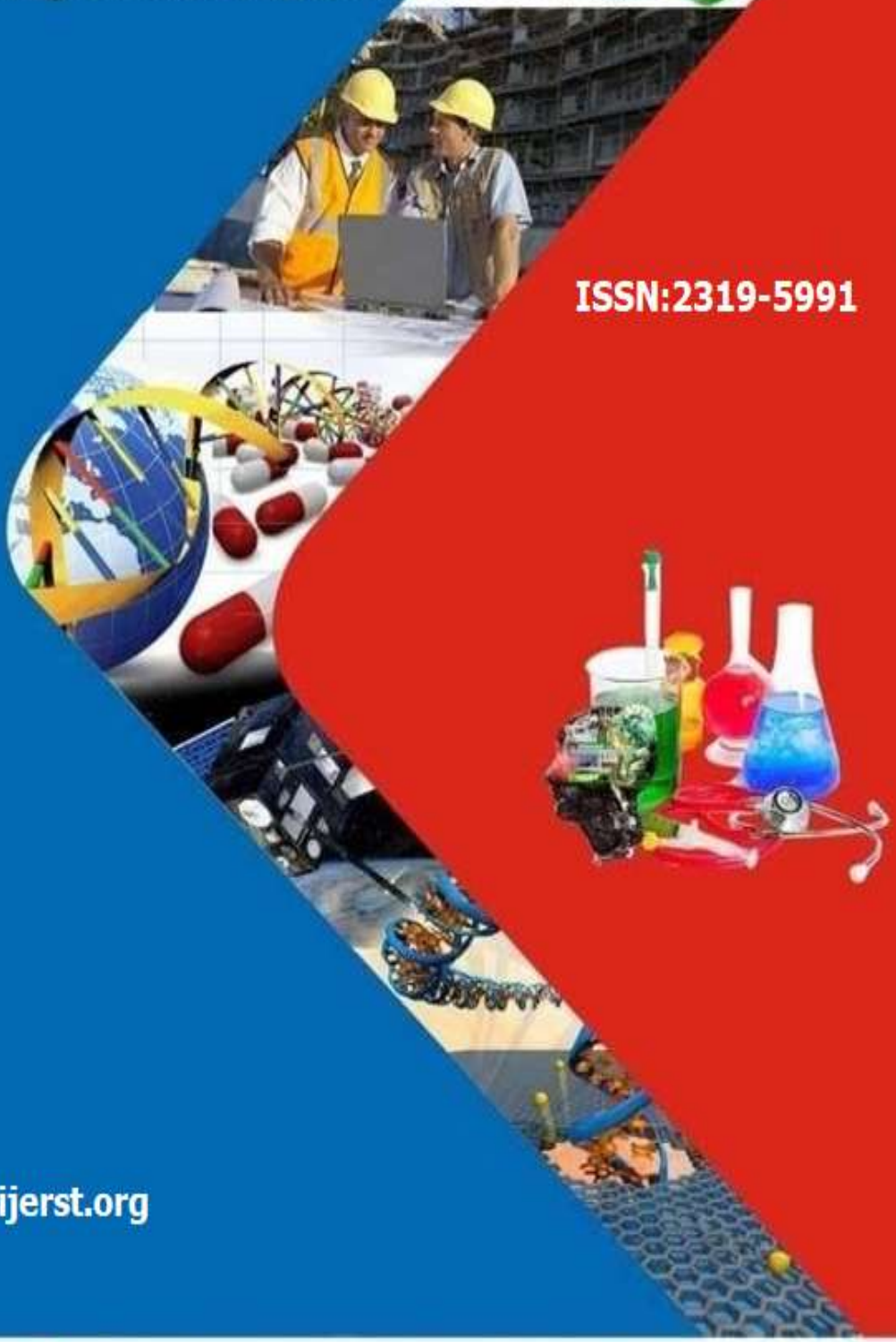
International Journal of Engineering Research and Science & Technology

IJERST

ISSN:2319-5991



www.ijerst.org

E-mail: editor@ijerst.org or ijerst.editor@gmail.com

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

PREDICTIVE MODELING FOR CROP YIELD ESTIMATION: MACHINE LEARNING CLASSIFIER COMPARISION

P. Manjulatha[1*], Y. Pavan Sai[2], G. Rakshit Kumar[2], K. Chandu Vara Prasad[2], M. Shiva Prasad[2]

[1]Assistant Professor, [2]UG Student ,[1,2]Department of Computer Science and Engineering (Information Technology), [1,2]Sree Dattha Institute of Engineering and Science, Sheriguda, Hyderabad, Telangana.

