Journal of King Saud University - Computer and Information Sciences 36 (2024) 101895

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University - Computer and Information Sciences

journal homepage: www.sciencedirect.com

Meta-knowledge guided Bayesian optimization framework for robust crop yield estimation

Muhammad Hanif Tunio [a], Jian Ping Li [a,*], Xiaoyang Zeng [a], Faijan Akhtar [a], Syed Attique Shah [b], Awais Ahmed [a], Yu Yang [a], Md Belal Bin Heyat [c]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China - UESTC, Sichuan, 611731, China
[b] School of Computing and Digital Technology, Birmingham City University, STEAMhouse, B4 7RQ, Birmingham, United Kingdom
[c] CenBRAIN Neurotech Center of Excellence, School of Engineering, Westlake University, Hangzhou, Zhejiang 310030, China

ARTICLE INFO

Keywords:
Crop yield estimation
Agricultural datasets
Meta-knowledge
Hyper-parameter optimization
Knowledge transfer

ABSTRACT

Accurate pre-harvest crop yield estimation is vital for agricultural sustainability and economic stability. The existing yield estimating models exhibit deficiencies in insufficient examination of hyperparameters, lack of robustness, restricted transferability of meta-models, and uncertain generalizability when applied to agricultural data. This study presents a novel meta-knowledge-guided framework that leverages three diverse agricultural datasets and explores meta-knowledge transfer in frequent hyperparameter optimization scenarios. The framework's approach involves base tasks using LightGBM and Bayesian Optimization, which automates hyperparameter optimization by eliminating the need for manual adjustments. Conducted rigorous experiments to analyze the meta-knowledge transformation of RGPE, SGPR, and TransBO algorithms, achieving impressive R^2 values (0.8415, 0.9865, 0.9708) using rgpe_prf meta-knowledge transfer on diverse datasets. Furthermore, the framework yielded excellent results for mean squared error (MSE), mean absolute error (MAE), scaled MSE, and scaled MAE. These results emphasize the method's significance, offering valuable insights for crop yield estimation, benefiting farmers and the agricultural sector.

1. Introduction

Accurate and timely crop yield estimation plays a vital role in agriculture, delivering substantial benefits to a spectrum of stakeholders encompassing farmers, agribusiness entities, and food producers. The significance of early crop yield estimation reverberates through multiple facets of agricultural operations, orchestrating enhanced precision in harvesting, optimized marketing strategies, streamlined labor allocation, and judicious resource management (Abbate et al., 2023; De Clercq et al., 2018). In the quest for precision and reliability in crop yield estimation, the agricultural landscape has been significantly influenced by the increasing availability of agricultural datasets (Reynolds et al., 2018). These datasets serve as a bedrock of empirical information, enriching predictive models and facilitating informed decision-making (Hegedus et al., 2023). Furthermore, integrating meta-knowledge, encompassing insights and expertise acquired from prior agricultural experiences, imparts a refined understanding of diverse crop yield determinants (Elnahal et al., 2022).

Computers, their hardware and software, and the supported devices significantly contribute to diverse aspects of human life (Diker, 2022), and one such aspect is yield estimation. In the continuous pursuit of elevating the performance of crop yield estimation models, hyperparameter optimization emerges as a crucial aspect that must be seriously considered (Padmapriya and Sasilatha, 2023). It stands as an anchor, enabling the refinement of model parameters through a meticulous and data-driven approach. The overarching goal is the augmentation of predictive accuracy and the fortification of model generalizability,

* Corresponding author.
E-mail addresses: 202014080107@std.uestc.edu.cn (M.H. Tunio), jpli2222@uestc.edu.cn (J.P. Li), 202011081605@std.uestc.edu.cn (X. Zeng), faijanakhtar98@gmail.com (F. Akhtar), syedattique.shah@bcu.ac.uk (S.A. Shah), 202014080105@std.uestc.edu.cn (A. Ahmed), augustus.yangyu@gmail.com (Y. Yang), belalheyat@gmail.com (M.B.B. Heyat).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2023.101895
Received 20 October 2023; Received in revised form 5 December 2023; Accepted 15 December 2023
Available online 4 January 2024

Nomenclature

R^2	R-Square
AKF-NAR	Alternative Kalman Filter-Nonlinear Autoregressive
AKF-SVR	Alternative Kalman Filter-Support Vector regression
ANFIS	Adaptive Neuro-fuzzy interface system
ANN	Artificial Neural Network
APSIM	Agricultural Production System sIMulator
AutoML	Automated Machine Learning
BMA	Bayesian Model Averaging
CDT	Crop Decision Tree
CK-NN	Crop k-Nearest Neighbor
CMLR	Crop Multivariate Logistic Regression
CNN	Convolutional Neural Network
CRF	Crop Random Forest
DT	Decision Tree
EDA	Exploratory Data Analysis
ELT	Ensemble Learning Tree
FFA	Firefly Algorithm
GA	Genetic Algorithm
LASSO	Least Absolute Shrinkage and Selection Operator
LightGBm	Light Gradient-boosting Machine
LSTM	Long-short term Memory
MAE	Mean Absolute Error
ML	Machine Learning
MLP-SMO	(Multi-Layer Perceptron-Spider Monkey Optimization
MLR	Multiple Linear Regression
MOA	Multi-verse Optimization Algorithm
MSE	Mean Squared Error
PSO	Particle Swarm Optimization
RF	Random Forest
RGPE	Ranking-Weighted Gaussian Process Ensemble
RMSE	Root Mean Square Scaled Error
SGPR	Sparse Gaussian Process Regression
SVM	Support Vector Machine
SVR	Support Vector Regression
TransBO	Transfer Bayesian Optimization

which are paramount for robust crop yield estimation (Cedric et al., 2022).

Hyperparameter optimization encapsulates a multifaceted process that systematically explores a model's hyperparameters, which are the external configurations governing the model's behavior. However, effectively optimizing many parameters is a challenge (Joseph et al., 2022). This process harmonizes empirical experimentation and algorithmic techniques, enhancing the model's ability to discern intricate patterns within agricultural datasets. Consequently, it empowers the model to provide more accurate and adaptable crop yield predictions, essential for informed agricultural decision-making (Chiu et al., 2022). Optimization involves finding a position that minimizes a real-valued fitness function. The Bayesian optimization model utilizes an internal Gaussian process model and an acquisition function to select the next evaluation point of the optimal hyperparameter (Loey et al., 2022). Various DL and ML approaches have also used Bayesian optimization algorithms for hyperparameter optimization. including Three-Way Decision-based Bayesian Deep Learning (TWDBDL) for uncertainty quantification (Abdar et al., 2021), scalable BO using Deep Neural Networks (Snoek et al., 2015), predicting the properties of recycled aggregate concrete through Bayesian optimization techniques (Zhang et al., 2023a; Al-Rawashdeh et al., 2023) conducted a comparative investigation employing Bayesian optimization to predict building damage grade caused by earthquakes across multiple machine learning algorithms.

Machine learning is a pivotal decision support tool in crop yield prediction (Van Klompenburg et al., 2020). Numerous studies have been conducted to predict crop yield using diverse methodologies, such as step-wise multi-linear regression (Drummond et al., 2003). However, choosing the ML method demonstrating superior performance for a given application is paramount (Aslan et al., 2022). The utilization of random forest RF models for regional and global crop yield estimation (Jeong et al., 2016), the application of neural networks for soybean and corn crop prediction (Kaul et al., 2005), assessments of the impact of climate change on agriculture (Crane-Droesch, 2018), the incorporation of Convolutional Neural Networks (CNNs) with satellite imagery (Russello, 2018), the development of regression models employing the weighted histogram technique for yield prediction across various crops (Marko et al., 2016), the utilization of recurrent neural networks (RNNs), long short-term memory networks (LSTMs), DT, A hybrid MLR-ANN, federated RF algorithm, lasso regression and gradient decent (You et al., 2017; Jhajharia et al., 2023; Gopal and Bhargavi, 2019; Zhang et al., 2023b), environment interaction-based modeling approaches (Ansarifar et al., 2020), and association rule mining-based models employing various algorithms to determine the optimal fit (Romero et al., 2013), furthermore, The paper optimized Agricultural Production System sIMulator (APSIM) using differential evolution adaptive Metropolis and Bayesian multiplication with Gaussian likelihood divides the growing season into five phases and used exploratory data analysis (EDA) and Random Forest to select variables for a hybrid model in maize yield prediction (Li et al., 2023), the SegNet and Support Vector Regression (SVR) were used for early grapevine yield prediction (Palacios et al., 2023), the integration of various machine learning and statistical methods, along with feature extraction techniques, has been explored for yield estimation purposes. These approaches encompass aggregated and disaggregated data, highlighting their potential to enhance the accuracy of yield predictions (Aworka et al., 2022).

Recent studies in crop yield estimation have brought to the forefront a recurring challenge for scientists in this field: manual hyperparameter tuning, and more specifically, the conundrum of selecting the most suitable algorithm when deploying machine learning-based solutions for crop yield estimation (Oikonomidis et al., 2022). This quandary is not unique to agricultural scientists. Still, it is also a pressing concern for machine learning professionals, as no single machine learning model universally outperforms all others for every problem (Yuen et al., 2016). Efforts to address these questions and offer solutions by leveraging diverse datasets rather than conducting extensive experimentation and results evaluation can mitigate several constraints. These constraints include the scarcity of machine learning and data science expertise and the cost associated with computational processing (Cunha et al., 2018).

One of the approaches to recommend the most suitable algorithm for crop yield estimation lies within the domain of AutoML, also known as Automated Machine Learning. AutoML focuses on automating the entire end-to-end process of developing machine learning solutions, thereby enabling more effective handling of emerging challenges (Hutter et al., 2019; He et al., 2021). The intersection of these research areas, hyperparameter tuning, algorithm selection, and AutoML, holds promise for advancing the field of crop yield estimation, ultimately contributing to more efficient and accurate predictions in agricultural contexts.

Meta-learning has found extensive utility in proposing methodologies for single-label classification and regression tasks within predictive

domains. These methodologies span diverse domains, encompassing image recognition (Aguiar et al., 2019; Bronskill et al., 2020), and text analytics (Dou et al., 2019). Notably, recent investigations in meta-learning have expanded to encompass multi-target regression and threshold prediction for identifying favorable instances (Ghaderi Zefrehi et al., 2023). Additionally, integrating Artificial Neural Networks (ANN) with meta-heuristic rules has yielded predictive models, exemplified in predicting biochar yield based on pyrolysis and biomass process conditions (Khan et al., 2022). While these advancements demonstrate notable progress, a conspicuous gap persists in the existing literature concerning yield prediction through meta-transfer learning. Although adapting meta-models with limited data has shown promise in model transfer learning, challenges persist, especially when dealing with extensive datasets encompassing diverse operations (Liu et al., 2023). The principal objective of this study was to address this critical need by providing an optimal approach for task adoption, particularly in organizing tasks into sub-task levels. Moreover, applying transfer techniques not only necessitates knowledge transfer to be tailored to the specific target task but also calls for the automation of hyperparameter optimization. These concurrent challenges underline the suggesting practical models for hyperparameter optimization, ultimately leading to improved accuracy and resilience in the realm of meta-knowledge transfer (Jeon et al., 2021).

