Technical Journal, University of Engineering and Technology (UET) Taxila, Pakistan Vol. 29 No. 1-2024
ISSN:1813-1786 (Print) 2313-7770 (Online)

Implementing Machine Learning Models – An Analysis of Agricultural Weather and Soil Data

S. Iqbal[1], S. Nizamani[2], I. Qasim[3], S. Siraj[4], M. A. Soomro[5], A. Ayaz[6]

[1]*Bahauddin Zakariya University, Multan, Pakistan*
[2,4,5]*Quaid-e-Awam University of Engineering, Sciences & Technology, Nawab Shah, Sindh, Pakistan*
[3]*Mehran University of Engineering and Technology Jamshoro, Sindh, Pakistan*
[6]*Shaheed Benazir Bhutto University, Shaheed Benazirabad, Sindh, Pakistan*

[1]shahid.iqbal@bzu.edu.pk

Abstract- Agricultural crop monitoring and data analysis has been challenging part in agronomics. The IoT based systems give a better solutions to monitor crop and collect data for further analysis. This study bases on soil and weather data collection and prepare a dataset. The dataset is constructed on the basis of four important parameters of weather and soil including temperature, soil pH level, humidity, soil moisture and smoke. The two major crops monitored named Wheat and Cotton for dataset construction. The data set attributes are selected by applying principal component analysis. The highest ranked attribute is selected for data analysis. Furthermore, machine learning models applied for the better results analysis comparatively. ML models parameters are Correlation coefficient, Mean Absolute error, Root Mean Squared Error, relative absolute error and Root relative squared error. The results are obtained by Applying WEKA Data Mining Tool. The results showed that Random Forest Classifier performed better results as compared to KNN and Decision Tree classifiers.

Keywords- Crop Yield, Smart Agricultural, Machine Learning

I. INTRODUCTION

In Pakistan agriculture is imperative industry for secondary economic and gross domestic production. The total GDP of the Pakistan agriculture is about 26% according to Pakistan bureau of statistics. Agriculture in Pakistan preserves prominence with effect of different economic and industrial growth. Many techniques introduced to improve productivity and quality production of the agriculture crop yield [1]. Computer based techniques imposed to apply rapid monitoring related to environment, soil and water parameters [2]. These all parameters have numerous implication and direct influence on the crop yield production. Technological methods for recording and monitoring real time parameters have been used to improve quality of products of agriculture crop [3]. This has been achieved by implementing WSN and IoT devices in the region where agriculture crops are cultivated. In the last few years, machine learning and computer vision methods have been applied to improve agricultural monitoring and predictions[4]. Semi-structure and raw data obtained by deployment of different sensors in agriculture field has been used to predict the important factors of crop yield.

Modern areas of research have improved methods of agricultural cultivation, monitoring of crop fields and crop yield prediction. Study plays important role in human life for improving reliability and efficiency of automated task based approach [5]. In Pakistan most focused area of research is agriculture due to its greater impact on GDP and import / export business. Variety of crop products are cultivated in Pakistan, however the secrops often lack expected yield production due to traditional methods of crop monitoring. The factors which cause low crop yield include dull soil, plant nutrition, environmental pollution and water pollution etc. By applying IoT trends of monitoring agricultural parameters, crop products can be significantly increased [6].

In agriculture, it is generally not trivial to predict the crop yield [7], as it depends on a number of factors such as daily mean temperature, humidity, green house effects, evaporation rate, wind speed, water quantity, solar radiation, etc. Studies have shown that determining crop yield prediction is beyond an intelligent guesswork and entails an in-depth analysis of many crucial factors and their underlying relationship. Recently, there is a growing interest in applying machine learning techniques to monitor a number of factors which play a vital role for improving crop yield [8]. In this paper, we first find out the parameters which play a key role in determining the crop yield. Subsequently, our aim was to gather the required dataset by deploying relevant sensors in crops. The dataset was used to train a machine learning algorithm to predict crop yield. State-of-the-art three machine learning approaches were used to evaluate the performance of each classifier on the collected dataset.

