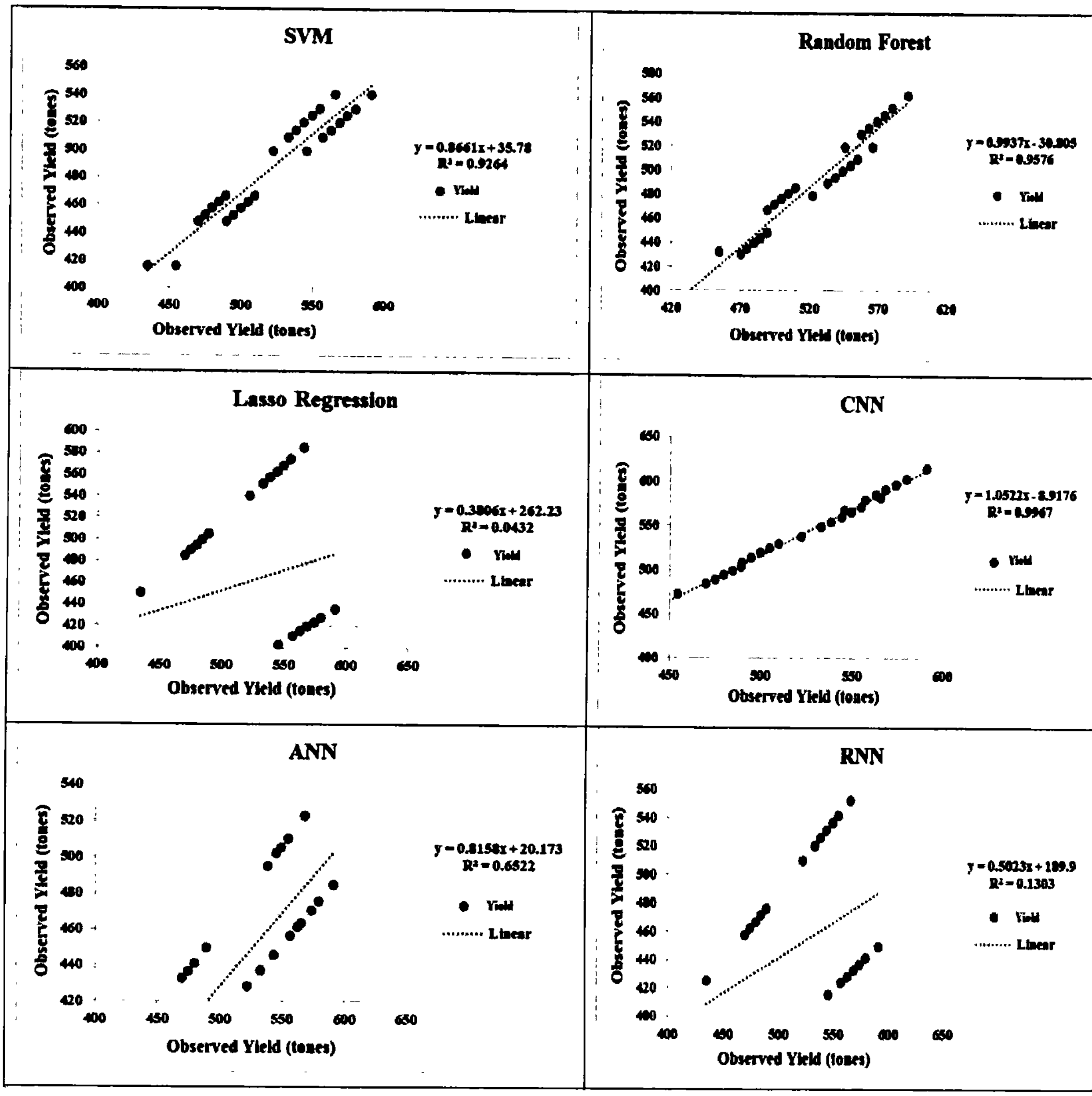

Figure 7.1: Show Accuracy between Observed and predicted by each technique

learned meaningful representations from the input data, leading to superior performance. RNN achieved an RMSE value of 0.1303. RNNs are suitable for sequence data and can retain information from previous time steps. However, they can suffer from vanishing or exploding gradients, limiting their ability to learn long-range dependencies. In the context of wheat yield estimation, RNN might not have effectively captured long-term dependencies in the data, leading to higher prediction errors compared to the top three methods [95].

Table 7.1: Overall comparison of all techniques SVM, RF, LASSO, CNN, ANN, RNN, and CRS Year's wise from 2017-2022

Year	Framework & CRS Results (Tones)						
	SVM	**RF**	**LASSO**	**CNN**	**ANN**	**RNN**	**CRS**
2017	848.172	658.3949	569.1468	799.1192	646.309	478.6387	616
2018	932.012	721.0227	623.5541	883.575	711.8301	532.196	612
2019	727.58	563.7632	487.0536	715.8042	557.7229	426.0078	582
2020	748.368	579.8707	500.1528	733.527	573.4643	439.5536	564
2021	807.12	624.5254	540.9163	780.9112	615.7477	467.7094	569
2022	639.804	496.0842	428.098	604.5954	481.0005	354.5905	642

Table 7.1 shows an overall comparison of all six techniques' total production in tones from 2017 to 2022 years. SVM yields relatively high results across all years, indicating its effectiveness in the given context.RF also performs well, particularly in 2018, with a result of 932.012. LASSO shows consistent performance over the years, with results ranging from 428.098 to 623.5541. CNN's performance is also consistent, with results ranging from 604.5954 to 883.575. ANN shows similar trends to CNN, with results ranging from 557.7229 to 711.8301. RNN generally performs lower compared to other frameworks, with results ranging from 354.5905 to 532.196. After all the results and compassion, we concluded that CNN and RF produced the best results in the study area within our area factors like temperature, precipitation, humidity, soil, etc. After CNN & RF, SVM produced close results but LASSO, RNN, and ANN deviated from the results. Table 7.1 shows overall comparisons of all techniques with CRS data of total production in tones whole district of Multan from 2017 to 2022.

