J. Integr. Sci. Technol. 2022, 10(2), 139-155 Article

Journal of Integrated
SCIENCE & TECHNOLOGY

The robust and efficient Machine learning model for smart farming decisions and allied intelligent agriculture decisions

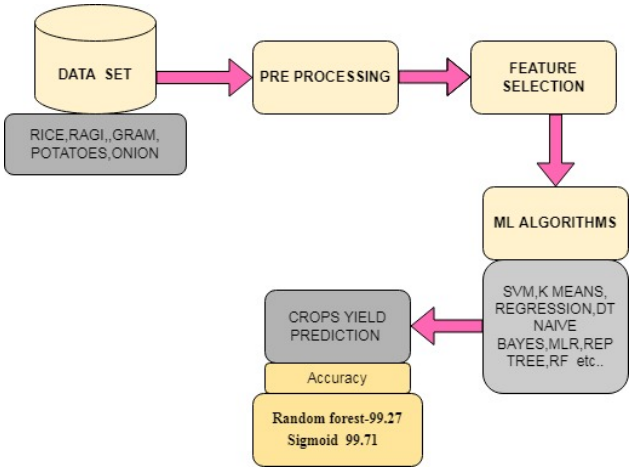
Shraban Kumar Apat[1*], Jyotirmaya Mishra[1], K. Srujan Raju[2], Neelamadhab Padhy[1]

[1]Deptartment of Computer Science and Engineering, Gandhi Institute of Engineering and Technology University, Gunupur, Odisha. India. [2]CMR Technical Campus, Jawaharlal Nehru Technological University, Hyderabad, India.

Received on: 07-May-2022, Accepted and Published on: 11-Aug-2022

ABSTRACT

Crop Yield Prediction is essential in today's rapidly changing agricultural market (CYP). Accurate prediction relies on machine learning algorithms and selected features. Any machine learning algorithm's performance might well be enhanced by introducing a diverse set of features into the same training dataset. Crop yield prediction includes parameters such as temperature, humidity, pH, rainfall, as well as crop name in forecasting the yield of the crop based on historical data. It offers us an indication of the best crop to expect in terms of weather conditions in the field. Crop prediction is a difficult task in the agricultural realm. The primary purpose of this research is to offer a novel machine learning approach for a heterogeneous data environment containing IoT-sensed data about the environment, agricultural conditions, plants' features, demands, etc. We have used data set of the following five crops (rice, ragi, gram, potato & onion) collected from Andhra Pradesh, Kaggle repository. In this work, we utilized diverse machine learning as well as deep learning algorithms such as Regression methods, Decision tree, Naive Bayes, SVM, K-Means, Expectation-Maximization (EM), and AI techniques (LSTM, RNN). It seems that among machine learning techniques, the Random Forest algorithm outperforms with 99.27% training accuracy for crop yield prediction. However, among sigmoid, ReLu, and tanh activation, sigmoid achieves 99.71 percent accuracy with four hidden layers for predicting the crop yield prediction.



Keywords: Agriculture decisions, smart farming decisions, IoT, Machine learning, Deep Learning

INTRODUCTION

In the last few years, the exponential rise in population has been witnessed globally. In parallel to such a high pace up surging population, the associated demands too have increased gigantically. Undeniably, the key demands about human beings are basic needs, including healthy foods, agro-products, efficient daily life, efficiency-oriented computing environments, efficient decision making, etc. However, amongst the significant demands, sufficient and healthy food has always been the dominant need. To meet such fundamental needs, the agriculture industry has gained irreplaceable significance. However, in practice, there have been numerous adverse conditions affecting or impacting the agriculture industry.

Numerous factors, including natural adverse conditions, plant disease, and a demand-supply disparity, can eventually change the overall agriculture system. Observing the circumstances above, it can be found that ensuring optimal yield and making an optimal farming decision by exploring at hand possibilities, demands, etc., can be vital to achieving higher agro-productivity to help all associated stakeholders.

Broadening the perceptual horizon, it can be found that the agriculture sector has a significant contribution to the global economy. Factually, with the expansion of population better farming practices should improve with precision farming and agricultural technology, this can have a strong impact on agriculture system growth prospects. It broadens the horizon for allied stakeholders to enhance farming practices, farming decisions, improved technology-assisted practices, demands-centric business decisions, etc. Different techniques have been proposed with such

Corresponding Author: Shraban Kumar Apat
GIET University, Gunupur, Odisha, India
Email: shraban.apat@giet.edu

Cite as: *J. Integr. Sci. Technol., 2022, 10(2), 139-155.*

 ISSN: 2321-4635 http://pubs.iscience.in/jist

motives encompassing heavy machinery, smart farming, optimal (controlled) fertilizer use, and enhanced decision-making.

Among the significant practices, the use of software computing systems, data analytics, sensor technologies, etc., have also gained widespread global attention. Recently developed techniques such as cloud computing, data analytics, and the Internet of Things (IoT) have gained widespread attention towards innovative farming decision-making practices. Technological advances such as data analytics, cloud computing, Big Data, as well as IoT have vital significance and the ability to support smart farming and intelligent agriculture practices. For example, IoT sensors can help retrieve potential or crucial environmental factors such as temperature, weather conditions, humidity, wind direction, speed, etc. At the same time, data science techniques can be vital for retrieving optimal decision variables to make an efficient decision. To assess such gigantically, massive datasets retrieved from the different sources or IoT sensors (called heterogeneous datasets) can be mined or processed using machine learning methods to make an optimal prediction and allied decision. In addition, the use of the different IoT collected (agricultural) data can help to make optimal decisions of modern farm operations. Summarily, the use of "IoT enabled Smart Sensors and Machine learning methods" can help to retrieve optimal decision-related outputs.

Consequently, it can help to make an intelligent or smart farming decision to gain a higher yield or productivity. It can help to achieve different stakeholders, increasing their profitability with better costs, time, and resources consumption. IoT and machine learning technologies can be of paramount significance for environmental sustainability, resource allocation, waste reduction, soil optimization, and data gathering and allied processing to make an optimal smart farming decision. Additionally, these technologies can also identify influential and non-influential factors that can make an optimal smart farming decision and market analysis and allied forecast. Farming remains the foundation of our financial system, notably in India. Agriculture offers food and raw materials to a large section of the population also providing work possibilities. In 2010, the global yearly emission was agriculture, forestry, and land-use change contributed roughly 20 to 25 percent. It is transformed both non-agricultural land and greenhouse gases that assist agriculture in coping with climate change.

Though numerable machine learning methods have been deployed to perform analysis on data mining. Amongst of the foremost prominent machine learning techniques include Decision Tree (DT), Support Vector Machine (SVM), Linear Regression (LR), Random Forest (RF), Support Vector Regression (SVR), Artificial Neural Network (ANN), Deep Learning (DL), Convolution Neural Network (CNN), Bayesian Network (BN), K-Means clustering, K-Nearest Neighbour (KNN), etc. The aforesaid machine learning methods have been analyzed to perform various kinds of smart farming decision-making practices for prediction alkies-animal welfare, Livestock production enhancement, yield forecasting, disease diagnosis, weed recognition, crop quality assessment as well as forecasting (analysis), species identification, plant disease detection, and prediction, water, and soil management practices. Undeniably, the facts above reveal that several efforts have been made to exploit machine learning methods and IoT technologies; however, differences in their respective performance often put a question mark over respective generalization ability and universal acceptance. It can also be found that the prediction accuracy of one machine learning method often varies from others, and in such cases generalizing their performance as optimal performance is questionable.

On the contrary, obtaining an optimal system and allied computing environment whose outcomes or prediction outputs could be considered optimal revitalizes academia-industries to achieve an optimal solution. With this motive, this paper focuses on developing a robust and efficient IoT data-assisted, and Advanced Machine Learning (AML) based agro-analysis and prediction model for smart farming and allied decision-making purposes. Unlike previous research, which focused on using a single or a few randomly selected machine learning methods for prediction, this study focuses on using the most effective ML algorithms and their efficacy to design a robust Heterogeneous Ensemble Learning Environment (HELE) for Agriculture Data analysis and prediction. As a research goal, in this research proposal, publically available and standard benchmark datasets have been proposed to be used. Though the allied data can be obtained from IoT sensor networks, considering scalability, depth of data, and suitability, benchmark datasets have been proposed for smart farming decisions in this research. Though in this overall process, machine learning methods have a vital impact, in this research, unlike the standalone machine learning algorithm as a solution, the HELE model is proposed. Machine learning-based prediction or analytics approaches require an optimal work environment such as data suitability, precision, outlier-less data environment, significant features, etc., which have not been considered significant existing research. Undeniably, enhancing machine learning-based predictive methods with the above-stated optimization (preprocessing) can be vital in achieving optimal prediction accuracy and reliability. Considering all these research scopes, a highly robust HELE machine learning concept-based Smart Farming and Intelligent Agriculture Prediction Model is proposed in this research. The proposed method encompasses multi-phase enhancement in the form of preprocessing, data sampling, feature extraction and selection, and HELE-based classification. The overall proposed model can accomplish the most efficient, reliable prediction output to make optimal Smart Farming and Agriculture decisions. Before discussing the research goals and allied implementation scopes, assessing existing research can be vital. In the subsequent section, a snippet of critical contribution has been mentioned.

CRITICAL CONTRIBUTION

i. A novel machine-learning algorithm used for heterogeneous data environment containing IOT-Sensed data about the environment, agricultural conditions, plant features, demands, etc.

ii. Applied different machine learning techniques as well as deep learning methods after an extensive survey over more than 40 papers.

iii. Proposed heterogeneous ensemble learning environment (HELE) & proved to be a more effective, reliable as well as accurate

machine learning model for smart farming and agro decision system.

iv. The study focused on five crops Onion, gram, potato, ragi & rice.

v. It has been seen that Random forest outperforms with 99.27% training accuracy for crop yield prediction.

vi. In the subsequent section, a snippet of the key research made in the last few years is discussed with this motive.

