Statistical Perspectives in Machine Learning for Crop Recommendations

Sricharani P
Professor, Department of Artificial Intelligence , Shri Vishnu Engineering College for Women , Bhimavaram, Andhra pradesh,India
charani.yashu@svecw.edu.in

Sai Sruthi A
Department of Artificial Intelligence , Shri Vishnu Engineering College for Women ,Bhimavaram,Andhra pradesh,India
saisruthiattanti@gmail.com

Sai Lakshmi Sirisha A
Department of Artificial Intelligence , Shri Vishnu Engineering College for Women , Bhimavaram, Andhra pradesh,India
lakshmiattili672@gmail.com

Sravya M
Department of Artificial Intelligence, Shri Vishnu Engineering College for Women ,Bhimavaram,Andhra pradesh,India
msravya2109@gmail.com

Gowthami Satya Sree A
Department of Artificial Intelligence, Shri Vishnu Engineering College for Women ,Bhimavaram,Andhra pradesh,India
anisettygowthami@gmail.com

Nandini D
Department of Artificial Intelligence , Shri Vishnu Engineering College for Women ,Bhimavaram,Andhra pradesh,India
dronavallinandini@gmail.com

Abstract--Agriculture stands as a vital component of human survival, contributing significantly to global economy. Even with its significance, there is a strong demand for advancements, especially with regard to crop recommendation systems. Accurate crop predictions play a vital role in boosting productivity, especially as climate change increasingly impacts crop production. Crop selection using traditional manual approaches based on soil and environmental characteristics has not worked well. Crop projections are highly dependent on a number of detailed factors, including soil properties, climatic patterns, temperature, rainfall, humidity, and geographic location. The creation of a recommendation system using different ML and DL approaches is the main goal of this chapter. Through the analysis of several criteria, the system seeks to recommend appropriate crops, assisting farmers in making well-informed decisions.This initiative holds promise for enhancing agricultural productivity and resilience in the face of evolving environmental challenges.

Keywords: *Crop recommendation, NaiveBayes(NB), Decision Trees, Random Forest(RF), Soil nutrients, Nitrogen-Phosphorous-potassium (NPK), Support Vector Machines (SVM), Correlation, Skewness, Kurtosis.*

I. INTRODUCTION

As the primary source of human nutrition, agriculture is essential to maintaining both global economic stability and food security. However, farmers face several difficulties due to the constantly shifting soil and climate patterns, which have an impact on crop sustainability and productivity. The creation of intelligent crop recommendation systems has surfaced as a viable approach to address these issues and enhance agricultural operations.The objective of this project is to design and implement a crop recommendation system that makes better use of weather and soil patterns to enhance agricultural decision-making. The system utilizes machine learning and data analytics to provide farmers customized crop selection recommendations based on soil properties like pH, pH levels, and soil type, as well as environmental factors like temperature, humidity, and rainfall.The integration of weather and soil data into the recommendation process enables a more comprehensive understanding of the agroecological conditions, thereby facilitating more accurate and informed decisions regarding crop choices. By tailoring recommendations to specific farm conditions, the system empowers farmers to optimize resource utilization, maximize yield, and adapt to changing environmental conditions. By giving farmers a trustworthy tool for crop selection based on weather and soil trends, through this project, we hope to promote environmentally friendly farming practices and smart agriculture.

II. LITERATURE REVIEW

Various machine learning techniques have been employed to address challenges such as weed and pest detection, crop production prediction, and plant leaf disease identification. Sharma et al. (2021) provided a comprehensive overview of agricultural yield globally, discussing prevalent traits and forecasting methodologies. They highlighted the potential for India to enhance agricultural output using ML and AI technologies. Ray et al. (2022) achieved remarkable accuracy rates of 99.54% and 98.52% by employing ensemble methods, majority voting, distribution analysis, and correlation analysis to classify 22 distinct crop types. Vashisht et al. (2022) proposed the use of extreme learning machines to forecast rice crop production based on factors like location, season, and cultivable area. Gupta et al. (2022) demonstrated the effectiveness of ML algorithms in segmenting large volumes of data for crop projections.