ABSTARCT

The aim is to enhance agricultural productivity through advanced predictive modeling. By leveraging machine learning techniques, the goal is to analyze and compare various classifiers to accurately estimate crop yields. Using historical data, weather patterns, soil conditions, and other relevant factors, a robust framework is to be developed for forecasting crop yields with high precision. Traditional methods of estimating crop yields often depend on manual observation and historical trends. These approaches are typically time-consuming, labor-intensive, and prone to errors due to the complex and dynamic nature of agricultural systems. Moreover, they often fail to fully utilize available data or adapt to changing environmental conditions, resulting in less accurate predictions. There is a clear need for a more efficient and accurate yield estimation system that takes advantage of machine learning algorithms. The focus is on overcoming the limitations of older methods by building predictive models capable of processing large datasets, recognizing patterns, and generating precise predictions based on variables such as weather, soil properties, and farming practices. The motivation stems from the potential impact on agricultural decision-making. Accurate predictions enable farmers to make better-informed choices regarding planting schedules, harvesting times, and resource distribution, ultimately improving productivity and optimizing resource use. By applying machine learning, the system offers valuable insights that support better crop management and contribute to sustainable agriculture and food security.

Keywords: Machine learning, Crop yield, Orthogonal Matching Pursuit (OMP), Calibrated Classifier, Agriculture.

1.INTRODUCTION

Crop yield estimation has long been a critical aspect of agricultural planning and management. Traditionally, farmers relied on manual observation, historical data, and basic statistical methods to predict crop yields. However, these approaches were often limited in accuracy and efficiency, as they could not fully capture the complex and dynamic nature of agricultural systems. With the advent of technology, particularly in the field of data science and machine learning, there emerged a new opportunity to revolutionize crop yield estimation.

In recent years, the application of machine learning techniques in agriculture has gained significant attention. Researchers began exploring the potential of leveraging advanced algorithms to analyze vast amounts of data and extract valuable insights for crop yield prediction. This marked a shift towards more data-driven and predictive approaches, moving away from reliance on traditional methods.

As machine learning algorithms evolved and computing power became more accessible, the feasibility of developing sophisticated predictive models for crop yield estimation increased. Researchers started experimenting with various machine learning techniques, such as decision trees,

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

random forests, support vector machines, and neural networks, to determine their effectiveness in predicting crop yields accurately.

Furthermore, the availability of large-scale agricultural datasets, including information on weather patterns, soil characteristics, crop types, and agricultural practices, provided researchers with valuable resources for training and validating predictive models. These datasets served as the foundation for building robust frameworks capable of generating precise crop yield forecasts.

2.LITERATURE SURVEY

Haque et al. [1] proposed two separate Machine Learning (ML) algorithms to evaluate the yield of crops. The algorithms Support Vector Regression (SVR) and Linear Regression (LR) have been well suited for validating variable parameters in continuous changeable estimation with 140 data points. The mentioned parameters are important determinants for crop yield. The error rate has been calculated for using Coefficient of Determination (R2) and Mean Square Error (MSE) with MSE yielding around 0.005 and R2 yielding around 0.86. The same dataset has been used to compare the performance of the algorithms quickly.

