Abstract

Enhancing crop yields remains a critical challenge confronting both farmers and governmental bodies. This study delves into the application and execution of Gradient Boosting , Random Forest, SVM and Dense layers with Adam optimiser algorithm. to forecast crop yields across numerous States in the USA. The selected tool for this investigation is scikit learn library. Agricultural data utilised in this research originated from the various historic data of 14 states of the USA of lint yield from 1964 to 2017. .

The input variables consisted mainly of the soil contents and the area which we harvested such as Nitrogen(%),Potash(%),Phosphorous(%). Through meticulous analysis, the best result was achieved using a Gradient boosting, leading to an MSE (Root Mean Square Error) value of 13069.8196 and an MAE (Mean Absolute Error) of 87.6631 on original variables.

The primary objective of this project is to conduct a comparative analysis of different Machine Learning techniques with the aim of enhancing yield prediction accuracy. By doing so, farmers gain invaluable insights, enabling them to optimise their cultivation practices. Moreover, the integration of predictive analytics empowers farmers to harness the potential benefits of technology, offering them a proactive approach towards crop management. This research serves as a stepping stone toward providing the agricultural sector with advanced tools and methodologies, fostering a more efficient and productive approach to crop cultivation. Ultimately, the findings of this study contribute to the broader goal of sustainable agriculture by equipping farmers with the knowledge and tools needed to make informed decisions and improve overall crop yield outcomes.

Contents

List of Figures

Chapter 1

Introduction

1.1 The Area of Work

The area of work for this project lies at the intersection of agricultural science, machine learning, and data analytics. The primary focus is on leveraging advanced predictive modelling techniques to forecast crop yields based on soil nutrient parameters. The study addresses the critical issue of optimising crop yield predictions through the analysis of soil nutrient parameters. This interdisciplinary approach holds the potential to revolutionise agricultural practices, offering more precise and data-driven strategies for crop management.

1.2 Problem Addressed

The agricultural landscape is fraught with challenges, and one of the critical issues faced by farmers and stakeholders is the difficulty in accurately predicting crop yields. Traditional methods often fall short in providing precise insights into the dynamic interactions between soil nutrient parameters and crop productivity, resulting in sub optimal resource management and decision-making. Conventional methods lack the precision required to predict crop yields with a high degree of accuracy, leading to uncertainty in planning and resource allocation.The inability to precisely anticipate crop outcomes based on soil nutrient levels contributes to inefficient use of resources, such as fertilisers and water, impacting both economic viability and environmental sustainability.

In summary, this study addresses the pressing challenges in agricultural practices related to uncertain crop yield predictions, inefficient resource utilisation, and the slow integration of technology. Through the application of advanced machine learning models, the project seeks to provide tangible solutions that empower farmers and enhance the overall sustainability and productivity of agriculture.

1.3 Motivation

The motivation behind undertaking this comprehensive study on crop yield prediction, leveraging machine learning models and soil nutrient parameters, stems from a deep-seated commitment to addressing critical challenges within the agricultural sector.

1. Enhancing Food Security:

The global demand for food is escalating, and ensuring food security is paramount. By improving the accuracy of crop yield predictions, this study contributes to the broader goal of sustaining a growing population.

2. Optimising Resource Utilisation:

Inefficient resource allocation in agriculture poses economic and environmental challenges. The study is motivated by the potential to optimise resource utilisation through precise predictions, reducing waste and promoting sustainability.

3. Empowering Farmers:

Farmers, as the backbone of agriculture, stand to benefit significantly from advanced technologies. The motivation is to empower farmers with tools and insights that enhance their decision-making capabilities, ultimately leading to increased productivity and economic resilience. [1]

4. Meeting Agricultural Challenges:

Agriculture faces multifaceted challenges, from climate change to evolving consumer demands. The study is motivated by a desire to equip the agricultural sector with modern solutions that address these challenges and foster adaptability.