Table 7.2 demonstrate the RMSE (kg/acre), MAE, and R2 of yield predictions made at the district level by the SVM, ANN, LASSO, RF, CNN, and RNN models. The outcomes were averaged runs to take into consideration the random initialization and dropout during DL model training [96–98]. Using a model trained on data from all previous years, each row corresponds to predictions made for that particular year. For winter wheat at the county scale, RMSE and R^2 were largely

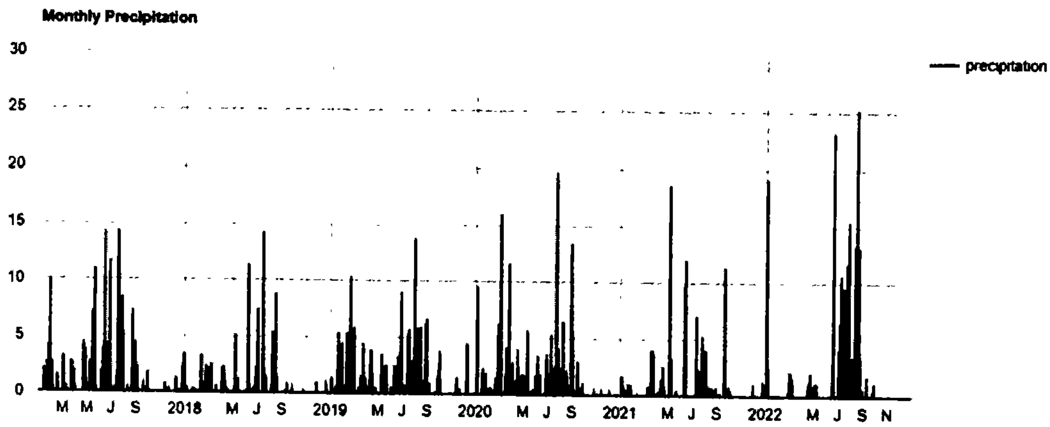

Figure 7.2: Monthly Precipitation pattern from 2017-2022

consistent throughout the analysis years, albeit in 2017 the CNN model had a rather high RMSE of 2.2667, 39.05 per acre (15.8% of the average yield). With an average R^2 of 0.85 and RMSE of

Table 7.2: RMSE, MAE, and R^2 for Train data of district level of all Model Evaluation Metrics

Models	R^2 score	MAE	RMSE
Random Forest	0.9576	0.5800	0.05617
SVM	0.9264	0.8217	0.0117
CNN	0.9967	0.6050	0.02667
ANN	0.6522	0.6520	0.4050
Lasso	0.0432	0.8183	0.6025
RNN	0.1303	0.7767	0.3917

768 kg/ha from 2017 to 2022, all six models had high prediction capacity for winter wheat yield at the district level, particularly the RF and SVM models (average R^2 of 0.582 and RMSE of 658.39 Tone and average R^2 of 0.8450 and RMSE of 848.17 Tone, respectively). For all of the models across the years, there was no performance forecast beyond the outlier. Comparing all six models with Punjab's Crop Report services is shown in Fig. 49. We also carried out a quadratic fit of three models.

Table 7.3 and Figure 7.3 shows in tabular and graphical, the accuracy in percentage in the scenario of our study area and with our factors, the results show CNN returns 98%, and RF is the best classifier that returns 97% accuracy and SVM shows 93% which is also well accuracy. Results show ANN, LASSO, and RNN show the lowest accuracy because of each classifier specification

Table 7.3: Overall Comparison of Accuracy of all Techniques.

Technique	Overall Accuracy
Convolutional Neural Network (CNN)	**98%**
Random Forest	**97%**
Support Vector Machine (SVM)	**93%**
Artificial Neural Network (ANN)	**90%**
Lasso Regression	**85%**
Recurrent Neural Network (RNN)	**80%**

in our study area.

The wheat yield in the Multan region has been increasing over the past few years, but several factors can affect it. These factors include: Wheat is a cool-season crop, so it is sensitive to high temperatures. The optimum temperature for wheat growth is between 15 and 25 degrees Celsius sow precipitation of Multan from 2017 to 2022 [99] in Figure 7.2. If temperatures are too high during the growing season, it can lead to reduced yields. For example, a study by the Pakistan Agricultural Research Council found that wheat yields decreased by 10% for every degree Celsius above 25 degrees Celsius. Wheat requires a moderate amount of rainfall to grow. The average annual rainfall in the Multan region is about 300 mm. If there is too little rain, the crop will not be able to produce a good yield. However, too much rain can also be a problem, as it can lead to flooding and waterlogging, which can damage the crop. For example, a study by the International Rice Research Institute found that wheat yields decreased by 10% for every 100 mm of rainfall above the optimum level of 500 mm. The monsoon is a seasonal wind system that brings rain to Pakistan. The monsoon pattern has been changing in recent years, with some years being wetter than others. A wetter year typically leads to higher yields. For example, a study by the Pakistan Meteorological Department found that wheat yields in the Multan region increased by 20% in the wettest year of the study period (2010) compared to the driest year (2009).