Nishant et al. [2] proposed web application to forecast the yield of nearly all types of crops grown in India. This study is unique because it uses simple parameters such as state, district, season, and region to predict crop yields in any year the user desires. To predict the yield, advanced regression techniques such as Kernel Ridge, Lasso, and ENet algorithms are used in the paper, as well as the principle of Stacking Regression to improve the algorithms. The root means the square error is the output metric that used in this work. The models when have been implemented individually, ENet had an error of about 4%, Lasso had an error of about 2% and Kernel Ridge had an error of about 1%, and after stacking it was less than 1%. Kalimuthu et al. [3] suggested a smartphone application for Android, which uses machine learning, which is one of the most advanced crops prediction technologies, to direct beginner farmers in sowing the appropriate crops. The naive Bayes algorithm proposes a method for doing so. The data seed of the crops is collected, along with the suitable parameters such as humidity, moisture, and temperature content, that is aids the crops' development. To begin the prediction process, users are encouraged to input parameters such as their location and temperature this application will take it automatically to start the process of prediction.

Y. J. N. Kumar et al. [4], implemented prediction system on crop production from the collecting of past data. Crop yield is estimated using data mining techniques. They used the Random Forest algorithm to forecast the highest yield crop as a product. Crops yield predictions are often appropriate in the agricultural sector. The higher the accuracy, the higher the benefit on the crop yield. Farmers will use the proposed technique to help them decide which crop to plant in their fields. Under this system would cover the widest range of crops possible. Farmers in India can benefit from accurate forecasting of various crops across various districts. Gupta et al. [5] used IoT and data mining in monitoring applications to make smart farming possible. Smart farming is a method of providing all of the services needed for a specific time. Soil moisture, light intensity, relative humidity, soil pH reading, and ambient temperature are all resources that are needed. They demonstrated the transformation of a device capable of gathering data from sensors using IoT in the agriculture field. This device successfully senses data and sends it locally to the thing speak cloud, which the user can then access via his or her custom website. The data mining techniques such as SVM, KNN and Random forest are used to crop-producing with the correct number of resources so that the farmer still has the upper hand, and by comparing the current pattern to the previous one, will be able to

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

determine whether or not a parameter is right. All of these agricultural process values are monitored on a user-defined platform.

Terliksiz & Altylar [6] used a 3D CNN model that leverages spatiotemporal features, a soybean yield prediction for Lauderdale County, Alabama, USA has been presented. From 2003 to 2016, the yield has been taken from the USDA NASS Fast Stat tool. Google Earth Engine was used to gather satellite data from (NASA's MODIS) land products surface reflectance and land surface temperature. For comparison of the results, the root means squared error (RMSE) has been used as the measurement metric. When 64x64 data frames with more than 20% cropland coverage were used without dropout, the best result for Lauderdale County, Alabama was obtained. The RMSE is 0.81 and the error is 2.70% in this situation. Khosla et al., [7] The Kharif crops yield were estimated in 2 steps; first, the rainfall has been estimated by using a modular artificial neural network MANN, and then the yield was predicted by using support vector regression SVR. The dataset of experimental included data from the year 2000 to 2016 and generated results for the years 2018 and 2019. Nigam et al. [8] utilized machine learning techniques such as KNN classifier, Random Forest, Artificial Neural Network, liner Regression and XGboost to predict yield crop. Based on the Mean Absolute Error MAE, the results of these techniques are compared. An algorithm of a machine learning algorithm can be help farmers to determine which crops planting to get the best yield by taking into account factors such as temperature, rainfall, area, …etc. When all parameters are combined, the results show that the best classifier is Random Forest.

S & R [9] determined most of the important points for accurate Crop's yield Prediction. For improved accuracy, the algorithm of machine learning namely Support Vector Machine, KNN, Regression, Random Forest and Artificial Neural Network have been proposed. The agricultural dataset contains 745 instances, 30% of data is used to test the prediction performance of the models, while 70% of the data is chosen at random and used to train the model. The results show that, by using the same farming training data, the RF algorithm achieves maximum precision employing its error analysis values for all the separated feature sub-sets. Kumar et al. [10] used the three supervised techniques SVM, KNN and Least Squared Support Vector Machine are. It is a comparative study that shows the training proposed model's accuracy and error rate. They have three different datasets: soil, rainfall, and yield. They then combined the datasets and used the techniques to determine the actual approximate cost as well as the 121 accuracy of the techniques used. The training model's accuracy will be higher and the error rate will be low. The technique with the highest accuracy rate (LS SVM =90%) and the lowest error rate (LS SVM=0,0362).