LITERATURE SURVEY

This section discusses critical literature discussing IoT and machine learning approaches for smart agriculture and allied decision-making. Van Kloppenburg et.al [1] used features like temp, rainfall, & type of soil, and applied ANN algorithm. in the prescribed models along with this few additional analyses performed on deep learning algorithms- such as CNN, LSTM, &DNN. It was noted that LSTM, DNN& CNN, algorithms are highly preferred DL techniques. PaudelDilli, et al [2] Along with ML used agronomic to build an ML base to forecast large-scale crop yield with consideration of the following features crop simulation outputs weather, soil data& remote sensing, and used MCYFS database. The experiment has been performed with various crops over different geographical areas. Wang, Y et al [3] used advanced ML techniques with seasonal data collected from different sources i.e. images from satellite, climatic data, soil, etc. The author used OLS, LASSO along with four well-known ML Techniques such as SVM, RF, Ada Boost, and DNN in the above study it was found that the Adaboost method (A reliable prediction ($R2 > 0.84$)) shows better results than other techniques. Guo, Y et al [4] In this study traditional MLR along with 03 advanced ML Methods such as BP, SVM& RF were considered with phenology, climate as well as geographical data to forecast rice yields. The results indicate that ML methods showed better results in contrast to the MLR method and the difference between RMSE (R2) prediction and observed rice yields were 800 (0.24), 737 (0.33), and 744 (0.31) kg/ha for BP, SVM and RF, respectively.

Bali, N. et.al [5] Survey over more than 80 research papers performed and reasonably gap has been identified. So a good no. of the hybrid model and DL methods were summarized as means of crop prediction, also analyzed how various factors like temp, humidity, etc. harm overall productivity of crops. ANN and ANFIS are capable to produce better results. Good accuracy has been shown by the hybridized model which used fuzzy and ANN so hence we might examine further hybridization strategies in the future for better to best predictions. Khaki et al [6] Corn and soybean production has been forecasted in his proposed model which used CNN-RNN techniques along with RF, DFNN, and LASSO, across the entire Corn Belt which includes13 states of USA for the three years. i.e. 2016 to 2018. The proposed model achieved an RMSE of 9% and 8% of their respective average yields and outperformed all other methods that were tested. Chu, Zheng, et al., [7] introduced the BBI-model, a three-sept forecasting model that integrates two BPNNs and an IndRNN. According to the findings of this framework, this BBI-model seemed to have the least MAE as well as RMSE for summer & winter rice forecast (0.0044, 0.0057) and (0.0074, 0.0192), correspondingly, when the network's layer count was fixed to 6. Those results demonstrated that the suggested methodology can reliably forecast summer as well as winter rice yields in 81 Chinese counties. As per Zhao et al. [8] in their work, the Internet of Things (IoT) could be significant for the agricultural industry, where different advanced technologies such as control networks and information networks can be strategically integrated to perform remote monitoring, control, and information management. Varghese et al., [9] too focused on exploiting IoT technology along with machine learning to design a cost-effective smart farming structure. The Author employed cross-related information, which was collected through different sensors deployed across the farm region. Machine learning was being employed to forecast the future state of the crops depending on the crop data obtained. With a similar motive, Zannouet al., [10] developed a Sorghum yield prediction model using machine learning. The authors performed their case study on a Sorghum field, where the TensorFlow method was applied in conjunction with the Convolution Neural Network (CNN) as well as Linear Regression to detect varied ears of Sorghum on an image and their respective weights. With the technology above, authors could achieve the average accuracy of 74.5%, while precision for weight estimation was obtained as 99%.

Fuady et al.,[11]In their research, they applied the ELM algorithm to control the temp & humidity of mushroom farmhouses. Considering optimal conditions for oyster mushroom growth, the authors found that the temp and humidity with 28° Celsius and of 80% respectively are suitable to achieve a good outcome. The authors applied an SLFN by modifying the H inverse matrix versus target matrix (often called ELM algorithm) to control the temp and humidity. The authors found that ELM can perform better than the classical backpropagation neural network and zero-order Fuzzy Logic Control.

Balducciet al.,[12] applied the IoT ecosystem-based smart farming model to employ IoT sensors that enabled information to understand the influential and non-influential factors while considering environmental, productive, and structural data from many sources. To predict values using and comparing innovative ML techniques authors exploited the real heterogeneous datasets. Gertpholet al.,[13] proposed smart hydroponic lettuce farms by exploiting IoTs. To achieve it, the authors collected environmental data, which was further utilized to manage the farm's activity in real-time. Experimented with the enormous amount of data set over light intensity, humidity, and temp with different regression techniques such as ANN, SVR & MLR as a prediction model. Doshi et al [14] proposed an intelligent crop recommendation model using machine learning methods. To achieve it, the authors applied the concept of Big Data Analytics and ML technologies and by Employing these methods, the authors also proposed the Agro Consultant model intended to aid Indian farmers to develop a knowledgeable and enthusiastic judgment on which crop to cultivate and when. Seasons, geographical region, soil conditions, plus ecological elements including temperature as well as rainfall also were taken into account in this research. Balducci et al., [15] focused on applying heterogeneous information and data as input from real-time agricultural datasets to make better decisions in farming. The author suggested applying the IoT concepts to collect

different real-time data, and also further suggested to process with an ML model (over historical data) towards time-series analysis for a smart decision system.

Muthusinghe et al.,[16] proposed a smart farming model by performing paddy harvest and rice demand prediction. Because Paddy crop yield and (rice) demand are affected by numerous factors like rainfall, humidity, citizen's lifestyles, etc., authors focused on harvest forecast and allied yield prediction. Authors have applied RNN and LSTM based learning models as prediction systems. Trebouxet al., [17] exploited the machine learning-assisted image recognition model for precision agriculture decisions. To achieve it, the authors applied machine learning models for the accurate detection of precise agricultural objects.

CN et al., [18] applied the agriculture dataset, which was processed with data mining techniques for better agriculture analysis. The authors considered precipitation, temp, evapotranspiration, area, production, and yield as a dataset. Authors applied different ML algorithms such as K-Means Clustering, SVM, KNN, and Bayesian network. Dholuet al., [19] developed IoT for precision agriculture applications. The authors applied Soil Moisture, Temperature & Relative Humidity around the plant to make optimal control decisions. Shindeet al., [20] reviewed the different crop disease prediction models using machine learning algorithms. The authors assessed different parameters such as humidity, temp, rainfall, wind flow, light intensity & soil Ph value to have significance for precision agriculture and productivity. The authors recommended that smart farming decisions and control can be realized better by deploying ML models with IoT. Parket al., [21] developed an image-based deep learning model for crop disease prediction. Intending to enhance productivity, the authors proposed a dynamic image analysis based on agriculture decision-making. Kitpoet al., [22], too, recommended the IoT-deep learning ecosystem for greenhouse monitoring and control. Their proposed model supports farmers with a better growth monitoring system by assessing temperature, humidity, water supply, and disease detection. Sarangdharet al., [23] Study on cotton leaf disease detection and control system with the help of machine learning regression-based model & the IoT ecosystem. Besides, the authors also proposed the soil quality mentoring model. The authors applied the SVM -based regression model that classified cotton leaf as Bacterial Blight, Alternaria, and Fusarium wilt as an ML model. The authors could have achieved a maximum classification accuracy of 83.26%. Sharma et al., [24] proposed a Big Data analytics model for the prediction of crops. To achieve it, the authors applied key information such as avg temp, avg humidity, and total rainfall that classified that production yield as a good yield or a bad yield. As a classifier, the authors used the SVM algorithm to make the two-class classification.

Liakoset al., [25] performed an in-depth analysis of the different IoT and machine learning approaches for smart farming or agriculture practices. Authors found that machine learning methods can be vital to predicting crop growth, yield, demand, etc. prediction, which can help achieve higher growth in the agriculture sector. Inyaemet al., [26] proposed an ML model for the prediction of rice. As a machine learning model, the authors applied the DTT and ANN techniques towards the prediction of rice crops productions. The use of N-fold cross-validation was found useful in increasing overall accuracy. Similarly, Shakooret al., [27] applied DTL-ID3 and K-NNR algorithms to frame a prediction model for agriculture production (for Bangladesh). As a prediction outcome, the authors considered six types of major crops like varieties of Rice, Potato, Jute, and Wheat, which were predicted by assessing or processing a data set using Decision Tree K-NN classifiers. Yahata et al., [28] developed a machine learning-based model of hybrid type for smart agricultural decisions. The researchers did this by integrating diverse image sensing technologies to enhance automated soybean flower as well as seedpod surveillance in real-world fields. Noticeably, in their developed image sensing methods, authors employed sensors configured with an agricultural cyber-physical system and allied Big Data platform to decide. Authors utilized weather, temperatures, humidity, solar radiation, soil properties, and other factors as input data to determine optimal growing selections. The proposed scheme as a whole integrates numerous image processing as well as ML approaches, with the CNN algorithm serving as an ML model.

Materneet al., [29] assessed the significance of IoT with cloud environment for the prediction of agricultural pests and diseases. To achieve it, the authors applied Amazon ML cloud-based services to find the pattern which was hidden and the LR model as a classifier to forecast the disease. Chopdaet al., [30] proposed a cotton crop disease prediction model using a decision tree classifier for smart farming purposes. Considering the emergence of IoT and advanced computing technologies such as data mining in this work, the authors suggested a decision tree classifier-dependent cotton crop disease prediction model. Gertpholet al.[31] proposed a predictive model for smart hydroponic lettuce farms. The authors applied IoT sensing technology for real-time environmental data collection followed by an RM-based prediction for optimal farm operation control. To perform it, authors have applied key features such as intensity of light, humidity, and temp. which generally affect the growth of the plants. Authors employed Root Mean Square Error as a metric for model evaluation, whereas other categorization methods such as LR, SVR, MLR, as well as ANN have been used.

Considering the intention to assess the freshness of the hydroponic produces, Wong patikaseree et al., [32] applied manual feature extraction by classical ML models along with a few ML techniques such as DT, NB, MLP, and DNN, focused more on deep neural networks to process raw images and authors found that the application of decision trees achieved a test accuracy of 98.12% which is proved to be better than DNN as a decision tree classifier. Alipio et al., [33] A smart farming proposal has been suggested where BN & exact inference being used where IoT sensors real-time farming environmental parameters were considered. Sensors, as well as actuators, were employed in their suggested model to track and regulate physical phenomena including illuminance, ph level, electrical conduction, the temperature of the water, & moisture. The authors also applied Bayesian Network to infer each parameter's optimum value by gathering the information received from various sensors. Prakash et al., [34] developed a soil moisture estimation model for agriculture optimization utilizing numerous

machine learning approaches including LR, SVM regression, as well as RNNs, to predict moisture in the soil in different days ahead. Verma et al., [35] The proposed IoT-assisted smart farming model suggested using IoT sensor networks comprising different sensor nodes for continuous soil acidity level, temperature, and other variables monitoring to make suitable farming decisions system. Mondal et al., [36] Proposed IoT-assisted Intelligent Agriculture Field Monitoring System primarily to monitor humidity in the soil and temp. for further processing of sensed data towards a smart farming decision. The authors applied ThingSpeak as a cloud model to store data for future data analysis. A similar effort was made by Srinivasulu et al., [37], where authors developed a cloud service-oriented architecture (CSoA) for agriculture decisions. Authors found that IoT and Big Data framework can be vital for the agriculture sector and allied GDP enhancement.

