Volume: 16 Issue: 06 Pages: 6621-6626 (2025) ISSN: 0975-0290

Data Visualization and Analytics for Crop Yield Estimation

Vidisha Tiwari
Masters in Technology Management, University of Illinois Urbana, Champaign, USA.
Hymavathi. M
Department of Computer Science and Engineering, N.B.K.R.I.S.T, Vidyanagar, Andhra Pradesh, India.
Bharath Reddy. M
School of Informatics, Computing & Cyber Systems
Northern Arizona University
S San Francisco St, Flagstaff, AZ 86011, USA.
Jahnavi. Y *
Department of Computer Science, Government Degree College - Naidupet,
Tirupathi (Dt), Andhra Pradesh, India.

----------**ABSTRACT**----------
The application of data visualization and analytics focus on to estimate crop yields. By leveraging historical datasets and employing machine learning techniques, this paper aims to provide accurate predictions of crop production based on various factors such as Area, Production, Yield. The insights derived from these analyses can assist policymakers and farmers in making informed decisions, ultimately contributing to more efficient agricultural practices and stable market prices. Interactive visualizations are utilized to represent complex data intuitively, making it easier to identify trends, correlations, and anomalies. The study compares multiple predictive models, highlighting the superiority of algorithms like Decision Tree, Random Forest in terms of accuracy, despite their computational complexity. The findings underscore the importance of integrating robust data analytics and visualization tools in modern agriculture to enhance productivity and sustainability.

Keywords - **Crop Yield Estimation, Data Visualization, Machine Learning Algorithms.**

Date of Submission : March 14, 2025 Date of Acceptance: April 21, 2025

I. INTRODUCTION

Crop yield prediction is a critical aspect of agricultural planning and management. Accurate yield predictions enable farmers and policymakers to make informed decisions regarding resource allocation, market strategies, and food security measures. Traditionally, crop yield estimation relied on empirical methods and historical data analysis, which often fell short in capturing the complex interactions between various agricultural factors.

With the advent of machine learning, a new era of precision agriculture has emerged. Machine learning algorithms can analyse vast amounts of data, uncover hidden patterns, and make predictions with greater accuracy. By leveraging data from diverse sources such as weather conditions, soil properties, crop characteristics, and historical yield records, machine learning models can provide timely and precise yield forecasts.

This approach involves using algorithms such as Linear Regression with Dummies, Random Forest with Dummies, Linear Regression, Random Forest, Support Vector Machine, Neural Network, Gradient Boosting, Decision Tree, among others, to build predictive models. These models are trained on historical datasets, enabling them to learn the relationships between different variables and crop yields. Factors like location, season, soil health, and crop type are taken into account to generate reliable predictions.

The integration of machine learning in crop yield prediction not only enhances the accuracy of forecasts but also provides valuable insights that can lead to better crop management practices. By anticipating yield outcomes, stakeholders can optimize planting schedules, irrigation plans, and fertilization strategies, ultimately contributing to increased agricultural productivity and sustainability.

This paper explores the implementation of machine learning techniques for crop yield prediction, discussing the methodologies, data requirements, and the comparative effectiveness of different algorithms. Through case studies and real-world applications, we demonstrate the potential of machine learning to revolutionize crop yield estimation and its impact on the agricultural sector.

II. LITERATURE REVIEW

The prediction of crop yields has long been a focus of agricultural research, aimed at improving food security and optimizing resource use. In recent years, machine learning (ML) has emerged as a powerful tool to enhance the accuracy and reliability of yield predictions. This literature review examines the advancements in crop yield prediction through the lens of machine learning, highlighting key methodologies, data sources, and findings from various studies.

Early research in crop yield prediction primarily relied on statistical methods and empirical models. Studies by Allen

Volume: 16 Issue: 06 Pages: 6621-6626 (2025) ISSN: 0975-0290

et al. (1998) and Doorenbos and Kassam (1979) used crop growth models based on historical yield data and environmental factors [1][2]. While these models provided a foundation for understanding crop yield determinants, they often lacked the ability to accurately predict yields under varying conditions due to their simplistic linear assumptions and limited data integration capabilities.

The application of machine learning (ML) has revolutionized crop yield prediction. Studies such as that by Jeong et al. (2016) have demonstrated the superior accuracy of ML models compared to traditional methods. Jeong et al. used Support Vector Machines (SVM) to predict corn yields, achieving notable improvements in prediction accuracy [3]. Similarly, Khaki and Wang (2019) employed deep learning techniques, highlighting their effectiveness in capturing complex non-linear relationships between yield determinants [4].