II. PROBLEM STATEMENT

The data of soil and weather is always important for crop cultivation and directly impacts on crop yield. Whereas the high temperature is very critical some time which is not easy to measure by human guesses. In this regard most of other parameters such as weather and soil parameters are crucial to manage time by time in whole crop cultivation process. In this regard measuring the data with time saving and economic way is hard to manage by traditional methods. Whereas, the automation of IoT based devices provide rapid, easy and accurate measure of data without matter of time or location. In this regard the obtained data in the form of dataset is also crucial to analysis impact of weather and soil data each other and how it can effect directly on the crop growth and yield. This can analyzed by applying machine learning approaches on the data.

III. OBJECTIVES

The major contributions of this research paper are asunder:

1. To find out the parameters affecting a crop's yield
2. To construct a dataset with respect to the discovered parameters
3. To train & Test and compare machine learning algorithms on constructed dataset

IV. LITERATURE REVIEW

Machine learning approaches divulge vital role in medical science applications for optimization of diagnostic activities. The diagnostic activities such as identifying diseases diagnostics, manufacturing of drugs, medical image diagnostics, personalized medicine, smart health records and clinical trial research are major part of machine learning research and development in health science [4]. KNN identify missing electrocardiogram signals (ECG) efficient method for missing data instances. Heart pulse scan be very sensitive to diagnose and identify the critical fluctuations in heart rate. The study [5] focused on KNN to necessitate missing instances data set on the way to increase effectiveness and accurate perceptions monitored missing electrocardiogram. Machine learning models compared on electrocardiogram data values specially missing values in dataset. Missing values are compared such as "zero method", "Mean method", "PC based method" and RPCA – Based method. After comparison of model finally KNN-based classification algorithm called modified weighted KNN classifier (MKDF-WKNN) fitted to handle imbalance electrocardiogram (ECG) data. Human body suffers for several biological factors by attacking viruses on different organs. Recently COVID – 19 global impacts was very confusing and alarming situation for human life [9]. The study is about COVID-19 detection strategy named as COVID patient detection strategy (CPDS). The hybrid features selection methodology is adopted (HFSM) used to extract features, captured by using computed tomography (CT) [10]. Features extractions methods named wrapper and filter methods are used to extract the features. Data imparts important countable deliberation for enhanced experiment results and effective predictions. The study compiled a data set for retrieval of the meaning full information. Agricultural research contributes many new challenges and application by implementation of machine learning methods. Machine learning contribution in agricultural fields likes pecies breading, species recognition, soil management, water management and irrigation controlled scheduling, yield prediction, crop quality, diseases detection, weed detection are remarkable advancement [11]. Machine learning in agriculture science growing interest entails new application sand challenges for there searchers. Machine learning model sand approaches are implemented to increase crop health assessment, crop yield, water quality monitoring, plant disease identification and classification for productive crop production[8]. Agriculture recommendation system for crop soil, fertilization, water requirements, weather update are introduced for better improvement in agricultural crop methods. KNN based recommendation system is proposed[12] and implemented for efficient management of agriculture quires for formers to make aware about government schemes, loan management and fertilizers management. Plant disease identification[13] is major are a concern for researchers and also important for better crop management. The tomato crop leaf disease classification[14] is proposed for identification and classification of leave disease. GLCM, Gabor and color features are used for classification whereas 600 hundred health and unhealthy plant leave are used for sampling. The classification of unhealthy and healthy are used by implementing PNN and KNN models. Crop yield prediction in major area of research in machine learning. It depend on weather, soil, water and plant disease factors. The study focused on the proximal sensing techniques and crop yield prediction based on KNN, Linear regression (LR), Elastic net (EN)and support vector machine [15]. The soil and crop properties are obtained from remote sensing techniques.