5. Technological Integration:

The slow integration of technology in agriculture has been a persistent issue. The motivation is to bridge this gap by demonstrating the practical applications of machine learning in agriculture, showcasing the transformative potential of cutting-edge technologies.

1.4 Existing System

Before the advent of machine learning, traditional methods for crop yield prediction relied on simpler statistical and empirical approaches. Some of the common traditional methods include:

1. **Historical Data Analysis**: Farmers and agricultural experts often rely on historical yield data for specific crops in a given region. By analysing trends and patterns over several years, they make predictions based on past performance.

2. **Expert Knowledge and Farmer Experience**: Local farmers and agricultural experts often possess extensive knowledge about the soil, climate, and crop characteristics in their region. They use this experiential knowledge to make predictions about crop yields.

3. **Soil Testing and Analysis**: Soil testing and analysis are crucial in traditional methods. Farmers assess soil fertility, nutrient levels, and other relevant factors to predict how well a particular crop will perform in a given area.

4. **Local Indicators and Phenology**: Traditional farmers often rely on local indicators and phenology (plant and animal life cycle events) to predict crop yields. For example, certain flowering or migration events may be used as cues for optimal planting times.

5. **Government Crop Forecasting Agencies**: In many countries, government agencies responsible for agriculture conduct surveys, gather data, and release crop forecasts based on traditional statistical methods. These forecasts help farmers and policymakers plan for the agricultural season.

6. **Local Agricultural Extension Services**: Agricultural extension services, provided by government agencies, offer advice and information to farmers based on traditional knowledge. These services disseminate information about best practices, pest control measures, and planting schedules.

While these traditional methods lack the sophistication and data-driven precision of AI and machine learning, they have been effective in many agricultural communities for generations. Combining traditional knowledge with modern technologies can often provide a comprehensive approach to crop yield prediction.

Chapter 2

Literature Review

A significant amount of research papers is available in the field of using of machine learning algorithms in predictions of crop yield.

The field of using machine learning algorithms to crop production prediction has produced a sizable number of research articles. A variety of developmental stage models and yield prediction models were incorporated into an open crop model that Su, Xu, and Yan [2] constructed using support vector machines (SVM). The model allowed for large-scale data integration since it had scale-independent factors and an open input mechanism. Finding hyperparameters, penalty coefficients, and ideal kernel functions for examining three different kinds of rice crops were the main goals.

Four supervised learning algorithms were investigated in a study by Saad and Rusli [3] in order to forecast rice productivity in Kedah, Malaysia, based on weeds, diseases, and pests. Comparing the Conjugate Gradient Descent algorithm to the Levenberg-Marquardt, Quick, and Back-propagation algorithms, the latter three performed worse.

Artificial neural network (ANN) models for forecasting rice yield in China's mountainous Fujian province were developed and assessed by Ji, Sun, and Yang [4]. After a comparison with linear regression models, it was found that neural network models performed more accurately than linear regression models but needed a large amount of data.

Ramesh, Vardhan [5] described crop density-based clustering and multiple linear regression techniques for predicting crop yield in the East Godavari district of Andhra Pradesh, India.

Support vector machines (SVM) were used by Gandhi, Petkar, and Armstrong [6] to forecast rice crop yield. They discovered that Multilayer perceptron, Bayesian, and Naïve Bayes-based networks consistently gave predictions that were precise, sensitive, and specific in comparison to other classifiers.

Using linear regression, Sellam and Poovammal [7] investigated the relationship between crop yield and Area Under Cultivation, Annual Precipitation, and the Food Price Index.

In order to predict wheat yields, Chemchem, Alin, and Michael [8] created a classifier model that used random forest. They pre-treated the model with SMOTE in order to improve its accuracy.

Machine learning models achieved 20% better accuracy in yield prediction than traditional non-machine learning techniques, according to Charoen and Pradit's [9] comparison of non-machine learning and machine learning predictive models. This difference was statistically significant.