3.PROPOSED METHODOLOGY

The research aims to develop a predictive model for crop yield estimation using machine learning classifiers. It begins with data import and preprocessing, where necessary libraries and the dataset are loaded. Initial data analysis is conducted by checking dataset information, descriptions, correlations, and null values. An unnamed column is dropped, and categorical variables are encoded. Data visualization is performed using Seaborn to generate a count plot of the target variable, yield class.

Next, the dataset is split into training and testing sets using train test split. Two classifiers are trained for comparison. The first is the Orthogonal Matching Pursuit (OMP) Classifier, implemented using Orthogonal Matching Pursuit from scikit-learn. The second is a calibrated version of Logistic Regression, trained using Calibrated Classifier CV from scikit-learn.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

For model evaluation, performance metrics such as accuracy, precision, recall, and F1-score are calculated using a custom function named performance metrics. In addition, classification reports and confusion matrices are generated for both classifiers to visually assess their performance.

To preserve the trained models, model persistence is implemented using NumPy's and save and reload model files if they already exist. The performance of both classifiers is compared and presented in a tabular format.

Finally, the model is used to make predictions on new data by loading another dataset. The same preprocessing steps are applied to this dataset, and the calibrated classifier is used to predict the yield class. The predicted results are printed alongside the corresponding input features.

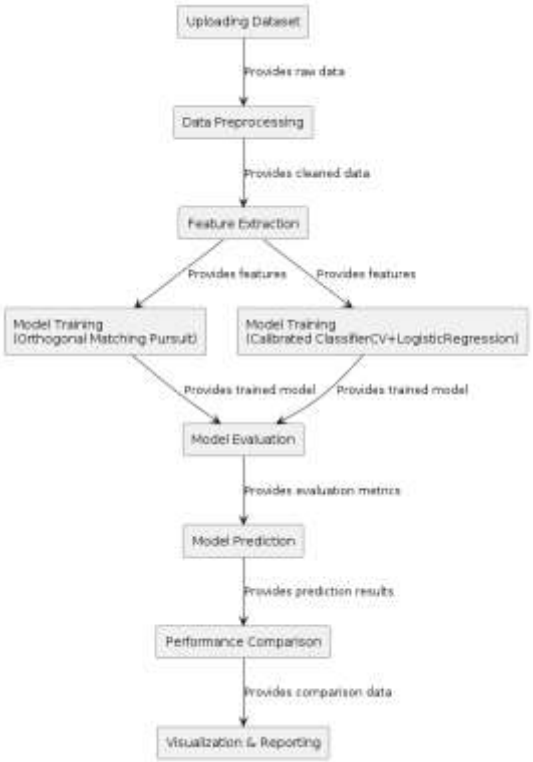


Fig. 1: Block Diagram of overall architecture of proposed system.

3.1 Orthogonal Matching Pursuit (OMP)

Orthogonal Matching Pursuit (OMP) is a greedy algorithm used in signal processing and statistics to solve sparse approximation problems. The goal of OMP is to represent a signal as a sparse linear combination of basis vectors (atoms) from a dictionary. OMP is widely used in applications such as compressed sensing, image processing, and machine learning due to its simplicity and effectiveness in handling high-dimensional data.

The OMP algorithm iteratively selects the dictionary atoms that best match the residual part of the signal. It begins with an initialization phase where the initial residual is set equal to the signal y , an

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

empty set is created to store the selected atoms, and the solution vector x is initialized to 0. The algorithm then proceeds through an iterative process to refine the solution.

In the first step of the iteration, correlations are computed between the residual and all the atoms in the dictionary D, resulting in a correlation vector given by $C = D$^T r, where r is the current residual. Next, the atom with the highest absolute correlation is selected, as it is the one most aligned with the residual, and its index is added to the set of selected atoms Λ. Following this, a least squares problem is solved to find the new coefficients for the selected atoms, formulated as x_Λ = arg min_z $\|y - D\_\Lambda z\|\_2$^2, where D_Λ is the submatrix of D containing the selected atoms, and x_Λ is the subvector of x corresponding to those atoms. The residual is then updated using the equation $r = y - D\_\Lambda x\_\Lambda$. This process checks a stopping criterion, which can be based on the number of iterations, the norm of the residual, or the sparsity level of the solution, and repeats the steps until the criterion is met.