In their analysis, Van et al. (2020) identified soil type, temperature, and rainfall as the most commonly utilized features in agricultural models, with artificial neural networks (ANNs) emerging as the predominant technique. Rashid et al. (2021) explored various ML methods for predicting agricultural productivity, focusing specifically on palm oil yields. Kalimuthu et al. (2020) employed the Naive Bayes algorithm in their approach. Additionally, Sharma et al. (2021) conducted a thorough examination of ML applications in agriculture, emphasizing the potential of machine learning and computer vision in improving cattle output through the identification and management of eating disorders, reproductive patterns, and behavioral predictions.

Decision trees, a popular ML and data mining technique, have been extensively utilized in agricultural research. For

979-8-3503-7339-4/24/$31.00 ©2024 IEEE

instance, the Iterating Dichotomizer 3 (ID3), a classic decision tree algorithm, was employed to classify land capacity based on soil data from the 38 Soil Series in Maharashtra's Wardha District. Factors such as depth, slope, drainage, texture, erosion, and permeability were considered in the classification of land capacity, showcasing the versatility and utility of decision tree algorithms in agricultural applications.

III. FLOW OF PROPOSED SYSTEM

A. DATA COLLECTION

With the help of the dataset that we downloaded from kaggle.com, we were able to build a prediction model that can provide recommendations about which crops would be best to grow on a certain farm based on a number of different factors. The current dimensions of the dataset are (2200, 8). Three elements of soil composition are considered: potassium, phosphorus, and nitrogen.

Data Collection

```
In [2]: df = pd.read_csv("C:/Users/hp/Downloads/Crop_recommendation (1).csv")
        df.head()
```

Out[2]:

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

```
In [3]: df.shape
Out[3]: (2200, 8)
```

Fig. 1. Small part of Dataset

B. FEATURES OF DATASET

In a crop recommendation system, the dataset typically includes various features that capture essential information about environmental conditions, soil characteristics, and crop performance. Some common features found in such datasets may include:

Nitrogen: Given its critical function in the growth and development of crops, nitrogen is an essential nutrient.In order to maximize crop development and output, nitrogen's function in a crop recommendation system is measuring and controlling soil nitrogen levels. This involves suggesting nitrogen fertilizer plans that are suitable for the crop type, growth stage, soil nutrient level, and environmental conditions. The system strives to optimize crop output while minimizing nitrogen losses to the environment, supporting sustainable agriculture practices, by precisely advising nitrogen application rates, timing, and techniques. Furthermore, nitrogen contributes to the overall structure and function of plants by serving as a vital building ingredient for proteins, amino acids, and DNA. Farmers frequently employ fertilizers that include nitrogen to increase crop output by ensuring that crops receive enough of this vital nutrient.

Phosphorous: For crops to function properly, phosphorus is essential because it promotes root development, enzyme activity, energy transfer, and the synthesis of DNA and RNA. Plant development, reproduction, and yield are all impacted by it. Crop maturity, seed formation, and overall production are all improved by adequate phosphorus levels. Utilizing fertilizers that contain phosphorus helps farmers guarantee that their crops receive an adequate amount of this vital ingredient for strong crop growth.

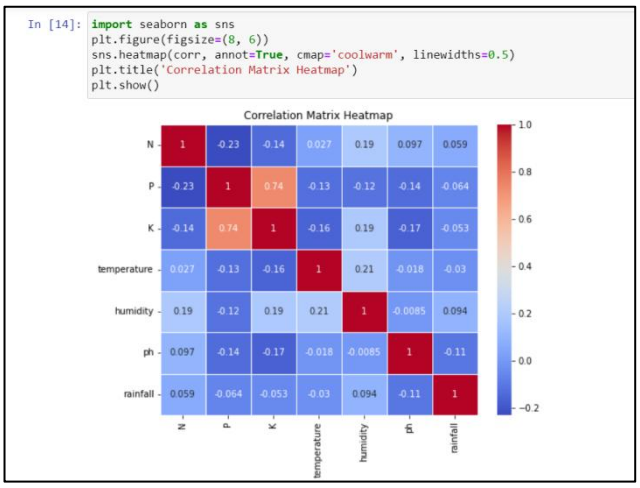
Potassium: Potassium is essential for crops because it helps with photosynthesis, increases stress tolerance, controls water balance, and activates enzymes that help with nutrient uptake. It influences the yield and quality of crops by being essential to the synthesis of proteins and carbohydrates. To guarantee that plants have enough potassium for healthy growth, farmers apply fertilizers containing potassium.