Several studies have conducted comparative analyses of different ML algorithms for crop yield prediction. Liakos et al. (2018) reviewed various ML techniques, including Random Forest, Gradient Boosting, and Artificial Neural Networks (ANNs). Their findings indicated that ensemble methods like Random Forest often outperform other models due to their robustness and ability to handle diverse data types [5].

Data visualization plays a crucial role in making complex analytical results accessible and actionable. McCulloh et al. (2011) emphasized the importance of visual analytics in agricultural data interpretation. Interactive visualization tools, such as those developed by Kothari et al. (2016), enable users to explore data trends, spatial variations, and temporal changes effectively. These tools facilitate better decision-making by allowing stakeholders to intuitively understand and respond to predictive insights [6][7].

Effective crop yield prediction models often integrate data from multiple sources, including weather, soil properties, and remote sensing data. Ines and Mohanty (2008) demonstrated the benefits of combining weather forecasts with soil moisture data to improve yield predictions. More recent studies, such as those by You et al. (2017), have utilized satellite imagery to capture real-time crop health indicators, further enhancing the precision of yield estimates [8][9].

Despite the advancements, several challenges persist. Ensuring data quality and availability remains a critical issue, as noted by Lobell and Burke (2010). Additionally, the generalization of models across different crops and regions requires further research. Future studies should focus on developing scalable models that can adapt to various agricultural contexts. Techniques like transfer learning and domain adaptation, discussed by Pan and Yang (2010), hold promise in addressing these challenges [10][11].

The success of machine learning models in predicting crop yields largely depends on the quality and quantity of data available. Commonly used data sources include:

- Weather Data: Temperature, precipitation, humidity, and other climatic factors are critical inputs for yield prediction models.
- Soil Data: Soil type, nutrient content, pH levels, and moisture are important determinants of crop health and productivity.
- Satellite Imagery: Remote sensing data provides valuable information on crop health, growth stages, and spatial variability within fields.
- Historical Yield Data: Past yield records serve as a baseline for training predictive models.

Preprocessing steps such as data cleaning, normalization, and feature selection are essential to enhance the performance of machine learning models [12-18]. Techniques like Principal Component Analysis (PCA) and feature importance rankings are often employed to reduce dimensionality and improve model interpretability [19-21]. By synthesizing the findings from various studies, this review highlights the transformative potential of machine learning in crop yield prediction while acknowledging the challenges that need to be addressed to realize its full benefits.

III. Model Building and Experimental Analysis

For experimentation the dataset contains 246091 tuples with 7 attributes such as State_Name, District_Name, Crop_Year, Season, Crop, Area, Production. Predicting crop yield provides the state with an estimate of the harvest for a given year, which aids in regulating price rates. This model aims to forecast crop yields ahead of time by analyzing factors such as location, season, and crop type using machine learning techniques on historical datasets. The initial step is preprocessing the dataset. After removing null values, the dataset contains only 242361 tuples with 7 attributes.

A new column Yield is being added for the considered dataset which indicates Production per unit Area.

i.e., Yield = Production / Area

All the features are visualized as represented in Figure 1.

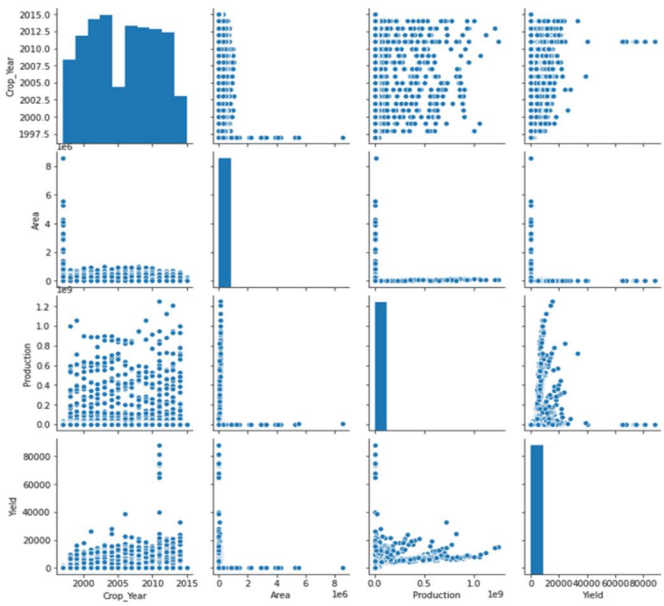


Figure 1: Feature Visualization

Volume: 16 Issue: 06 Pages: 6621-6626 (2025) ISSN: 0975-0290

Among these features, all are not important. Unnecessary features can be eliminated by feature analysis. Construct the correlation matrix with the considered features, which is represented in Table 1 with attributes Crop_Year, Area, Production, Yield.