Climate change is expected to significantly impact wheat yields in Pakistan. The country is already experiencing more extreme weather events, such as droughts and floods, which can damage crops and reduce yields. For example, a study by the World Bank found that due to climate change, wheat yields in Pakistan could decrease by up to 20% by 2050.

Fertilizers are essential for increasing wheat yields. However, too much fertilizer can pollute the environment and damage the soil. For example, a study by the Food and Agriculture Organization of the United Nations found that wheat yields in Pakistan could decrease by up to 10% if the current

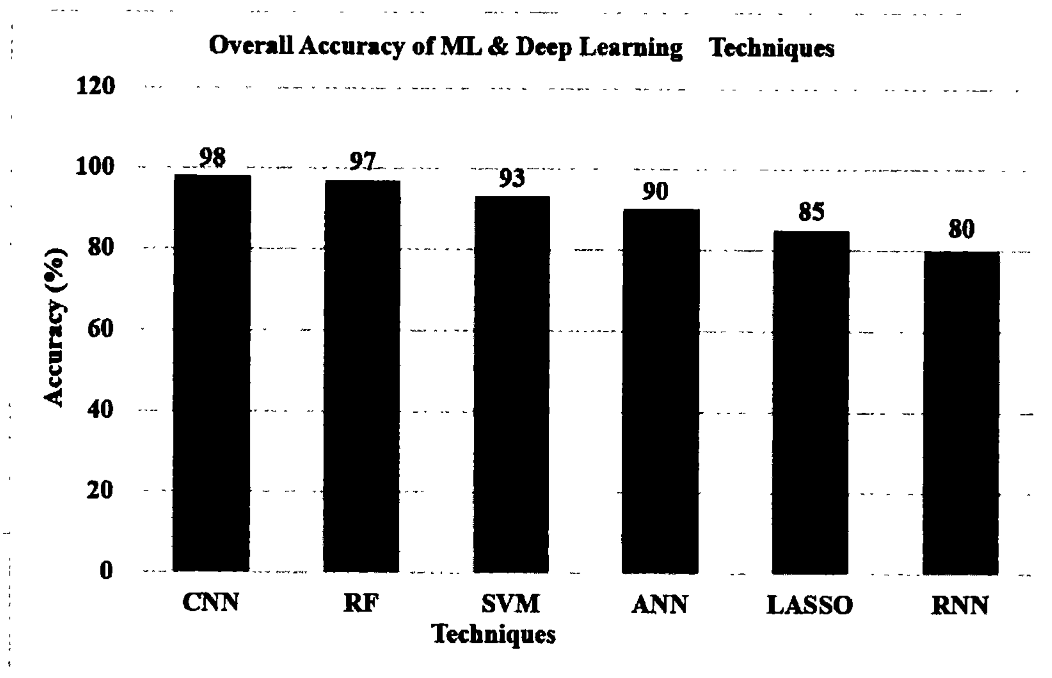

Figure 7.3: Overall comparison of the accuracy of all techniques in graphical

rate of fertilizer use continues. In summary, each machine learning algorithm has its strengths and weaknesses in the context of crop yield prediction. CNN and RF performed exceptionally well, likely due to their ability to capture spatial patterns and complex relationships in the data. SVM and ANN also performed well, demonstrating their effectiveness in capturing nonlinear relationships. Lasso Regression, while performing adequately, may not capture as many complex relationships as other models. RNN, although designed for sequential data, achieved the lowest accuracy in this scenario, indicating potential challenges in capturing temporal dependencies effectively. The practices used by farmers can also have a significant impact on wheat yields. For example, if farmers do not use good irrigation practices, the crop may not receive enough water. If farmers do not use good crop rotation practices, the soil may become depleted of nutrients. For example, a study by the International Maize and Wheat Improvement Center found that wheat yields in Pakistan could increase by up to 15% if farmers adopted better irrigation practices.

Some farmers are reluctant to adopt new agricultural technologies, such as improved seeds and irrigation systems. This can lead to lower yields [94]. For example, a study by the Pakistan Agricultural Research Council found that wheat yields in Pakistan could increase by up to 10% if farmers adopted improved seeds.

We analyze Why 2018 has a maximum yield and 2022 minimum yield, the reason is that in 2022 due to more rainfall in March and increased precipitation, each classifier showed low wheat crop yield in Multan shown in Figure 7.2. Some farmers may not be aware of the latest agricultural practices or technologies. This can also lead to lower yields. For example, a study by the World Bank found that wheat yields in Pakistan could increase by up to 5% if farmers were more aware of the latest agricultural practices.