Chapter 3

Proposed Work

Dataset Acquisition and Preprocessing

The most responsible factors for the variation in crop production are uncertainties in the weather conditions, depletion of the nutrition level of soils, fertilizer availability, and cost, pest control, and other factors. It is very difficult to collect old data on soil nutrient levels, pest outbreaks, and various plant physiological characteristics, as there is no regular collection system and managing this information season-wise. Therefore, we restrict to the main variables as nutrients which can be the fertiliser used because Fertilisers are used in the pre-sowing stage of crops or very early stage of cultivation as predictors for crop yield.

For Indian agriculture, there is not an ample amount of data on various government sites such as data.gov.on, directorate of economics and statistics, soil and land use survey of India, and various state's soil and agriculture departments sites.

We are here using the data set of cotton fields of 14 states of the USA from 1964-2017.

We combine the various nutrients, fertiliser data, and cotton crop yield over that period into a single data set.

References to Dataset (column wise):

1. State - Name of the 14 states in USA
2. Year - Data from 1964-2017 (Historical occurrences and records)
3. *Nutrient* - nutrient in soil
4. *Name of Fertilizer* (Pounds/Acre) - Amount of the fertiliser being given to that acreage
5. Area Planted - As mentioned
6. Harvested Area - As mentioned

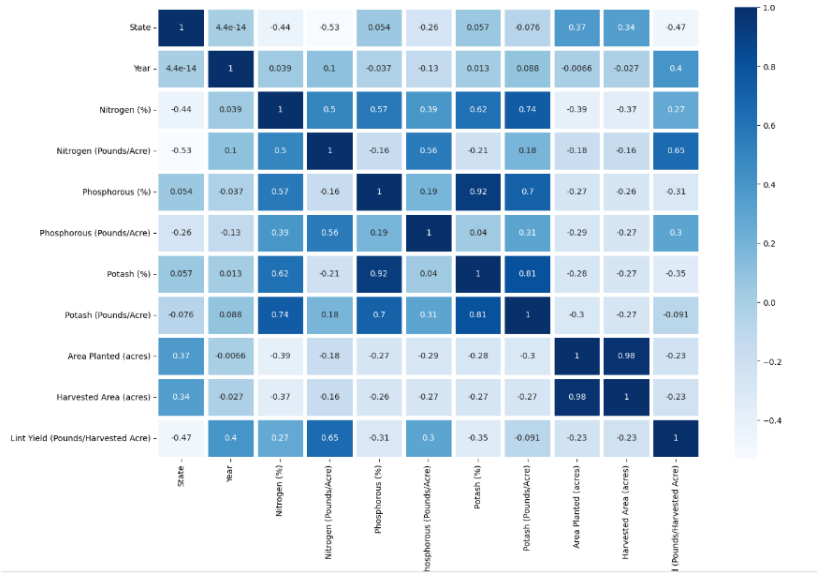

FIGURE 3.1: Correlation Heat map

7. Lint Yield - Yield of the cotton crop

As part of pre-processing, the missing values were replaced with the mean values but we replaced not with mean but with the median as their are outliers which will hindered the accuracy if we replace it with mean. Further to increase accuracy, the values were later replaced with the mean values of that feature, corresponding to the state; since states were showing varied values that weren't in correlation with each other.Here

Unique values of States have been noted and were mapped to corresponding integral values, to get a dataset that was capable to be trained on the regression model used.In Adam Optimiser we dropped year and state column as it is not that much important for prediction.

Pearson's correlation used to check for redundant features. Fortunately, all the features show a value greater than —0.2— which shows a medium-high relation between the features and the target variable (Lint Yield).Figure(3.1)

Training and Testing Models

The transformed data is then split into two sets namely, training sets and testing sets before applying the machine learning classifiers. This Split is majorly around 80 percent for training and 20 percent for Testing.