Automation industry is purely basis on accurate decision making with smart controlled system. The factory digitization, inventory management safety and security, quality control, packing optimization and logistic and supply chain optimization are major applications of IoT. Whereas, as machine learning in combination with IoT can improve the pattern recognition, self-manufacturing, automated data optimization and efficient predictions can improve substantial performance of decision making. Industrialization and home automation is one of most focused area of study in these days. The daily routine activities automation are widely commercialized for providing new opportunities of business. The world leading companies and firms are providing automation services which bases on IoT devices models. Machine learning with efficient predictive analysis provides rapid solution of rational activities on regular basis [16]. The trending activities in development of smart home automation and industrialization is now an open industry for both researchers and commercialization. Home automation is very effective to manage daily routine activities despite of this it creates challenges for maintained and controlled system such as intelligence decision making, authentication problems, secure identification, data security, content uploading, end-to-end encryption, protocol communication and many network security problems [17]. The waste Management is also hot topic for researchers in smart cities. The population growing rate creates problem to manage waste in large cities. Machine learning provide opportunity with collaboration of IoT technology can make efficient waste management solution. Human guided systems intensified due to automated tasks activities topographies and maps in densely populated areas. IoT based system with smart Machine learning decision making is very helpful in tourisms [18]. Study is observed on the IoT with modeling of machine learning models by implementation of 5G technology. The smart human – guided system based on classification of tourism selection behaviors. KNN based machine learning classification is used to help tourists for selection of tour spots. The data is important aspect of machine learning in shape of huge dataset. The data set contains many complexities and irregularities which can create dissimilarities and unwanted outcome by predicting any critical output. The outliers can create many vulnerable output if it is majority form [19]. The study [20] to detect outlier by applying supervised and unsupervised machine learning models. Isolation forest (iForest) is used as an unsupervised method whereas Support vector machine (SVM), K-Nearest Neighbor, and random forest is used as supervised learning models. The number of parameters are such as precision, recall, F1 Score, ROC-AUC curved are used for evaluation of supervised and unsupervised machine learning models. The trend of business is changed due to millions of smart devices and online web services. The automation industry brought many new challenge as well as drawback which can be fulfilled by applying IoT based solution. The water quality assurance, fast food, quality manufacturing of daily routine products, horizontal and vertical systems, autonomous robots, cyber security technology, the cloud and additive manufacturing are major applications of industry 4.0. The automation challenges and risk factor are needed to make fault diagnostic capability[4]. The study proposed a framework named ensemble sparse supervised learning model (ESSM). This model base on typical deep learning model as two parts featured learning and model learning. The traditional manufacturing conversion (industry1.0) to smart advanced manufacturing (industry 4.0) is remarkable contribution of computational methods [21]. Combined machine vision and machine learning methods are generalized for manufacturing and material handling. These different effects can be monitored and controlled by applying IoT and machine learning smart paradigm [22].The waste management techniques can provide productive results to conversion from waste to plant energy production.

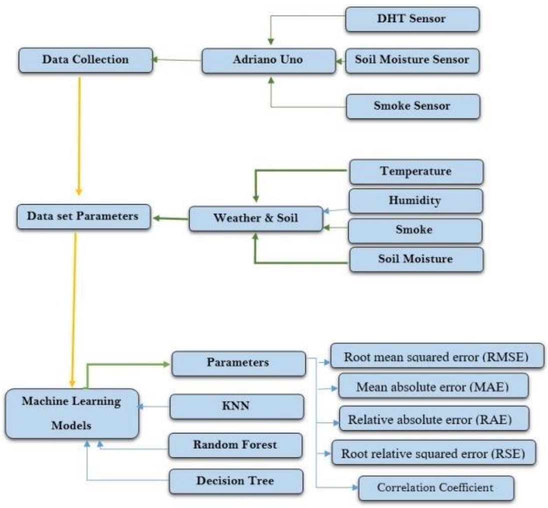


Figure 1 Research Methodology of research

V. RESEARCH METHODOLOGY

Data Collection

Data was collected by deploying IoT based sensor module. The module was developed by using Adriano Uno with four sensors connected for the weather and soil data collection. Moreover, four nodes were developed with Adriano board and sensors. The sensors used were DHT for temperature and humidity, Q11 for smoke, and soil moisture. The sensor nodes were deployed at different places of agricultural field. These sensors are connected with Wi-Fi module which connect edin each Adriano Model to send data remotely over cloud.

Data Set Statistics

Dataset is obtained at the real time from the sensors deployment during crop and wheat seasons. The dataset consists of temperature, humidity, smoke, and soil moisture. Table 1 illustrates the parameters of the dataset.