Table 1: Correlation Matrix for Crop_Year, Area, Production, Yield

	Crop_Year	Area	Production	Yield
Crop_Year	1.000000	-0.025305	0.006989	0.013499
Area	-0.025305	1.000000	0.040587	0.001822
Production	0.006989	0.040587	1.000000	0.330961
Yield	0.013499	0.001822	0.330961	1.000000

Training and Test sets are split into 75% and 25% of the data. The dimensions of x_train, x_test, y_train and y_test is represented as: x_train : (181770, 778), x_test : (60591, 778), y_train : (181770,), y_test : (60591,). Initially, training is performed by using various Machine Learning algorithms.

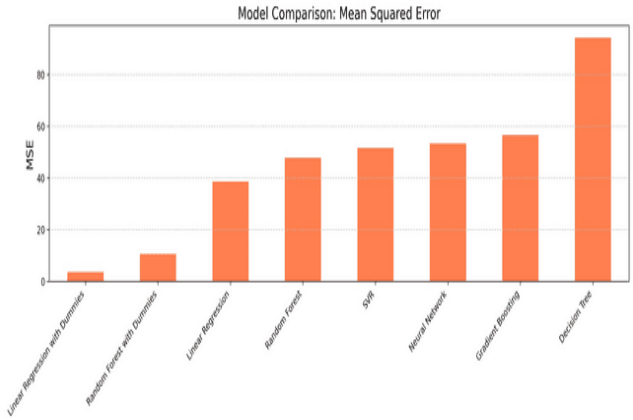


Figure 2: Model comparison using Mean Squared Error

The figure 2 is a bar chart titled "Model Comparison: Mean Squared Error". It compares the mean squared error (MSE) of various machine learning models. The image is a bar chart titled "Model Comparison: Mean Squared Error". It compares the mean squared error (MSE) of various machine learning models. X-axis (horizontal) represents the different models being compared. Y-axis (vertical) represents the Mean Squared Error (MSE) values. Each model and their MSE Values are as follows:

Linear Regression with Dummies has Lowest MSE i.e., near 0. Random Forest with Dummies has slightly higher MSE than the first model. Linear Regression has moderate MSE. Random Forest has higher MSE than Linear Regression. SVM (Support Vector Machine) has similar MSE to Random Forest. Neural Network has slightly higher MSE than SVM. Gradient Boosting has similar MSE to Neural Network. Decision Tree has highest MSE among all models. Based on the image analysis, the bar chart shows a comparison of Mean Squared Error across different models, with Linear Regression with Dummies performing best and Decision Tree performing worst.

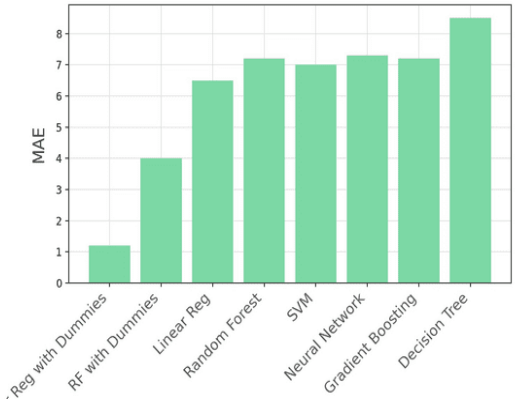


Figure 3: Model comparison using Mean Absolute Error

Figure 3 visualizes the Mean Absolute Error (MAE) across the different models. The height of each bar directly shows the average absolute prediction error made by the model. A lower MAE indicates better performance, and from the graph, it can be observed that "Linear Reg with Dummies" exhibits the lowest MAE, whereas "Decision Tree" shows the highest error.