Jumatet al.[38] proposed an IoT-assisted plant disease detection, diagnosis, and treatment model. Their proposed model detects plant diseases in advance and informs farmers to initiate early remedial measures. The authors used a raspberry pi based hardware model to collect the data, while a random forest classifier with a color histogram was used to perform plant disease detection and classification. Farhat Abbas. et al., [39] proposed a study in which soil data, crop properties were collected thru proximal sensing for two years(2017&2018)thereafter four data sets were formed and experimented with different machine learning algorithms like LR, EN, KNN&SVR in the 06 fields. The study showed that SVR performed better with RMSE of 5.97, 4.62, 6.60, and 6.17 for the choosen04 datasets. It was recommended that large datasets may be considered to make better predictions using different machine learning models. X.E. Pantaziet al., [40] collected crops growth rate characteristics and soil as parameters by using NIR spectroscopy sensor and then they proposed models such as CPANN, SKN, and XY-fusion network to predict wheat yield. The image preprocessing and analysis was made by orthorectification, in-band reflectance calibration, as well as NDVI computation. The input to the models was 8798 fusion vectors. The study showed that the best result was obtained from the SKN network, with an overall cross-validation accuracy of 81.65%. With the limitation of not modeling continuous output relations, we can enhance the proposed architecture with smooth interpolating kernels.

Mohsen Shahhosseini. et al., [41] predicted pre-growing season maize yield and nitrate loss using four machine learning algorithms, i.e., LASSO Regression, Ridge Regression, random forest, and Extreme Gradient Boosting (XGBoost) as meta-models. The dataset for this prediction includes the site's experimental data obtained from Sustainable Corn CAP Research Database, soil information obtained from SSURGO database, and daily weather from the year 1987-2016 is obtained from Dayment. The XGBoost outperformed all the other Ml techniques, with an RMSE of 13.4 percent. Numerous different ML approaches may be employed for assessment in the future.

Ahmed Kayad et al. [42] suggested a theory for tracking within-field fluctuation in corn yield utilizing MR, RF, as well as SVM and comparing the approaches, and identifying the best prediction model. The study was carried out on 22 hectares field from the year 2016 to 2018. The model used different vegetation indices, like NDVI, NDRE1,2, GNDVI, EVI, WDRVI, mWDRVI, GCVI, etc, obtained from Sentinel-2 images. This model predicted that Green Normalized Difference Vegetation Index (GNDVI) showed the highest r2 value of 0.48, and the best periods of monitoring is the crop age between 105-135 days from plantation date, and RF proved to be the best machine learning model for predicting the yield with R2 of 0.6. Future tests can be done to find out the best economical combination of the volume of training data for accurate prediction rather than surveying training data for the model.

RESEARCH GAP AND MOTIVATION

Machine learning technologies have a decisive impact on mining gigantically huge datasets to make optimal decisions. Correlating "AIoT and Machine learning" as composite technology, it can be found that this paradigm can be of utmost significance to enable optimal farming or allied decision. Though numerous efforts have been made to use machine learning methods for prediction and analysis; however, diversity in respective performance over the same dataset puts a question on the generalization of these at hand methods. In other words, different machine learning methods that are often selected randomly by researchers perform differently and give diverse output. Therefore, it makes accepting such standalone machine learning-based methods suspicious and confines (at least in terms of generalization or as a universal solution). In such a case, developing a robust and efficient machine learning environment can be of utmost significance to Prediction over AIoT collected data. It motivates academic industries to build a powerful and efficient machine learning environment whose outputs could be generalized as an optimal solution. With this clear aim, an extremely robust as well as efficient Heterogeneous Ensemble Learning Environment (HELE) was being developed in this research, which can guarantee to exploit the efficacy of the maximum possible machine learning algorithms to accomplish optimal prediction outputs. Furthermore, for any data analytics model or machine learning model, its effectiveness significantly depends on the data suitability. With this motive, in this research proposal, a "Multi-Phase Optimization (MPO)" concept has been introduced to enhance data quality, practicality, and scalability to achieve higher accuracy and reliability by the proposed HELE machine learning model.

Summarily, the following inferences can be stated:

RESEARCH GAP

1. Lack of Generalization ability by various machine learning techniques,
2. Use of random machine learning techniques over heterogeneous datasets,
3. Lack of heterogeneity and data sufficiency,
4. Lack of optimal data availability and learning environment, and
5. Use of single machine learning methods as standalone analytics tools or classifiers.

RESEARCH MOTIVATION

1. The use of heterogeneous datasets from different sources can make better decision making, especially for intelligent farming or agriculture.

2. The use of Multi-Phase-Optimization (MPO) can also enhance the overall computing environment to serve optimally enriched and reliable predictions towards intended smart farming and allied intelligent agriculture decision-making processes.
3. Unlike classical research, where authors have applied small-scale single-type data for analysis, multiple-sourced data with common decision intent can also be applied to make a more efficient (real-time) farming decision.
4. Unlike a single machine learning-based standalone classification system, using multiple classifiers (even heterogeneous) to behave as an ensemble learning method can be vital to achieving optimal prediction results.
5. Heterogeneous Ensemble Learning Environment (HELE) can be the most effective, reliable, and accurate machine learning model for smart farming prediction and analysis systems.
6. The aforementioned discussion as well as related implications can be regarded as the principal motivating reasons behind this work.

RESEARCH OBJECTIVES

Given the vital necessity of a reliable and efficient smart farming and intelligent agriculture system that allows for better decision-making, this research proposal primarily intends to exploit the efficacy of sophisticated systems such as IoT, sensor technologies, neural networks, etc. With this motive, in this proposal, a few research objectives have been identified. The overall research objectives are categorized into two broad types; primary aims and secondary objectives. Noticeably, primary objectives signify the overall research goal or intent, while specific objectives state the methodological paradigms to be applied to achieve primary objectives. A snippet of these key research objectives is given as follows:

RESEARCH METHODOLOGY

This section discusses the suggested method and the allied implementation schematic. The overall proposed prediction model for smart farming and intelligent agriculture is performed through the following stages.

1. Heterogeneous Data Acquisition
2. Preprocessing
3. Feature Extraction and Selection, and
4. HELE-based prediction.
5. The detailed discussion of the above-stated phased-implementation model is given as follows:

PHASE-1: HETEROGENEOUS DATA ACQUISITION

Since this research emphasizes employing IoT sensed agriculture data to make certain targeted predictions or analyses for better (intelligent) farming or agriculture decisions, data has been proposed to be sourced from different benchmark sources such as Kaggle and numerous others. Undeniably, in the majority of the existing works, authors have focused merely on using limited data sources such as environmental data, weather data, crop, or plant data with local weather conditions to make a localized prediction. Therefore, these can't be sufficient to make more informative analytics or prediction. Factually, possessing large variables or research goal-oriented information can make predictions more efficient and constructive. With this motive, this research intends to exploit benchmark data obtained from different possible sources, including IoT, collected data, standard or specific data prepared for smart farming prediction, etc. Here, the prime objective is to employ large-scale heterogeneous data to predict crop type, plant growth, etc., which can help stakeholders make the optimal decisions. And therefore, the focus is on retrieving the maximum possible (associated) information to make predictions. Since the agriculture dataset in question is derived from various sources, it is referred to as a "Heterogeneous Agriculture Dataset."

PHASE-2: PREPROCESSING

Because the information was acquired from several resources, it needed to be filtered to be appropriate for further analysis. With this motive, in this research proposal, multi-phase pre-processing has been recommended. Noticeably, the data can be in different formats, nature, size, etc. At first, data augmentation has been processed to make it uniform. Then, it can be done using one-hot coding, Word2Vec, or similar methods. Moreover, Min-Max Normalization, as well as outlier recognition, have been presented for the gathered information. Uniform data specimens were taken to manage these phases, which would then be employed to examine for data imbalance. However, at this phase, the emphasis is also on getting ideal samples. Therefore, different sampling methods, such as random sampling, up-sampling, etc., can be applied to alleviate such issues. At this point, we executed data pre-processing procedures to replace lacking data with the matching feature average as well as to normalize the information to the range [0, 1], because certain algorithms cannot tolerate negative values. The generated data was then utilized to train, validate, & test the machine learning classifiers utilized in this work. An adequate amount of data samples was required for the training process to adequately acquire the data pattern behavior but also guarantee a high-quality learning process, whilst the validation procedure was being used to adjust the classifier hyper-parameters by choosing the proper values that satisfy the machine learning problem. Lastly, this machine learning model can be evaluated during the testing step, which provides information about the trained classifier's performance.

PHASE-3: FEATURE EXTRACTION AND SELECTION

To make any analytics model or machine learning efficient, feeding it with sufficient data input is always challenging, especially when it requires a balance between the computational cost, several variables, feature-sets, and allied performance. Bulky data input can force a machine learning method to undergo convergence and local minima, eventually reducing computational efficiency and accuracy. On the other hand, it can impose huge computational overheads. This research proposal proposes different feature selection methods to mitigate such unwanted computation, such as WR Sum Test, PCT, PCA, GA-based feature selection, etc. Once obtaining the optimal data elements and allied features it has been processed for machine learning-based classification. A detailed discussion of the proposed machine learning classification or allied prediction model is given in the subsequent section.

