COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR CROP YIELD PREDICTION

Dr. N. Kiranmai[1],*, V. Ashwini[2], P. Karan[2], V. Rajesh[2], P. Uday[2]
[1]Associate Professor, Department of CSE (DS), TKR College of Engineering & Technology, Meerpet, Telangana 500097
[2]B.Tech (Scholar), Department of CSE (DS), TKR College of Engineering & Technology, Meerpet, Telangana 500097
Correspondence: nkiranmai@tkrcet.com

ABSTRACT

Crop yield prediction plays a critical role in agricultural planning, food security, and economic stability. Accurate forecasting enables farmers, policymakers, and agricultural stakeholders to make informed decisions regarding crop selection, resource allocation, and risk management. With the increasing availability of agricultural data and advancements in computational techniques, machine learning has emerged as an effective approach for improving prediction accuracy. This study presents a comparative analysis of multiple machine learning algorithms for predicting crop yield based on historical agricultural data. The dataset used includes key parameters such as rainfall, temperature, soil characteristics, humidity, and previous yield records. These features are selected due to their significant influence on crop productivity. Three widely used machine learning models—Linear Regression, Decision Tree, and Random Forest—are implemented and evaluated for their predictive performance. The models are trained and tested using standard preprocessing techniques, including data cleaning, normalization, and feature selection, to ensure data quality and consistency. Performance evaluation is conducted using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2 score). The experimental results demonstrate that ensemble-based approaches, particularly Random Forest, outperform traditional models in terms of accuracy, robustness, and generalization capability. The findings highlight the importance of selecting appropriate algorithms and relevant features to enhance prediction performance. The proposed approach provides a data-driven framework that can assist farmers and agricultural planners in optimizing crop productivity and improving decision-making processes in modern agriculture.

KEYWORDS: Crop Yield Prediction, Machine Learning, Random Forest, Data Analytics, Agriculture, Predictive Modeling.

1. INTRODUCTION

Agriculture remains one of the most vital sectors globally, directly influencing food security, economic growth, and rural development. Accurate prediction of crop yield is essential for effective agricultural planning, supply chain management, and minimizing risks associated with uncertain environmental conditions. Reliable yield forecasting enables farmers and policymakers to make informed decisions regarding crop selection, irrigation planning, and resource allocation. However, traditional methods of yield estimation rely heavily on manual observation and historical trends, which often lack precision, scalability, and the ability to adapt to dynamic environmental changes [1-3].

With the rapid growth in data availability and advancements in computational power, machine learning techniques have emerged as powerful tools for predictive analytics in agriculture. These techniques are capable of analyzing complex relationships among multiple variables such as climatic conditions, soil properties, and farming practices. By leveraging large datasets and advanced algorithms, machine learning models can uncover hidden patterns and generate accurate predictions, thereby improving decision-making processes and reducing uncertainty in agricultural operations.

This study focuses on comparing different machine learning algorithms to identify the most effective approach for crop yield prediction. The selected algorithms include Linear Regression, Decision Tree, and Random Forest. Each of these models has distinct characteristics and advantages. Linear Regression serves as a simple and interpretable baseline model for identifying linear relationships between input features and yield. Decision Tree models are capable of capturing non-linear patterns through hierarchical decision rules, making them suitable for complex datasets. Random Forest, an ensemble learning method, combines multiple decision trees to improve prediction accuracy and reduce overfitting, resulting in more robust performance.

The primary objective of this research is to evaluate and compare the performance of these algorithms using real-world agricultural data. The study also highlights the significance of preprocessing techniques, including data cleaning, normalization, and feature selection, in enhancing model accuracy and reliability [4].

By conducting a comprehensive comparative analysis, this research aims to provide valuable insights into selecting the most suitable machine learning model for crop yield prediction. The findings are expected to contribute to the development of data-driven agricultural systems, enabling smarter farming practices, improved productivity, and better resource management in modern agriculture.