Table1Parametersof the dataset

B	C	D	E
Soil Moisture	**Smoke**	**Humidity '%'**	**Temperature 'C'**
232	664	35.6	33.1
232	662	36.7	33.1
232	663	36.7	33.1
231	663	36.8	33.1
231	662	36.8	33.1
232	664	36.9	33.1
230	663	37	33.1
231	662	36.9	33.1
231	662	37	33.1
231	662	37	33.1
231	663	37	33.1
234	663	37	33.1
233	663	37	33.1
233	663	37.1	33.1
234	663	37	33.1
239	663	37	33.1
234	663	37.1	33.1
234	661	37.1	33.1
234	662	37.1	33.1
237	661	37.1	33.1
244	662	37.1	33.1
245	661	37.1	33.1
243	663	37.1	33.1
242	661	37.1	33.1
241	663	37.1	33.1
242	661	37.1	33.1
242	663	37.1	33.1
242	662	37.1	33.1
245	662	37.3	33.1
244	663	37.4	33.1
247	662	37.3	33.1

Wednesday | thursday | Friday | Saturday | Sunday | **Monday**

Data Limitation

This study is bases on the following limitations

- Data is obtained from the real time scenario of wheat and cotton crop
- The IoT based module is used for auto data collection
- The data is collected in day time
- The data is in number format
- The data comprises weather and soil parameters values

Machine Learning Classifiers

To predict the yield crops based on the weather and soil data, four different machine learning classifiers named KNN, RF and DT were used to train and evaluate the statistical data by different parameters such as correlation coefficient, mean absolute error (MAE), root means quared error (RMSE), relative absolute error(RAE) and root relative squared error(RSE).

3.5 Attribute Selection using Principle Component Analysis (PCA)

Table 2 shows the attribute details in the dataset where simple parameters are used to evaluate their minimum, maximum, mean and standard deviation. The attributes shown in Table 2 are the dataset parameters which were used in the dataset for further analysis and evaluation of crop yield prediction.

Table. 3 Parameters of dataset Ranked attribute

Parameters	Min	Max	Mean	StdDev
Soil moisture	250	368	292.582	41.85
Smoke	611	864	691.509	42.513
Humidity	32	61.8	45.244	9.793
Temperature	28	45	36.417	3.683

Table 3 presents the parameters detail and their measure of tendencies. In this table the parameters values are considered as Min, Max, and Mean and Standard deviation.

Table. 4 Parameters of dataset and Measure of Tendencies

Parameters	V1	V2	V3
Soil Moisture	-0.6373	-0.0252	-0.2454
Smoke	-0.4735	0.317	0.8114
Humidity	-0.6077	-0.1934	-0.3856
Temperature	0.0178	0.9281	-0.3642

Table 4 presents the parameters ranked attribute which is given as V1, V2 and V3. The ranked attributes shows variable ranking for better attribute selection.

Ranked Attributes

When working with data it is important to rank attributes or variables based on their importance or relevance to the problem at hand. The specific attributes that should be ranked depend on the nature of data and goal of the analysis. There are some common ranking attribute methods such as correlation with target variables, feature importance, domain knowledge, information gain and variance. The Table 4 shows the soil moister, humidity, smoke and temperature values according to their rankings.

Table 4: Attributes based on their ranks

0.4259	**1**	**Soil Moisture (-0.637)**	**Humidity-0.608**	**Smoke-0.474**	**Temperature+0.018**
0.158	**2**	Temperature 0.928	Smoke+0.317	Humidity-0.193	Soil Moisture-0.025
0.0151	**3**	Smoke0.811	Humidity-0.386	Temperature -0.364	Soil Moisture-0.245

Evaluation of Machine Learning Models using WEKA Tool

For data analysis and prediction we used ML models named Random forest, Decision tree and KNN. Each model is used for 699 instances and 4 attributes. A 10-fold cross validation was used to validate the dataset. For model evaluation the correlation coefficient, mean absolute error(MAE), root mean squared error (RMSE), relative absolute error (RAE) and root relative squared error (RSE) were applied. The WEKA tool was used to find the model parameters and their characteristics. Figure 2 (a) illustrates the evaluation results of error values using decision tree classifier, while Figure 2 (b) and Figure 2 (c) illustrates the evaluation results of error values using KNN and random forest classifiers.