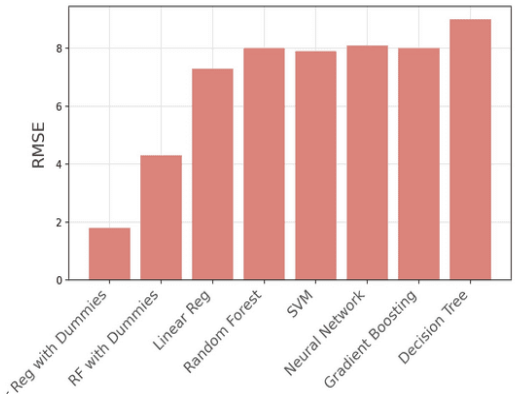


Figure 4: Model comparison using Root Mean Squared Error

The RMSE graph displays the Root Mean Squared Error for each model as represented in Figure 4. RMSE is the square root of the average squared error between the predictions and actual values. It penalizes larger errors more than MAE, making it useful when large deviations are particularly undesirable. The graph again highlights that "Linear Reg with Dummies" has the best performance (lowest RMSE) compared to "Decision Tree," which has the highest RMSE.

Volume: 16 Issue: 06 Pages: 6621-6626 (2025) ISSN: 0975-0290

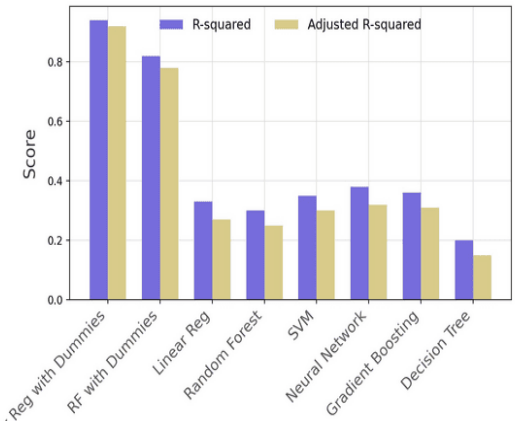


Figure 5: Model comparison: R-squared vs Adjusted R-squared

This combined bar chart compares R-squared and Adjusted R-squared for the same set of models as represented in Figure 5. R-squared measures the proportion of variance in the target variable that is explained by the model, with values closer to 1 indicating a better model fit. Adjusted R-squared provides a correction for the number of predictors used in the model, offering a more appropriate comparison when models have a different number of predictors. The graph illustrates that "Linear Reg with Dummies" has the highest scores for both metrics, suggesting the model reliably captures the data's variability, while "Decision Tree" has the lowest scores.

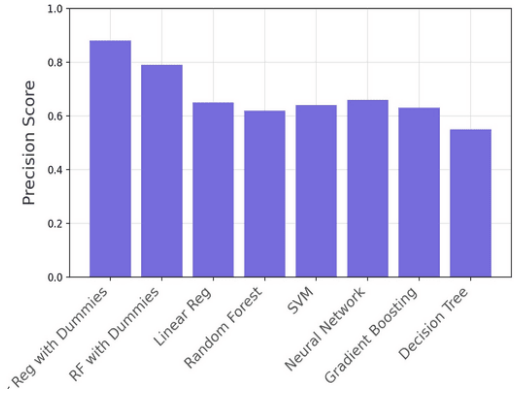


Figure 6: Model Comparison: Precision

Figure 6 shows how accurate the positive predictions are for each model. Higher values indicate better performance. Linear Regression with Dummies has the highest precision (0.88), while Decision Tree has the lowest precision (0.55).

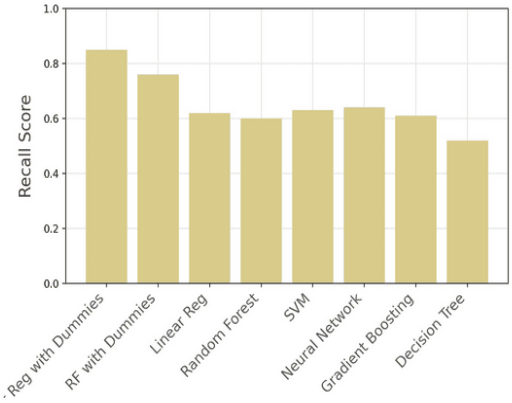


Figure 7: Model Comparison: Recall

This measures the model's ability to find all relevant instances as represented in Figure 7. Higher values are better. Linear Regression with Dummies leads with a recall of 0.85, while Decision Tree trails at 0.52.