Humidity: There are various ways that humidity affects crops. Plant transpiration, the process by which water travels from roots to leaves and then evaporates into the atmosphere, benefits from adequate humidity levels. This promotes the uptake of nutrients from the soil and helps plants chill down. While extremes can put crops under stress or increase their risk of disease, maintaining a balanced humidity level is essential for optimum growth.

Rainfall: Rainfall is very important to crops since it gives them a natural and vital supply of water. Water is essential for several plant functions, such as cell division, photosynthesis, and nutrient uptake. Crops get the moisture they need for a healthy growth and development when there is enough rainfall. Additionally, rainfall helps maintain soil moisture levels, creating optimum circumstances for seed germination. Additionally, it helps recharge groundwater, which is necessary to keep crops growing during dry spells.

Temperature: A crop's ability to grow and develop is influenced by temperature, which makes it an important factor. Temperature affects many aspects of a plant's life cycle, including enzyme activity, photosynthesis, seed germination, and reproduction. Farmers take these elements into account when choosing crop kinds and scheduling plantings since different crops have different temperature requirements. While variations in temperature can affect crop pest activity and disease prevalence, proper temperatures are necessary for healthy growth. In general, cultivating crops successfully depends on an awareness of and ability to control temperature.

PH: Crops depend heavily on the pH of the soil because it affects microbial activity and nutrient availability. To guarantee that key nutrients remain in a form that plants can easily absorb, it is imperative to maintain the proper pH range. It also affects how well the rhizosphere functions, which has an effect on how well plant roots absorb nutrients. Farmers frequently use compounds like lime or sulfur to modify the pH of their soil in order to achieve the ideal conditions for a particular crop. Furthermore, controlling the pH of the soil helps avoid problems like aluminum toxicity in acidic soils, which eventually improves crop productivity and health.



.

Fig.2.Correlation Matrix(Heat map)

C. METHODOLOGY

i)KNN:

The KNN algorithm classifies a new instance into the most similar category to the existing cases based on how similar it is thought to be to the data and previous examples.A new data point is classified using the KNN approach by comparison with all of the previously collected data. Put another way, the KNN method streamlines the process of classifying recently discovered data.Although it is mainly used for classification problems, the KNN approach can also be used to solve regression problems. Once the user provides input, the trained KNN model may identify the most appropriate crops based on how close these input features are to the training data.

ii)Naive Baye's

Algorithm for probabilistic machine learning Bayes theorem serves as the foundation for Naive Bayes. The probabilistic classifier generates predictions by considering the probability of an event occurring. Because it presumes that the existence of one attribute is unrelated to the frequency of other attributes, it is known as naive. An apple, for example, is recognized as a red, spherical, sweet fruit if fruit definitions are based on color, form, and taste. In order to assist identify that it is an apple, each attribute functions independently of the others. Crop recommendation systems are among the applications that can make use of it because to its simplicity and capacity to work well with categorical data. It is possible to predict the likelihood that a specific crop would be suitable for a specific set of environmental conditions in the context of a crop recommendation system.

iii)SVM:

SVM is a potent algorithm utilized for tasks involving organizing data into groups and predicting continuous outcomes. It operates by determining the optimal hyperplane that effectively separates data points associated with different groups within a multi-dimensional space. SVM seeks to maximize the margin between these groups, enhancing its ability to generalize and resist overfitting. Moreover, SVM can address complex relationships within data by leveraging kernel functions to transform input features into higher-dimensional spaces where linear decision boundaries can be established. SVM's versatility and effectiveness have led to its widespread adoption across various domains, including bioinformatics, image analysis, text processing, and financial modeling. SVM play a crucial role in crop recommendation systems by aiding in the classification of soil types, crop varieties, and optimal fertilizer recommendations. In this context, SVM analyzes various input features such as soil nutrient levels, climate data, crop characteristics, and historical yield data to predict the most suitable crops for a particular season.

iv)Decision trees:

In supervised learning, decision tree algorithms are a flexible and popular technique that may be applied to both regression and classification problems. Its capacity to methodically assess input attributes at decision nodes, which enables it to make successive conclusions based on the properties of the data, is its main strength. The structure of the algorithm is tree-like, where nodes stand for decisions and branches for various attribute values. The algorithm iterates through the tree, fine-tuning its predictions until it reaches the leaf nodes, which determine the final results or suggestions. Decision trees are prized for their interpretability, which offers a clear and simple decision-making process. In agriculture, their utility is particularly notable as they enable the analysis of factors such as soil composition, climatic conditions, and past crop performance to identify the most suitable crops for specific circumstances.

v)Random Forest:

By combining several separately built decision trees, the Random Forest method, a decision tree extension, excels in improving prediction accuracy. To encourage variation across the component models and reduce overfitting issues, it deliberately makes use of random subsets of both data instances and features. Intricate patterns within the data are captured by this ensemble approach, which also enhances applicability to new cases. Because of its adaptability, Random Forest is used in many different fields, including finance, healthcare, and ecology. Accurate and stable forecasts for complicated situations are provided by this powerful tool for real-world decision-making, which can handle high-dimensional datasets and is easy to deploy. Because of its distinctive combination of ensemble intelligence and randomness, the method is a highly effective and popular machine learning solution.

vi)Logistic Regression:

The primary supervised learning method, LR is skilled at solving problems involving two distinct outcomes; it calculates the probability that an instance will be associated with a specific category.The output of LR, in contrast to linear regression, is converted into a probability range between 0 and 1. This is accomplished by using the logistic function. For applications ranging from medical diagnosis to customer churn prediction, logistic regression is a popular option since it is easily interpreted and simple to use. To help identify the critical elements impacting the projected outcomes, Logistic Regression also offers insights into the significance of features. The benefit of Logistic Regression is that it is easy to understand, efficient, and simple.

```
In [43]: models = {
             'Logistic Regression': LogisticRegression(),
             'K-Nearest Neighbors': KNeighborsClassifier(),
             'Decision Tree': DecisionTreeClassifier(),
             'Random Forest': RandomForestClassifier(),
             'Naive Bayes'   : GaussianNB(),
             'Support Vector Machine': SVC()
         }

         for name, md in models.items():
             md.fit(x_train,y_train)
             ypred = md.predict(x_test)
             print(f"{name}  with accuracy : {accuracy_score(y_test,ypred)}")

Logistic Regression  with accuracy : 0.9636363636363636
K-Nearest Neighbors  with accuracy : 0.9590909090909091
Decision Tree  with accuracy : 0.9840909090909091
Random Forest  with accuracy : 0.9954545454545455
Naive Bayes  with accuracy : 0.9954545454545455
Support Vector Machine  with accuracy : 0.9681818181818181
```

Fig.3.Various machine leraning models

D. RELATION BETWEEN FEATURES AND LABELS:

- Rice, Maize, and Jute share similar values in terms of
- Nitrogen (N), Phosphorus (P), temperature, and pH. However, Maize exhibits lower levels of potassium and humidity compared to Jute and Rice. In terms of rainfall, Rice requires the highest amount, Jute needs a moderate level, and Maize can thrive with less rainfall.
- For the cultivation of watermelon and muskmelon, an average temperature of 24 degrees Celsius with high humidity is essential. These crops can flourish in low rainfall conditions and require lower amounts of potassium and phosphorus, but a higher level of nitrogen for optimal growth.
- Pulses and lentils necessitate the same levels of NPK, temperature, pH, and rainfall. However, the humidity requirements vary. Hence, based on humidity conditions, any of these pulses can be grown in the same land.
- Grapes and apple crops demand higher percentages of potassium and phosphorus for successful cultivation.
- It's a common requirement for all crops to have an average temperature of 25 degrees Celsius .
- All Pulses have same range of feature values.
- For optimal growth of crops an average of 6.5 pH is needed.