3.0.1 Random Forest

Random Forest is a powerful algorithm for crop yield prediction, offering robustness, feature importance analysis, and the ability to handle complex datasets. Its implementation involves

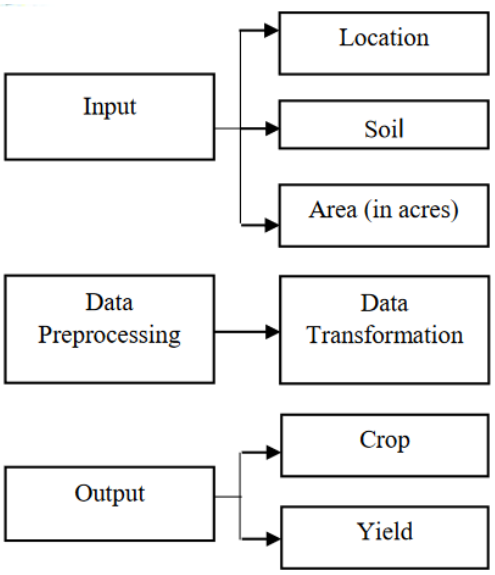

FIGURE 3.3: Architecture

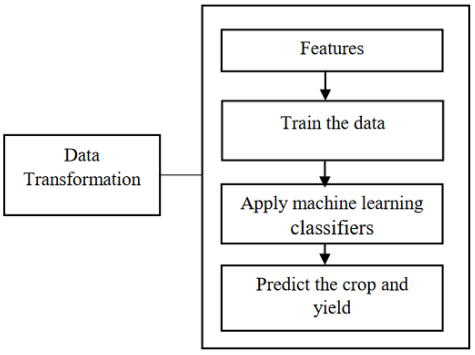

FIGURE 3.4: Model Design

training multiple decision trees on bootstrapped samples with random feature selection, leading to an ensemble model that provides accurate and stable predictions.

Here's an elaboration on the implementation of Random Forest for crop yield prediction:

1. Ensemble Learning: - Random Forest is an ensemble learning method, meaning it builds multiple decision trees during training and merges them together to get a more accurate and stable prediction.

2. Decision Trees: - The basic building block of a Random Forest is a decision tree. Decision trees are constructed by recursively splitting the data based on features, aiming to create subsets with similar outcomes.

3. Random Feature Selection:- In each decision tree of the Random Forest, only a random subset of features is considered at each split. This introduces diversity among the trees and helps prevent overfitting to specific features.

4. Bootstrapped Sampling: - During the training phase, each tree is trained on a bootstrap sample of the original dataset. This involves randomly selecting samples with replacement. This process ensures diversity among the trees.

5. Aggregation of Predictions:- Once the individual trees are trained, the predictions from each tree are aggregated to make the final prediction. For regression tasks like yield prediction, the average prediction from all trees is often used.

Evaluate the performance of the Random Forest model using metrics such as Mean Squared Error (MSE) or Mean Absolute Error (MAE) for regression tasks.

3.0.2 Gradient Boosting

Gradient Boosting is another powerful machine learning algorithm commonly used in crop yield prediction. It belongs to the ensemble learning family and is known for its ability to build robust predictive models.It leverages the strengths of ensemble learning to create accurate and robust models. Its implementation involves training weak learners sequentially, with a focus on correcting errors made by the ensemble. Here's an elaboration on the implementation of Gradient Boosting for crop yield prediction:

1. Sequential Training:- Gradient Boosting builds an ensemble of weak learners (typically decision trees) sequentially. Each new learner corrects the errors made by the previous ones.

2. Gradient Descent: - The algorithm minimizes the errors by optimizing a loss function. It uses gradient descent to find the direction in which the loss function decreases the fastest and updates the model accordingly.

3. Weak Learners:- Decision trees are commonly used as weak learners in Gradient Boosting. These are often shallow trees, referred to as "stumps," to avoid overfitting.