2. LITERATURE SURVEY

Crop yield prediction has been widely studied due to its critical role in ensuring food security, improving agricultural productivity, and supporting economic planning. Accurate estimation of crop yield enables better decision-making for farmers, policymakers, and agricultural industries [5]. Earlier approaches to crop yield prediction primarily relied on traditional statistical techniques such as linear regression, multiple regression, and time-series analysis [6]. These methods were simple to implement and computationally efficient, providing baseline results for yield estimation. However, they were limited in their ability to capture complex, non-linear relationships between environmental factors and crop productivity [7]. As agricultural data became more diverse and high-dimensional, the shortcomings of these conventional techniques became increasingly evident [8].

With advancements in computational capabilities and the availability of large datasets, machine learning techniques have emerged as powerful tools for agricultural prediction tasks [9]. Researchers have explored various machine learning algorithms, including Decision Trees, Support Vector Machines (SVM), and k-Nearest Neighbors (k-NN), to improve prediction accuracy [10-12]. Decision Trees have gained popularity due to their interpretability and ability to handle both numerical and categorical data. They model decision-making processes in a hierarchical structure, making them easy to understand and implement [13]. However, studies have shown that Decision Trees tend to overfit when trained on large or noisy datasets, resulting in reduced generalization performance [14].

Support Vector Machines have also been widely applied in crop yield prediction due to their effectiveness in handling high-dimensional data and complex relationships [15]. By using kernel functions, SVM models can transform data into higher-dimensional spaces where linear separation becomes possible. Despite their advantages, SVM models require careful parameter tuning and are computationally expensive, especially when dealing with large-scale agricultural datasets [16]. Similarly, k-Nearest Neighbors is a simple and intuitive algorithm that has been used in some prediction studies [17]. While it can produce reasonable results for smaller datasets, its performance is highly dependent on the choice of distance metrics and decreases significantly as dataset size increases [18].

In recent years, ensemble learning methods have gained significant attention due to their improved predictive performance and robustness. Techniques such as Random Forest and Gradient Boosting combine multiple base models to enhance overall accuracy and reduce variance [19]. Random Forest, in particular, constructs multiple decision trees and aggregates their outputs to generate more stable predictions. This approach effectively reduces overfitting and improves generalization capability [20]. Several studies have demonstrated that Random Forest consistently outperforms individual models in crop yield prediction tasks, especially when dealing with noisy or incomplete data. Gradient Boosting methods further enhance prediction performance by sequentially improving weak learners, although they are more sensitive to hyperparameter tuning and may require careful optimization [21].

In addition to selecting appropriate machine learning models, feature selection and data preprocessing have been identified as critical factors influencing prediction accuracy [22]. Agricultural datasets often contain numerous variables, not all of which contribute significantly to yield prediction. Researchers emphasize the importance of selecting relevant features such as rainfall, temperature, humidity, soil nutrients, and historical yield data [23]. Irrelevant or redundant features can introduce noise and reduce model performance. Techniques such as correlation analysis, feature importance evaluation, and Principal Component Analysis (PCA) are commonly used to identify and retain significant attributes [24]. Proper preprocessing steps, including handling missing values, normalization, and outlier removal, further improve the quality of input data and enhance model performance [25].

Recent research has also explored the application of deep learning techniques, including Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN), for crop yield prediction. These models are capable of capturing highly complex and non-linear relationships within large datasets. Deep learning approaches can integrate multiple data sources, such as satellite imagery, weather data, and soil information, providing a comprehensive framework for agricultural analysis [26-27].

However, these models require large volumes of data and significant computational resources, which may not always be feasible in practical scenarios, particularly in resource-constrained environments.

Comparative studies have become increasingly important in evaluating the performance of different machine learning algorithms under consistent conditions. These studies assess multiple models using standardized evaluation metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2 score). The results of such analyses indicate that no single algorithm consistently outperforms others across all datasets. Instead, model performance depends on various factors, including dataset size, feature quality, data preprocessing techniques, and parameter tuning [28].