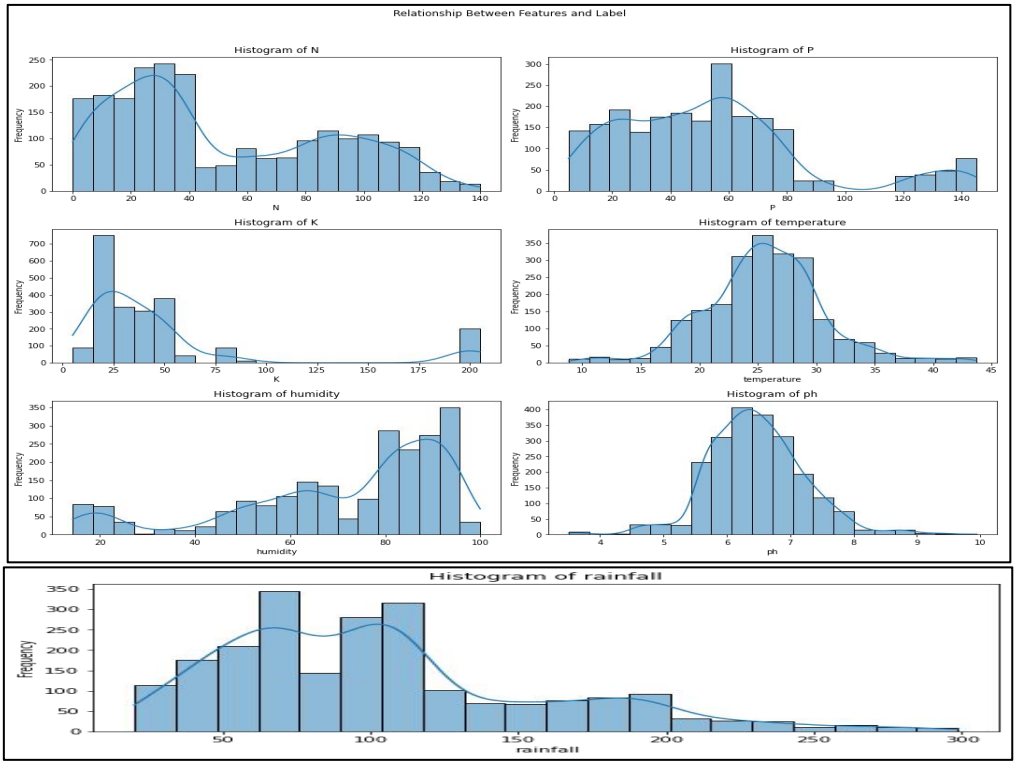


Fig.4.*Histograms representing relation between features and labels*

E. RESULT

The DecisionTree(DT), RF, and NB classifiers demonstrate impressive accuracy scores, with RF and NB achieving the highest accuracy rate of 99.54%. These models stand out as exceptional options for precise crop recommendation.

- 'P', 'K', and 'rainfall' show a positive skew, suggesting more instances with lower values and a few with higher values.
- 'N', 'Temperature', and 'pH' display near-zero skewness, implying a more balanced spread of values.
- 'Humidity' demonstrates negative skewness, indicating more instances with higher humidity and fewer with lower humidity levels.
- N', 'humidity', 'pH' and 'rainfall' exhibit values close to zero kurtosis, suggesting these features have distributions similar to a normal distribution, neither too peaked nor too flat.
- 'temperature' demonstrates a moderate kurtosis value of 1.227029, suggesting a moderately peaked distribution.
- 'K' stands out with a high kurtosis value of 4.436523, indicating a distribution significantly more peaked or heavy-tailed compared to a normal distribution. This suggests that 'K' might have extreme values or exhibit more outliers.