PHASE-4: HETEROGENEOUS ENSEMBLE LEARNING ENVIRONMENT BASED PREDICTION

The emphasis throughout this phase was also on constructing a new HELE model leveraging a range of cutting-edge neural network models, such as Decision Trees, SVM, Linear Regressions, and Random Forests, as well as ANN, DL, and K-Means clustering and K-Nearest Neighbourhood Analysis (KNN), HELE-MVE and HELE-BTE are the two acronyms for these drugs. The eventual model developed in this research proposal intends to predict different aspects of agriculture practices such as crop types, weather conditions, plant disease, market conditions, etc. This eventually can help in making optimal smart-farming and intelligent agriculture decisions. Overall performance assessment has to be done thoroughly for all the machine learning methods as standalone and ensemble learning-based approaches. The proposed model is expected to be compared or assessed concerning certain standard research methods available to infer the effectiveness and acceptability of the research contributed. A snippet of the proposed method is given as shown in Table 1.

Table 1: Proposed Methods

Dataset	Agriculture datasets containing KAGGLE, etc. (about smart farming, crop-yield production, and plant assessment)
Pre-processing	Significant test (Wilkinson test), Min-max Normalization Outlier Analysis
Feature Selection	Rank Sum Test Principle Component Analysis Pearson correlation Analysis Evolutionary Computing assisted feature selection
Machine Learning	Decision Trees, LR, LOGR,NB
Output	Predicted Ouput/Category/etc.(Depend on data and intend)
Performance Parameters	Accuracy, Precision, F-Measure, F1-Score, etc.

PHASE 4.1: MODEL SELECTION WITH HYPER PARAMETER OPTIMIZATION

As illustrated in the upcoming section, we utilized Weka as well as the Auto-Weka plugin to construct, modify model parameters as advised, train, assess, and pick the best-fit categorization models. The Waikato Environment for Knowledge Analysis is a well-known Java-dependent machine learning (ML) package. Its workbench offers a gathering of visualization tools including algorithms for information analytics and forecasting modeling, as well as graphical user interfaces for accessing this capability. Auto-Weka is a Weka package plugin that leverages Bayesian optimization to provide such a highly optimized parametric machine learning framework without the requirement for user intervention. ROC presents us with a two-dimensional curve with a (1, 1). The finest model has a curve near y = 1 and an AUC near 1.0. Random-guessing provides an AUC of 0.5. This allows comparing a model to a random prediction without regard for thresholds or the proportion of individuals in each class.

PROPOSED METHODOLOGY

Below shown Fig.1 is a proposed model for crop-yield prediction. Five different crops dataset has been collected from various districts of Andhra Pradesh such as Onion, Gram, Potato, Ragi, and Rice. Every column in the dataset is named after a state or district. The dataset for Gram, Onion, Potatoes, Ragi, and rice consists of 296, 453, 66, 356 & 622 no of rows respectively. These strategies helped us to identify the top features for our investigation. In Data Pre-processing, dependent and independent variables were separated then the missing data were filled with the mean of all the data in that column.

EXPERIMENTAL OBSERVATION AND RESEARCH SIGNIFICANCE

Few of the important research significances hypothesized in this study were summarized as follows: Unlike classical efforts, this research has considered heterogeneous datasets, which can provide better insight and multiple decision variables to assist optimal smart-farming prediction, decision, and allied analytics systems. In several existing efforts, a restricted data set was employed when conducting this research IoT information collected from the benchmark and publically available sources have been considered. As a result, it can make comprehensive analytics more productive and constructive to decide optimal smart-farming and/or intelligent agriculture.

For this research, efforts have been made to improve data preprocessing, data sampling (in case of possible data imbalance condition), enhanced (multiple phased) feature extraction, and selection, this might serve an active role in boosting the overall computational efficacy as well as reliability of the findings.

- This article has applied different ML approaches such as decision trees, regression techniques, neural networks, advanced neural networks, enhanced pattern mining models such as SVM, SVR, etc., as base classifiers. Eventually, with such heterogeneity, the final prediction outcome can be more efficient and reliable. Thus, this approach can be more computationally efficient, accurate, and constructive in making optimal smart-farming and/or intelligent agriculture decisions.
- The overall approach, which exploits the NN approach with an enhanced and computationally enriched model, can be vital to achieving optimal smart-farming and/or intelligent agriculture decisions.
- Thus, the overall proposed model can significantly achieve optimal smart-farming and/or intelligent agriculture decisions.

K-MEANS ALGORITHM

This k means algorithm is indeed a prototype-dependent partitioned approach that specifies prototypes in terms of a centroid in an n-dimensional ongoing space. This methodology employed by K-means has been outlined in depth. The very first stage in the procedure is the initialization of K centroids (Line 3), where K signifies the count of preferred clusters supplied by the consumer.

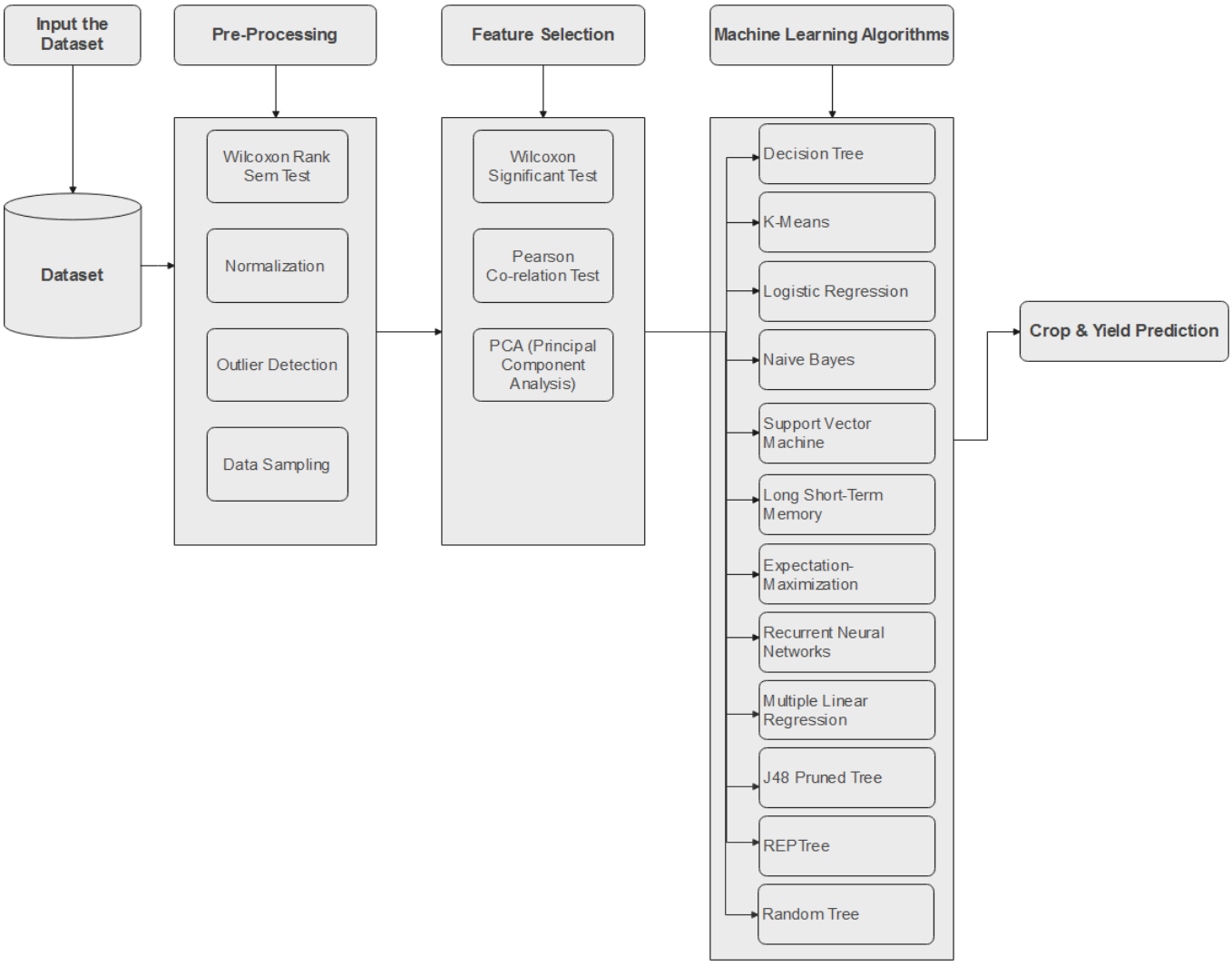


Figure 1: Presents the graphical presentation of the proposed model.

Subsequently, for every point, the centroid closest to it is assigned, culminating in the establishment of groups (Line 5). The centroid of every cluster is typically updated at the end of the procedure (Line 6). These steps were continued till the centroid does not change.

Algorithm for K-Mean

1. Input: K
2. Start
3. Initial centroids should be chosen from a set of K locations.
4. repeat

   Create K clusters by assigning each point to the centroid that is closest to it;

   Recalculate each cluster's centroid;
5. until Centroids remain constant.

K-means was using a proximity function to compute the proximity between each location as well as the centroids. While Euclidean Distance, Manhattan, and Cosine are the most often utilized functions, there are more accessible for various data types. Additionally, K-means employs an objective function to define the clustering objective mathematically. Typically, the target was to minimize the Euclidean Distance between an item as well as the cluster centroid.

Each predictor's data is saved in one of the nine clusters. These groups appear to be more precise and effective than the preceding ones, chosen at random. If a new observation is received, the distance from the clusters will predict the outcome. The resulting K-means clustering sizes are displayed below. The cluster with K = 4 has the largest size, whereas the cluster with K = 2 has the smallest. These can justify the decrease of the principal component analysis features.

Table2: Distance measure through K-Mean

Cluster#	1	2	3	4	5	6	7	8	9
Cluster Size	39	28	47	117	96	83	77	105	57

Table 3: No. of 0's and 1's in training set

% of cluster in 0's and 1's	1	2	3	4	5	6	7	8	9
0's	46	73.51	67.6	71.53	**77.92**	70.51	68.37	54.15	76.48
1's	**43**	34.489	52.42	38.46	33.07	37.48	38.62	37.84	29.51

Now we need to check how much percentage of 0's and 1's available in each and every cluster during the training set

The above Table 3 indicates that cluster#5th, i.e., 77.92% % of 0's. As a result, the yield is unproductive. As a result, the smallholder farmer must avoid possessing characteristics comparable to those found in the fifth cluster.

Accuracy: According to the prior projections on the training set, a point in the first cluster has a 43 % chance of producing a productive yield. In the testing set, it would have a 67.56% chance of making a productive crop. As a result, we must design a loss function to quantify the model's accuracy.