Overall, the existing literature highlights that ensemble methods, particularly Random Forest, offer a balanced trade-off between accuracy, robustness, and computational efficiency. These models are well-suited for handling the complexity and variability inherent in agricultural data. This study builds upon previous research by conducting a systematic comparative analysis of multiple machine learning algorithms for crop yield prediction. The objective is to identify the most effective model and provide practical insights that can support data-driven decision-making in modern agriculture.

3. PROPOSED METHODOLOGY

3.1 Data Collection

The proposed system begins with the collection of agricultural data from reliable and publicly available sources. The dataset consists of multiple attributes influencing crop yield, including climatic factors such as rainfall, temperature, and humidity, along with soil-related parameters and historical yield records. These variables are selected based on their direct and measurable impact on agricultural productivity. The collected data serves as the primary input for building predictive models.

3.2 Data Preprocessing

The collected dataset is subjected to preprocessing to ensure data quality and consistency. Missing values are handled using suitable imputation techniques to avoid information loss. Noisy and inconsistent data points are identified and removed to improve reliability. Numerical features are normalized to a common scale, which enhances the performance of machine learning algorithms. Categorical attributes are transformed into numerical representations using encoding methods, enabling their use in model training.

3.3 Feature Selection

Feature selection is performed to identify the most relevant variables contributing to crop yield prediction. Correlation analysis is used to evaluate the relationship between input features and the target variable. Features that exhibit strong correlation are retained, while redundant and less significant attributes are eliminated. This process reduces dimensionality, improves computational efficiency, and enhances the overall performance of the predictive models.

3.4 Model Implementation

The methodology involves the implementation of three machine learning algorithms: Linear Regression, Decision Tree, and Random Forest. Linear Regression is used as a baseline model to establish a fundamental relationship between input variables and crop yield. Decision Tree is employed to capture non-linear relationships through a hierarchical decision-making structure. Random Forest, an ensemble learning technique, combines multiple decision trees to improve prediction accuracy and reduce overfitting, making it more robust compared to individual models.

3.5 Model Training and Testing

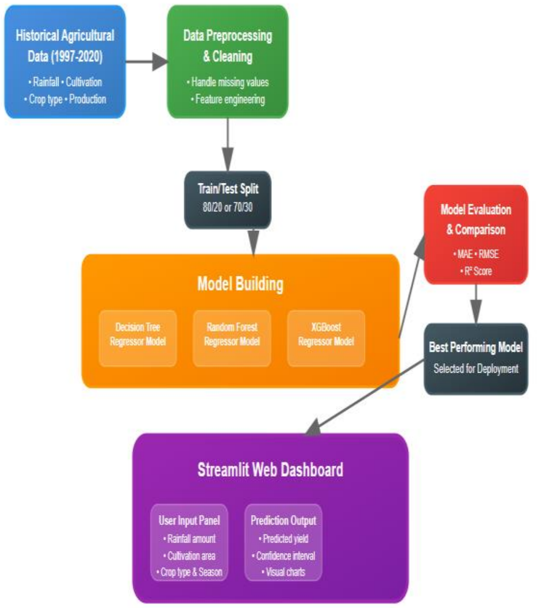
The dataset is divided into training and testing subsets to evaluate model performance. A standard split ratio of 80:20 is used, where the majority of the data is allocated for training and the remaining portion is reserved for testing. The models are trained using the training dataset and evaluated on the testing dataset to measure their ability to generalize to unseen data. Cross-validation techniques may also be applied to ensure stability and consistency in performance.

3.6 Performance Evaluation

The performance of the implemented models is assessed using standard evaluation metrics. Mean Absolute Error (MAE) is used to measure the average magnitude of prediction errors. Root Mean Square Error (RMSE) provides insight into the variance of errors and penalizes larger deviations. The coefficient of determination (R^2 score) is used to evaluate how well the model explains the variability in crop yield. These metrics collectively provide a comprehensive evaluation of model accuracy and reliability.