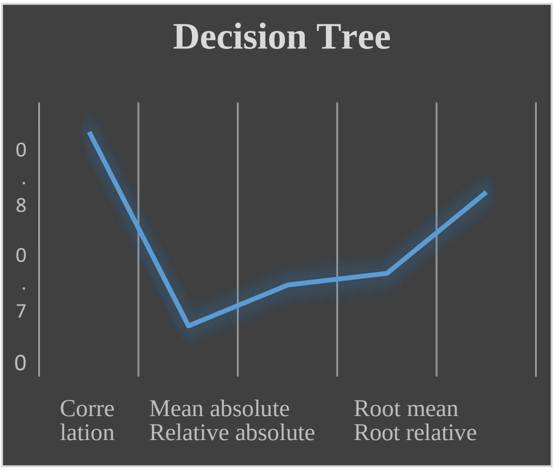


Figure2 (a): Evaluation of error values of Decision Tree

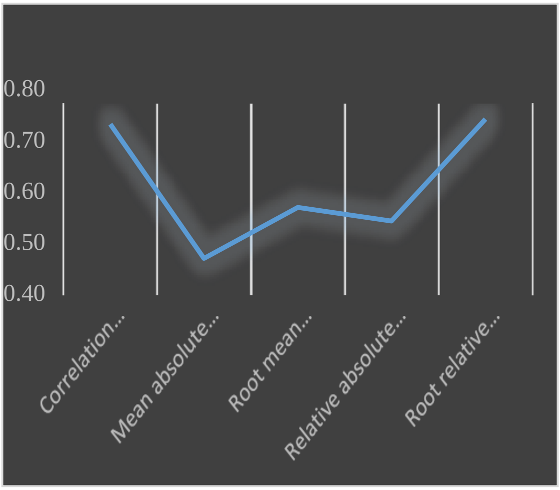
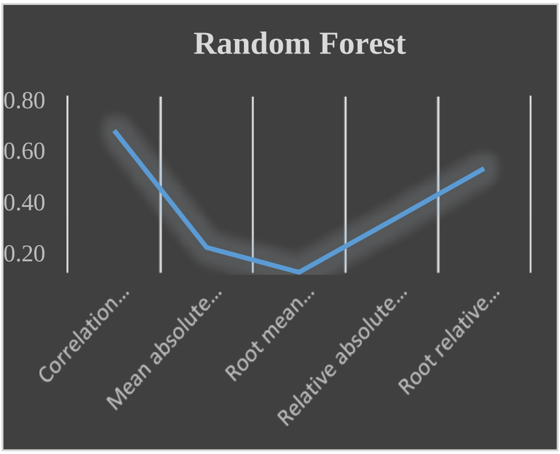


Figure 2(b): Evaluation of error values of KNN



c) Evaluation of error values of random forest

Figure 2(a), (b) and(c) illustrates model performance on different parameters.

The above table 5 shows comparative analysis of KNN, RF and DT model. The comparison is bases on MAE, RMSE, RAE and RSE. The results shows KNN gives 0.7163 of correlation coefficient, with 0.1529, 0.3669, 30.9863 and 73.8068 % of MAE, RMSE, RAE and RSE respectively. The KNN shows low correlation coefficient as compared to RF and DT model which shows good results. In this regard MAE of RF is low as compared to KNN an d and DT. Last RAE of the KNN is lowest than RF and DT. Whereas RSE of RF is lowest other two KNN and DT. Random forest classifier shows better results as it has lowest error values and is on the top of the threemodels.Whereasdecisiontreehasslightlymoreerrorvalueascomparedtotherandomforestand KNN gives highest error values than the other classifiers according to the parameters.

Table 5. Models Comparative summary

Models	**Correlation coefficient**	**MAE**	**RMSE**	**RAE**	**RSE**
KNN	0.7163	0.1529	0.3669	30.9863 %	73.8068 %
RF	0.8085	0.1415	0.2949	29.4044%	59.1931 %
DT	0.7985	0.1664	0.2993	33.722 %	60.2075 %

The table 5 shows over all summer of the KNN, DT and RF. The three models are compared according to the selected value parameters. Table5showstheevaluation results of each machine learning model, while Figure 3 illustrates the comparative results.