To predict biochar yield, the investigation conducted in the study by Haq et al. (2022) employed a diverse set of machine learning techniques, encompassing Ensembled Learning Tree (ELT), Support Vector Machines (SVM), Gaussian Process Regression (GPR), and Decision Trees (DT). Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) were used in conjunction with these methods. The predictive model incorporated key pyrolysis reaction parameters, including heating rate, temperature, particle size, time, N2 flow, and biomass attributes comprising proximate and structural characteristics. Furthermore, the study harnessed machine learning models in tandem with state-of-the-art optimization techniques, feature selection methodologies, and hyperparameter tuning processes, all aimed at enhancing the predictive performance of the framework.

The effectiveness and robustness of crop yield estimation can be significantly enhanced by integrating transferable meta-models alongside automated hyperparameter tuning. Hyperparameter optimization in the context of yield estimation serves to refine model complexity, regularization, robustness, and learning rates. As demonstrated by Charoen-Ung and Mittrapiyanuruk (2019), a notable precedent in the field involved training a random forest classifier for sugarcane yield prediction. This endeavor incorporated forward feature selection in combination with hyperparameter tuning. However, it is essential to acknowledge that such hyperparameter tuning methods often demand a substantial degree of specialized knowledge and expertise, as Wu et al. (2019) highlighted. The fusion of agricultural yield estimation with meta-learning presents an intriguing avenue for research. Addressing the existing gap in this domain and developing a robust hyperparameter optimization approach can significantly enhance prediction frameworks' overall efficacy and resilience. This research investigates the establishment of a relationship between machine learning model performance and optimal hyperparameters using LightGBM as the base learner and Bayesian optimization. The proposed solution centers on creating a framework seamlessly incorporating domain-specific knowledge from various crops into automated hyperparameter optimization. This integration ensures the robustness and effectiveness of the crop yield estimation framework. Notably, the effects of data preparation techniques in conjunction with hyperparameter tuning have received limited attention in prior research (Meola et al., 2023).

In laying the foundation for our Meta knowledge-based framework for robust crop yield estimation, this study contributes to ongoing initiatives to revolutionize crop yield estimation. This effort involves transferring knowledge between optimization tasks, leveraging meta-model principles integrated with Bayesian Optimization to facilitate the transfer of meta-knowledge for hyperparameter optimization in target tasks.

The proposed study has significantly contributed by implementing a novel meta-knowledge-guided framework for precise crop yield estimation. Firstly, a comprehensive analysis was undertaken to assess the influence of hyperparameters on the framework's performance. This study systematically explored different hyperparameter configurations and meticulously evaluated their impact on yield estimation accuracy. This rigorous process enabled us to identify the optimal hyperparameter settings, providing crucial insights and guidelines for fine-tuning the framework to achieve exceptional results.

The study introduced a novel mechanism for transferring meta-knowledge from meta-models to target tasks to enhance the framework's performance. Moreover, this study seamlessly integrated an automated hyperparameter selection mechanism, empowering the framework to autonomously choose the most optimal hyperparameters for each specific target task. These innovative advancements significantly improve the framework's adaptability and efficiency, creating a comprehensive solution that effectively harnesses meta-knowledge while requiring minimal manual intervention. This approach also streamlined a set of knowledge transfer algorithms for target task adoption and comparative analysis, specifically focusing on their performance within yield estimation.

The main contributions of this study are summarized as follows:

- A novel framework was developed, seamlessly integrating a Bayesian Optimization model with a meta-framework. This novel approach facilitates dynamic hyperparameter optimization for enhanced performance.
- The study delved into the crucial aspect of hyperparameters within the context of the developed framework. A comprehensive investigation was carried out to discern their impact on model performance.
- Comparative Analysis of Transferable Meta-Models was conducted by simplifying the target task's RGPE, SGPR, and TransBO methods. This assessment provided valuable insights into the strengths and weaknesses of each approach.
- The study designed a robust and efficient framework capable of automatically transferring meta-knowledge derived from base tasks to select the most optimal hyperparameters for target tasks.
- To assess the framework's generalizability, three diverse agricultural crop datasets were utilized. This analysis offers insights into the framework's performance across varied agricultural contexts, emphasizing its versatility and applicability.

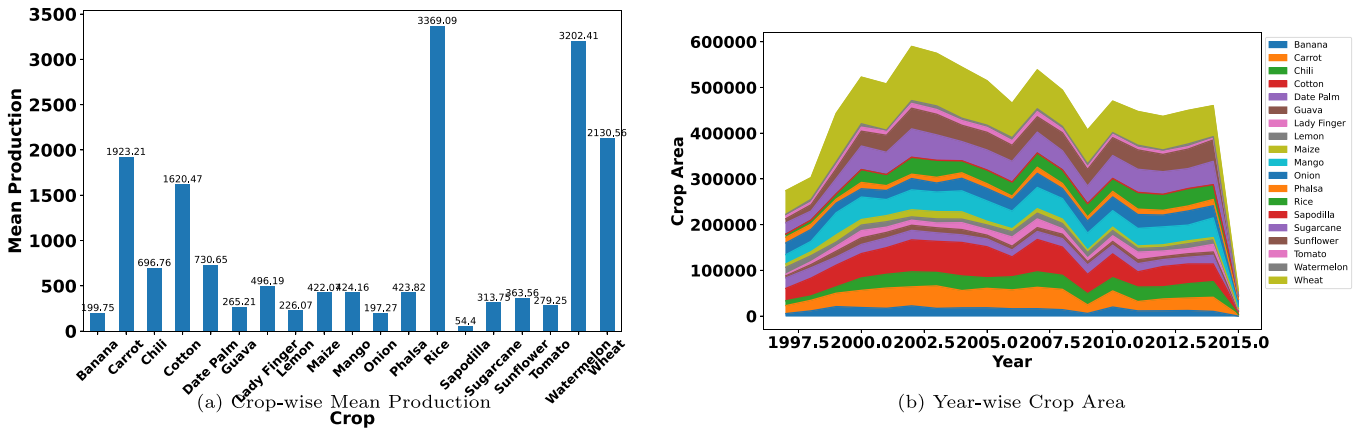
The rest of this study is structured as follows: Section 2 presents the materials and methods, and Section 3 delves into an in-depth exploration of meta-data processing and the intricate knowledge transfer mechanism of the proposed meta-knowledge-guided framework. Section 4 presents detailed results and findings. Finally, Section 5 encapsulates the study's conclusions and key takeaways.

2. Materials and methods

This section explores the fundamental components of the methodology, data collection, and analysis method, beginning with the stages of data gathering and preparation. Subsequently, this paper provides a thorough presentation of our exploratory data analysis (EDA) efforts, encompassing a diverse range of visualizations and analytical methods applied to reveal latent insights and identifiable patterns within a private dataset.

2.1. Data collection and preprocessing

The scarcity of real-world datasets for yield estimation poses a significant challenge for researchers in this domain. Due to a lack of



(a) Crop-wise Mean Production (b) Year-wise Crop Area

Fig. 1. Comparison of crop-wise mean production (a) and year-wise crop area (b).

real-world data, the researchers use the synthetic dataset (Korneva and Blockeel, 2020). To address this issue, three diverse datasets have been utilized in this study, including one of the synthetic (Wild blueberry Yield Prediction) datasets that simulate real-world conditions alongside publicly available and privately sourced datasets to enhance the reliability of the framework. This subsection provides an overview of the data collection methodology, encompassing the utilization of these diverse datasets for yield estimation.

1. *Sindh Crops Dataset*, a private dataset comprises a diverse collection of real-world data, meticulously gathered and preprocessed to ensure its relevance and reliability. The resulting dataset contains the yield records for 19 different crops with 35,364 samples for a period ranging from 1997 to 2015. This dataset was compiled from several district-level administrative institutions in the Sindh Province of Pakistan. The dataset encompasses various attributes and features for a comprehensive investigation into our research objectives, which essentially required this framework for robust crop yield estimation. The private dataset was employed to validate the disparities and evaluate the impact of incorporating the original dataset on the framework's performance with two other public datasets. This investigation aimed to assess variations and determine whether integrating the original (private) dataset for training improves the framework's efficacy.
2. *Wild blueberry Yield Prediction Dataset:* another publicly available dataset, sourced[1] a synthetic dataset was generated using an open-source Wild Blueberry Pollination Simulation Model (Obsie et al., 2020), a spatially-explicit computer simulation program specifically designed to explore the effects of various factors on pollination efficiency and yield within the wild blueberry agroecosystem, those factors include out-crossing and self-pollination, plant spatial arrangement, weather conditions, and bee species compositions, in isolation combination, that accurately represent the dynamics of wild blueberry pollination. This dataset contains various features, and it contains 15,289 samples. The validity of the simulation model for the synthetic data was established through field observation and experimentally collected data over the last 30 years (Qu and Drummond, 2018) in the USA, Maine, and Canadian Maritimes.
3. *Food Agriculture Organization and World Bank Dataset* The third publicly available dataset was sourced from the Food Agriculture Organization[2] and World Bank Data[3], the renowned institutions known for their authoritative and reliable agricultural data. The Data on crop yield and pesticides were obtained from the FAO, and data on rainfall and temperature were collected from the World Bank. Further, it involves extensive data gathering of 28,244 samples of 10 crops collected for 94 countries. It also involves the verification and framework-based validity assurance procedures to ensure accuracy and consistency.

2.2. Exploratory data analysis (Sindh Crops Dataset - our private dataset)

Using visualizations, descriptive statistics, and other methods, EDA analyses and summarizes the key features of a dataset (Gupta and Sharma, 2023). The EDA plays a crucial role in uncovering patterns, relationships, behavior, and co-relation within the datasets. This paper employs EDA techniques to understand a private dataset comprehensively and lay the foundation for our subsequent analyses and findings. Further, the study presents exploratory data analysis for a primary dataset to uncover meaningful insights such as crop-wise mean production, year-wise crop area, and scatter matrix for the Sindh Crops Dataset; however, the other two datasets are public.

2.2.1. Crop-wise mean production

Fig. 1(a) illustrates the crop-wise mean production values. The bar plot shows the y-axis of the figure represents the mean production values for the crops and the names of different crops used in the dataset, including bananas, carrots, sapodillas, etc, represented as the x-axis of the plot. The bar figure shows the average production status of various crops. The rice crop has the highest crop mean production value of 3369.09, watermelon has 3202.41, wheat has 2130.56, carrot has 1923.21, and cotton has 1620.47. Conversely, onion, sapodilla, Sunflower, and Tomato have lower mean production values.