Chapter 8

Conclusions and Future Directions

In this study, we used six models, including three Deep Learning models (CNN,ANN and RNN) and three Machine Learning ML models (RF, SVM, and LASSO) to predict winter wheat yield from 2017 to 2022. First, framework processed all factors (including climate, satellite, soil properties, and coordinates data) on the GEE platform, Arc Map,USGS,Python (Anaconda) and Remote sensing. Finally, we evaluated generalization and accuracy for all years at the Multan district level. Overall, only the CNN and RF models performed well at the district level out of the six models, the other four did not. The results showes a new scalable, straightforward, and affordable framework for estimating winter wheat yield on a regional scale using openly available data and GEE platforms, Arc Map, and Remote Sensing, which may be used to estimate crop yield globally and in regions with sparsely observed data. For crop yield estimation, predictions, and catastrophe monitoring across vast regions, the framework may be further enhanced by fusing crop models with more specialized EVI data. Additionally, we looked at the possibility of timely in-season yield prediction of wheat and discovered that in Pakistan, good prediction performance may be obtained around one month before harvest. This study develops a strong modeling framework that combines climate and satellite data to forecast agricultural output at large regional scales, and this framework is made to be adaptable to various crops and geographical situations. The current study was limited to one district Multan in the south Punjab region of Pakistan, but it might be expanded to include other provinces of Pakistan as well as other more agriculturally dependent regions like central Punjab and the Pothohar region. Future additions might include the effects of floods, water logging, and salinity on wheat crops. The government may obtain forecasts for the next years from various regions of Pakistan, which will aid in planning and developing policies to satisfy the nation's food needs. Further research can also examine other deep learning and machine learning models to enhance the effectiveness and analysis.

8.1 Thesis Contribution

The Framework provided an important contribution to agricultural research under 20 factors, including meteorology, types of chemical fertilizer, land use categories, types of soil, soil moisture, soil consistency, soil texture, and soil reaction [99]. First processed the raw data, cleaned the data, and normalized the data to reduce the percentage of error to maximize the accuracy of our forecast, which assisted our farmers in increasing their crop yield with the least amount of work. Additionally, we used Machine Learning, deep neural network approach [100], and Remote sensing techniques, which eventually set a new standard for the performance of our entire system. With an error rate of under 10%, we have seen a very positive result. Communities will eventually gain from this level of high precision as they make better agricultural decisions in the future. The farmers, the government, agricultural stakeholders, policymakers, and the society in our country ultimately need to know about this to monitor food security and to specify crop yield and business. It is a hybrid framework that uses deep learning, machine learning techniques, remote sensing, and software engineering to autonomously estimate wheat crop production before one month harvesting on a big scale in Pakistan. Before harvest, the framework estimates wheat crop yield. This framework is not region-specific. It evaluates the efficacy of current Deep Learning and machine learning techniques on our dataset in the context of Pakistan's South Punjab Districts. Instead of manual collection, the framework uses GIS and remote sensing methods to gather agricultural yield datasets. It helps our ranchers increase crop yield prediction with the least amount of work.

Appendix A

A Framework for Crop Yield Prediction of Wheat in Pakistan using Remote Sensing and Deep Learning. A Case of South Punjab

A.1 A questionnaire followed to collect the data from the farmers, List to collect the data from the farmers (Ethnographic Study

A questionnaire followed to collect the data from the farmers	
Farmer details Name and address Contact no	
Location of Plot	
Area of land holding	
Previous crop sown	
Soil type	
Soil nutrient status	
Variety name and duration	
Date of transplanting /sowing	
Irrigation details No. of irrigations Stages of irrigation	
Fertilizer details Rate of application Stage of application with quantity	
Pest and disease attack (if any) Name and quantity of insecticides/ pesticides used	
Date of harvesting	
Yield (Kg ha^{-1})	
Soil health card Details	
Other Comments	

List to collect the data from the farmers (Ethnographic Study)		
Sr.#	**Area**	**# of Persons**
1	Bosan Road Multan	10
2	Sher Shah Multan	5
3	Shujabad	10

Bibliography

[1] S. Khaki, L. Wang, and S. V. Archontoulis, "A cnn-rnn framework for crop yield prediction," *Frontiers in Plant Science*, vol. 10, p. 492736, 2020.

[2] S. Khaki and L. Wang, "Crop yield prediction using deep neural networks," *Frontiers in plant science*, vol. 10, p. 452963, 2019.

[3] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, "Software engineering for machine learning: A case study," in *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 2019, pp. 291–300.

[4] M. Qiao, X. He, X. Cheng, P. Li, H. Luo, L. Zhang, and Z. Tian, "Crop yield prediction from multi-spectral, multi-temporal remotely sensed imagery using recurrent 3d convolutional neural networks," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102436, 2021.

[5] M. Giannopoulos, A. Aidini, A. Pentari, K. Fotiadou, and P. Tsakalides, "Classification of compressed remote sensing multispectral images via convolutional neural networks," *Journal of Imaging*, vol. 6, no. 4, p. 24, 2020.

[6] K. Lashari, "Land use atlas of pakistan," *Ministry of Environment Government of Pakistan. https://wedocs. unep. org/bitstream/handle/20.500*, vol. 11822, p. 9393, 1974.

[7] P. R. Shewry and S. J. Hey, "The contribution of wheat to human diet and health," *Food and energy security*, vol. 4, no. 3, pp. 178–202, 2015.

[8] S. Hao, D. Ryu, A. Western, E. Perry, H. Bogena, and H. J. H. Franssen, "Performance

of a wheat yield prediction model and factors influencing the performance: A review and meta-analysis," *Agricultural Systems*, vol. 194, p. 103278, 2021.

[9] A. Heidari, N. Jafari Navimipour, M. Unal, and G. Zhang, "Machine learning applications in internet-of-drones: Systematic review, recent deployments, and open issues," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–45, 2023.

[10] B. Panigrahi, K. C. R. Kathala, and M. Sujatha, "A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models," *Procedia Computer Science*, vol. 218, pp. 2684–2693, 2023.

