IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 February 2024, accepted 10 March 2024, date of publication 18 March 2024, date of current version 22 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3376735

RESEARCH ARTICLE

Accurate Wheat Yield Prediction Using Machine Learning and Climate-NDVI Data Fusion

MUHAMMAD ASHFAQ[1], IMRAN KHAN[2], ABDULRAHMAN ALZAHRANI[3], MUHAMMAD USMAN TARIQ[4], (Member, IEEE), HUMERA KHAN[5], AND ANWAR GHANI[2,6]

[1]Department of Software Engineering, International Islamic University Islamabad, Islamabad 44000, Pakistan
[2]Department of Computer Science, International Islamic University Islamabad, Islamabad 44000, Pakistan
[3]Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah 21493, Saudi Arabia
[4]Marketing, Operations, and Information System, Abu Dhabi University, Abu Dhabi, United Arab Emirates
[5]Department of Information Systems, Faculty of Computing and Information Technology, Northern Border University, Rafha 76312, Saudi Arabia
[6]Big Data Research Center, Jeju National University, Jeju-si, Jeju-do 63243, South Korea

Corresponding author: Anwar Ghani (anwar.ghani@iiu.edu.pk)

ABSTRACT Due to exponential population growth, climate change, and an increasing demand for food, there is an unprecedented need for a timely, precise, and dependable assessment of crop yield on a large scale. Wheat, a staple crop worldwide, requires accurate and prompt prediction of its output for global food security. Traditionally, the development of empirical models for crop yield forecasting has relied on climate data, satellite data, or a combination of both. Despite the enhanced performance achieved by integrating satellite and climate data, the contributions from various sources (Climate, Soil, Socioeconomic, and Remote sensing) remain unclear. The lack of well-defined comparisons between the performance of regression-based approaches and different Machine Learning (ML) methods in yield prediction necessitates further investigation. This study addresses the gaps by combining data from multiple sources to forecast wheat yield in the Multan region in the Punjab province of Pakistan. The findings are compared to the benchmark provided by Crop Report Services (CRS) Punjab, with three widely used ML techniques (support vector machine (SVM), Random Forest (RF), and Least Absolute Shrinkage and Selection Operator (LASSO)) by integrating publicly available data within the GEE (Google Earth Engine) platform, including climate, satellite, soil properties, and spatial information data to develop alternative empirical models for yield prediction using data from 2017 to 2022, selecting the best attribute subset related to crop output. The district-level simulated yield data set was analyzed with three ML models (SVM, RF, and LASSO) as a function of seasonal weather, satellite, and soil. The results indicate that combining all datasets using three ML algorithms achieves better yield prediction performance (R^2: 0.74-0.88). Incorporating spatial information and other properties into benchmark models can improve the prediction from 0.08 to 0.12. Random forest outperformed the competitor models with a Root Mean Square Error (RMSE) of 0.05 q/ha and R^2 of 0.88. Comparative analysis shows that random forest with 97% and SVM with 93% yielded better results in the study area.

INDEX TERMS Machine learning, RF, LASSO, remote sensing, SVM, CNN, crop yield prediction.

I. INTRODUCTION

Wheat, one of the three principal crops farmed globally, is a substantial source of calories, protein, and vital micronutrients for people [1], [2]. Wheat continued high-yield potential, on the other hand, faces significant threats due to a variety of production restrictions, including rising temperatures, increased precipitation unpredictability, and frequent extreme weather events [3], [4]. Accurate crop output forecasts before harvest, thus ensuring food security and trade. Traditional agricultural yield evaluation methods include conducting field surveys during the growing season or relying on prior

The associate editor coordinating the review of this manuscript and approving it for publication was Mouloud Denai.

knowledge of the crop-growing environment. Conventional yield estimates, however, are difficult due to problems such as small sample sizes, insufficient staff for required sampling frequency and size, and the impact of inter-annual climate variability. Data processing concerns such as sampling and non-sampling mistakes can also cause dependability issues and inaccuracies [5].

ML techniques are increasingly used in various fields like health [6] and education [7]. To address these challenges, researchers increasingly turn to nonlinear models for agricultural yield estimation [1], [8]. The application of ML for estimating agricultural yields has gained significant attention in recent years [9], [10]. Various ML techniques have been extensively utilized to achieve precise predictions for the yields of diverse crops. The efficacy of deep learning network frameworks has been underscored by deploying several ML models for estimating agricultural yields. Common examples of these models include SVM, RF, and LASSO [10].