2.2.2. Year-wise crop area

Fig. 1(b) illustrates the crop area for each year, in which a y-axis label represents the crop area and the x-axis indicates the year from 1997 to 2015. Every shaded region in the figure symbolizes a different hue corresponding to the crop names listed on the right side of the figure.

2.2.3. Scatter matrix of crop yield dataset

The scatter matrix of the crop yield dataset is shown in Fig. 2. Each color-shaded point in the scatter plot represents a specific crop name shown on the right side of the scatter plot. Each row and column of the scatter matrix's scatter plot grid represents a different variable (Year, Area, Rain, Production). Each variable's unique distributions are displayed as histograms or kernel density plots in the matrix's diagonal elements. The scatter matrix's off-diagonal cells each contain a

[1] https://www.kaggle.com/competitions/playground-series-s3e14/data
[2] http://www.fao.org/home/en/
[3] https://data.worldbank.org/

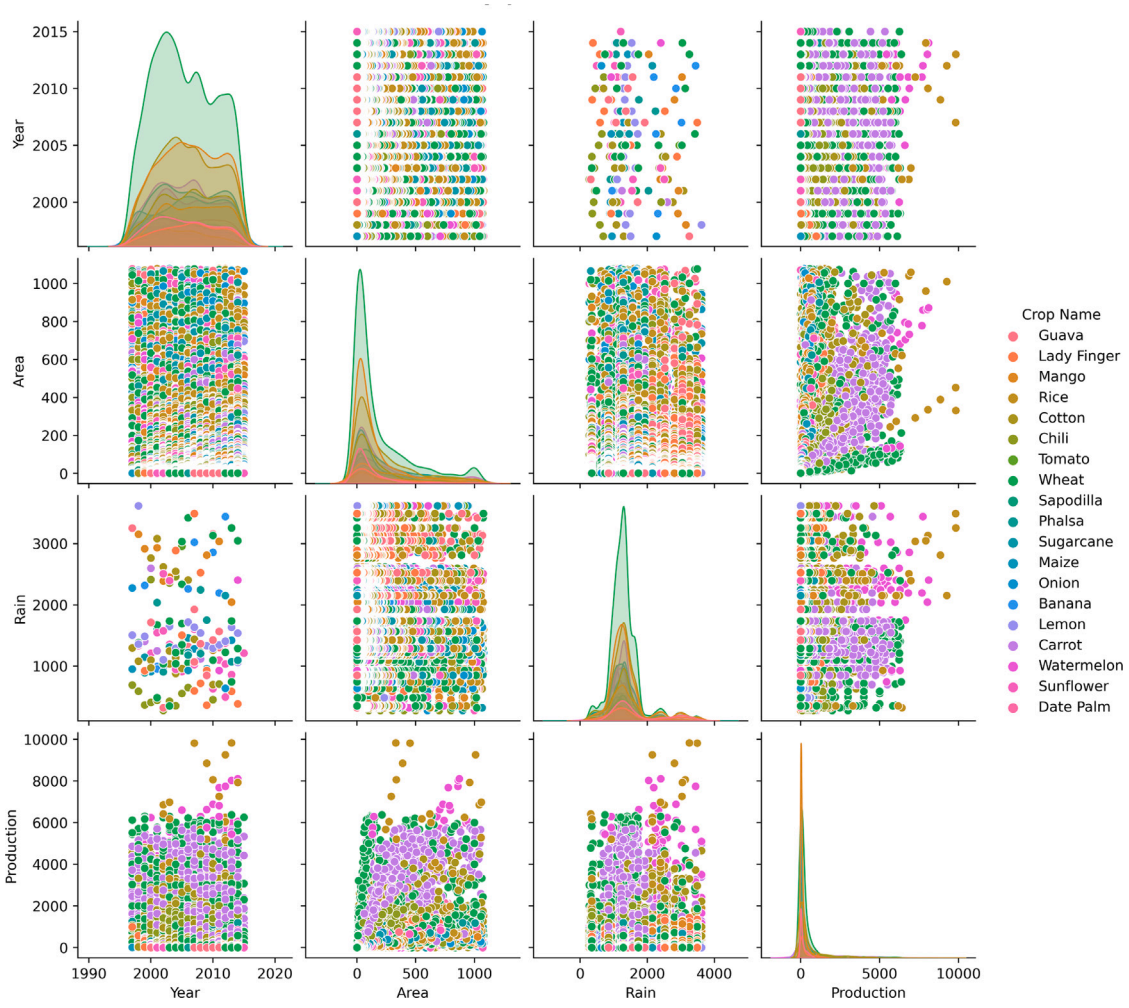


Fig. 2. Scatter Matrix Plot for Yield Estimation with regards to Dataset's Variables by Crops.

scatter plot comparing two variables, and these display the correlation between two variables while differentiating points by the "Crop Name" through the use of color as indicated in the list form on the right side of the figure.

- *Year vs. Year (Diagonal):* The histogram shows each crop's "Year" variable, which displays the distribution of data points for each crop over various years.
- *Area vs. Area (Diagonal):* This histogram illustrates the distribution of the "Area" variable for each specific crop.
- *Rain vs. Rain (Diagonal):* This histogram displays the distribution of the "Rain" variable for every crop.
- *Production vs. Production (Diagonal)* The histograms show how each crop's "Production" variable is distributed.
- *Year vs. Rain (Top-Right Cell):* This scatter plot indicates the correlation between rainfall and the year. The plot shows the more or less rainfall for certain years.
- *Year vs. Production (Bottom-Left Cell):* This diagonal map demonstrates how crop output has evolved. Long-term trends can be seen through patterns in the figure.
- *Area vs. Rain (Second Row, First Column Cell):* This scatter plot determines whether the area under cultivation and the amount of rainfall are related.
- *Area vs. Production (Second Row, Second Column Cell):* This diagram explores the relationship between cultivated area and crop production.
- *Rain vs. Production (Second Row, Third Column Cell):* This scatter plot shows if there is a relationship between rainfall and agricultural productivity.

Various crops were distinguished using these colored points to analyze the interconnections with multiple variables. These observed trends provide valuable insights into the relationship between crop yield and the Sindh Crops dataset variables.

3. Meta-knowledge guided framework

In the context of working with limited data and meta-models, one of the challenges is the need to automate hyperparameter optimization dynamically. The novel framework was developed based on the principles of meta knowledge transfer, which leverages a machine learning base model to iterate autonomously through a self-improvement cycle for each base-task and target task. However, the best hyperparameter optimization is performed dynamically during each iteration of this self-improvement cycle. The resulting performance outcomes are then utilized to train a Bayesian optimization model integrated with the central meta-model in the framework. This approach streamlines the hyperparameter tuning process and enhances overall efficiency in handling small amounts of datasets to increase the robustness of the model for yield estimation. Fig. 3 illustrates our meta-knowledge guided framework, which represents a meta-model learning from various base tasks and assimilating knowledge from an integrated Bayesian optimization model to perform well with new target tasks. To optimize

Algorithm 1 Meta-Knowledge-Transfer with Bayesian Optimization model

```
Initiate Model with Data_Name
Load Data and split data into base tasks N based on Crop as a Feature
Initialize Bayesian Optimization mode model {#For more comprehensive information on Bayesian Optimization, refer to a separate algorithm as in Algorithm 2.}
Fit Bayesian Optimization mode model on the training data for the last base task N-1
for each base task do
    Generate history data for transfer learning
    Create ConfigurationSpace for base task
    Initialize empty list for auxiliary loss values
    Create History object for the base task
    for each result in base task do
        Sample configuration from ConfigurationSpace
        Evaluate objective function using sampled configuration
        Update History with observation
        Append auxiliary loss to the list
    end for
    Append History object to transfer learning history list
end for
Initialize Advisor object for target task optimization
for each iteration do
    Get suggestions from the Advisor
    Evaluate the objective function for the target task using the suggested configuration
    Log the result
    Update Advisor with observation
end for
```

the hyperparameters and parameters of target tasks by transferring the meta-knowledge model, which incorporates the observations from the target tasks via Bayesian optimization. These optimizations are further influenced by insights gained from meta tasks, ensuring robust and adaptable results. This process enhances the meta model's ability to generalize its learning and efficiently adapt to new tasks.

The left-top rectangular block in the diagram represents different initial tasks (Initial Task-1, Initial Task-2, Initial Task-3, and Initial Task-n), each contributing to the meta model's learning process. Data preprocessing, which handles missing values, outliers, and inconsistencies, ensures accuracy and appropriateness. Duplicate removal reduces redundancy and potential biases. The term "preprocessed Initial task" denotes standardized, high-quality data for use with a meta-knowledge-guided model, which enhances the model's learning rate and generalization. The dynamic best hyperparameter optimization process for robust yield prediction includes the following subsections.

3.1. Meta model

Our framework's core component is the Meta Model, which leverages data and information from base tasks to acquire meta-knowledge of underlying data patterns, relationships, and representations. It focuses on extracting relevant features, learning task-specific representations, and discerning relationships between base and target tasks. The Meta Model optimizes its hyperparameters, including the model's behavior (e.g., learning rate, regularization parameters, network architecture), to achieve optimal performance on base tasks. This study employs the Meta Model in conjunction with an integrated Bayesian optimization model, as depicted in Fig. 3, for efficient hyperparameter optimization. Through iterative learning and refinement of parameters and hyperparameters based on observations from base tasks, the meta-model continuously adapts and improves its internal representations and predictions, enhancing generalization to target tasks. This process integrates Bayesian optimization with the Meta Model to optimize hyperparameters using observed data, facilitating robust knowledge transfer to target tasks.

3.2. Bayesian optimization model

The primary goal of Bayesian Optimization (BO) is to ascertain the global minimum, aiming to pinpoint the optimal solution or configuration within the given parameter space, denoted as $\mathbf{x_{opt}}$, of a function with unknown internal workings within a constrained domain $\mathcal{X}$. This is accomplished through an iterative process of evaluating the function at different input configurations $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ and observing the associated outputs $b_1, b_2, \ldots, b_n$. In each iteration, a probabilistic model f is trained on the gathered data $D = (\mathbf{a}k, b_k)k = 1^n$, enabling the model to capture the underlying patterns and relationships within the observations.