Now we need to check how much percentage of 0's and 1's available in each and every cluster during the training set.

Table 4: No. of 0's and 1's in the testing set

% of cluster in 0's and 1's	1	2	3	4	5	6	7	8	9
0's	32.43	70	70	69.24	78.5	70	78.125	73.33	62.5
1's	**67.56**	30	30	38.46	21.5	30	21.875	26.66	37.5

We have estimated the number of zeros and ones available from the 9th clusters from the above Table 4

Table 5: Percentages of 0's in the training as well as testing set

% of 0's in both (Training and Testing)	1	2	3	4	5	6	7	8	9
Training set	46	73.51	67.6	71.53	78.5	70.51	68.37	54.15	76.48
Testing set	32.43	70	70	69.24	78.5	70	78.125	73.33	62.5

In the above table-5 Percentage of 0's has been computed in the training and testing set.

ABNORMALITY

Apart from the first cluster, the percentages of the training, as well as testing sets in the table above, were similar. This has shown in the diagram below. After specifying the ideal count of K-mean clusters, which would be 9, we discovered that, except for the first cluster, most unproductive yield percentages were equal. In training and testing sets, the first cluster gave distinct outcomes.

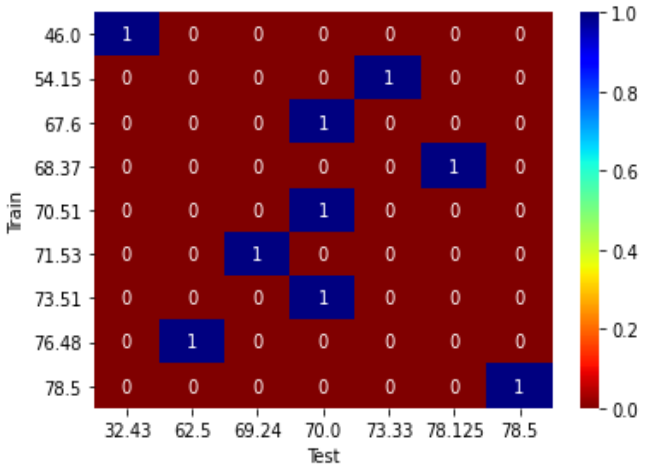


Figure 2: K- Mean Clustering

By considering our crop prediction's training and testing data, we have plotted the above graph of K-Means clustering algorithms. Based on cluster value (k=2), the mean accuracy evaluated is 76.48%.

EXPECTATION-MAXIMIZATION (EM)

Expectation-Maximization seems to be another prominent clustering technique. EM seems to be a statistical method that emerged as a parameter estimator and has since been used in the field of clustering. While EM serves comparable services to K-means, it's also categorized as a fuzzy clustering approach since it employs a probabilistic model for estimating the likelihood of every item belonging to each cluster. For the entire population of objects, the EM algorithm provides a mixed distribution. A mixture distribution is made up of multiple independent distributions. The parameters were computed via Maximum Likelihood Estimation (MLE). It estimates the values of numerous parameters, including mean and variance, to increase the probability that the items belong to any certain distribution. As a consequence, every cluster gets depicted by a distinct distribution, and also its parameter values match the characteristics of the cluster. These parameters were originally unknown; however, utilizing EM, the coefficients being computed, and items that are much more plausible to suit the distribution have been deemed anticipated. The estimated probability distribution of the items is derived during the E-step (Line 4) of the method. The M-step (Line 5) then determines the attributes of the individual distributions to boost the desired probability as much as feasible. Both processes are repeated until convergence is reached (i.e., the parameters do not change) or until a predetermined number of interactions is reached.

The Expectation-Maximization Algorithm is represented as pseudocode.

1. begin
2. Choose a starting set of parameters;
3. repeat

3.1. Calculate the probability of each object occurring in each distribution;

3.2. Maximum Likelihood Estimation (MLE) new parameter values for each distribution taking P into account are calculated.

4. Until The parameters remain unchanged.

Implementation Details

Step 1: Load the crop dataset, which comprises numerous parameters.

Step 2: Install the necessary libraries as well as packages.

Step 3: Data preprocessing was performed accordingly.

Step4: To prepare the dataset, the data gets split into two sets: training as well as testing sets.

LINEAR REGRESSION

The Linear Regression technique is applied to a dataset (secondary data) containing different features such as date, minimum temperature, maximum temperature, humidity levels (both min and max), precipitation, etc. Calculated the mean and standard deviation for this dataset, but the difference is greater, which indicates the data is noisy. The dataset was cleaned using various approaches. First, we have plotted the graph using humidity and precipitation since these two variables are favorably connected to other variables. Following that, we drew a graph between humidity levels and precipitation, as shown below. The mean value of humidity was calculated for this graph, and maximum humidity values showed zero precipitation. It received an accuracy rating of 89.72 percent in the end.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta^3 x_{i3} + \beta_4 x^{i4} \quad (1)$$

Where β seems to be the regression coefficient, y would be the predicted value, while x represents the data set.

Table 6: Accuracy of crop prediction using linear regression for a variety of crops

Crops	Accuracy	
	Training Set	Testing Set
Rice	86.58	93.87
Ragi	92.29	93.77
Gram	94.29	93.67
Potato	94.39	92.39
Onion	**95.27**	92.29

In Figure 3, we plotted the accuracy of the crop prediction by using simple LR. The graph is plotted by considering humidity and

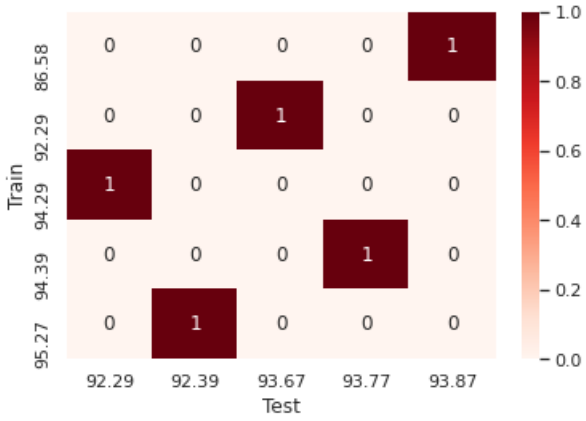


Figure 3. Simple Linear Regression

precipitation parameter. The accuracy rate of the potato crop seems to be 95.27 percent, which is high in this case.

The Linear Regression Hypothesis Function is

$$Y=\theta 1+X.\theta 2 \quad (2)$$

While training the model we were provided:

X: input training data

Y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of Y for a given value of X. The model gets the best regression fit line by finding the best θ1 and θ2 which deals.

θ1: Intercept

θ2: coefficient of X.

MULTIPLE LINEAR REGRESSION

In MLR the population regression line for p explanatory variables x1, x2,..., xp is defined to be μy=β0+β1x1+β2x2+……+βpxp

This line depicts how mean response μy fluctuates when the explanatory factors vary. The recorded values of y range around their means μy and are considered to have the same SD. The fitted values b0, b1,.....bp compute the population regression line parameters β0,β1, β2,….βp.

Formally, the model for multiple linear regressions, given n observations, is

$$Yi = \beta 0+\beta 1xi1+\beta 2xi2+……..+\beta pxip+€ \quad (3)$$

for i=1,2,3……n.

Table 7: Shown accuracy of crop prediction using Multiple linear regression for a variety of crops

Crops	Accuracy	
	Training Set	Testing Set
Rice	87.58	88.87
Ragi	92.39	91.67
Gram	93.39	92.57
Potato	95.26	93.55
Onion	**96.27**	94.39

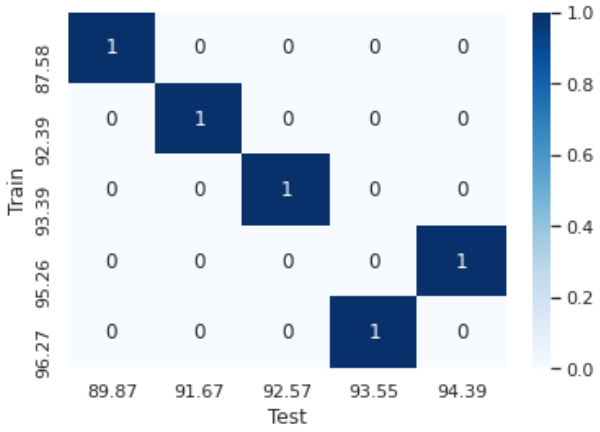


Figure 4: Multiple Linear Regression

By considering more than two variables, we have evaluated the accuracy of our hybrid model by using multiple linear regression. The above Figure 4. is plotted between the training and test data used for crop prediction. In this case, the onion crop gives the highest accuracy of 96.27%.

LOGISTIC REGRESSION (LOGR)

LOGR is an alternative ML technique from the field of statistics. It is a must for binary classification problems. Logistic regression is named for the core method used as a logistic function.

It's a curve in S-shape that can take any real-valued no. to map between 0 and 1 but rarely remain at those limits.

$$1 / (1 + e^{\wedge}\text{-value}) \quad (4)$$

Where e is the base of the natural logarithms

REPRESENTATION USED FOR LOGISTIC REGRESSION

Like linear regression, logistic regression employs an equation as its representation.

To anticipate an output value (y), input values (x) are linearly mixed with weights or coefficient values (referred to as the Greek capital letter Beta). The modeled output value was binary (0 or 1) rather than numeric, which is a fundamental distinction from linear regression.

Below is an example logistic regression equation:

$$y = e^{\wedge} (b0 + b1*x) / (1 + e^{\wedge}(b0 + b1*x)) \quad (5)$$

Table 8 Using LR to predict different crops

Crops	Accuracy	
	Training Set	Testing Set
Rice	87.68	90.37
Ragi	91.29	91.27
Gram	86.69	87.67
Potato	**91.56**	89.19
Onion	91.47	72.35

The below-mentioned table implemented using Logistic regression for different crops prediction. We have applied LR and observed that potato gives the highest accuracy during the training period, i.e., 91.56, and Ragi gives the highest testing accuracy i.e. 91.27.