4. Boosting Process: - During the boosting process, each new tree focuses on the mistakes made by the previous ones. The final prediction is a weighted sum of the predictions from all trees.

5. Training the Gradient Boosting Model:- Sequentially train a series of weak learners (trees) on the training data. Each tree corrects the errors made by the ensemble so far.

6. Hyperparameter Tuning through cross validation:- Adjust hyperparameters such as the learning rate, tree depth, and regularisation parameters to optimise the model's performance. This often involves using cross-validation techniques.We have 5 folds of cross validation in this

Evaluate the model's performance using appropriate metrics for regression tasks, such as Mean Squared Error (MSE) or Mean Absolute Error (MAE).

3.0.3 Support Vector Machine

Support Vector Machines (SVM) is a supervised machine learning algorithm that can be employed for crop yield prediction.In the context of crop yield prediction, SVM can be used for regression tasks where the goal is to predict a continuous variable like crop yield.it is leveraging the strengths of the kernel trick to handle both linear and non-linear relationships. Its implementation involves choosing a suitable kernel function, training the model, and evaluating performance using regression metrics.

Here's an elaboration on the implementation of SVM for crop yield prediction:

1. Kernel Trick:- SVM uses a "kernel trick" to transform the input data into a higher-dimensional space, making it easier to find a hyperplane that separates different classes. Common kernel functions include linear, polynomial, and radial basis function (RBF).

2. Support Vectors:- Support Vectors are the data points that lie closest to the decision boundary (hyperplane). These vectors are crucial in determining the optimal hyperplane for separation.

3. Hyperplane:- In a regression task like crop yield prediction, the hyperplane represents the best fit to the data, with the goal of minimizing the error between predicted and actual yield values.

Evaluate the model's performance using appropriate metrics for regression tasks, such as Mean Squared Error (MSE) or Mean Absolute Error (MAE).

3.0.4 Adam Optimiser Using Dense Layer Neural Network Model

The combination of dense layers in a neural network with the Adam optimizer provides a powerful approach for crop yield prediction. This implementation involves defining the architecture, compiling the model, training, tuning hyperparameters, evaluating, and making predictions. Below is an elaboration on the implementation of this approach:

Neural Network Architecture: - Choose a neural network architecture suitable for regression tasks. A simple architecture may consist of one or more dense layers, depending on the complexity of the problem.

5. Model Compilation:- Compile the neural network model, specifying the loss function (e.g., mean squared error for regression), the Adam optimizer, and evaluation metrics. The Adam optimizer can be instantiated with specific hyperparameters such as learning rate.

6:-Splitting Data:- Divide the dataset into training and testing sets. The training set is used to train the neural network, and the testing set is used for evaluation.

7. Training:- Train the neural network using the training dataset. During training, the Adam optimizer will adjust the weights of the dense layers to minimize the chosen loss function.

8. Hyperparameter Tuning:-Fine-tune hyperparameters such as the learning rate, the number of hidden units in the dense layers, and the number of epochs based on the model's performance on the validation set.

9. Evaluation:- Evaluate the model's performance using relevant metrics for regression tasks, such as Mean Squared Error (MSE) or Mean Absolute Error (MAE).

Neural networks with dense layers can capture complex non-linear relationships in the data, which can be beneficial for predicting crop yields influenced by multiple factors.

The Adam optimizer adapts the learning rate for each parameter, allowing for more efficient convergence and avoiding the need for manual tuning.

Neural networks are flexible and can be adapted to handle various data types and complexities in crop yield prediction.

Neural networks can learn hierarchical representations directly from raw data, making them capable of end-to-end learning without extensive feature engineering.

Chapter 4

Simulation and Results

We simulate 4 different algorithms and models for this comparitive analysis.

1- Random Forest

2-Gradient Boosting

3-Scalar Vector Machine (Non-Linear).