[11] W. W. Guo, H. Xue *et al.*, "Crop yield forecasting using artificial neural networks: A comparison between spatial and temporal models," *Mathematical Problems in Engineering*, vol. 2014, 2014.

[12] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and electronics in agriculture*, vol. 147, pp. 70–90, 2018.

[13] H. Russello, "Convolutional neural networks for crop yield prediction using satellite images," *IBM Center for Advanced Studies*, 2018.

[14] P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," *Computers and electronics in agriculture*, vol. 163, p. 104859, 2019.

[15] H. B. Dias and P. C. Sentelhas, "Sugarcane yield gap analysis in brazil–a multi-model approach for determining magnitudes and causes," *Science of the total environment*, vol. 637, pp. 1127–1136, 2018.

[16] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (cnn) in vegetation remote sensing," *ISPRS journal of photogrammetry and remote sensing*, vol. 173, pp. 24–49, 2021.

[17] R. Tanabe, T. Matsui, and T. S. Tanaka, "Winter wheat yield prediction using convolutional neural networks and uav-based multispectral imagery," *Field Crops Research*, vol. 291, p. 108786, 2023.

[18] J. Sun, L. Di, Z. Sun, Y. Shen, and Z. Lai, "County-level soybean yield prediction using deep cnn-lstm model," *Sensors*, vol. 19, no. 20, p. 4363, 2019.

[19] H. Tian, P. Wang, K. Tansey, J. Zhang, S. Zhang, and H. Li, "An lstm neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the guanzhong plain, pr china," *Agricultural and Forest Meteorology*, vol. 310, p. 108629, 2021.

[20] O. Sadak, F. Sadak, O. Yildirim, N. M. Iverson, R. Qureshi, M. Talo, C. P. Ooi, U. R. Acharya, S. Gunasekaran, and T. Alam, "Electrochemical biosensing and deep learning-based approaches in the diagnosis of covid-19: A review," *Ieee Access*, vol. 10, pp. 98 633–98 648, 2022.

[21] I. Ali, F. Greifeneder, J. Stamenkovic, M. Neumann, and C. Notarnicola, "Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data," *Remote Sensing*, vol. 7, no. 12, pp. 16 398–16 421, 2015.

[22] M. Tejas, L. Santhosh, V. Potdar, V. Hegde, T. Sumukha, and T. V. KT, "Techniques for the identification of early blight and late blight in potato leaves: A comparative review."

[23] Y. Chen, Z. Zhang, and F. Tao, "Improving regional winter wheat yield estimation through assimilation of phenology and leaf area index from remote sensing data," *European journal of agronomy*, vol. 101, pp. 163–173, 2018.

[24] A. X. Wang, C. Tran, N. Desai, D. Lobell, and S. Ermon, "Deep transfer learning for crop yield prediction with remote sensing data," in *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, 2018, pp. 1–5.

[25] Y. Zhu, S. Wu, M. Qin, Z. Fu, Y. Gao, Y. Wang, and Z. Du, "A deep learning crop model for adaptive yield estimation in large areas," *International Journal of Applied Earth Observation and Geoinformation*, vol. 110, p. 102828, 2022.

[26] T. Wang, W. Zhou, J. Xiao, H. Li, L. Yao, L. Xie, and K. Wang, "Soil organic carbon prediction using sentinel-2 data and environmental variables in a karst trough valley area of southwest china," *Remote Sensing*, vol. 15, no. 8, p. 2118, 2023.

[27] Y. Di, M. Gao, F. Feng, Q. Li, and H. Zhang, "A new framework for winter wheat yield prediction integrating deep learning and bayesian optimization," *Agronomy*, vol. 12, no. 12, p. 3194, 2022.

[28] K. Gavahi, P. Abbaszadeh, and H. Moradkhani, "Deepyield: A combined convolutional

neural network with long short-term memory for crop yield forecasting," *Expert Systems with Applications*, vol. 184, p. 115511, 2021.

[29] U. Hayat, S. Ali, A. Mateen, and H. Bilal, "The role of agriculture in poverty alleviation: Empirical evidence from pakistan," *Sarhad Journal of Agriculture*, vol. 35, no. 4, pp. 1309–1315, 2019.

[30] J. Wang, P. Wang, H. Tian, K. Tansey, J. Liu, and W. Quan, "A deep learning framework combining cnn and gru for improving wheat yield estimates using time series remotely sensed multi-variables," *Computers and Electronics in Agriculture*, vol. 206, p. 107705, 2023.

[31] R. Li, D. Wang, B. Zhu, T. Liu, C. Sun, and Z. Zhang, "Estimation of nitrogen content in wheat using indices derived from rgb and thermal infrared imaging," *Field Crops Research*, vol. 289, p. 108735, 2022.

[32] S. S. Olofintuyi, E. A. Olajubu, and D. Olanike, "An ensemble deep learning approach for predicting cocoa yield," *Heliyon*, vol. 9, no. 4, 2023.

[33] M. U. Ahmed and I. Hussain, "Prediction of wheat production using machine learning algorithms in northern areas of pakistan," *Telecommunications policy*, vol. 46, no. 6, p. 102370, 2022.

[34] Y. Pan, Z. Pan, Y. Wang, and W. Wang, "A new fast search algorithm for exact k-nearest neighbors based on optimal triangle-inequality-based check strategy," *Knowledge-Based Systems*, vol. 189, p. 105088, 2020.