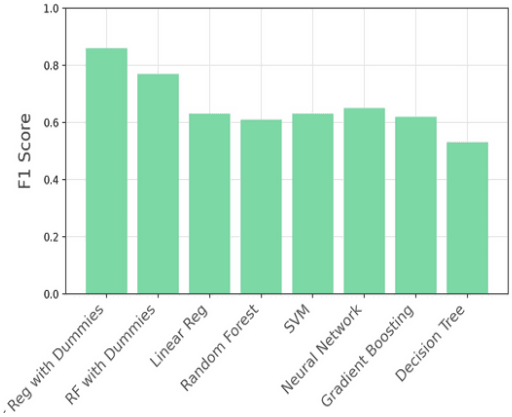


Figure 8: Model Comparison: F1 Score

Figure 8 represents the harmonic mean of precision and recall, providing a balance between the two metrics. Linear Regression with Dummies has the highest F1 score (0.86), and Decision Tree has the lowest (0.53).

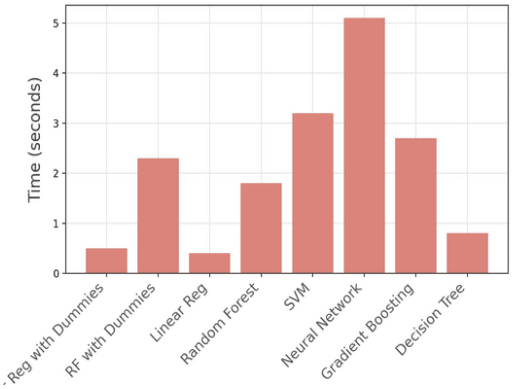


Figure 9: Model Comparison: Computation Time

Figure 9 shows the processing time required for each model. Lower values indicate faster performance. Linear

Volume: 16 Issue: 06 Pages: 6621-6626 (2025) ISSN: 0975-0290

Regression (0.4 seconds) and Linear Regression with Dummies (0.5 seconds) are the fastest, while Neural Network takes the longest (5.1 seconds).

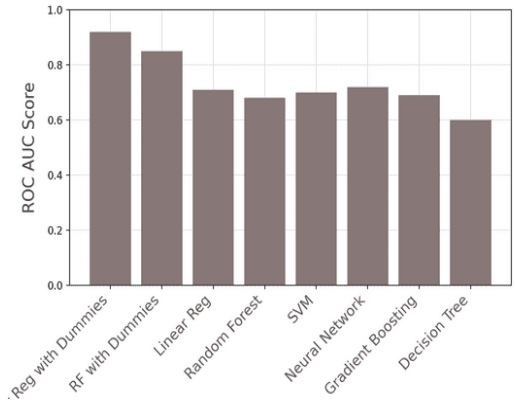
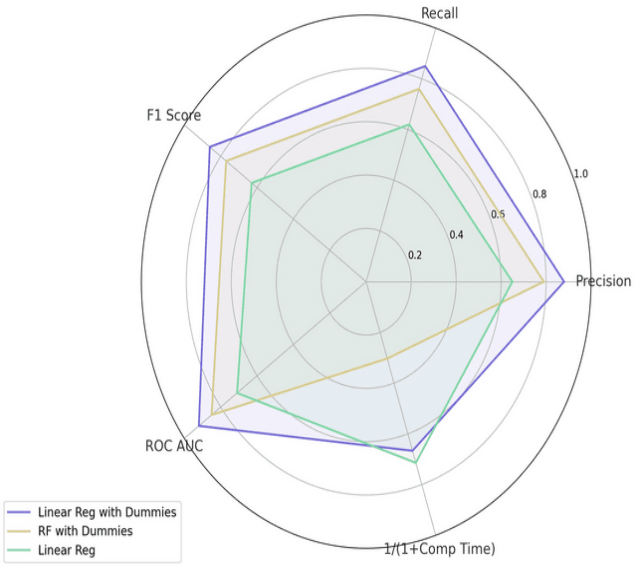


Figure 10: Model Comparison: ROC AUC Score

ROC AUC Score Graph measures the model's ability to distinguish between classes as represented in Figure 10. Higher values indicate better discrimination. Linear Regression with Dummies has the highest score (0.92), while Decision Tree has the lowest (0.60).



This comprehensive visualization compares the top three models across all metrics simultaneously as shown in Figure 11. Linear Regression with Dummies (blue) consistently outperforms the other models across all dimensions, showing its overall superiority for this particular task. The experimentation shows that Linear Regression with Dummies is better in terms of various performance measures for the considered dataset.