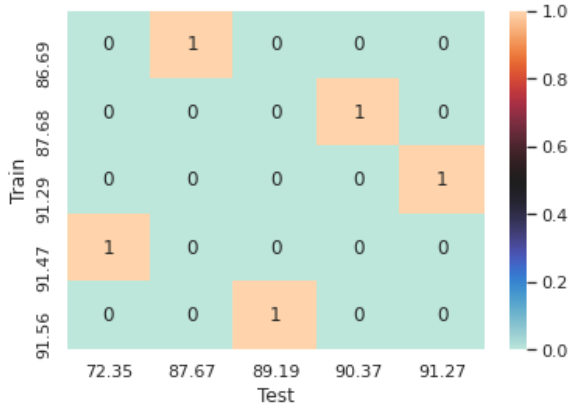


Figure 5: Logistic Regression

To enhance the data accuracy prediction based on binary classification, we have used logistic regression by considering the trained and testing dataset. As a result, it is observed that in the above Fig.5, the mean accuracy rate is high which is 91.56 %, in the case of the potato crop.

NAÏVE BAYES(NB)

A supervised learning technique that's also predicated on the Bayes theorem and is utilized to tackle categorization problems. Assumption A- is a response variable and B- is an input attribute from ML As a result of the equation, P(A|B): Given the input attributes, the conditional probability of a response variable has a certain value. The posterior probability is another name for this.

P(A)---> Response variable.

P(B) --->likelihood of the training data or evidence.

The likelihood of the training data is known as P(B|A).

$$\text{Posterior} = \text{Prior} * \text{likelihood} / \text{evidence} \quad (6)$$

The evidence can be broken down into its component elements. Now, if A and B are both independent events,

$$P(A,B) = P(A)P(B) \quad (7)$$

Hence, the equation become,

$$P(A|b1...bn) = P(A)P(b1\ldots bn|A)/ (P(b1...bn)) \quad (8)$$

Bayes theorem provides a way to calculate posterior probability P(c|x) from P(c), P(x), and P(x|c).

Look at the equation below:

P(c), P(x), as well as P(x|c) may be utilized to compute the posterior probability P(c|x) by using the Bayes theorem.

Consider the following formula:

$$P(c|x)=P(x \mid c)P(c)/P(x) \quad (9)$$

Where P (c |x) is the posterior probability of the class (c, target) given predictor (x, attributes).

P(c) represents the class prior probability.

P(x|c) denotes the likelihood, which is the probability of predicting a particular class.

P(x) represents the predictor's prior probability.

$$P(c|x)= P(x1|c)*P(x2|c)*P(x3|c)*….* P(xn|c)*P(c) \quad (10)$$

Table 9: Using Naïve Bayes to predict different crops

Crops	Accuracy	
	Training Set	Testing Set
Rice	84.58	86.77
Ragi	**90.32**	**90.12**
Gram	76.59	77.77
Potato	88.22	88.19
Onion	79.77	80.25

The above table mentioned the accuracy of the different crops. We have estimated the accuracy of the training and testing set and found that the NB algorithm yields a training accuracy of 90.32 percent and just a testing accuracy of 90.12 percent.

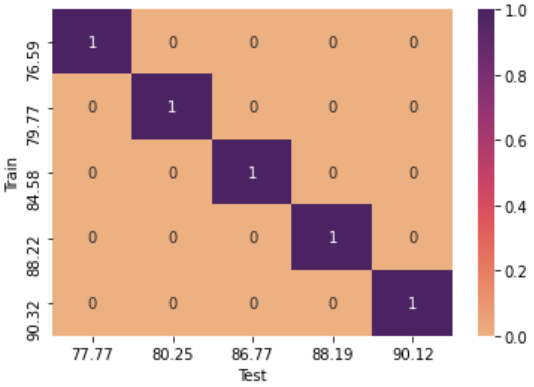


Figure 6: Naive Bayes

In the case of the Naïve Bayes classifier, the accuracy range varies from 70-90%. It is observed from the above graph that the maximum accuracy for the ragi crop is highest which is 90.12 %, and for Gram is minimum which is 76.59%.

SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) is indeed a machine learning technique that would be used to handle categorization as well as regression problems. Here each data item is plotted as a point in an-dimensional space where every feature represents the value of a specific coordinate. As a consequence, categorization may be conducted by locating the hyper-plane which distinguishes the two classes.

The below-mentioned table represents the prediction of different crops using SVM algorithms. This algorithm aims to divide the provided data into the decision surface using a decision surface splitter. The decision surface is divided into two classes of the hyperplane. One way to determine if the classification is correct is to use training data from five different crops, then use testing data from those same crops to see if the separating hyper-plane was found and applied correctly.

Table 10: Using support vector regression to predict different crops

Crops	Accuracy	
	Training Set	Testing Set
Rice	89.78	91.47
Ragi	94.39	93.97
Gram	96.29	95.67
Potato	92.26	90.19
Onion	90.47	72.35

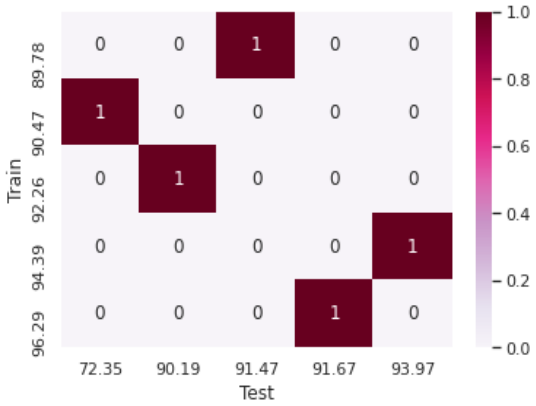


Figure 7: SVM

The above graph illustrates the accurate predictions for five different types of crops. Based on the performance evaluation metrics, it is highly recommended to go for gram crop cultivation as it results from a maximum accuracy rate of 96.29%.

The Table mentioned above uses the SVM regressor. We have estimated the different crops, and finally, we conclude that Gram's training and testing accuracy (96.29,95.67) is more than others.

DECISION TABLE

In this approach, we have considered the 100 instances and made the 31 rules. Out of these instances, 68 subsets evaluated and merit the best subset found is 85.

Table 11: Decision table

Correctly classify instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
87%	13%	0.704	0.31	0.34	70.28 %	73.71 %

In above Table 11, we have done the decision table construction algorithm where we found only 87 percent of instances were precisely categorized, while 13 percent were erroneously classified. In addition, we have estimated the different statistics measurements.

In Table-12, we have estimated the accuracy by class. Apart from this, true positive and false positive rate and their weighted average are calculated.

Table 12: Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
0.924	0.235	0.884	0.924	0.904	0.706	0.939	0.962
0.765	0.076	0.839	0.765	0.800	0.706	0.939	0.862
Weighted Avg 0.870	0.181	0.869	0.870	0.868	0.706	0.939	0.928

DECISION TREE

The decision tree is amongst the most often utilized categorization techniques for various kinds of prediction. It generates different meaningful rules with less computation. Decision trees provide a clear indication for selecting the required and important fields for classification(s)/prediction(s).

Table 13: Using DT to predict different crops

Crops	Accuracy	
	Training Set	Testing Set
Rice	87.58	92.87
Ragi	92.39	91.87
Gram	90.29	89.87
Potato	95.26	91.29
Onion	**96.27**	**94.25**

In above-mentioned Table-13 has been estimated using decision tree construction and found Onion having highest accuracy rate i.e. 96.27%, and during testing, it has obtained 94.25%. accuracy.

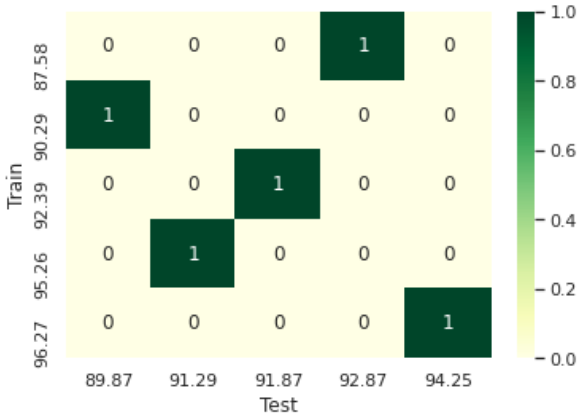


Figure 8. Decision Tree

In the case of the decision tree classifier onion crop gives the best result of 96.27%. Figure 7. predicts categorical and continuous data by considering their accurate value based on classification. To increase the efficiency of our model, we have used a pruning mechanism.

J48 PRUNED TREE

To classify different applications and get accurate classification results, the J48 method is utilized. J48 algorithm is an excellent machine learning technique for use when you want to review categorical data continuously. When provided a list of training data, C4. 5 identically develops decision trees as ID3 does, employing the idea of information entropy as the foundation. The training data is comprised of a variety of items.

```
a5 = false: c0 (44.0/2.0)
a5 = true
|  a8 = false
|  |  a9 = false
|  |  |  a2 = false: c0 (4.0/1.0)
|  |  |  a2 = true
|  |  |  |  a0 = false
|  |  |  |  |  a4 = false: c1 (2.0)
|  |  |  |  |  a4 = true: c0 (2.0)
|  |  |  |  a0 = true: c1 (5.0)
|  |  a9 = true: c1 (15.0/1.0)
|  a8 = true
|  |  a1 = false
|  |  |  a2 = false
|  |  |  |  a0 = false: c1 (4.0/1.0)
|  |  |  |  a0 = true: c0 (3.0)
|  |  |  a2 = true
|  |  |  |  a4 = false: c1 (5.0)
|  |  |  |  a4 = true: c0 (2.0)
|  |  a1 = true: c0 (14.0/2.0)
```

The count of leaves developed is 11, as well as the tree, reaches 21 in size. The table below indicates how many crops correctly classified and incorrectly classified instances during the training as well as a testing phase. 93% found correctly classify the instances, and only 7% incorrectly classified instances. Apart from these, the statistical measurement is done and represented in the tabular format as shown below.

Table 14: Statistical measurement representation

Correctly classify instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
93%	07%	0.8406	0.1211	0.2461	26.9304 %	51.9514 %

Table 15: Detailed Accuracy By Class using **J48 pruned tree**

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Weighted Avg 1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

We used the J48 pruned tree in the above table and estimated the true positive and false-positive rates. Both TP and FP give 1.00. We also found the precision experimentally, recall F-Measure, MCC, ROC curve PRC gives 1.00.