[35] J. W. Rouse, R. H. Haas, J. A. Schell, D. W. Deering *et al.*, "Monitoring vegetation systems in the great plains with erts," *NASA Spec. Publ*, vol. 351, no. 1, p. 309, 1974.

[36] S. Janarthanan, T. Ganesh Kumar, S. Janakiraman, R. K. Dhanaraj, M. A. Shah *et al.*, "An efficient multispectral image classification and optimization using remote sensing data," *Journal of Sensors*, vol. 2022, 2022.

[37] J. M. Deines, R. Patel, S.-Z. Liang, W. Dado, and D. B. Lobell, "A million kernels of truth: Insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the us corn belt," *Remote sensing of environment*, vol. 253, p. 112174, 2021.

[38] D. B. Lobell, "The use of satellite data for crop yield gap analysis," *Field crops research*, vol. 143, pp. 56–64, 2013.

[39] K. M. Laleh, M. Ghorbani Javid, I. Alahdadi, E. Soltani, S. Soufizadeh, and J. L. González-Andújar, "Wheat yield gap assessment in using the comparative performance analysis (cpa)," *Agronomy*, vol. 13, no. 3, p. 705, 2023.

[40] S. I. Hassan, M. M. Alam, M. Y. I. Zia, M. Rashid, U. Illahi, and M. M. Su'ud, "Rice crop counting using aerial imagery and gis for the assessment of soil health to increase crop yield," *Sensors*, vol. 22, no. 21, p. 8567, 2022.

[41] C. Uma, Y. Bekele, and T. Hirko, "Determinants of technical efficiency of wheat production in ethiopia: a review," *Journal of Economics and Sustainable Development*, vol. 8, no. 19, pp. 11–15, 2017.

[42] N. K. Newlands, D. S. Zamar, L. A. Kouadio, Y. Zhang, A. Chipanshi, A. Potgieter, S. Toure, and H. S. Hill, "An integrated, probabilistic model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty," *Frontiers in Environmental Science*, vol. 2, p. 17, 2014.

[43] M. Alagurajan and C. Vijayakumaran, "Ml methods for crop yield prediction and estimation: an exploration," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 3, 2020.

[44] A. Hassanzadeh, *On the Use of Imaging Spectroscopy from Unmanned Aerial Systems (UAS) to Model Yield and Assess Growth Stages of a Broadacre Crop*. Rochester Institute of Technology, 2022.

[45] M. F. Celik, M. S. Isik, O. Yuzugullu, N. Fajraoui, and E. Erten, "Soil moisture prediction from remote sensing images coupled with climate, soil texture and topography via deep learning," *Remote sensing*, vol. 14, no. 21, p. 5584, 2022.

[46] S. Yang, L. Hu, H. Wu, H. Ren, H. Qiao, P. Li, and W. Fan, "Integration of crop growth model and random forest for winter wheat yield estimation from uav hyperspectral imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 6253–6269, 2021.

[47] D. Wang, W. Cao, F. Zhang, Z. Li, S. Xu, and X. Wu, "A review of deep learning in multi-scale agricultural sensing," *Remote Sensing*, vol. 14, no. 3, p. 559, 2022.

[48] C. F. Sánchez Valdés, M. Arreguin Hernandez, A. Dzubinska, M. Reiffers, J. Sanchez, and R. Varga, "Magnetostructural transition and magnetocaloric effect in thermally annealed mn0. 5fe0. 5nisi1-xalx melt-spun ribbons (x= 0.055 and 0.060)," *Instituto de Ingeniería y Tecnología*, 2022.

[49] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (cnn) in vegetation remote sensing," *ISPRS journal of photogrammetry and remote sensing*, vol. 173, pp. 24–49, 2021.

[50] Y. Li, H. Zeng, M. Zhang, B. Wu, Y. Zhao, X. Yao, T. Cheng, X. Qin, and F. Wu, "A county-level soybean yield prediction framework coupled with xgboost and multidimensional feature engineering," *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, p. 103269, 2023.

[51] C. Müller, J. Elliott, J. Chryssanthacopoulos, A. Arneth, J. Balkovic, P. Ciais, D. Deryng, C. Folberth, M. Glotter, S. Hoek *et al.*, "Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications," *Geoscientific Model Development*, vol. 10, no. 4, pp. 1403–1422, 2017.

[52] L. Gong, M. Yu, S. Jiang, V. Cutsuridis, and S. Pearson, "Deep learning based prediction on greenhouse crop yield combined tcn and rnn," *Sensors*, vol. 21, no. 13, p. 4537, 2021.

[53] A. Hassanzadeh, *On the Use of Imaging Spectroscopy from Unmanned Aerial Systems (UAS) to Model Yield and Assess Growth Stages of a Broadacre Crop.* Rochester Institute of Technology, 2022.

[54] D. Gómez, P. Salvador, J. Sanz, and J. L. Casanova, "Potato yield prediction using machine learning techniques and sentinel 2 data," *Remote Sensing*, vol. 11, no. 15, p. 1745, 2019.

[55] L. J. Young, "Agricultural crop forecasting for large geographical areas," *Annual review of statistics and its application*, vol. 6, pp. 173–196, 2019.

[56] H. Sun, X. Zhang, S. Chen, D. Pei, and C. Liu, "Effects of harvest and sowing time on the performance of the rotation of winter wheat–summer maize in the north china plain," *Industrial Crops and Products*, vol. 25, no. 3, pp. 239–247, 2007.