Subsequently, an acquisition function $\alpha(\mathbf{a})$ is employed in this framework to select the next best hyperparameter, balancing the trade-off between exploration and exploitation. In this context, f_i denotes a noisy approximation of the actual function value. Gaussian Process (GP) regression is employed to estimate the underlying function, resulting in a posterior probability $f(\mathbf{a}|D)$ with a calculated mean $\mu(\mathbf{x})$ and variance $\sigma^2(\mathbf{a})$, which are expressed in Eqs. (1) and (2) as follows:

$$\mu(\mathbf{A}_*) = \mathbf{k}_*{}^\top(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{b}, \tag{1}$$

$$\sigma^2(\mathbf{a}) = k(\mathbf{a}_*, \mathbf{a}_*) - \mathbf{k}_*{}^\top(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{k}_*, \tag{2}$$

Within this framework, $k(\cdot,\cdot)$ represents a kernel function with associated hyperparameters $\boldsymbol{\sigma}$. The vector $\mathbf{k}$ denotes the co-variances between the new point $\mathbf{x}$ and all training points, while $\mathbf{K}$ corresponds to the kernel matrix of size $k \times k$. Each entry K_{ij} of the kernel matrix is defined as $k\theta(\mathbf{A}_i, \mathbf{A}_j)$.

The expected improvement (EI) is the commonly used acquisition function, known for its closed-form computation and typically yielding good results. Let $f(\mathbf{x_{best}})$ represent the current best function value, defined as $f(\mathbf{a_{best}}) = \min_{k\in1,\ldots,k} f(\mathbf{a}_i)$. The EI is expressed in Eq. (3) and was computed as follows:

$$\begin{aligned}\alpha(\mathbf{A}|D) &= \mathbb{E}\mathbb{I}_{f(\mathbf{A}|D)}[\max(0, f(\mathbf{A}_{\text{best}})\\ &-f(\mathbf{A}))] = \sigma(\mathbf{A}(z)\Phi(Z) + \sigma(\mathbf{A})\phi(Z))\end{aligned} \tag{3}$$

Where $z = (f(\mathbf{x_{best}}) - \mu(\mathbf{x}))/\sigma(\mathbf{x})$, $\Phi(\cdot)$ represents the cumulative distribution function, and $\phi(\cdot)$ represents the probability density function of the normal distribution.

These steps constitute a critical part of our research methodology, allowing us to adapt the Bayesian Optimization model to various base tasks and leverage historical data for more effective learning and generalization of this framework.

The Algorithm 1 depicts a framework for knowledge-transfer and hyperparameter optimization, which leverages two surrogate types gp and prf, for each of three algorithms (RGPE, SGPR, and TransBO) for analyzing the efficiency of knowledge-transfer of the target task using the information from multiple base tasks. However, Algorithm 2 depicts the Bayesian Optimization model.

3.3. Base model and objective method

In this framework, our first step entails initializing the Bayesian Optimization model to simplify further processing. The model is then meticulously trained using the training data related to the base task. For each base task, our study executes the following steps:

Generating historical data that can be used for transfer learning purposes is crucial in facilitating the adaptation and enhancement of model performance on new tasks. To achieve this, the configuration Space is created for the base task. This space encompasses a range of hyperparameters and configurations the model can explore during its learning process. A history object was also established dedicated

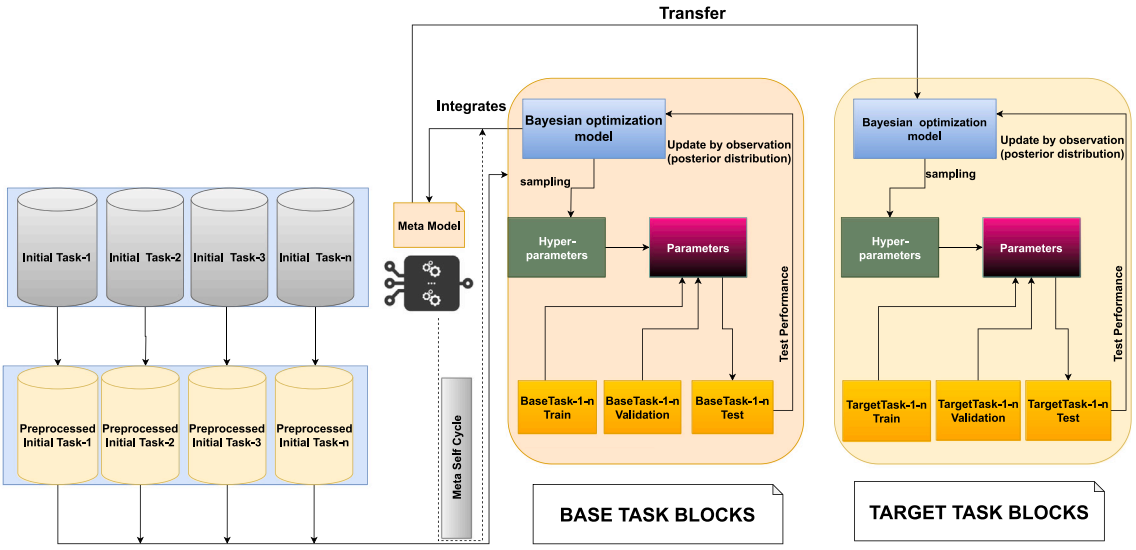


Fig. 3. Meta-Knowledge Guided Framework for Robust Yield Estimation.

Algorithm 2 Bayesian Optimization

1: Initialize Bayesian Optimization model

- Define the parameter space X

2: Iterative Steps:

- Evaluate the function at different input configurations a_1, a_2, ..., a_n and observe the associated outputs b_1, b_2, ..., b_n.
- Train a probabilistic model f on data $\mathbb{D} = (a_k, b_k)k = 1^n$ to capture underlying patterns and relationships.
- Calculate the mean $\mu(x)$ and variance $\sigma^2(a)$ of the posterior probability $f(a|D)$ (GP) regression:
  - $\mu(A_*) = k_*^T(K + \sigma^2 I)^{-1} b$
  - $\sigma^2(a) = k(a_*, a_*) - k_*^T(K + \sigma_N^2 I)^{-1} k_*$, where k(·, ·) represents the kernel function, k is the vector of co-variances between the new point x and all training points, K is the kernel matrix of size $k \times k$, and σ^2 is a hyperparameter.
- Select the next best hyperparameter using an acquisition function α(a) to balance exploration and exploitation. The expected improvement $(\mathbb{EI})$ is a commonly used acquisition function:
  - $\alpha(A|D) = \mathbb{EI}_{f(A|D)}[max(0, f(A_{best}) - f(A))] = \sigma(A(z)\Phi(Z) + \sigma(A)\phi(Z)$, where A_{best} represents the current best function value.
- Update the model and continue iterating.

3: Return the optimal solution/configuration $xopt$.

to capturing and storing the model's past observations, configurations, and associated outcomes of the current base task. For each result obtained during the base task's execution, the system performs the following sub-steps:

The random sampling technique was employed to select a configuration from the Configuration Space, effectively determining a specific set of hyperparameters and settings for the model. Subsequently, an objective function is defined to convert and evaluate the chosen configuration, resulting in the derivation of model parameters. This process involves training a LightGBMRegressor on the base tasks and making predictions on the target task, with performance metrics used to assess the model's prediction capabilities. As the transfer learning process progresses, the History object is continually updated with the observed configuration and its corresponding outcome. After each base task, the History object associated with that task is appended to a transfer learning history list. This cumulative list is a valuable resource, accumulating historical data from all base tasks and facilitating the transfer of meta-knowledge across tasks.

Advisor Object for transferable meta-model and Target Task Optimization used an essential component called the Advisor object, specifically designed and integrated with optimizing the target task. This Advisor is crucial in guiding the meta-knowledge-transfer and optimization process for the target task, ensuring efficient and effective learning.

Before delving into Section 3.4, Transferable meta-model, it is important to introduce the Advisor object, a crucial component integral in facilitating meta-knowledge transfer and optimizing the target task. Subsequently, Section 3.4 explores the transferable Meta-model in detail, which complements the functionalities of the Advisor object. They constitute a comprehensive framework for effective knowledge transfer and adaptive target task optimization. Here's a breakdown of the steps associated with the Advisor object.

Initialize the advisor object, which is pivotal in providing valuable recommendations for fine-tuning the model's configuration tailored to the target task. The advisor continuously improves suggestions throughout the iterative optimization process by leveraging insights from the model's performance and observations. During each iteration, the advisor is consulted to retrieve configuration suggestions, encompassing relevant hyperparameters and configurations for the model to consider in the context of the target task. By incorporating these suggestions, the model effectively explores the parameter space, optimizing its performance and achieving desirable outcomes. The model's performance on the target task is evaluated using the suggested configuration, and the objective function measures the model's performance under the current recommendation. To track the progress of the optimization process, the results of each evaluation are logged, providing a comprehensive record of the model's performance at each iteration. Following the evaluation, the Advisor object is updated with the observed performance data, enabling the advisor to adapt and refine its suggestions for subsequent iterations. This iterative optimization approach empowers the model to continuously improve and fine-tune its configuration, ultimately enhancing its performance on the target task.

The model repeats this process for each iteration, allowing the Advisor to provide refined suggestions. Further, it helps the model to adapt and improve its performance on the target task over time.

3.4. Transferable meta-model

To facilitate meta-knowledge transfer for the target task within our framework, the three distinct meta-knowledge transfer algorithms, Ranking-Weighted Gaussian Process Ensemble (RGPE), Sparse Gaussian Process Regression (SGPR), and Transfer Bayesian Optimization (TransBO)-topov3, were employed. For each algorithm, two different surrogate types were utilized: (1) Gaussian Process (GP) for capturing complex relationships within the meta-data and (2) Random Forest (RF), referred to as "prf" for practical problems like hyperparameter optimization (HPO). These algorithms are employed for hyperparameter optimization based on the historical data relevant to the target task. The following sections describe each algorithm and its associated mechanisms and simplifications.