Confusion Matrix for **J48 pruned tree**

```
a  b<-- classified as
66  0 |  a = c0
0 34 |  b = c1
```

REPTree

Random forests are machine learning techniques capable of performing both classification and regression using a method of bootstrap aggregation on several decision trees. Random forests are decision trees that aggregate the decisions of many decision trees. It makes use of the supervised learning approach. Every tree in the random forest voted. The total count of votes cast by all of the forest's trees decides the ultimate decision. Trees are generated at random by looking for and picking the optimum split for each node. In decision trees, regression is implemented by modeling each class's estimated probability against the case number. In random forests, the probability of a sample is computed using the discriminant function, as shown in equation(10).During experiment we have taken 100 observations

$$gc(x)=1/t/\Sigma\ ti=1\ \hat{p}(ci\ ,\ vi\ (x)) \qquad (11)$$

$\hat{p}$ (c|vi (x)) indicates the probability of x belonging to class c. The tree is built by estimating the probability of x belonging to class c.

Table 16: Statistical measurement of REPTree

Correctly classified instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
74%	26%	0.4692	0.3125	0.3953	69.46%	83.43%

Table 17: Detailed Accuracy by Class using REPTree

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
0.636	0.059	0.955	0.636	0.764	0.551	0.551	0.847
0.941	0.364	0.571	0.941	0.711	0.551	0.789	0.558
Weighted Avg 0.740	0.162	0.824	0.740	0.746	0.551	0.789	0.749

Confusion Matrix For **REPTree**

42 24 |a = c0
2 32 |b = c1

Random Forest Tree

Random Forest Regression is a supervised EL mechanism basically to address classification and regression problems to improve the accuracy rate in prediction. Each tree in the RF produces a class prediction, and also the category with the highest votes becomes our model's prediction. In our work, we utilized scikit-learn to implement the RF model to experiment over 05 types of crops represented by the input characteristics for area and productivity. Hence observed production grades classification accuracy being accurate on both the training and testing among these 05 different crops, as shown in Table 18.

Table 18: Random forest regression has been utilized to forecast the yield of diverse crops

Crops	Accuracy	
	Training Set	Testing Set
Rice	97.98	92.87
Ragi	96.39	92.87
Gram	91.29	89.87
Potato	96.26	92.29
Onion	**99.27**	**98.25**

After predicting the accuracy of five different crops, we have estimated the statistical measurement. Here we have estimated MAE, RMSE, RAE, RRSE, and kappa statistics.

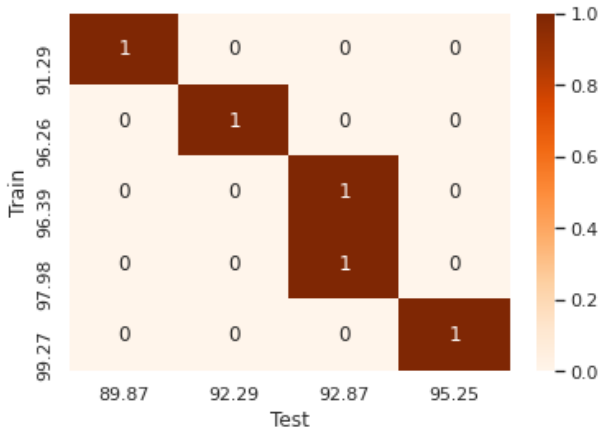


Figure 9: Random Forest

The above Figure 9 emphasizes crop prediction by using random forest classifiers. Depending on the algorithm's effectiveness, it is discovered that onion crop results in high accuracy of 99.27%, which is plotted by considering the train and test data

Table 19: Statistical measurement for Random Forest tree

Correctly classify instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
100%	0%	1	0	0	0	0

In the above table, the correctly classified instances are 100%, and none of them are incorrectly classified.

Table 20: Detailed Accuracy By Class using Random Forest Tree

TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Weighted Avg 1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The above table represents the detailed accuracy by class using Random Forest Tree where different accuracy measurements are done.

Confusion Matrix for RandomTree

a b<-- classified as
66 0 | a = c0
0 34 | b = c1

7.13 Neural Network

The Artificial Neural Network (ANN) has indeed been regarded as a real learning strategy, with its architecture inspired by the HBNS. It has is its novel information processing architecture. The fundamental building block of an ANN is a neuron that is widely connected structurally like a biological neuron in human beings.it stimulates the human brain like learning from previous experiences or examples. Sophisticated approach. However, computers can be far more useful whether they can learn from experience instead of demanding explicit direction. In essence, systems will become more beneficial if they're being trained effectively. ANNs cannot be designed to perform a specific task. The training data should be carefully picked, or else the network may perform erroneously. One negative point may be the lack of interpretability, as the network solves the problem autonomously.

RNN-(Recurrent Neural Networks) –This network is just like human brain very usefull in short term memory context solve sequence handling in a great way. .RNN generallymemorise things for short durations of time. That might be replicable, however information may be lost if a huge quantity of words are given into it.

$$h(t)=fc\ (h(t-1),x(t)) \quad (12)$$

Where h(t)-→ a new state
fc->function with parameter c
h(t-1)→old state,
x(t))→input vector at time step t

Deep Learning

Deep learning is an approach that can assess raw information to evaluate if it is adequate for categorization or regression. Various deep learning methods are available, such as CNN, the PNN, the Recurrent Neural Network (RNN), etc.[43]

7.14.1 LSTM: It is a subtype of the RNN network which used to process longer inputs. Typically, an LSTM has four gates: forget, input, cell state, and output. With distinct activity in which the cell state stores all of the input sequence information, others are utilized to regulate the input and output activities. An RNN is appropriate for time-series data because RNN models the linkages between previous data and possible future. Fig. 2a shows an RNN with a "long dependency" problem. As a result, it is decided that the LSTM network, which can learn long-term dependencies, shall be used in the model. In Fig. 2b, LSTM architecture is depicted, in which there are three hidden states, three gates, and one vanilla RNN cell (as well as the vanilla RNN cell in Fig. 2a). LSTMs are a promising solution to sequence and time series related problems LSTM makes modest changes to the data by multiplying and adding. A paradigm for information flows emerges in LSTM cell states. In this manner, LSTMs may preferentially recall or forget data. There are 3 distinct dependencies on the information at a specific cell state.

Such dependencies might well be deployed to any issue as follows:

1. The preceding time stamp step's cell state.
2. The previous cell's hidden state
3. the current time step input

Table 21: For the Execution procedure of LSTM

1. Keras defines a neural network as a sequence of layers.
2. Compile the network that has several parameters that must be met.
3. Fit the network that requires both an input patterns matrix X and a corresponding output patterns array y as well as the given training data.
4. Fit the network with the supplied input patterns X as well as output patterns y.
5. Predict in a format specified by the network's output layer.

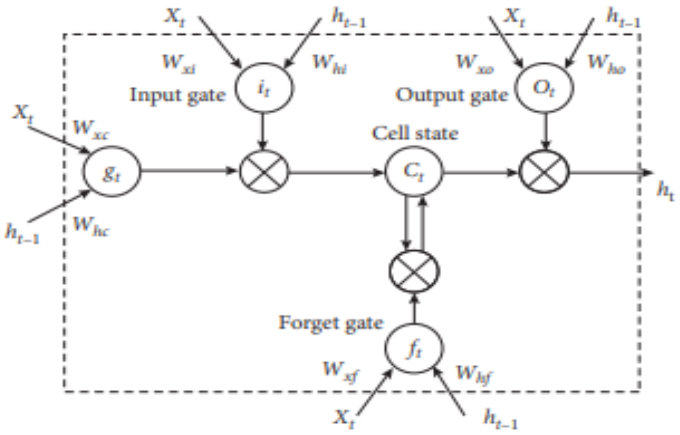


Figure 10: Illustrates the structure of a basic LSTM unit.

Table 22: Compares distinct activation functions including their test accuracy depending on the hidden layer.

Hidden Layer	Activate Function	Hidden Neurons	Test accuracy
2	Sigmoid	6	90.53%
3	Sigmoid	7	92.85%
4	Sigmoid	8	99.71%
5	Sigmoid	11	93.71%
2	ReLU	6	80.47%
3	ReLU	7	82.46%
4	ReLU	8	85.68%
5	ReLU	11	82.72%
2	Tanh	6	69.34%
3	Tanh	7	71.86%
4	Tanh	8	77.5%
5	Tanh	11	78.29%

The dataset is read using pandas after the packages are imported, and then it is submitted to analysis. Preprocessing entails data encoding, missing value filling, and feature scaling, as explained in the methodology section's data preprocessing section. Following the completion of preprocessing, the dataset gets partitioned into two sections: training as well as testing. It is divided such that 80 percent of the data has been utilized for testing, whereas the remaining 20 percent has been used for neural network training. To carry out the implementation, the Neural Network incorporates many classes of algorithms. One such example is multi-layer perceptron is the class utilized in this example (MLP).

ACTIVATION FUNCTION

Activation function plays a major role in artificial neuron n/w.when x1 and x2 are inputs, w1 and w2 are the neuron's weights, Neuron's output is mentioned as.

$$Y= \text{sum(weights *inputs)} + \text{bias} \quad (13)$$

Where y signifies the output of a weighted input value, weights represent the value assigned to a neuron during processing, bias represents the adjustment parameter. The neuron's output can range between - and + infinity. Each layer uses an activation function to transfer the input value to the (0, 1) range and generate the required output. The activation function reveals whether or not a neuron is activated. Tanh, ReLU, and Sigmoid are the three most frequently used activation functions.

TEST ACCURACY ACTIVATION FUNCTION

In the prototype model, three activation functions have been chosen for comparison. The sigmoid, ReLu, as well as Tanh activation functions, with hidden layers of 2, 3, 4, as well as 5, are examples of test accuracy. While contrasting the 3 activation functions, sigmoid, ReLU, as well as tanh, it is discovered that sigmoid achieves the maximum accuracy of 99.71 percent with four hidden layers. The sigmoid function can be applied when a requirement to forecast a result based on a probability measure arises. For example, the sigmoid curve has a value between 0 and 1. When the value is less than 0.5, the sigmoid activation function predicts that the output will be zero, and when the value is larger than 0.5, the outcome will be one.

Table 23: Comparison of Each Model

Name of the model	Mean Absolute Error
RNN	22.16
LSTM	34.15

In the above table, RN N obtained the MAE is 22.16 and LSTM is about 34.15.