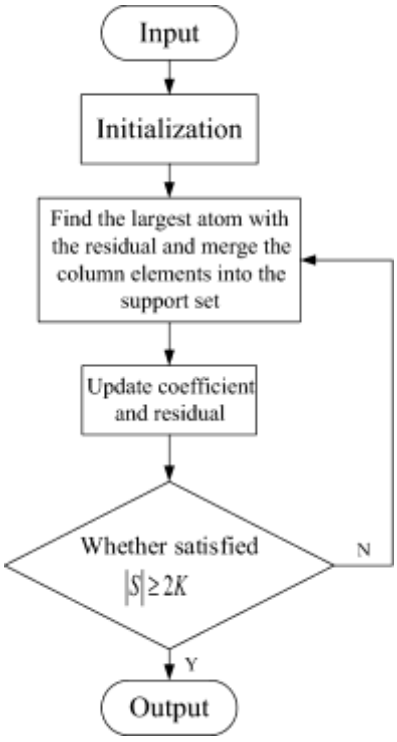


Fig. 2: Orthogonal Matching Pursuit (OMP) algorithm.

The output of the OMP algorithm is the sparse coefficient vector x, which approximates the signal y as $y \approx D x$, providing a sparse representation of the original signal based on the selected dictionary atoms.

3.2 Calibrated Classifier

A CalibratedClassifier is a method used in machine learning to improve the probability estimates provided by a classifier. Many classifiers, such as decision trees, support vector machines, or neural networks, do not naturally produce well-calibrated probabilities. Well-calibrated probabilities are those that reflect the true likelihood of an event occurring. Calibrated classifiers adjust these probability outputs so that they can be interpreted as reliable probabilities.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

Principle of Probability Calibration: The goal of probability calibration is to ensure that the predicted probabilities match the true probabilities. For instance, if a classifier predicts a probability of 0.8 for an event occurring in 100 cases, then ideally, the event should occur in approximately 80 of those cases.

3.2.1 Logistic regression

Logistic regression predicts the probability of an outcome that can only have two values (i.e., a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

1. A linear regression will predict values outside the acceptable range (e.g., predicting probabilities outside the range 0 to 1)
2. Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

4.RESULTS AND DISCUSSION

This research aims to develop a robust system for predicting crop yields by employing and comparing various machine learning classifiers. This implementation involves several crucial steps, each contributing to the overall accuracy and effectiveness of the predictive model.

4.1 Dataset description

The dataset contains a detailed collection of information related to crop yield and environmental factors. It is useful for creating prediction models and for studying agriculture to help make better decisions. By looking at how these factors are connected, researchers and people working in farming can understand what affects crop growth and find ways to improve farming and support food production.

The Unnamed: 0 column seems to be an index or ID number for each row. It helps tell the records apart. The Area column shows the region or location where the crops were grown. This helps explain how different areas affect crop yield and conditions. The Item column tells us which crop or product was grown in each area. Since different crops grow in different ways and react differently to the environment, this information is important. The Year column shows when the data was collected. This helps in studying changes in crops and conditions over time. The hg/ha_yield column shows how much crop was produced, measured in hectograms per hectare. It is a key measure of how successful the farming was. The average_rain_fall_mm_per_year column shows the yearly average rainfall in millimeters. Rainfall affects how well crops grow. The pesticides_tonnes column shows how much pesticide (in tonnes) was used. Pesticides affect both crop yield and the environment. The avg_temp column shows the average temperature during the growing season. Temperature affects how crops grow and develop. The yield_class column groups crop yield into different levels or categories. It helps compare how well crops did under different conditions.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

4.2 Results description

Fig. 3 displays a count plot for an output variable categorized into three yield levels: average, low, and high. The plot shows that the "average yield" category has the highest count at 9602.0, represented by a teal bar, followed closely by the "low yield" category with a count of 9202.0 in a yellow bar, and the "high yield" category with a count of 9202.0 in a purple bar. The counts for low and high yields are identical, indicating a balanced distribution between these two categories, while the average yield category has a slightly higher frequency, suggesting it is the most common outcome in the dataset.