[57] K. Zeleke and C. Nendel, "Analysis of options for increasing wheat (triticum aestivum l.) yield in south-eastern australia: The role of irrigation, cultivar choice and time of sowing," *Agricultural Water Management*, vol. 166, pp. 139–148, 2016.

[58] M. Cai, C. An, C. Guy, C. Lu, and F. Mafakheri, "Assessing the regional biogenic methanol emission from spring wheat during the growing season: A canadian case study," *Environmental Pollution*, vol. 287, p. 117602, 2021.

[59] A. Joshi, B. Mishra, R. Chatrath, G. Ortiz Ferrara, and R. P. Singh, "Wheat improvement in india: present status, emerging challenges and future prospects," *Euphytica*, vol. 157, pp. 431–446, 2007.

[60] A. Mucherino, P. Papajorgji, and P. M. Pardalos, "A survey of data mining techniques applied to agriculture," *Operational Research*, vol. 9, pp. 121–140, 2009.

[61] J. Majumdar, S. Naraseeyappa, and S. Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data," *Journal of Big data*, vol. 4, no. 1, p. 20, 2017.

[62] T. Iizumi and N. Ramankutty, "How do weather and climate influence cropping area and intensity?" *Global food security*, vol. 4, pp. 46–50, 2015.

[63] T. N. Liliane and M. S. Charles, "Factors affecting yield of crops," *Agronomy-climate change & food security*, p. 9, 2020.

[64] J. W. Hansen, S. J. Mason, L. Sun, and A. Tall, "Review of seasonal climate forecasting for agriculture in sub-saharan africa," *Experimental agriculture*, vol. 47, no. 2, pp. 205–240, 2011.

[65] S.-y. ZHANG, X.-h. ZHANG, X.-l. QIU, T. Liang, Z. Yan, W.-x. CAO, and L.-l. LIU, "Quantifying the spatial variation in the potential productivity and yield gap of winter wheat in china," *Journal of integrative agriculture*, vol. 16, no. 4, pp. 845–857, 2017.

[66] B. Wang, X. Gu, L. Ma, and S. Yan, "Temperature error correction based on bp neural network in meteorological wireless sensor network," *International Journal of Sensor Networks*, vol. 23, no. 4, pp. 265–278, 2017.

[67] N. Alexandratos and J. Bruinsma, "World agriculture towards 2030/2050: the 2012 revision," 2012.

[68] M. Ashraf, M. S. A. Ahmad, M. Öztürk, and A. Aksoy, "Crop improvement through different means: Challenges and prospects," *Crop production for agricultural improvement*, pp. 1–15, 2012.

[69] G. P. Miriyala and A. K. Sinha, "Prediction of crop yield using deep learning techniques: a concise review," *Recent Advances in Computer Based Systems, Processes and Applications*, pp. 145–159, 2020.

[70] A. Shelestov, M. Lavreniuk, N. Kussul, A. Novikov, and S. Skakun, "Exploring google earth engine platform for big data processing: Classification of multi-temporal satellite imagery for crop mapping," *frontiers in Earth Science*, vol. 5, p. 232994, 2017.

[71] H. Russello, "Convolutional neural networks for crop yield prediction using satellite images," 2018.

[72] I. Ahmad, U. Saeed, M. Fahad, A. Ullah, M. Habib ur Rahman, A. Ahmad, and J. Judge, "Yield forecasting of spring maize using remote sensing and crop modeling in faisalabad-punjab pakistan," *Journal of the Indian Society of Remote Sensing*, vol. 46, pp. 1701–1711, 2018.

[73] T. Islam, T. A. Chisty, and A. Chakrabarty, "A deep neural network approach for crop selection and yield prediction in bangladesh," in *2018 IEEE region 10 humanitarian technology conference (R10-HTC)*. IEEE, 2018, pp. 1–6.

[74] A. Nigam, S. Garg, A. Agrawal, and P. Agrawal, "Crop yield prediction using machine learning algorithms," in *2019 Fifth International Conference on Image Information Processing (ICIIP)*. IEEE, 2019, pp. 125–130.

[75] Y. Cai, K. Guan, D. Lobell, A. B. Potgieter, S. Wang, J. Peng, T. Xu, S. Asseng, Y. Zhang, L. You *et al.*, "Integrating satellite and climate data to predict wheat yield in australia using machine learning approaches," *Agricultural and forest meteorology*, vol. 274, pp. 144–159, 2019.

[76] V. Joshua, S. M. Priyadharson, and R. Kannadasan, "Exploration of machine learning approaches for paddy yield prediction in eastern part of tamilnadu," *Agronomy*, vol. 11, no. 10, p. 2068, 2021.

[77] N. Kim and Y.-W. Lee, "Machine learning approaches to corn yield estimation using satellite

images and climate data: A case of iowa state: A case of iowa state," *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, vol. 34, no. 4, pp. 383–390, 2016.

[78] E. Bélisle, Z. Huang, S. Le Digabel, and A. E. Gheribi, "Evaluation of machine learning interpolation techniques for prediction of physical properties," *Computational Materials Science*, vol. 98, pp. 170–177, 2015.

[79] K. Kuwata and R. Shibasaki, "Estimating crop yields with deep learning and remotely sensed data," in *2015 IEEE international geoscience and remote sensing symposium (IGARSS)*. IEEE, 2015, pp. 858–861.