1. Ranking-Weighted Gaussian Process Ensemble: The advances in transfer learning for Bayesian optimization are a common approach used to represent the current task (denoted as 't') as a linear combination of surrogate models. This involves merging a task-specific surrogate model with surrogate models for all previous tasks (from 1 to t−1). If surrogate models for base tasks are available from previous runs, they can be used without retraining, simplifying the process and eliminating the need for meta-features. However, determining the optimal weights for combining these models is a crucial challenge. This section explores the ranking method used for the two-stage transfer surrogate for learning linear combinations and uses an acquisition function to enhance ensemble performance. The Eq. (4) illustrates the linear combination of mean predictions from surrogate models.

$$\mu(x_*) = \sum_{i=1}^{t} w_i \mu_i(x_*) \tag{4}$$

In evaluation, this method assesses the combination weights of base models based on their predictions for the target task, evaluating their generalization capability. This model employs cross-validation, leave-one-out models to gauge the target model's generalization. This entails the target model omitted to calculate the loss accordingly to account for this cross-validation approach as shown in Eq. (5)

$$L(f, D^t) = \sum_{k=1}^{n_t} \sum_{l=1}^{n_t} \left(1\left((f(X_k^t) < fX_l^t) \oplus (y_k^t < y_l^t)\right)\right) \tag{5}$$

Once all models adeptly arrange data points in the correct order, the subsequent step involves assigning weights to each model based on the probability of it exhibiting the lowest ranking loss within the ensemble. This probability is estimated using a Monte Carlo approximation, where bootstrap samples are drawn for each model, and their alignment with observed data is assessed to compute model weights as below in Eq. (6).

$$w_i = \frac{1}{S} \sum_{s=1}^{S} \frac{\left[I(i \in \arg\min_i {}' l_{i',s})\right]}{\left[\sum_{j=1}^{t} I(j \in \arg\min_i {}' l_{i',s})\right]} \tag{6}$$

2. Sparse Gaussian Process Regression: This algorithm was used with a sequence of tasks $\{S_i\}_{i=1}^k$ on unknown objective functions $\{f_i\}_{i=1}^k$, where the current task is S_k, and build two sequences of regressors $\{R_i\}_{i=1}^k$ and $(R_i')_{i=1}^k$ having posterior mean functions $\{\mu_i\}_{i=1}^k$ also $(\mu_i')_{i=1}^k$, In the same order, and posterior standard deviation functions $\{\sigma_i\}_{i=1}^k$ and $(\sigma_i')_{i=1}^k$, final predictions, in the given order, will be μ_k and σ_k. Here $D_i = \{(x_t^i, y_t^i)\}_t$ serve as the dataset for the task S_i. Let R_i' be a regressor trained using data $\{((x_t^i, y_t^t - \mu_{i-1}(x_t^i)))_t\}$, which calculates μ_i' and also σ_i'. Then locate and find the posterior means at level i as $\mu_i(x) := \mu_i'(x) + \mu_{i-1}(x)$. To calculate the posterior standard deviations. i, $\sigma_i(x)$, to be a weighted geometric mean of $\sigma_i'(x)$ and $\sigma_{i-1}(x)$, where the weights are a function of the amount of data (i.e., completed trials) in S_i and S_{i-1}.
3. Two-Phase Transfer Bayesian Optimization (TransBO): Here, the roles of two parameters, w and p, within the framework of TransBO (Transfer Bayesian Optimization). w is tasked with merging k source base surrogates to best fit with target observations, while p balances among the two surrogates, $\mathcal{M}^S$, and $\mathcal{M}^T$. The main overarching goal of TransBO is to maximize the generalization performance of $\mathcal{M}^{TL}$.

The strategy is to obtain the source surrogate $\mathcal{M}^S$ as a weighted combination of the predictions of source base surrogates $\{\mathcal{M}^1, \ldots, \mathcal{M}^K\}$ as expressed in Eq. (7).

$$\mathcal{M}^S(\mathbf{x}) = \sum_{i=1}^{K} w_i \mathcal{M}^i(\mathbf{x}) \tag{7}$$

To systematically obtain $\mathcal{M}_S$, this method uses a principled approach by employing a differentiable pairwise ranking loss function to assess the alignment between the predictions of $\mathcal{M}_S$ and the given observations in $\mathcal{D}$. In the context of Hyperparameter Optimization (HPO), This ranking loss function was computed as in Eq. (8).

$$L(W, \mathcal{M}^S; \mathcal{D}) = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1, y_j < y_k}^{n} \phi(\mathcal{M}^S(x_k) - \mathcal{M}^S(x_j)) \tag{8}$$

$$\phi(z) = \log(1 + e^{-z}) \tag{9}$$

The framework extends the process of acquiring the source surrogate $\mathcal{M}^S$, specifically the optimization of W, by formulating it as the subsequent constrained optimization problem:

Minimize W: $L(W, \mathcal{M}^S; \mathcal{D})$ subject to: $1^T w = 1, \quad w \geq 0$, hence the goal is the ranking loss of $\mathcal{M}_S$ on $\mathcal{D}$

In order to capture the generalization properties of $\mathcal{M}^{TL}$, as depicted in Eq. (10). Concurrently, this approach also used the partial surrogate model $\mathcal{M}_{-i}^T$ directly on $\mathcal{D}_{-i}^T$. Subsequently, framework combines the surrogates $\mathcal{M}_{-i}^S$ and $\mathcal{M}_{-i}^T$ linearly to yield $\mathcal{M}_{-i}^{TL}$.

$$\mathcal{M}_{-i}^{TL} = (p^S)(M_{-i}^S) + (p^T)(M_{-i}^T) \tag{10}$$

3.5. Model evaluation metrics

To assess and evaluate the performance of a meta-knowledge-guided model by employing training testing splits on datasets. After training the Bayesian Optimization mode model, this study utilizes widely accepted metrics in machine learning regression problems (Yang et al., 2023) to evaluate the model performance. The model predicts the crop yield ('y-pred') based on the test data ('x-test') and calculates evaluation metrics such as MSE (mean squared error), MAE (mean absolute error), scaled-MSE (scaled mean squared error) divided by the standard deviation of 'y-test'), scaled-MAE (Scaled mean absolute error) and the R^2 score to evaluate the model's fit or appropriateness. A higher R^2 value indicates a stronger alignment between the predicted values and the ground truth, highlighting the superior fit of the predictive model.

RMSE (Root Mean Square Error):

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Y_{\text{true},i} - Y_{\text{pred},i})^2} \tag{11}$$

RMSSE (Root Mean Square Scaled Error):

$$RMSSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\frac{Y_{\text{true},i} - Y_{\text{pred},i}}{MASE} \right)^2} \tag{12}$$

MASE (Mean Absolute Scaled Error)

$$\text{MASE} = \frac{1}{N} \sum_{i=1}^{N} \frac{\left| Y_{\text{true},i} - Y_{\text{pred},i} \right|}{\frac{1}{N-1} \sum_{j=2}^{N} \left| Y_{\text{true},j} - Y_{\text{true},j-1} \right|} \tag{13}$$

R-Square (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (Y_{\text{true},i} - Y_{\text{pred},i})^2}{\sum_{i=1}^{N} (Y_{\text{true},i} - \bar{Y}_{\text{true}})^2} \tag{14}$$

Table 1
Performance of transferable RGPE, SGPR, and TOPOv3 algorithms on diverse agricultural data.

Model type	MSE	MAE	Scaled MSE	Scaled MAE	R_Square	AvgTime/Epoch
Sindh Crops Dataset						
rgpe_prf	0.0020	0.0157	0.0176	0.2578	**0.8437**	5.93
sgpr_prf	0.0116	0.0540	0.1028	0.8836	0.0877	5.61
topov3_prf	0.0021	0.0163	0.0182	0.2671	0.8383	5.25
rgpe_gp	0.0020	0.0158	0.0179	0.2584	0.8415	5.27
sgpr_gp	0.0122	0.0555	0.1077	0.9087	0.0446	5.91
topov3_gp	0.0022	0.0167	0.0196	0.2742	0.8257	5.54
Blueberry						
rgpe_prf	0.0142	0.0853	0.0138	0.1033	**0.9865**	2.85
sgpr_prf	0.5078	0.5497	0.4937	0.6654	0.5176	2.67
topov3_prf	0.0176	0.0893	0.0171	0.1081	0.9833	2.35
rgpe_gp	0.0235	0.1054	0.0229	0.1276	0.9776	3.5
sgpr_gp	0.1743	0.3222	0.1695	0.3900	0.8344	2.95
topov3_gp	0.0189	0.0924	0.0184	0.1119	0.9821	2.41
FAO & World_Bank Dataset						
rgpe_prf	0.0294	0.0787	0.0293	0.1037	**0.9708**	3.68
sgpr_prf	0.5753	0.5528	0.5738	0.7285	0.4276	2.89
topov3_prf	0.0300	0.0784	0.0299	0.1034	0.9702	4.46
rgpe_gp	0.1570	0.2844	0.1566	0.3749	0.8438	3.12
sgpr_gp	0.2583	0.3736	0.2577	0.4923	0.7430	3.65
topov3_gp	0.0638	0.1632	0.0636	0.2151	0.9366	3.83

Table 2
Parameter importance and identified optimal hyperparameters for a diverse set of datasets.

Dataset	L_Rate	N_Esti	NL	MD	MCS	Subsample	OOV	TrialsNo
Parameter Importance								
Sindh	0.5726	0.1975	0.0246	0.0144	0.0096	0.0046	N/A	N/A
Blueberry	0.4282	0.4045	0.0167	0.0088	0.0181	0.0056	N/A	N/A
FAO	0.7261	0.1644	0.0291	0.0033	0.0018	0.0018	N/A	N/A
Optimal Hyperparameters								
topov3-prf	0.09912	36	412	31	16	0.9000	0.0019	50
sgpr_prf	0.0489	98	476	90	30	1	0.0116	50
rgpe_gp	0.3	101	209	101	30	0.7000	0.0236	50

Table 3
Comparative analysis of proposed framework with current SOTA.

RefStudy	Dataset	Method(s)	MSE	MAE	RMSE	R_Square
Obsie et al. (2020)	Wild Blueberry	XGBoost	0.0639	0.1519	0.2527	0.9380
Seireg et al. (2022)	Wild Blueberry	Stacking	0.01756	–	0.1325	0.9840
Borrero and Borrero-Domínguez (2023)	Wild Blueberry	Hybrid AKF-SVR Hybrid AKF-NAR	–	7.86, 7.93	13.37, 16.73	0.973, 0.958
Proposed Framework	**Wild Blueberry**	**rgpe_prf**	**0.0142**	**0.0853**	**0.119**	**0.9865**
Aworka et al. (2022)	FAO and World Bank	CRF	0.1176	–	0.3430	0.9227
Mariadass et al. (2022)	FAO and World Bank	XGBoost	–	0.0936	0.1812	0.9800
Cedric et al. (2022)	FAO and World Bank	Ck-NN, CDT, CMRL	–	0.0160, 0.088, 0.0315	–	0.9503, 0.9465, and 0.8380
Ahmed (2023)	FAO and World Bank	MLP-SMO	–	–	0.13	0.96
Proposed Framework	**FAO and World Bank**	**rgpe_prf**	**0.0294**	**0.0787**	**0.1714**	**0.9708**
Proposed Framework	**Sindh Crops**	**rgpe_prf**	**0.0020**	**0.0157**	**0.0447**	**0.8437**

4. Results and discussion

This section discusses the results obtained by the proposed framework for three diverse agriculture crop datasets employed in this study.