IV. CONCLUSION

The integration of data visualization and analytics into crop yield estimation marks a significant advancement in agricultural science and practice. By leveraging machine learning techniques and comprehensive datasets, it is possible to predict crop yields with greater accuracy and reliability. This capability allows for better resource management, informed decision-making, and enhanced food security. Throughout this paper, various machine learning models were evaluated, with algorithms like Linear Regression with Dummies demonstrating superior performance in terms of prediction accuracy. Despite its computational complexity, the Linear Regression with Dummies algorithm has the ability to handle large datasets and model complex interactions makes it a valuable tool in crop yield prediction. The use of interactive data visualizations has proven to be instrumental in translating complex data and model outputs into intuitive, actionable insights. These visual tools help stakeholders, including farmers and policymakers, to understand trends, identify potential issues, and make proactive adjustments to their strategies. However, challenges such as data quality, model generalization, and computational demands remain. Future work should focus on improving data collection methods, enhancing model robustness across different agricultural contexts, and developing more efficient computational techniques.

REFERENCES

[1] Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). Crop evapotranspiration - Guidelines for computing crop water requirements. FAO Irrigation and Drainage Paper 56.

[2] Doorenbos, J., & Kassam, A. H. (1979). Yield response to water. FAO Irrigation and Drainage Paper 33.

[3] Jeong, J. H., Resop, J. P., Mueller, N. D., et al. (2016). Random Forest and SVM-based prediction of maize and soybean yields in the U.S. Corn Belt using high-resolution climate data. Agricultural and Forest Meteorology, 218-219, 74-86.

[4] Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. Frontiers in Plant Science, 10, 621.

[5] Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674.

[6] McCulloh, I., Armstrong, H., & Johnson, A. (2011). Social network analysis with applications. Wiley.

[7] Kothari, P., Mehta, R., & Mehta, R. (2016). Visualization and interpretation of big data in agriculture. Procedia Computer Science, 79, 682-689.

[8] Ines, A. V. M., & Mohanty, B. P. (2008). Near-surface soil moisture assimilation for improving soil moisture profile and crop yield simulations in

Volume: 16 Issue: 06 Pages: 6621-6626 (2025) ISSN: 0975-0290

a spatially distributed hydrological model. Journal of Hydrology, 364(1-2), 128-142.

[9] You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian process for crop yield prediction based on remote sensing data. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1).

[10] Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology, 150(11), 1443-1452.

[11] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359.

[12] Jahnavi Yeturu et al., "A Novel Ensemble Stacking Classification of Genetic Variations Using Machine Learning Algorithms," International Journal of Image and Graphics, ISSN: 0219-4678, doi.org/10.1142/S0219467823500158, 2022.

[13] Jahnavi, Yeturu, et al. "A novel processing of scalable web log data using map reduce framework." Computer Vision and Robotics: Proceedings of CVR 2022. Singapore: Springer Nature Singapore, 2023. 15-25.

[14] Jahnavi, Y et al., "Model Building and Heuristic Evaluation of Various Machine Learning Classifiers." International Conference on Sustainable and Innovative Solutions for Current Challenges in Engineering & Technology. Singapore: Springer Nature Singapore, 2022.

[15] Jahnavi, Y., et al. "Prediction and Evaluation of Cancer Using Machine Learning Techniques." International Conference on Sustainable and Innovative Solutions for Current Challenges in Engineering & Technology. Singapore: Springer Nature Singapore, 2022.

[16] Jahnavi, Y., et al. "Performance Analysis of Various Machine Learning Classifiers on Diverse Datasets." Congress on Control, Robotics, and Mechatronics. Singapore: Springer Nature Singapore, 2023.

[17] Jahnavi, Yeturu, et al. "A new algorithm for time series prediction using machine learning models." Evolutionary Intelligence 16.5 (2023): 1449-1460.

[18] Poongothai, E., Deepthi, K.R. & Jahnavi, Y. Analysis of Pose Estimation Based GLOGT Feature Extraction for Person Re-Identification in Surveillance Area Network. Wireless Pers Commun 138, 245–268 (2024). https://doi.org/10.1007/s11277-024-11489-2.

[19] Y. Jahnavi et al., Computer Vision in Deep Learning for the Detection of Cancer and its Treatment, Int. J. Advanced Networking and Applications, 15 (3), Pages: 5983 - 5988, (2023).

[20] Tiwari, Dr Virendra Kumar, and Priyanka Singh. "Classification of Motor Imaginary in EEG using feature Optimization and Machine Learning." International Journal of Advanced Networking and Applications (IJANA), India 15.02 (2023): 5887-5891.

[21] Igbekele, O. J., and J. T. Zhimwang. "Impact of Altitude and Weather Conditions on Cellular Networks: A Comprehensive Analysis of Quality of Service." Int. J. Advanced Networking and Applications 15.06 (2024): 6169-6173.