4- Dense Layer using Adam optimiser Model

Main metrics which we considered for this is Mean Absolute Error(MAE) and MSE of all the models.

Here after analysing the results we found out that SVM is having highest Mean Absolute Error so we are not going forward with this. Most efficient model was of Gradient boosting algorithm based which have MAE value around 87.663. We here attached the original v/s predicted value of crop yeild graphs of Random Forest and Gradient Boosting

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|$$

FIGURE 4.1

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (S_i - O_i)^2}$$

FIGURE 4.2

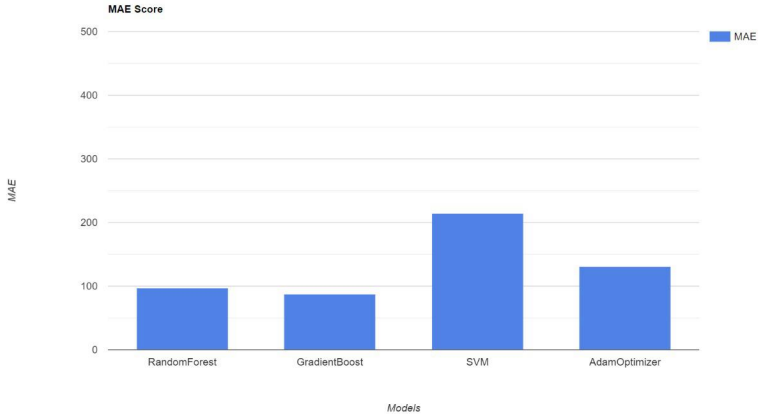

FIGURE 4.3: Comparitiv Analysis

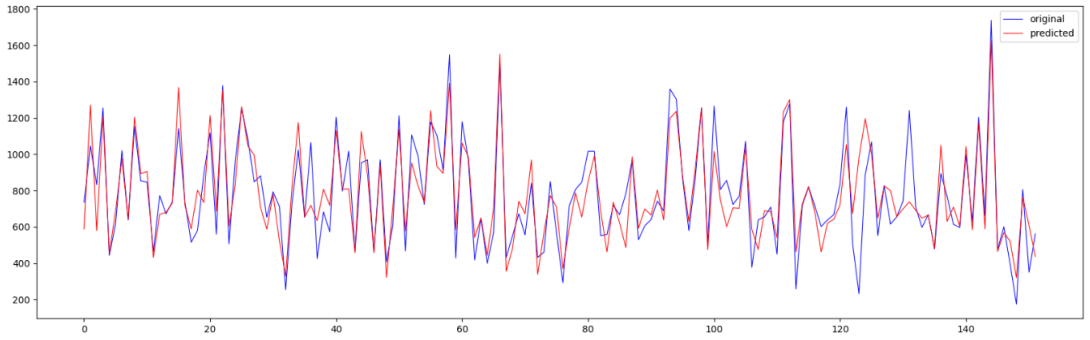

FIGURE 4.4: Random Forest

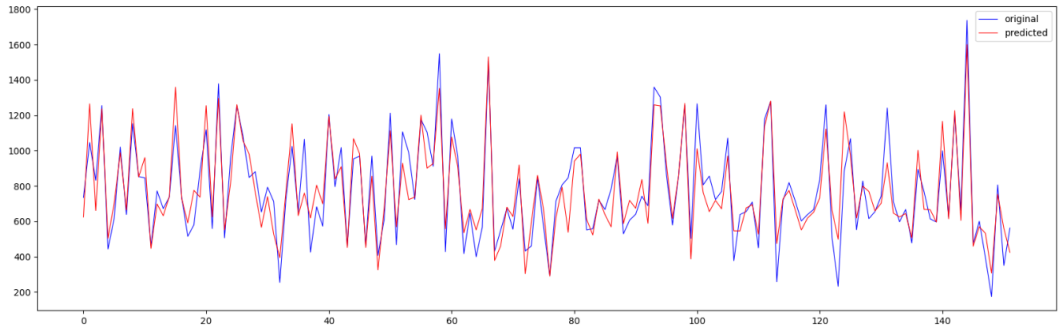

FIGURE 4.5: Gradient Boosting

We also tune the cross fold validation hyper parameter in gradient boosting. We did 5 fold cross validation gradient boosting. In dense layer we used adam optimiser to optimise the results.