Fig. 4 illustrates the confusion matrix for the Orthogonal Matching Pursuit (OMP) Classifier model used to predict crop yield classes: high, low, and average. The matrix reveals a fairly balanced spread of predictions, with notable counts in both correct and incorrect classifications across all categories, suggesting moderate accuracy but highlighting issues with misclassifications, particularly between high and low yield predictions, indicating the model struggles to differentiate these classes effectively.

Fig .5 displays the confusion matrix for the Calibrated CV + Logistic Classifier model, also predicting the same yield classes. This matrix shows a stronger performance, with higher counts along the diagonal, representing correct predictions for each class, and fewer misclassifications overall, especially for high and low yield classes, though some errors remain in identifying average yield, demonstrating that this calibrated model is generally more accurate than the OMP model in classifying the true yield categories.

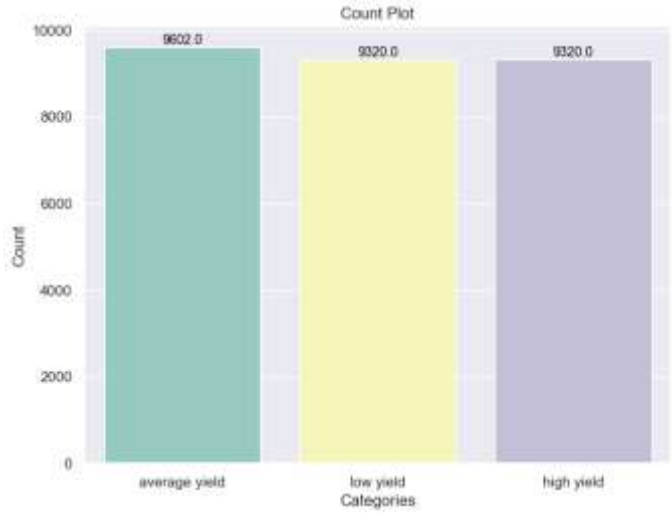


Fig. 3: Count plot for the output variable.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

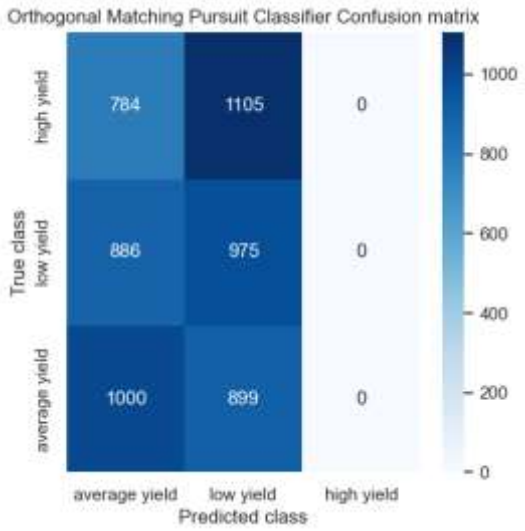


Fig. 4: Confusion Matrix obtained using OMP Classifier model.