Table 1 comprehensively evaluates three meta-knowledge-transfer algorithms using performance metrics. Using the Sindh Crops Dataset, this framework demonstrated remarkable results in terms of MSE, MAE, scaled MSE, and MAE, as well as R-Square. Particularly for the Ranking-weighted Gaussian Process Ensemble ('rgpe_prf') algorithm, it recorded MSE of 0.0020 and MAE of 0.0157, indicating high accuracy and a well fit of (R^2 = 0.8437). Furthermore, in the first row of the table for the Sindh Crops Dataset, the value of "AvgTime/Epoch" for the rgpe_prf model is 5.93. This represents the average time unit in seconds taken per epoch during the training process of the proposed model. Conversely, sparse Gaussian process regression ('sgpr_prf') performs less effectively, displaying higher MSE and MAE values and a low R^2 of 0.0877. Furthermore, using the Wild Blueberry dataset, 'rgpe_prf' excels with a remarkably low MSE of 0.0142 and MAE of 0.0853, accompanied by a maximum R^2 of 0.9865, and the value of "AvgTime/Epoch" is 2.85 s that indicates the excellent fit. However, 'sgpr_prf' falls behind with significantly higher MSE and MAE values (R^2 = 0.5176). The framework on FAO and World_Bank Dataset showcases 'rgpe_prf' as the superior performer with a low MSE of 0.0294, MAE of 0.0787, and

a commendable R^2 of 0.9708, the value of average time per epoch is 3.68 s while other models exhibit varying levels of performance. Hence, the Sindh Crops and FAO & World Bank Datasets contain a larger number of samples, which is why the average time per epoch is higher.

Whereas Table 2 presents the importance of parameters in the proposed meta-knowledge guided framework. The parameters are listed along with their corresponding importance scores, which indicate their impact on the model's performance. This study maintained this feature's importance for all the algorithms used by the proposed framework. Here, it presents the usage of the hyperparameter's importance score. As evident from Table 1 the Sindh dataset learning_rate (L_Rate) parameter has the highest importance of (0.5726), followed by n_estimators (N_Esti) as (0.1975), Num Leaves (NL) as of (0.0246), and Other parameters, such as max_depth (MD) min_child_samples (MCS), and sub-samples, have lower importance scores, suggesting a relatively lesser impact on the model's performance. However, the Blueberry dataset's learning_rate parameter has the importance of (0.4282), n_estimators of (0.4045), NL of (0.0167), and Other parameters, such as max_depth, min_child_samples and subsample of (0.0056) have lower importance scores showing minor influence on the model's performance. The FAO and World Bank dataset, learning_rate parameter contains the importance of (0.726011), n_estimators of - (0.1644), NL of (0.0291), and Other parameters, such as max_depth, min_child_samples and a subsample of (0.0018) implying a less impact on the model's performance.

Furthermore, the second part of Table 2 shows the optimal hyperparameters and best-performing models and their corresponding optimal hyperparameter values for diverse datasets. For the Sindh Agriculture Crops dataset, the topov3-prf model achieved optimal performance for best hyperparameter optimization with a learning rate of 0.09912, N Estimators of 36, num_leaves of 412, Max depth of 31, Minimum Child Samples of 16, the subsample of 0.9000, and the optimal objective value is 0.0019. In the case of the Blueberry dataset, the Sgpr-prf model showcased superior performance with a learning rate of 0.048895, N estimators of 98, num leaves of 476, max depth of 90, min child samples of 30, and a subsample of 1.0. and the optimal objective value is 0.011635. For the FAO and World Bank dataset, the rgpe_gp model emerged as the top performer with a learning rate of 0.3000, N estimators of 101, num leaves of 209, max depth of 101, min child samples of 30, subsample of 0.7000, and optimal objective value of 0.0236. All these optimal hyperparameter values were obtained through an optimization process involving 50 trials, emphasizing their significance in maximizing model performance for each dataset. These numeric results empower researchers and practitioners in agriculture with valuable guidance for selecting appropriate models and optimizing hyperparameters, ultimately enhancing the precision of agricultural yield estimation, data analysis, and decision-making processes.

This study conducted a comprehensive performance analysis of a meta-knowledge transfer, including numerical evaluations outlined in Table 1. Additionally, we plotted the convergence and scatter plots of the RGPE algorithm for three distinct datasets employed in this study. The performance of these meta-knowledge transfer algorithms is visually depicted in Figs. 4 to 6, where the individual data points are represented as small dots. The plots' x-axis corresponds to observation iterations from 0 to 50, while the y-axis indicates the respective objective values. These visualizations provide a clear representation of the algorithms' behavior and their impact on the datasets.

To analyze the convergence behavior, we divided the observation time into intervals of ten epochs. We calculated the mean and variance of the best optimal observation within each interval. The convergence curve plots in three Figs. 4 to 6 show the decreasing trend of the objective value over time, indicating the optimization process's development over iterations. The convergence curve provides insights into the effectiveness and efficiency of the RGPE algorithm in identifying optimal hyperparameters. The shape of the curve and its convergence rate demonstrate the algorithm's ability to locate the optimal hyperparameters. Additionally, each scatter plot for RGPE displays the predicted yield production observations on the y-axis and the true yield production on the x-axis. These scatter plots effectively depict the relationship between the true and predicted yield production observations.

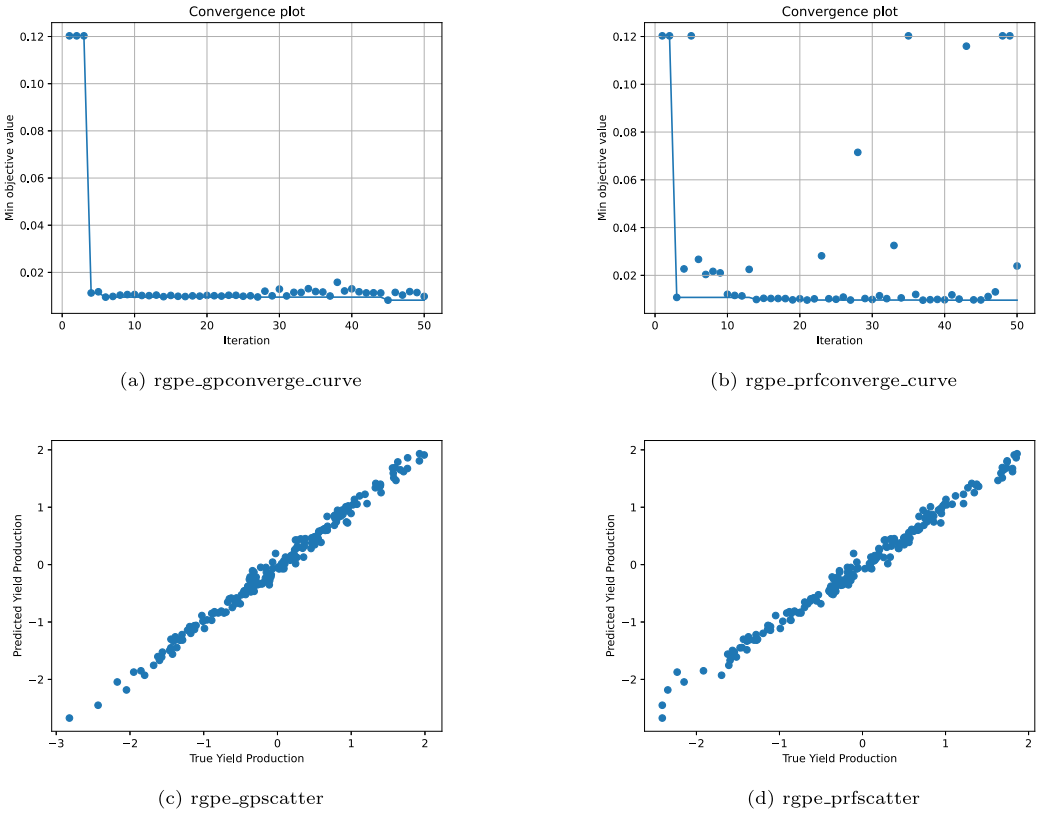
4.1. Comparative analysis

This section presents a Table 3 that discusses a comparative analysis of state-of-the-art crop yield estimation methods. However, all these SOTA methods typically use a single dataset. This model has distinct advantages compared to addressing limitations in current state-of-the-art methods, like automatic hyperparameter optimization and meta-knowledge transferability for yield estimation. This framework achieves robustness and better performance by adapting hyperparameters based on acquired meta-knowledge. Additionally, the transfer of meta-knowledge from base tasks enhances the model's ability to leverage prior insights, improving overall accuracy in target task adaption for yield estimation. The performance of the proposed framework and other recent methods, including Obsie et al. (2020) with XGBoost, and Seireg et al. (2022) utilizing Stacking with multiple models, applied to the Wild Blueberry dataset, was evaluated using multiple evaluation metrics. It is worth noting that some of the recent methods did not use normalized values. However, In this study, the values were normalized using a derived standard value obtained by dividing it by the corresponding metric value mentioned in the referenced paper. This normalization was performed by taking the standard deviation of the original Wild Blueberry, FAO, and World Bank datasets, ensuring consistency and improved comparability across different methods. Such as, study (Obsie et al., 2020) presented an un-normalized value of RMSE 343.026, and with our derived normalized standard method, this value is presented as 0.2527 (343.026/1356.9552 = 0.2527).

Obsie et al. (2020) achieved an MSE of 0.0639, an MAE of 0.1519, an RMSE of 0.2527, and an R^2 of 0.9380, indicating a good fit between the predicted and true yield values.

Seireg et al. (2022), reported an MSE of 0.01756, an RMSE of 0.1325, and an R^2 of 0.9840, suggesting a strong correlation between the predicted and true yield values in terms of MSE and RMSE. Furthermore, The approach by Borrero and Borrero-Domínguez (2023) involved a Hybrid Alternative Kalman Filter-support vector regression (AKF-SVR) and Hybrid Alternative Kalman Filter Nonlinear Autoregressive (AKF-NAR), yielding MAE values of 7.86 and 7.93, RMSE values of 13.37 and 16.73, and R^2 values of 0.973 and 0.958, respectively. This approach demonstrated reasonable performance and correlation between the predicted and true yield values. However, it is important to note that this approach still has the above-mentioned limitations. In comparison, the proposed Meta-Knowledge-Guided framework, utilizing the RGPE approach, demonstrated exceptional performance on the Wild Blueberry dataset, achieving an MSE of 0.0142, MAE of 0.0853, RMSE of 0.119, and an R2 value of 0.9865, demonstrating its superiority over the other methods.