Crop output is subject to the influence of climate, management strategies, and genotypic factors. Large-scale meteorological events can significantly impact crop growth and yields by altering regional climatic patterns, as highlighted in studies such as [11] and [12]. Given the pivotal role of environmental conditions in crop development across various stages, utilizing a diverse set of environmental parameters for predicting crop yield becomes crucial, as emphasized by [13]. Furthermore, satellite imagery can detect biotic variables, including diseases and insects, which can impact crop development and manifest in leaf traits, as discussed by [14]. Incorporating remote sensing measures that enable real-time crop growth status monitoring can enhance yield forecasts' accuracy.

Crop production predictions at regional, national, and global scales have commonly employed meteorological data, remote sensing data, or a combination of both, as evidenced by studies such as [15], [16], and [17]. Prediction models typically integrate primary meteorological and satellite-based input variables, including temperature, precipitation, solar radiation, and the Normalized Difference Vegetation Index (NDVI), as outlined in works [18], [19]. Despite its importance as a climatic factor affecting plant growth through its influence on foliar gas and heat exchange, modification of foliar boundary layers, and alteration of water status, wind speed has received comparatively less attention in these prediction models [20], [21].

Process-based models demand extensive data inputs and calibration criteria, as emphasized by [22]. As an illustration, the development of satellite-based light use efficiency models, assuming that gross primary production is solely determined by the amount of photosynthetically active radiation, aimed to estimate vegetation gross primary production [23]. In a comparative study assessing statistical performances, [16] discovered that Random Forest outperformed multiple linear regression in all evaluated metrics, showcasing its potential for predicting agricultural productivity [16]. Although ML has demonstrated success in numerous large-scale agricultural yield prediction studies in China, as highlighted by [24], its suitability for national-scale wheat yield prediction remains unexplored [25].

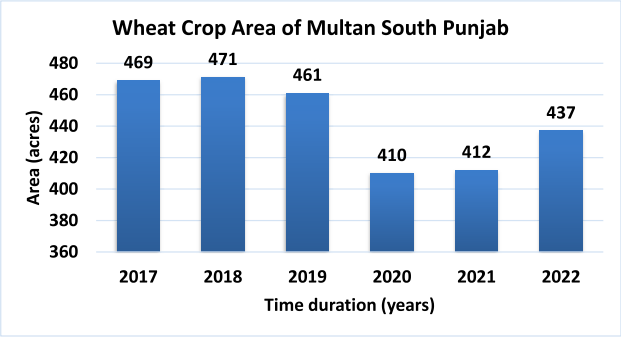

FIGURE 1. Multan wheat crop area from 2017 to 2022.

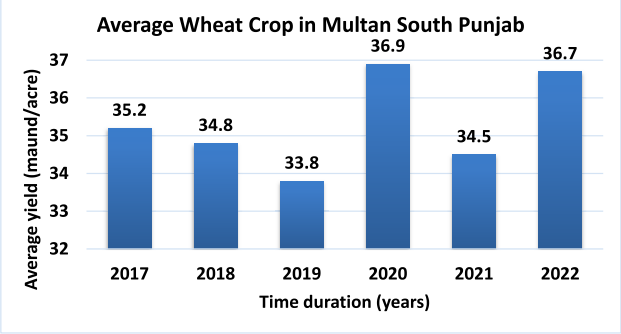

FIGURE 2. Multan wheat crop area from 2017 to 2022.

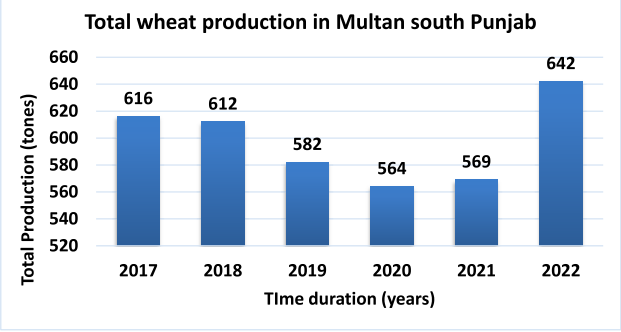

FIGURE 3. Wheat crop production in Ton from 2017 to 2022.