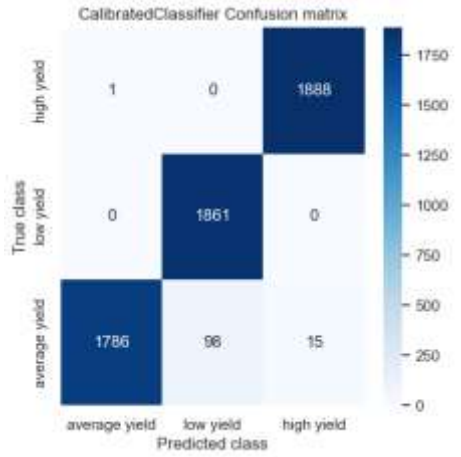


Fig. 5: Confusion Matrix obtained using Calibrated cv+ logistic Classifier model.

```
Area                                0.00
Item                                4.00
Year                             1991.00
hg/ha_yield                     20698.00
average_rain_fall_mm_per_year    1485.00
pesticides_tonnes                 121.00
avg_temp                           15.36
Name: 0, dtype: float64
Model Predicted of Row 0 Test Data is---> high yield
Area                                0.00
Item                                0.00
Year                             1992.00
hg/ha_yield                     24876.00
average_rain_fall_mm_per_year    1485.00
pesticides_tonnes                 121.00
avg_temp                           16.06
Name: 1, dtype: float64
Model Predicted of Row 1 Test Data is---> average yield
Area                                0.00
Item                                1.00
Year                             1992.00
hg/ha_yield                     82920.00
average_rain_fall_mm_per_year    1485.00
pesticides_tonnes                 121.00
avg_temp                           16.06
Name: 2, dtype: float64
Model Predicted of Row 2 Test Data is---> low yield
```

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

Fig. 6: Model Prediction on Uploaded Test Data.

Fig. 6 shows a model's predictions for crop yield across three test data rows, each with features like area, year, rainfall, pesticide use, and average temperature. The model predicts "high yield" for the first row, "average yield" for the second, and "low yield" for the third, indicating how different combinations of environmental and agricultural factors influence the predicted yield categories.

5.CONCLUSION

This research on predictive crop yield estimation using machine learning classifiers has demonstrated promising results in accurately predicting crop yield classes based on various input features. Through the utilization of advanced machine learning algorithms such as Orthogonal Matching Pursuit (OMP) and Calibrated ClassifierCV with logistic regression, the models have been trained and evaluated on a comprehensive dataset containing information on factors like area, crop type, year, yield, rainfall, pesticides usage, and average temperature. The performance evaluation of these models has shown substantial precision, recall, F1-score, and accuracy, indicating their effectiveness in classifying crop yield classes. The confusion matrices provide detailed insights into the models' predictive capabilities and their ability to differentiate between different yield classes.

REFERENCES

[1] Haque, F. F., Abdelgawad, A., Yanambaka, V. P., &Yelamarthi, K. (2020, June). Crop yield analysis using machine learning algorithms.In 2020 IEEE 6th World Forum on Internet of Things (WF-IoT) (pp. 1-2). IEEE.

[2] Nishant, P. S., Venkat, P. S., Avinash, B. L., & Jabber, B. (2020, June). Crop Yield Prediction based on Indian Agriculture using Machine Learning. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-4). IEEE.

[3] Kalimuthu, M., Vaishnavi, P., & Kishore, M. (2020). Crop Prediction using Machine Learning. 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 926–932. https://doi.org/10.1109/ICSSIT48917.2020.9214190

[4] Kumar, Y. J. N., Spandana, V., Vaishnavi, V. S., Neha, K., & Devi, V. G. R. R. (2020, June). Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector.In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 736- 741). IEEE.

[5] (Gupta et al., 2020)Gupta, G., Setia, R., Meena, A., &Jaint, B. (2020, June). Environment Monitoring System for Agricultural Application using IoT and Predicting Crop Yield using Various Data Mining Techniques. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 1019-1025). IEEE.

[6] Terliksiz, A. S., &Altýlar, D. T. (2019, July). Use of deep neural networks for crop yield prediction: A case study of soybean yield in Lauderdale county, Alabama, USA. In 2019 8th International Conference on Agro-Geoinformatics (Agro- Geoinformatics) (pp. 1-4). IEEE.

[7] Khosla, E., Dharavath, R., &Priya, R. (2019). Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression. Environment, Development and Sustainability, 1-22.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

[8] Nigam, A., Garg, S., Agrawal, A., &Agrawal, P. (2019, November). Crop yield prediction using machine learning algorithms.In 2019 Fifth International Conference on Image Information Processing (ICIIP) (pp. 125-130). IEEE.

[9] Maya Gopal P. S. &Bhargavi R. (2019) Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms, Applied Artificial Intelligence, 33:7, 621 642, DOI: 10.1080/08839514.2019.1592343

[10] Kumar, A., Kumar, N., & Vats, V. (n.d.). 2018 EFFICIENT CROP YIELD PREDICTION USING MACHINE LEARNING ALGORITHMS. International Research Journal of Engineering and Technology (IRJET) 05(06), 9.