3.7 Comparative Analysis

A comparative analysis is conducted to evaluate the effectiveness of the implemented models. The results obtained from each algorithm are analyzed based on the defined performance metrics. The strengths and limitations of each model are examined to determine the most suitable approach for crop yield prediction. The analysis helps in identifying the model that offers the best balance between accuracy, robustness, and computational efficiency.

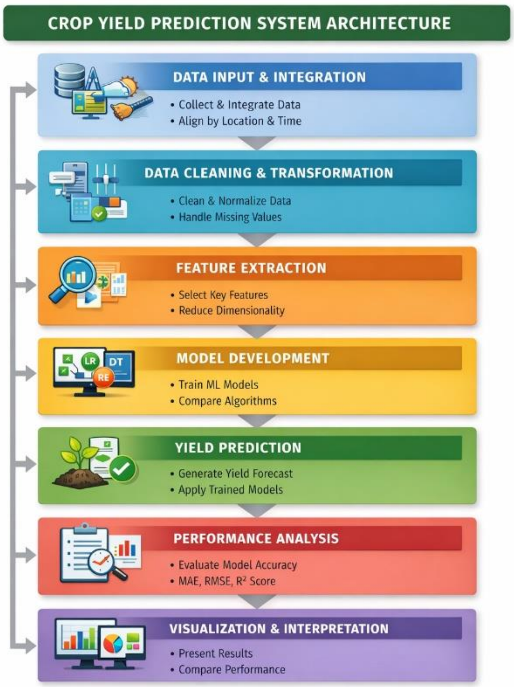


4. ARCHITECTURE

The proposed crop yield prediction system is designed using a layered architecture that integrates data handling, machine learning models, and analytical evaluation to generate accurate predictions. The system processes agricultural data through multiple stages, ensuring effective transformation from raw input to meaningful output.

4.1 Data Input and Integration Layer

This layer focuses on collecting and integrating agricultural data from multiple sources. The dataset includes environmental and soil-related attributes such as rainfall, temperature, humidity, and historical crop yield. Additional parameters such as soil nutrients and seasonal variations may also be included to improve prediction accuracy. Data is gathered from reliable sources including agricultural databases and public repositories. Since data from different sources may vary in format, integration techniques are applied to combine them into a unified structure. The data is aligned based on common attributes such as location and time, ensuring consistency and enabling effective analysis in subsequent stages.



4.2 Data Cleaning and Transformation Layer

In this stage, raw data is processed to remove inconsistencies and improve overall quality. Missing values are handled using suitable imputation techniques to avoid data loss and maintain dataset completeness. Noisy and inconsistent data points are identified and removed to enhance reliability. Numerical features are normalized or scaled to ensure uniformity across all variables, which improves model performance and convergence. Categorical attributes are converted into numerical form using encoding methods, making them compatible with machine learning algorithms. This layer ensures that the dataset is consistent, structured, and suitable for accurate model training and analysis.

4.3 Feature Extraction Layer

This layer is responsible for identifying the most significant features that influence crop yield prediction. Statistical techniques such as correlation analysis are applied to evaluate the relationship between input variables and the target output. Features with strong influence are retained, while redundant, irrelevant, or highly correlated attributes are removed to avoid multicollinearity. This process reduces dimensionality and enhances computational efficiency. By focusing only on meaningful features, the model achieves better generalization and improved prediction accuracy. Effective feature selection also minimizes overfitting and ensures that the learning algorithms capture the most relevant patterns within the dataset.

4.4 Model Development Layer

The core functionality of the system is implemented in this layer, where multiple machine learning algorithms are trained using the processed dataset. Models such as Linear Regression, Decision Tree, and Random Forest are employed to learn underlying patterns and relationships between input features and crop yield. Each model is trained independently to capture different types of data behavior, including linear and non-linear relationships. Hyperparameters may be tuned to optimize model performance and improve prediction accuracy. This layer enables comparative analysis by generating multiple predictive models, facilitating the identification of the most effective algorithm for crop yield prediction.