This study compared multitemporal Moderate Resolution Imaging Spectroradiometer (MODIS) enhanced vegetation index (EVI) and NDVI data to estimate rice crop yields in the Mekong River Delta (MRD) of Vietnam [26]. In this study, the author tried decision trees and Random Forests for crop yield prediction and a Decision Support System for Agro-Technology Transfer (DSSAT) simulation for crop yield prediction [27]. The primary goal of this study was to

assess the applicability of monthly composites from Sentinel-2 images for rice yield prediction at the field size in Taiwan using ML approaches [28]. This work used Landsat 8 surface reflectance products from 2017, 2018, and 2019 to map the satellite-based NDVI, leaf area index (LAI), and normalized difference water index (NDWI) [29].

This study aims to create an ML approach for predicting rice crop yields in Taiwan using time-series Moderate Resolution Imaging Spectroradiometer (MODIS) data [30].

The main aims are (a) to evaluate the prediction accuracy of the four ML algorithms, ANN, SVR, KNN, and RF, and (b) to evaluate the impact of different feature sets on ML algorithms. The following feature selection algorithms are used to identify unique feature sets: Forward Feature Selection (FFS), Correlation-based Feature Selection (CBFS), Variance Inflation Factor (VIF), and Random Forest Variable Importance (RFVarImp) [31].

To create a county-level corn yield prediction model based on Bayesian Neural Network (BNN) employing numerous publically available data sources, such as time-series satellite products, sequential climatic measurements, soil property maps, and historical maize yield records [32].

Remote sensing (RS) systems are increasingly used to develop decision support tools for modern farming systems, aiming to improve yield output and nitrogen control while lowering operating costs and environmental effects. However, RS-based systems require massive amounts of remotely sensed data from many platforms. Therefore, more attention is increasingly being paid to ML methods. This is owing to the ML system's ability to process many inputs and handle nonlinear tasks [33]. As a result, the scholarly literature was screened to identify a wide range of essential features for capturing current progress and trends, such as (a) the research areas most interested in ML techniques (RF, SVM, LASSO) in agriculture, as well as the geographical distribution of the contributing organizations, (b) the most efficient ML models, (c) the most investigated crops and animals, and (d) the most implemented features and technologies [34].

Deep learning techniques are typically unsuitable for general-purpose applications since they require vast data. Tree ensembles typically outperform them for traditional ML issues. Furthermore, they are computationally intensive to train and need significantly more expertise to tune (i.e., setting the architecture and hyper-parameters) [35].

Conventional or outdated approaches to yield estimation are both labor-intensive and time-consuming. Additionally, the collection of yield data from a limited number of villages fails to adequately represent the broader agricultural landscape [3]. Although past research has effectively utilized various models for predicting wheat and other crop yields, there is a clear emphasis on carefully selecting base learners and predictors to enhance model performance. For instance, a study employed a step-wise multiple regression method to develop models for predicting departmental-level wheat yield in France [36]. Predicting wheat crop yield in Pakistan poses challenges due to its dependence on a multitude of internal factors (such as seed, disease, and plant health), external factors (including weather, soil, irrigation, and socio-economics), and the reliance on manual methodologies, leading to imprecise estimates before harvesting.

This study aimed to investigate meteorological constraints on winter wheat output and develop a yield forecast model based on meteorological data [37]. This study proposes an approach using three ML models for predicting crop yield. This technique utilizes remote sensing images and various factors as input, and the input data undergoes validation in a software engineering process [8]. The methodology presented in this study holds significant value for policy-makers in improving the identification of import strategies and making decisions to address large-scale food security challenges in Pakistan [38].

The general contributions of the article are as follows.

- Exploring the viability of ML-based yield forecasting using historical yield data.
- Showcasing the significance of yield detrending and the influence of climate, soil, and socioeconomic factors on yield estimation
- Minimize Pakistan's local conventional manual wheat crop yield estimation process.
- This study would make possible accurate crop yield prediction of wheat before the harvesting season.

The remainder of the paper is structured as follows: Section II covers the Materials and Methods employed in the study, while Section III delves into the proposed approach. Section II-C presents results and analysis of the proposed approach, while section V concludes the article.

II. MATERIALS AND METHODS

This section describes the meteorological characteristics, Datasets, and data sources.

A. STUDY REGION

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. A physically based HYDRUS-1D and linear regression models were used to analyze the correlation between meteorological parameters and temperature at different depths in silt loam soil [39]. Here are some weather statistics for the wheat-growing season. High-density raster photographs of the target region, located at latitude 29.848212 N and longitude 71.263367 at 423 feet (129 m) above sea level, were among the data 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 of 437 acres according to Crop Report Services Punjab shown in Fig. 4. The terrain is flat and alluvial, making it suitable for agriculture. The irrigation network of canals makes the