Comparative analysis of KNN, RF and DT model. The comparison is bases on MAE, RMSE, RAE and RSE. The results shows KNN gives 0.7163 of correlation coefficient, with 0.1529, 0.3669, 30.9863 and 73.8068 % of MAE, RMSE, RAE and RSE respectively. The KNN shows low correlation coefficient as compared to RF and DT model which shows good results. In this regard MAE of RF is low as compared to KNN an d and DT. Last RAE of the KNN is lowest than RF and DT. Whereas, RSE of RF is lowest other two KNN and DT.

VI. Limitation of Study

- Limited area of data collection due to stern temperature conditionin Nawab Shah District, Sindh, Pakistan
- Limited Soil Conditions and reading .i.e. having two soil types
- Limited number of parameters have been recorded and given to the classifiers for the training and testing
- Only three machine learning classifiers have been selected for the classification

VII. Conclusion

This research paper focused on the crop yield prediction based on the obtained statistics by deploying sensors in the crop field. The area of the crop covered was about two acres i.e. 87120Sq. foot. The dataset is constructed on the basis of four important parameters of weather and soil such as temperature, soil pH level, humidity, soil moisture and smoke. The two major crops selected to monitor were Wheat and the Cotton. The time duration for the data collection was about two seasons, which was around twenty four months. The parameters selected for the machine learning models were correlation coefficient, mean absolute error, root mean squared error, relative absolute error and root relative squared error. The results were obtained by applying WEKA data mining tool.

VIII. Future Work

This study can further be conducted for extension as

- The more data parameters can be added
- The models can be applied latest as used in this paper are little bit old
- The image data can also be added for crop disease monitoring, classification and detection

References

[1] R. Akhter and S. A. Sofi, “Precision agriculture using IoT data analytics and machine learning,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 8, pp. 5602–5618, 2022, doi: 10.1016/j.jksuci.2021.05.013.

[2] I. H. Sarker, “Machine Learning for Intelligent Data Analysis and Automation in Cybersecurity: Current and Future Prospects,” *Ann. Data Sci.*, vol. 10, no. 6, pp. 1473–1498, 2023, doi: 10.1007/s40745-022-00444-2.

[3] O. Folorunso *et al.*, “Exploring Machine Learning Models for Soil Nutrient Properties Prediction: A Systematic Review,” *Big Data Cogn. Comput.*, vol. 7, no. 2, 2023, doi: 10.3390/bdcc7020113.

[4] A. S. Budiman *et al.*, “Smart Dome 4.0: Low-Cost, Independent, Automated Energy System for Agricultural Purposes enabled by Machine Learning,” *J. Phys. Conf. Ser.*, vol. 2224, no. 1, pp. 0–11, 2022, doi: 10.1088/1742-6596/2224/1/012118.

[5] R. Punithavathi *et al.*, “Computer Vision and Deep Learning-enabled Weed Detection Model for Precision Agriculture,” *Comput. Syst. Sci. Eng.*, vol. 44, no. 3, pp. 2759–2774, 2023, doi: 10.32604/csse.2023.027647.

[6] D. Paudel *et al.*, “Machine learning for large-scale crop yield forecasting,” *Agric. Syst.*, vol. 187, no. June 2020, p. 103016, 2021, doi: 10.1016/j.agsy.2020.103016.

[7] H. Burdett and C. Wellen, “Statistical and machine learning methods for crop yield prediction in the context of precision agriculture,” *Precis. Agric.*, vol. 23, no. 5, pp. 1553–1574, 2022, doi: 10.1007/s11119-022-09897-0.

[8] Y. Zhang, M. Wang, D. Zhao, C. Liu, and Z. Liu, “Early weed identification based on deep learning: A review,” *Smart Agric. Technol.*, vol. 3, no. October 2022, p. 100123, 2023, doi: 10.1016/j.atech.2022.100123.