[80] I. Ali, F. Greifeneder, J. Stamenkovic, M. Neumann, and C. Notarnicola, "Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data," *Remote Sensing*, vol. 7, no. 12, pp. 16 398–16 421, 2015.

[81] H. Tian, P. Wang, K. Tansey, J. Zhang, S. Zhang, and H. Li, "An lstm neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the guanzhong plain, pr china," *Agricultural and Forest Meteorology*, vol. 310, p. 108629, 2021.

[82] A. González Sánchez, J. Frausto Solís, W. Ojeda Bustamante *et al.*, "Predictive ability of machine learning methods for massive crop yield prediction," 2014.

[83] D. B. Lobell, "The use of satellite data for crop yield gap analysis," *Field crops research*, vol. 143, pp. 56–64, 2013.

[84] F. Affholder, C. Poeydebat, M. Corbeels, E. Scopel, and P. Tittonell, "The yield gap of major food crops in family agriculture in the tropics: Assessment and analysis through field surveys and modelling," *Field Crops Research*, vol. 143, pp. 106–118, 2013.

[85] Z. Liu, X. Yang, K. G. Hubbard, and X. Lin, "Maize potential yields and yield gaps in the changing climate of northeast china," *Global change biology*, vol. 18, no. 11, pp. 3441–3454, 2012.

[86] J. R. Romero, P. F. Roncallo, P. C. Akkiraju, I. Ponzoni, V. C. Echenique, and J. A. Carballido, "Using classification algorithms for predicting durum wheat yield in the province of buenos aires," *Computers and electronics in agriculture*, vol. 96, pp. 173–179, 2013.

[87] X. Zhang, Z. Nan, Y. Sheng, L. Zhao, G. Zhou, G. Yue, and J. Wu, "Analysis of time-series modis 250m vegetation index data for vegetation classifiation in the wenquan area over the qinghai-tibet plateau," in *2010 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2010, pp. 2059–2062.

[88] A. Li, S. Liang, A. Wang, and J. Qin, "Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques," *Photogrammetric Engineering & Remote Sensing*, vol. 73, no. 10, pp. 1149–1157, 2007.

[89] L. Benos, A. C. Tagarakis, G. Dolias, R. Berruto, D. Kateris, and D. Bochtis, "Machine learning in agriculture: A comprehensive updated review," *Sensors*, vol. 21, no. 11, p. 3758, 2021.

[90] P. Q. Khang, K. Kaczmarczyk, P. Tutak, P. Golec, K. Kuziak, R. Depczyński, M. Hernes, and A. Rot, "Machine learning for liquidity prediction on vietnamese stock market," *Procedia Computer Science*, vol. 192, pp. 3590–3597, 2021.

[91] J. F. Progga, M. N. H. Khan, and M. M. AminN, "Meteorological parameters–soil temperature relations in a sub-tropical summer grassland: Physically-based and data-driven modeling," *Atatürk Üniversitesi Ziraat Fakültesi Dergisi*, vol. 54, no. 2, pp. 48–56, 2023.

[92] S. Hussain, M. Mubeen, A. Ahmad, N. Masood, H. Hammad, M. Amjad, and M. Waleed, "Satellite-based evaluation of temporal change in cultivated land in southern punjab (multan region) through dynamics of vegetation and land surface temperature. open geosci 13 (1): 1561–1577," 2021.

[93] P. Lemenkova and O. Debeir, "Multispectral satellite image analysis for computing vegetation indices by r in the khartoum region of sudan, northeast africa," *Journal of imaging*, vol. 9, no. 5, p. 98, 2023.

[94] A. K. Srivastava, N. Safaei, S. Khaki, G. Lopez, W. Zeng, F. Ewert, T. Gaiser, and J. Rahimi, "Winter wheat yield prediction using convolutional neural networks from environmental and phenological data," *Scientific reports*, vol. 12, no. 1, p. 3215, 2022.

[95] S. K. Sahu, A. Mokhade, and N. D. Bokde, "An overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance: Recent progress and challenges," *Applied Sciences*, vol. 13, no. 3, p. 1956, 2023.

[96] F. Rundo, F. Trenta, A. L. di Stallo, and S. Battiato, "Machine learning for quantitative finance applications: A survey," *Applied Sciences*, vol. 9, no. 24, p. 5574, 2019.

[97] P. Liashchynskyi and P. Liashchynskyi, "Grid search, random search, genetic algorithm: a big comparison for nas," *arXiv preprint arXiv:1912.06059*, 2019.

[98] S. T. Arab, R. Noguchi, S. Matsushita, and T. Ahamed, "Prediction of grape yields from time-series vegetation indices using satellite remote sensing and a machine-learning approach," *Remote Sensing Applications: Society and Environment*, vol. 22, p. 100485, 2021.

[99] N.-T. Son, C.-F. Chen, Y.-S. Cheng, P. Toscano, C.-R. Chen, S.-L. Chen, K.-H. Tseng, C.-H. Syu, H.-Y. Guo, and Y.-T. Zhang, "Field-scale rice yield prediction from sentinel-2 monthly image composites using machine learning algorithms," *Ecological informatics*, vol. 69, p. 101618, 2022.

[100] S. Arshad, J. H. Kazmi, M. G. Javed, and S. Mohammed, "Applicability of machine learning techniques in predicting wheat yield based on remote sensing and climate data in pakistan, south asia," *European Journal of Agronomy*, vol. 147, p. 126837, 2023.