Code snippets of this model :-

adam = tf.keras.optimizers.Adam(

learning$_rate = 0.000001$,

beta$_1 = 0.9$,

beta$_2 = 0.999$,

epsilon=1e-07,

amsgrad=False,

weight$_decay = 0.1$,

clipnorm=None,

clipvalue=None,

global$_clipnorm = None$,

use$_ema = False$,

ema$_momentum = 0.99$,

ema$_overwrite_frequency = None$,

jit$_compile = True$,

name='Adam'

)

In this model we also use the early stopping , checkpoint and batch normalisation in it with 5 dense layers of neural networks.

Chapter 5

Conclusions and Future Work

Crop yield prediction is essential for formulating a country's food policies. Timely and accurate predictions will be of great help to economic strategists. It would lead to proper export and import formulation based on the agricultural products. As we have identified in the literature survey, Machine learning models give consistently better results than non-machine learning models. Based on the experiments done and the research methodology employed in this project Gradient Boosting, was found to be the best model for prediction of yield. This study was able to achieve MSE value of 13069.819 , MAE of 87.663. Random Forest Regression also showed decent performance characteristics and could also be expanded on further to obtain better results. For future research work, the aim would be to create improved models for different crops and try to enhance model accuracy by using larger amounts of descriptive data. The approach used in this study is intended for application on a variety of crops and also we can use this prediction to predict the final price of the yield that will be given to the farmers by using the demand and supply laws of market..

Bibliography

[1] M. G. PS, "Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms," *Applied Artificial Intelligence*, vol. 33, no. 7, pp. 621–642, 2019.

[2] Y.-x. Su, H. Xu, and L.-j. Yan, "Support vector machine-based open crop model (sbocm): Case of rice production in china," *Saudi journal of biological sciences*, vol. 24, no. 3, pp. 537–547, 2017.

[3] P. Saad, M. R. M. Juhari, N. K. Jamaludin, S. S. Kamarudin, A. Bakri, and N. Rusli, "Backpropagation algorithm for rice yield prediction," in *Proc. of the Ninth Int. Symp. on Artificial Life and Robotics (AROB 9 th'04) Beppu, Oita, Japan*, 2004.

[4] B. Ji, Y. Sun, S. Yang, and J. Wan, "Artificial neural networks for rice yield prediction in mountainous regions," *The Journal of Agricultural Science*, vol. 145, no. 3, pp. 249–261, 2007.

[5] D. Ramesh and B. V. Vardhan, "Analysis of crop yield prediction using data mining techniques," *International Journal of research in engineering and technology*, vol. 4, no. 1, pp. 47–473, 2015.

[6] N. Gandhi, O. Petkar, and L. J. Armstrong, "Rice crop yield prediction using artificial neural networks," in *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, pp. 105–110, IEEE, 2016.

[7] V. Sellam and E. Poovammal, "Prediction of crop yield using regression analysis," *Indian Journal of Science and Technology*, vol. 9, no. 38, pp. 1–5, 2016.

[8] A. Chemchem, F. Alin, and M. Krajecki, "Combining smote sampling and machine learning for forecasting wheat yields in france," in *2019 IEEE second international conference on artificial intelligence and knowledge engineering (AIKE)*, pp. 9–14, IEEE, 2019.

[9] P. Charoen-Ung and P. Mittrapiyanuruk, "Sugarcane yield grade prediction using random forest and gradient boosting tree techniques," in *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 1–6, IEEE, 2018.