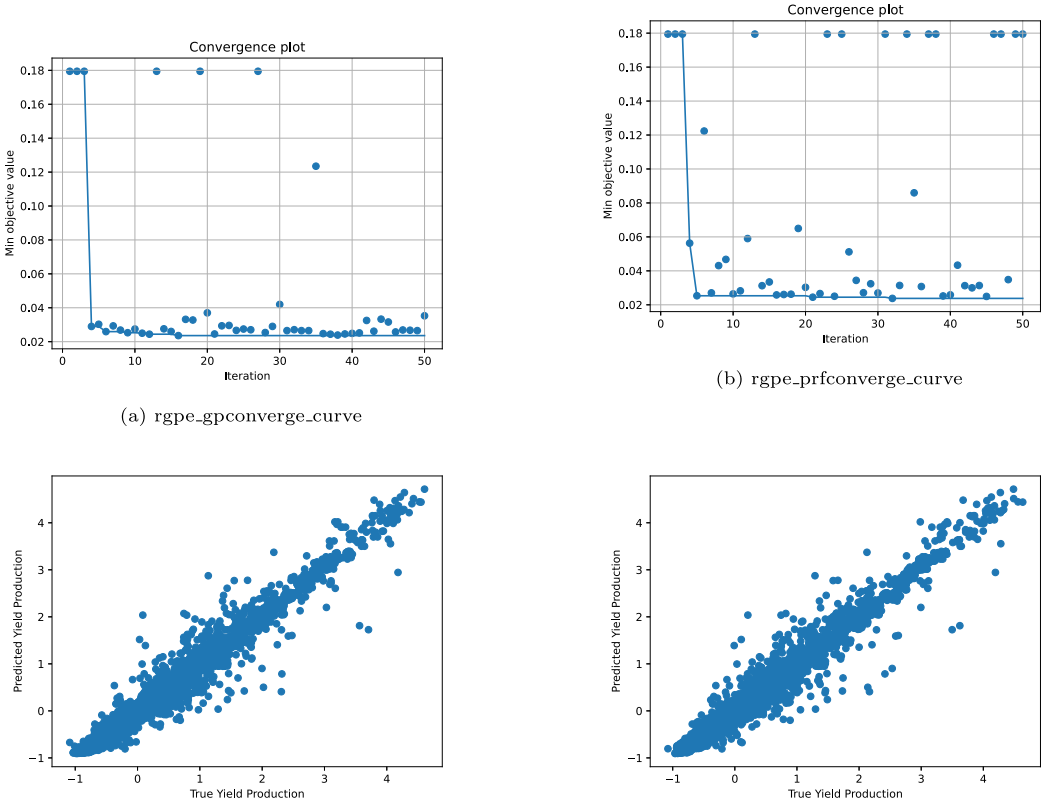
For the FAO and World Bank dataset, Aworka et al. (2022) implemented. Crop Random Forest (CRF), resulting in an MSE of 0.117649, an RMSE of 0.1325, and an R^2 of 0.9272. However, a specific MAE value was not reported. Mariadass et al. (2022) employed XGBoost with data on crop yield and pesticides obtained from the FAO and data on rainfall and temperature collected from World Bank Open Data. Although the exact MSE and RMSE values were not disclosed, they achieved an R^2 of 0.98, indicating a reasonably accurate fit to the data. Cedric et al. (2022) utilized crop k-Nearest Neighbor (Ck-NN), Crop Decision Tree (CDT), and Crop Multivariate Logistic Regression (CMRL) with the FAO and World Bank datasets. They reported MAE values of 0.0160, 0.088, and 0.0315 and R^2 values of 0.9503, 0.9465, and 0.8380, respectively. However, specific MSE and RMSE values were not provided for their approach. Ahmed (2023) employed Multi-Layer



(a) rgpe_gpconverge_curve (b) rgpe_prfconverge_curve

(c) rgpe_gpscatter (d) rgpe_prfscatter

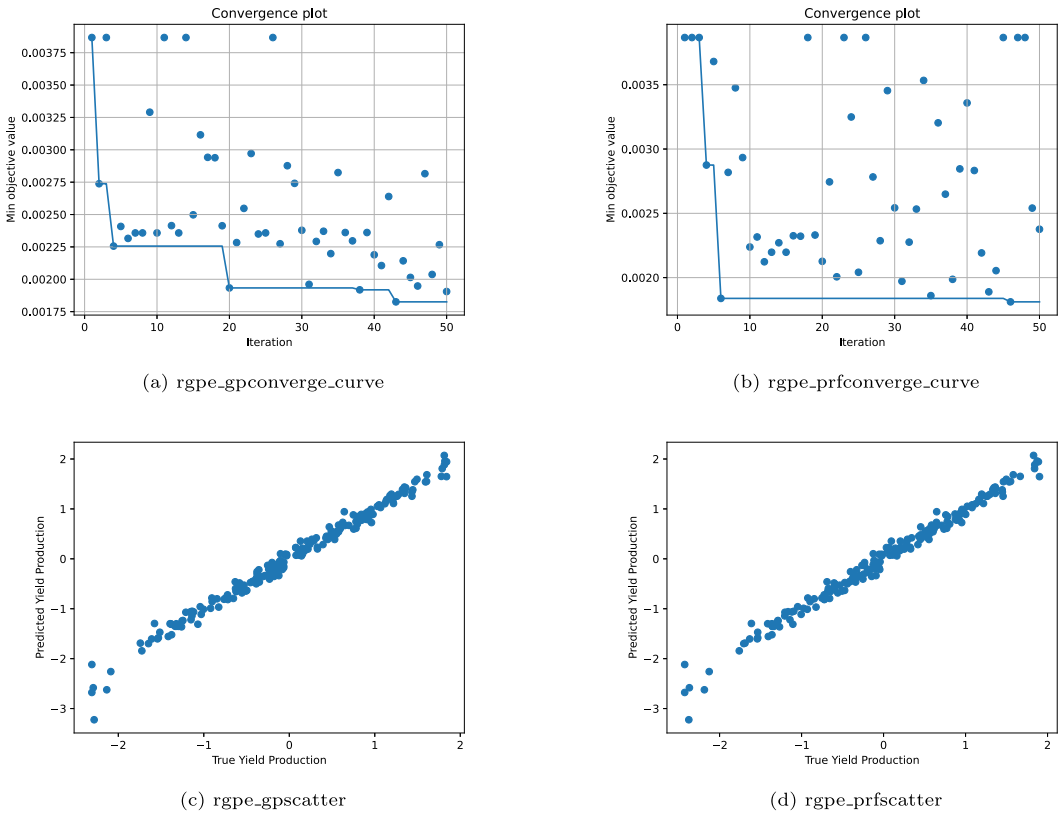
Fig. 4. Performance Analysis: Convergence and Scatter Plots of Meta-Knowledge Transfer on Wild Blueberry Dataset.



(a) rgpe_gpconverge_curve (b) rgpe_prfconverge_curve

(c) rgpe_gpscatter (d) rgpe_prfscatter

Fig. 5. Performance Analysis: Convergence and Scatter Plots of Meta-Knowledge Transfer on FAO & World Bank Dataset.



(a) rgpe_gpconverge_curve (b) rgpe_prfconverge_curve

(c) rgpe_gpscatter (d) rgpe_prfscatter

Fig. 6. Performance Analysis: Convergence and Scatter Plots of Meta-Knowledge Transfer on Sindh Crops Dataset.

Perceptron-Spider Monkey Optimization (MLP-SMO) with the FAO and World Bank datasets. They achieved an RMSE of 0.13 and R^2 value of 0.96.

In comparison to the OSTA methods, the proposed framework, applying the RGPE method to the FAO and World Bank dataset, achieved a competitive performance with MSE of 0.0294, MAE of 0.0787, RMSE of 0.1714, and R^2 of 0.9708, indicating accurate predictions and a strong relationship between the predicted and true yield values. It is important to note that this study normalized the values in the proposed work using a specific standard value for the FAO dataset, ensuring consistency and comparability with other methods.

Lastly, for the Sindh crops, which is a private dataset, our proposed Meta-Knowledge Guided framework, by using the RGPE method, yielded an impressive MSE of 0.0020, an MAE of 0.0157, an RMSE of 0.0447, and an R^2 of 0.8437, demonstrating accurate predictions and a satisfactory fit. The comprehensive analysis of the comparative results highlights the effectiveness of our proposed framework, showcasing superior performance in terms of MSE, MAE, RMSE, and R^2 across different datasets. These findings contribute to advancing crop yield estimation methods, emphasizing the potential and significance of the RGPE transferable meta-model in accurately predicting and estimating crop yields.

5. Conclusion

This study introduces a novel meta-knowledge-guided framework for precise crop yield estimation, addressing hyperparameter optimization, meta-knowledge transferability, robustness, and model generalizability across diverse agricultural datasets. The framework utilizes LightGBM for base tasks and integrates Bayesian Optimization to extract knowledge insights for the meta-model. This integration enables dynamic hyperparameter optimization, ensuring optimal performance and effective transfer of meta-knowledge for the target tasks using RGPE, SGPR, and TOPOV3 algorithms. The extensive experimental results underpin the effectiveness of the framework with the rgpe_prf algorithm consistently delivering strong performance, as evidenced by a maximum R^2 of (0.9865, 0.9708, and 0.8437) and exhibit excellent accuracy as indicated by low MSE, MAE, and RMSE values in comparison with SOTA. However, sgpr_gp and sgpr_prf algorithms potentially lower R^2 and higher MSE, MAE, and RMSE values than RGPE and TOPOV3. Our findings contribute to advancing the field of crop yield estimation and provide a novel, practical, and robust solution. This study could extend the meta-knowledge transferable approach to incorporate additional private diverse datasets to enhance and evaluate its performance across different crop types and geographic regions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors extend their gratitude to the funding organizations.

This work was partially supported by the National Natural Science Foundation of China, the National High Technology Research and Development Program of China, the New Century Excellent Talent Support Project of the Chinese Ministry of Education, China, the project of Science and Technology Department of Sichuan Province (Grant No. 2021YFG0322). The project of the Science and Technology Department of Chongqing Municipality, China (Grant No. CSTC2022JXJL0214). The Science and Technology Research Program of Chongqing Municipal Education Commission, China (Grant No. KJZD-K202114401), Chengdu Chengdian Network Technology Co., Ltd., China, Chongqing Qinchengxing Technology Co., Ltd., Chengdu Haitian Digital Technology Co., Ltd., China, Chengdu Civil-military Integration Project Management Co., Ltd., China, and Sichuan Yin Ten Gu Technology Co., Ltd., China

References

Abbate, S., Centobelli, P., Cerchione, R., 2023. The digital and sustainable transition of the agri-food sector. Technol. Forecast. Soc. Change 187, 122222.

Abdar, M., Samami, M., Mahmoodabad, S.D., Doan, T., Mazoure, B., Hashemifesharaki, R., Liu, L., Khosravi, A., Acharya, U.R., Makarenkov, V., et al., 2021. Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning. Comput. Biol. Med. 135, 104418.

Aguiar, G.J., Mantovani, R.G., Mastelini, S.M., de Carvalho, A.C., Campos, G.F., Junior, S.B., 2019. A meta-learning approach for selecting image segmentation algorithm. Pattern Recognit. Lett. 128, 480–487.

Ahmed, S., 2023. A software framework for predicting the maize yield using modified multi-layer perceptron. Sustainability 15 (4), 3017.

Al-Rawashdeh, M., Al Nawaiseh, M., Yousef, I., Bisharah, M., Alkhadrawi, S., Al-Bdour, H., 2023. Predicting building damage grade by earthquake: a Bayesian optimization-based comparative study of machine learning algorithms. Asian J. Civ. Eng. 1–12.