TABLE 1. *Performance metrics across diverse models*

Model Name	Accuracy	Precision	Recall	F1-Score
LogisticRegression	96.36%	96.36	96.36	96.36
K-Nearest Neighbours	95.90%	95.90	95.90	95.90
Decision Tree	98.40%	98.63	98.63	98.63
Random Forest	99.54%	99.31	99.31	99.31
Naive Bayes	99.54%	99.54	99.54	99.54
Support Vector Machine	96.81%	96.81	96.81	96.81

TABLE 2 .Skewness and Kurtosis Values

FEATURES	SKEWNESS	KURTOSIS
N	0.509374	-1.058562
P	1.010083	0.855599
K	2.373547	4.436523
TEMPERATURE	0.184807	1.227029
HUMIDITY	-1.090963	0.298722
PH	0.283736	1.649095
RAINFALL	0.965098	0.602974

IV. SUMMARY

- For crop recommendation, the effectiveness of several machine learning algorithms was assessed.
- Significant correlations between crop types and environmental characteristics were found by correlation analysis.
- Temperature, phosphorus, and nitrogen have all been shown to have significant relationships with particular crop varieties.
- Comprehending these interrelationships can assist farmers in making knowledgeable choices about crop selection and farming techniques.
- The results have real-world applications in agriculture, giving farmers insightful knowledge on the best crops to plant depending on the local environment.
- Farmers are able to increase yields, optimize resource consumption, and adjust to shifting environmental dynamics by utilizing machine learning algorithms.
- By choosing crops that are suited to the soil and climate of their region, farmers can minimize the amount of fertilizer and water they use while also supporting environmental sustainability.

V. CONCLUSION

The culmination of our research underscores the fusion of advanced machine learning methodologies with rigorous statistical analyses, offering a comprehensive framework for crop recommendation systems. As we navigated through the process, meticulous attention was paid to every stage, beginning with the meticulous preprocessing of data to ensure its integrity and relevance.Integral to our approach was the utilization of statistical techniques to scrutinize the data distribution's goodness of fit and discern its underlying shape. Through the lens of skewness and kurtosis metrics, we gained invaluable insights into the asymmetry and peakedness of the data, augmenting our understanding of its intrinsic characteristics.Simultaneously, we engaged in a thorough exploration of machine learning algorithms, honing in on RF and NB classifiers as the most adept tools for crop recommendation. Their superior accuracy, validated through rigorous testing, solidified their position as the cornerstone of our recommendation system.Furthermore, our statistical analyses served not only to validate the efficacy of our machine learning models but also to provide a deeper understanding of the dataset's dynamics. By visualizing data distributions through various charts and metrics, we gained a nuanced perspective, empowering us to make informed decisions in agricultural contexts.In summation, the synergy between machine learning and statistical analyses offers a robust framework for agricultural decision-making. By combining predictive power with a deep understanding of data dynamics, our approach not only enhances crop recommendation accuracy but also fosters a more profound comprehension of agricultural systems, paving the way for sustainable and informed practices in the field.

VI. OUTLINING THE PLANNED METHODOLOGY

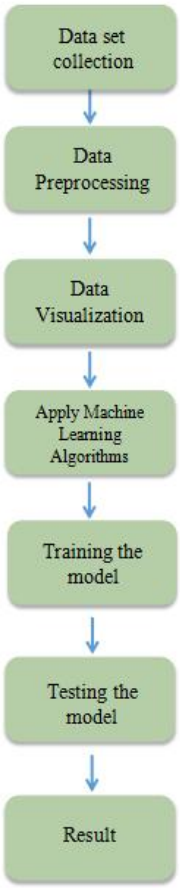


Fig.6.Flow diagram for the suggested methodology

REFERENCES

[1] Abbaszadeh, P., Gavahi, K., Alipour, A., Deb, P., & Moradkhani, H. (2022). Bayesian multi-modeling of deep neural nets for probabilistic crop yield prediction. *Agricultural and Forest Meteorology, 314*, 108773. doi:10.1016/j.agrformet.2021.108773