CONCLUSION

This study addresses one of the most significant issues that Indian farmers face: determining which crop will produce the best results. Using machine learning processes, the system will plan and grow a recommendation model to provide crop recommendations based on geological and climatic parameters. The recommendation crop system has been designed in such a way that it takes into account. The collection contains data on five distinct crops, including rice, ragi, Gram, potato, and Onion. There are different machine learning as well as deep learning techniques were utilized and found random forest 99.27 % at the training set and 98.25% testing set. The difference between training and testing set results is 1.02. Similarly, when compared with sigmoid, ReLu, and tanh activation it is found that sigmoid achieves the maximum accuracy of 99.71% with four hidden layers. Previously, Random Forest models were proven to be the most accurate in predicting crop yield. We saw the same thing. Sigmoid achieves 99.71 percent accuracy with four hidden layers compared to ReLu and tanh.

REFERENCES

1. T. Van Klompenburg, A. Kassahun, C. Catal. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture* **2020**, 177, 105709.
2. D. Paudel, H. Boogaard, A. de Wit, S. Janssen, S. Osinga, C. Pylianidis, I. N. Athanasiadis. Machine learning for large-scale crop yield forecasting. *Agricultural Systems* **2021**, 187, 103016.
3. Y. Wang, Z. Zhang, L. Feng, Q. Du, T. Runge. Combining multi-source data and machine learning approaches to predict winter wheat yield in the conterminous united states. *Remote Sensing* **2020**, 12(8), 1232.
4. Y. Guo, Y. Fu, F. H ao, X. Zhang, W. Wu, X. Jin, J. Senthilnath, (2021). Integrated phenology and climate in rice yields prediction using machine learning methods. Ecological Indicators, 120, 106935.
5. N. Bali, A. Singla. Emerging Trends in Machine Learning to Predict Crop Yield and Study Its Influential Factors: A Survey. *Archives of Computational Methods in Engineering* **2021**, 1-18.
6. S. Khaki, L. Wang, S. V. Archontoulis. A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science* **2020,** 10, 1750.
7. Z. Chu, & J. Yu. An end-to-end model for rice yield prediction using deep learning fusion. *Computers and Electronics in Agriculture* **2020**, 174, 105471.
8. J. C. Zhao, J. F. Zhang, Y. Feng, J. X. Guo. The study and application of the IOT technology in agriculture. In *2010 3rd international conference on computer science and information technology* **2010,** 2, 462-465.
9. R. Varghese, S. Sharma, Affordable Smart Farming Using IoT and Machine Learning, *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India* **2018**, 645-650.
10. J. G. N. Zannou, V. R. Houndji. Sorghum Yield Prediction using Machine Learning. *2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART), Paris, France* **2019**, 1-4.
11. G. M. Fuady, A. H. Turoobi, M. N. Majdi, M. Syaiin, R. Y. Adhitya, I. Rachman, R. T. Soelistijono. Extreme learning machine and back propagation neural network comparison for temperature and humidity control of oyster mushroom based on microcontroller. *In 2017 International Symposium on Electronics and Smart Devices (ISESD)* **2017**, 46-50.
12. F. Balducci, D. Fomarelli, D. Impedovo, A. Longo, G. Pirlo. Smart Farms for a Sustainable and Optimized Model of Agriculture. *2018 AEIT International Annual Conference, Bari* **2018**, 1-6.
13. S. Gertphol, P. Chulaka, T. Changmai. Predictive models for Lettuce quality from the Internet of Things-based hydroponic farm. *2018 22nd International Computer Science and Engineering Conference (ICSEC), Chiang Mai, Thailand* **2018**, 1-5.
14. Z. Doshi, S. Nadkarni, R. Agrawal, N. Shah. AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India* **2018**, 1-6.
15. B. Fabrizio, I. Donato, P. Giuseppe. Machine Learning Applications on Agricultural Datasets for Smart Farm Enhancement. *Machines* **2018**, 6, 38.
16. M. R. S. Muthusinghe, S. T. Palliyaguru, W. A. N. D. Weerakkody, A. H. Saranga, & W. H. Rankothge. Towards smart farming: accurate prediction of paddy harvest and rice demand. In *2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)* **2018**, 1-6.
17. J. Treboux, D. Genoud. Improved Machine Learning Methodology for High Precision Agriculture. *2018 Global Internet of Things Summit (GIoTS), Bilbao* **2018**, 1-6.
18. C. N. Vanitha, N. Archana, R. Sowmiya. Agriculture analysis using data mining and machine learning techniques. In *2019 5th international conference on advanced computing & communication systems (ICACCS)* **2019**, 984-990.
19. M. Dholu, K. A. Ghodinde. Internet of Things (IoT) for Precision Agriculture Application. *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli* **2018**, 339-342.

20. S. S. Shinde, M. Kulkarni. Review Paper on Prediction of Crop Disease Using IoT and Machine Learning. *2017 International Conference on Transforming Engineering Education (ICTEE), Pune* **2017**, 1-4.
21. H. Park, E. JeeSook, S. Kim. Crops Disease Diagnosing Using Image-Based Deep Learning Mechanism. *2018 International Conference on Computing and Network Communications (CoCoNet), Astana* **2018**, 23-26.
22. N. Kitpo, Y. Kugai, M. Inoue, T. Yokemura, S. Satomura. Internet of Things for Greenhouse Monitoring System Using Deep Learning and Bot Notification Services. *2019 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA* **2019**, 1-4.
23. A. Sarangdhar, V. R. Pawar. Machine learning regression technique for cotton leaf disease detection and controlling using IoT. *2017 International conference of Electronics, Communication, and Aerospace Technology (ICECA), Coimbatore* **2017**, 449-454.
24. S. Sharma, G. Rathee, H. Saini. Big Data Analytics for Crop Prediction Mode Using Optimization Technique. *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan Himachal Pradesh, India* **2018**, 760-764.
25. Konstantinos G. Liakos, B. Patrizia, M. Dimitrios, P. Simon, B. Dionysis. Machine Learning in Agriculture: A Review, *Sensors* **2018**, 18, 2674.
26. U. Inyaem. Construction Model Using Machine Learning Techniques for the Prediction of Rice Produce for Farmers. *2018 IEEE 3rd International Conference on Image, Vision, and Computing, Chongqing* **2018**, 870-874.
27. M. T. Shakoor, K. Rahman, S. N. Rayta, A. Chakrabarty. Agricultural production output prediction using Supervised Machine Learning techniques. *2017 1st International Conference on Next Generation Computing Applications (NextComp), Mauritius* **2017**, 182-187.
28. S. Yahata, T. Onishi, K. Yamaguchi, S. Ozawa, J. K itazono, T. Ohkawa, H. Tsuji. A hybrid machine learning approach to automatic plant phenotyping for smart agriculture. In *2017 International Joint Conference on Neural Networks (IJCNN)* **2017**, 1787-1793.
29. N. Materne, M. Inoue. Potential of IoT System and Cloud Services for Predicting Agricultural Pests and Diseases. *2018 IEEE Region Ten Symposium (Tensymp), Sydney, Australia* **2018**, 298-299.
30. J. Chopda, H. Raveshiya, S. Nakum, and V. Nakrani. Cotton Crop Disease Detection using Decision Tree Classifier. *2018 International Conference on Smart City and Emerging Technology (ICSCET), Mumbai* **2018**, 1-5.
31. S. Gertphol, P. Chulaka, T. Changmai. Predictive models for Lettuce quality from the Internet of Things-based hydroponic farm. *2018 22nd International Computer Science and Engineering Conference (ICSEC), Chiang Mai, Thailand* **2018**, 1-5.
32. K. Wongpatikaseree, N. Hnoohom, and S. Yuenyong. Machine Learning Methods for Assessing Freshness in Hydroponic Produce. *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Pattaya, Thailand* **2018**, 1-4.
33. M. I. Alipio, A. E. M. Dela Cruz, J. D. A. Doria, R. M. S. Fruto. A smart hydroponics farming system using exact inference in Bayesian network. *2017 IEEE 6th Global Conference on Consumer Electronics (GCCE), Nagoya* **2017**, 1-5.
34. S. Prakash, A. Sharma, S. S. Sahu. Soil Moisture Prediction Using Machine Learning. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore* **2018**, 1-6.
35. S. Verma, R. Gala, S. Madhavan, S. Burkle, S. Chauhan, C. Prakash. An Internet of Things (IoT) Architecture for Smart Agriculture. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India* **2018**, 1-4.
36. M. AshifuddinMondal, Z. Rehena. IoT Based Intelligent Agriculture Field Monitoring System. *2018 8th International Conference on Cloud Computing, Data Science & Engineering, Noida* **2018**, 625-629.
37. P. Srinivasulu, M. S. Babu, R. Venkat, and K. Rajesh. Cloud service-oriented architecture (CSoA) for agriculture through the internet of things (IoT) and big data. *2017 IEEE International Conference on Electrical, Instrumentation, and Communication Engineering (ICE ICE), Karur* **2017**, 1-6.
38. M. H. Jumat, M. S. Nazmudeen, A. T. Wan. Smart farm prototype for plant disease detection, diagnosis & treatment using IoT device in a greenhouse. *7th Brunei International Conference on Engineering and Technology 2018 (BICET 2018), Bandar Seri Begawan, Brunei* **2018**, 1-4.
39. F. A bbas, H. Afzaal, A. A. Farooque, & S. Tang,. Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy* **2020**, 10(7), 1046.
40. X. E. Pantazi, D. Moshou, T. Alexandridis, R. L.Whetton, A. M. Mouazen. Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and electronics in agriculture* **2016**, 121, 57-65.
41. M. Shahhosseini, R. A. Martinez-Feria, G. Hu, S. V. Archontoulis. Maize yield and nitrate loss prediction with machine learning algorithms. *Environmental Research Letters* **2019**, 14(12), 124026.
42. A. Kayad, M. Sozzi, S. Gatto, F. Marinello, F. Pirotti. Monitoring within-field variability of corn yield using Sentinel-2 and machine learning techniques. *Remote Sensing* **2019**, 11(23), 2873.
43. K. Bhosle, B. Ahirwadkar. Deep learning Convolutional Neural Network (CNN) for Cotton, Mulberry and Sugarcane Classification using Hyperspectral Remote Sensing Data. *J. Integr. Sci. Technol.* **2021**, 9 (2), 70–74.