[9] W. Peng and O. Karimi Sadaghiani, “A review on the applications of machine learning and deep learning in agriculture section for the production of crop biomass raw materials,” *Energy Sources, Part A Recover. Util. Environ. Eff.*, vol. 45, no. 3, pp. 9178–9201, 2023, doi: 10.1080/15567036.2023.2232322.

[10] S. Gudadhe, A. Thakare, and A. M. Anter, “A novel machine learning-based feature extraction method for classifying intracranial hemorrhage computed tomography images,” *Healthc. Anal.*, vol. 3, no. January, p. 100196, 2023, doi: 10.1016/j.health.2023.100196.

[11] X. Pei *et al.*, “Robustness of machine learning to color, size change, normalization, and image enhancement on micrograph datasets with large sample differences,” *Mater. Des.*, vol. 232, p. 112086, 2023, doi: 10.1016/j.matdes.2023.112086.

[12] W. Wang, C. Zhang, J. Song, and D. Xu, “The impact of target price policy on cotton cultivation: Analysis of county-level panel data from china,” *Agric.*, vol. 11, no. 10, pp. 1–18, 2021, doi: 10.3390/agriculture11100988.

[13] P. Vasavi, A. Punitha, and T. Venkat Narayana Rao, “Crop leaf disease detection and classification using machine learning and deep learning algorithms by visual symptoms: A review,” *Int. J. Electr. Comput. Eng.*, vol. 12, no. 2, pp. 2079–2086, 2022, doi: 10.11591/ijece.v12i2.pp2079-2086.

[14] M. Shoaib *et al.*, “Deep learning-based segmentation and classification of leaf images for detection of tomato plant disease,” *Front. Plant Sci.*, vol. 13, no. October, pp. 1–18, 2022, doi: 10.3389/fpls.2022.1031748.

[15] I. H. Khan *et al.*, "Early detection of powdery mildew disease and accurate quantification of its severity using hyperspectral images in wheat," *Remote Sens.*, vol. 13, no. 18, 2021, doi: 10.3390/rs13183612.

[16] S. Qazi, B. A. Khawaja, and Q. U. Farooq, "IoT-Equipped and AI-Enabled Next Generation Smart Agriculture: A Critical Review, Current Challenges and Future Trends," *IEEE Access*, vol. 10, pp. 21219–21235, 2022, doi: 10.1109/ACCESS.2022.3152544.

[17] V. K. Quy *et al.*, "IoT-Enabled Smart Agriculture: Architecture, Applications, and Challenges," *Appl. Sci.*, vol. 12, no. 7, 2022, doi: 10.3390/app12073396.

[18] H. Hua, Y. Li, T. Wang, N. Dong, W. Li, and J. Cao, "Edge Computing with Artificial Intelligence: A Machine Learning Perspective," *ACM Comput. Surv.*, vol. 55, no. 9, 2023, doi: 10.1145/3555802.

[19] R. Wazirali, "An Improved Intrusion Detection System Based on KNN Hyperparameter Tuning and Cross-Validation," *Arab. J. Sci. Eng.*, vol. 45, no. 12, pp. 10859–10873, 2020, doi: 10.1007/s13369-020-04907-7.

[20] M. A. Bhatti, R. Riaz, S. S. Rizvi, S. Shokat, F. Riaz, and S. J. Kwon, "Outlier detection in indoor localization and Internet of Things (IoT) using machine learning," *J. Commun. Networks*, vol. 22, no. 3, pp. 236–243, 2020, doi: 10.1109/JCN.2020.000018.

[21] M. K. Sott *et al.*, "Precision Techniques and Agriculture 4.0 Technologies to Promote Sustainability in the Coffee Sector: State of the Art, Challenges and Future Trends," *IEEE Access*, vol. 8, pp. 149854–149867, 2020, doi: 10.1109/ACCESS.2020.3016325.

[22] K. Khosravi, F. Rezaie, J. R. Cooper, Z. Kalantari, S. Abolfathi, and J. Hatamiafkoueieh, "Soil water erosion susceptibility assessment using deep learning algorithms," *J. Hydrol.*, vol. 618, no. September 2022, p. 129229, 2023, doi: 10.1016/j.jhydrol.2023.129229.