[2] Agarwal, S., & Tarar, S. (2021). A hybrid approach for crop yield prediction using machine learning and deep learning algorithms. *Journal of Physics: Conference Series*, *1714*(1), 012012. doi:10.1088/1742- 6596/1714/1/012012

[3] Aghighi, H., Azadbakht, M., Ashourloo, D., Shahrabi, H. S., &Radiom, S. (2018, December). Machine learning regression techniques for the silage maize yield prediction using time-series images of landsat 8 OLI. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(12), 4563–4577. doi:10.1109/JSTARS.2018.2823361

[4] P.Sricharani and.B. SrinivasaRao."Variable control charts based on Dagum distribution",*Research Journal of Mathematical and Statistical Sciences* 2019, Vol. 7(3),pp: 43-50, www.iscamaths.com , www.isca.in , www.isca.m eissn: 2320-6047

[5] Goel, L., & Mishra, A. (2022). A survey of recent deep learning algorithms used in Smart farming. *IEEE Region 10 Symposium (TENSYMP),* (pp. 1–6). IEEE. 10.1109/TENSYMP54529.2022.9864477

[6] Gupta, M. V, S. K. B., B, K., Narapureddy, H. R., Surapaneni, N., & Varma, K. (2022). Various crop yield prediction techniques using machine learning algorithms. *IEEE Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, (pp.273–279).IEEE. 10.1109/ICAIS53314.2022.9742903

[7] Gupta, R., Sharma, A. K., Garg, O., Modi, K., Kasim, S., Baharum, Z., Mahdin, H., & Mostafa, S. A. (2021). WB-CPI: Weather based crop prediction in India using big data analytics. *IEEE Access : Practical Innovations, Open Solutions*, *9*, 137869–137885. doi:10.1109/ACCESS.2021.3117247

[8] Haque, F. F., Abdelgawad, A., Yanambaka, V. P., & Yelamarthi, K. (2020). Crop yield prediction using deep neural network. *6th World Forum on Internet of Things (WF-IoT),* (pp. 1–4). IEEE Publications. 10.1109/WF-IoT48130.2020.9221298

[9] Iniyan, S., Varma, V. A., & Naidu, C. T. (2023). Crop yield prediction using machine learning techniques. *Advances in Engineering Software*, *175*, 103326. doi:10.1016/j.advengsoft.2022.103326

[10] Kalaiarasi, E., & Anbarasi, A. (2022). Multi-parametric multiple kernel deep neural network for crop yield prediction. *Materials Today: Proceedings, 62*(7), 4635–4642. doi:10.1016/j.matpr.2022.03.115

[11] Kalimuthu, M., Vaishnavi, P., & Kishore, M. (2020). Crop prediction using machine learning. *Third IEEE International Conference on Smart Systems and Inventive Technology (ICSSIT),* (pp. 926–932). IEEE. 10.1109/ICSSIT48917.2020.9214190

[12] Kumar, M., Kumar, A., & Palaparthy, V. S. (2021). Soil sensors-based prediction system for plant diseases using exploratory data analysis and machine learning. In IEEE Sensors Journal, 21(16). doi:10.1109/ JSEN.2020.3046295

[13] Li, Z. Ding, L., & Xu, D. (2022). Exploring the potential role of environmental and multi-source satellite data in crop yield prediction across Northeast China. *Science of the Total Environment, 815*. doi:10.1016/j. scitotenv.2021.152880

[14] Liu, Z., Bashir, R. N., Iqbal, S., Shahid, M. M. A., Tausif, M., & Umer, Q. (2022). Internet of things (IoT) and machine learning model of plant disease prediction–blister blight for tea plant. *IEEE Access : Practical Innovations, Open Solutions*, *10*, 44934–44944. doi:10.1109/ACCESS.2022.3169147