Ansarifar, J., Akhavizadegan, F., Wang, L., 2020. Performance prediction of crosses in plant breeding through genotype by environment interactions. Sci. Rep. 10 (1), 11533.

Aslan, M.F., Sabanci, K., Durdu, A., Unlersen, M.F., 2022. COVID-19 diagnosis using state-of-the-art CNN architecture features and Bayesian optimization. Comput. Biol. Med. 142, 105244.

Aworka, R., Cedric, L.S., Adoni, W.Y.H., Zoueu, J.T., Mutombo, F.K., Kimpolo, C.L.M., Nahhal, T., Krichen, M., 2022. Agricultural decision system based on advanced machine learning models for yield prediction: Case of East African countries. Smart Agric. Technol. 2, 100048.

Borrero, J.D., Borrero-Domínguez, J.D., 2023. Enhancing short-term berry yield prediction for small growers using a novel hybrid machine learning model. Horticulturae 9 (5), 549.

Bronskill, J., Gordon, J., Requeima, J., Nowozin, S., Turner, R., 2020. Tasknorm: Rethinking batch normalization for meta-learning. In: International Conference on Machine Learning. PMLR, pp. 1153–1164.

Cedric, L.S., Adoni, W.Y.H., Aworka, R., Zoueu, J.T., Mutombo, F.K., Krichen, M., Kimpolo, C.L.M., 2022. Crops yield prediction based on machine learning models: Case of West African countries. Smart Agric. Technol. 2, 100049.

Charoen-Ung, P., Mittrapiyanuruk, P., 2019. Sugarcane yield grade prediction using random forest with forward feature selection and hyper-parameter tuning. In: Recent Advances in Information and Communication Technology 2018: Proceedings of the 14th International Conference on Computing and Information Technology. IC2IT 2018, Springer, pp. 33–42.

Chiu, M.C., Yan, W.M., Bhat, S.A., Huang, N.F., 2022. Development of smart aquaculture farm management system using IoT and AI-based surrogate models. J. Agric. Food Res. 9, 100357.

Crane-Droesch, A., 2018. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environ. Res. Lett. 13 (11), 114003.

Cunha, T., Soares, C., de Carvalho, A.C., 2018. Metalearning and recommender systems: A literature review and empirical study on the algorithm selection problem for Collaborative Filtering. Inform. Sci. 423, 128–144.

De Clercq, M., Vats, A., Biel, A., 2018. Agriculture 4.0: The future of farming technology. In: Proceedings of the world government summit. Dubai, UAE, pp. 11–13.

Diker, A., 2022. An efficient model of residual based convolutional neural network with Bayesian optimization for the classification of malarial cell images. Comput. Biol. Med. 148, 105635.

Dou, Z.Y., Yu, K., Anastasopoulos, A., 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. arXiv preprint arXiv:1908.10423.

Drummond, S.T., Sudduth, K.A., Joshi, A., Birrell, S.J., Kitchen, N.R., 2003. Statistical and neural methods for site–specific yield prediction. Trans. ASAE 46 (1), 5.

Elnahal, A.S., El-Saadony, M.T., Saad, A.M., Desoky, E.-S.M., El-Tahan, A.M., Rady, M.M., AbuQamar, S.F., El-Tarabily, K.A., 2022. The use of microbial inoculants for biological control, plant growth promotion, and sustainable agriculture: A review. Eur. J. Plant Pathol. 162 (4), 759–792.

Ghaderi Zefrehi, H., Sheikhi, G., Altınçay, H., 2023. Threshold prediction for detecting rare positive samples using a meta-learner. Pattern Anal. Appl. 26 (1), 289–306.

Gopal, P.M., Bhargavi, R., 2019. A novel approach for efficient crop yield prediction. Comput. Electron. Agric. 165, 104968.

Gupta, U., Sharma, R., 2023. Analysis of criminal spatial events in india using exploratory data analysis and regression. Comput. Electr. Eng. 109, 108761.

Haq, Z.U., Ullah, H., Khan, M.N.A., Naqvi, S.R., Ahad, A., Amin, N.A.S., 2022. Comparative study of machine learning methods integrated with genetic algorithm and particle swarm optimization for bio-char yield prediction. Bioresour. Technol. 363, 128008.

He, X., Zhao, K., Chu, X., 2021. Automl: A survey of the state-of-the-art. Knowl.-Based Syst. 212, 106622.

Hegedus, P.B., Maxwell, B.D., Mieno, T., 2023. Assessing performance of empirical models for forecasting crop responses to variable fertilizer rates using on-farm precision experimentation. Precis. Agric. 24 (2), 677–704.

Hutter, F., Kotthoff, L., Vanschoren, J., 2019. Automated Machine Learning: Methods, Systems, Challenges. Springer Nature.

Jeon, M., Noh, Y., Jeon, K., Lee, S., Lee, I., 2021. Data gap analysis of ship and maritime data using meta learning. Appl. Soft Comput. 101, 107048.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.M., Gerber, J.S., Reddy, V.R., et al., 2016. Random forests for global and regional crop yield predictions. PLoS One 11 (6), e0156571.

Jhajharia, K., Mathur, P., Jain, S., Nijhawan, S., 2023. Crop yield prediction using machine learning and deep learning techniques. Procedia Comput. Sci. 218, 406–417.

Joseph, L.P., Joseph, E.A., Prasad, R., 2022. Explainable diabetes classification using hybrid Bayesian-optimized TabNet architecture. Comput. Biol. Med. 151, 106178.

Kaul, M., Hill, R.L., Walthall, C., 2005. Artificial neural networks for corn and soybean yield prediction. Agric. Syst. 85 (1), 1–18.

Khan, M., Ullah, Z., Mašek, O., Naqvi, S.R., Khan, M.N.A., 2022. Artificial neural networks for the prediction of biochar yield: a comparative study of metaheuristic algorithms. Bioresour. Technol. 355, 127215.

Korneva, E., Blockeel, H., 2020. Towards better evaluation of multi-target regression models. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 353–362.

Li, J., Li, G., Wang, L., Li, D., Li, H., Gao, C., Zhuang, M., Zhuang, J., Zhou, H., Xu, S., et al., 2023. Predicting maize yield in Northeast China by a hybrid approach combining biophysical modelling and machine learning. Field Crops Res. 302, 109102.

Liu, Z., Liu, L., Xie, Y., Jin, Z., Jia, X., 2023. Task-adaptive meta-learning framework for advancing spatial generalizability. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 12. pp. 14365–14373.

Loey, M., El-Sappagh, S., Mirjalili, S., 2022. Bayesian-based optimized deep learning model to detect COVID-19 patients using chest X-ray image data. Comput. Biol. Med. 142, 105213.

Mariadass, D.A., Moung, E.G., Sufian, M.M., Farzamnia, A., 2022. Extreme gradient boosting (XGBoost) regressor and Shapley additive explanation for crop yield prediction in agriculture. In: 2022 12th International Conference on Computer and Knowledge Engineering. ICCKE, IEEE, pp. 219–224.

Marko, O., Brdar, S., Panic, M., Lugonja, P., Crnojevic, V., 2016. Soybean varieties portfolio optimisation based on yield prediction. Comput. Electron. Agric. 127, 467–474.

Meola, A., Winkler, M., Weinrich, S., 2023. Metaheuristic optimization of data preparation and machine learning hyperparameters for prediction of dynamic methane production. Bioresour. Technol. 372, 128604.

Obsie, E.Y., Qu, H., Drummond, F., 2020. Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. Comput. Electron. Agric. 178, 105778.

Oikonomidis, A., Catal, C., Kassahun, A., 2022. Hybrid deep learning-based models for crop yield prediction. Appl. Artif. Intell. 36 (1), 2031822.

Padmapriya, J., Sasilatha, T., 2023. Deep learning based multi-labelled soil classification and empirical estimation toward sustainable agriculture. Eng. Appl. Artif. Intell. 119, 105690.

Palacios, F., Diago, M.P., Melo-Pinto, P., Tardaguila, J., 2023. Early yield prediction in different grapevine varieties using computer vision and machine learning. Precis. Agric. 24 (2), 407–435.

Qu, H., Drummond, F., 2018. Simulation-based modeling of wild blueberry pollination. Comput. Electron. Agric. 144, 94–101.

Reynolds, C., Fletcher, R.J., Carneiro, C.M., Jennings, N., Ke, A., LaScaleia, M.C., Lukhele, M.B., Mamba, M.L., Sibiya, M.D., Austin, J.D., et al., 2018. Inconsistent effects of landscape heterogeneity and land-use on animal diversity in an agricultural mosaic: a multi-scale and multi-taxon investigation. Landsc. Ecol. 33, 241–255.

Romero, J.R., Roncallo, P.F., Akkiraju, P.C., Ponzoni, I., Echenique, V.C., Carballido, J.A., 2013. Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires. Comput. Electron. Agric. 96, 173–179.

Russello, H., 2018. Convolutional neural networks for crop yield prediction using satellite images. In: IBM Center for Advanced Studies.

Seireg, H.R., Omar, Y.M., Abd El-Samie, F.E., El-Fishawy, A.S., Elmahalawy, A., 2022. Ensemble machine learning techniques using computer simulation data for wild blueberry yield prediction. IEEE Access 10, 64671–64687.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, M., Adams, R., 2015. Scalable bayesian optimization using deep neural networks. In: International Conference on Machine Learning. PMLR, pp. 2171–2180.

Van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agric. 177, 105709.

Wu, J., Chen, X.Y., Zhang, H., Xiong, L.D., Lei, H., Deng, S.H., 2019. Hyperparameter optimization for machine learning models based on Bayesian optimization. J. Electron. Sci. Technol. 17 (1), 26–40.

Yang, F., Zhang, D., Zhang, Y., Zhang, Y., Han, Y., Zhang, Q., Zhang, Q., Zhang, C., Liu, Z., Wang, K., 2023. Prediction of corn variety yield with attribute-missing data via graph neural network. Comput. Electron. Agric. 211, 108046.

You, J., Li, X., Low, M., Lobell, D., Ermon, S., 2017. Deep gaussian process for crop yield prediction based on remote sensing data. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, No. 1.

Yuen, S.Y., Chow, C.K., Zhang, X., Lou, Y., 2016. Which algorithm should I choose: An evolutionary algorithm portfolio approach. Appl. Soft Comput. 40, 654–673.

Zhang, X., Dai, C., Li, W., Chen, Y., 2023a. Prediction of compressive strength of recycled aggregate concrete using machine learning and Bayesian optimization methods. Front. Earth Sci. 11, 1112105.

Zhang, Q., Zhao, X., Han, Y., Yang, F., Pan, S., Liu, Z., Wang, K., Zhao, C., 2023b. Maize yield prediction using federated random forest. Comput. Electron. Agric. 210, 107930.