4.5 Yield Prediction Layer

After the training phase, the system utilizes the trained machine learning models to generate predictions for new input data. The learned relationships between input features and crop yield are applied to estimate yield values based on environmental and soil parameters. This layer processes unseen data and produces prediction outputs that reflect the expected crop productivity. The effectiveness of this layer depends on the model's ability to generalize patterns from training data to real-world scenarios. It serves as the primary output stage of the system, providing actionable insights for agricultural planning and decision-making.

4.6 Performance Analysis Layer

This layer evaluates the effectiveness of each machine learning model using standard performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2 score). These metrics provide quantitative measures of prediction accuracy, error distribution, and model reliability. The evaluation process enables a systematic comparison of different models under consistent conditions. Based on the obtained results, the most accurate and stable model is identified for crop yield prediction. This layer plays a crucial role in validating model performance and ensuring the selection of an optimal predictive approach.

4.7 Visualization and Interpretation Layer

The final layer presents the prediction results and comparative performance of the machine learning models in a structured and interpretable format. Visualization techniques such as graphs and performance plots are used to represent model accuracy and error metrics clearly. These visual representations facilitate easier analysis of differences among models and highlight the most effective approach. By transforming complex numerical outputs into understandable formats, this layer improves interpretability and usability of the system. It supports informed decision-making by providing clear insights into crop yield predictions and overall model performance.

5. RESULT

The experimental results obtained from the implementation of multiple machine learning algorithms demonstrate clear differences in prediction performance and model efficiency. The models considered in this study include Linear Regression, Decision Tree, and Random Forest, all trained and evaluated on the same pre-processed dataset to ensure consistency in comparison.

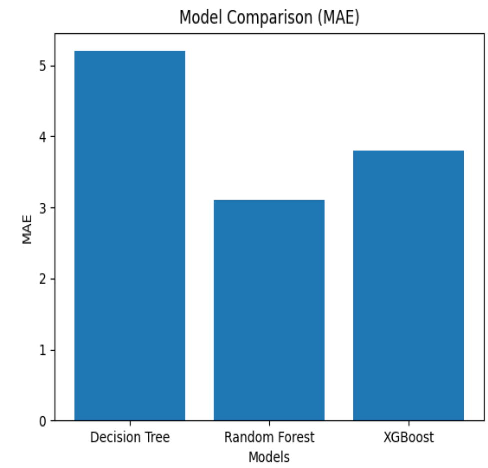
Linear Regression, used as a baseline model, provided moderate prediction accuracy. While it was effective in capturing simple linear relationships between input variables and crop yield, its performance was limited when dealing with complex and non-linear patterns present in agricultural data. As a result, it showed higher error values compared to other models.

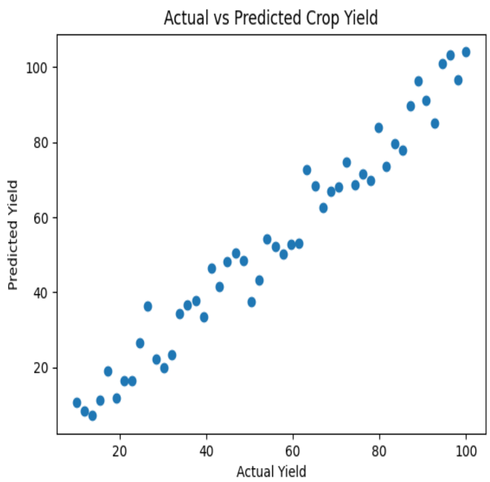
The Decision Tree model demonstrated improved performance over Linear Regression by effectively capturing non-linear relationships within the dataset. However, it exhibited signs of overfitting, particularly when the model became too sensitive to variations in the training data. This reduced its ability to generalize well to unseen data, leading to inconsistencies in prediction accuracy.

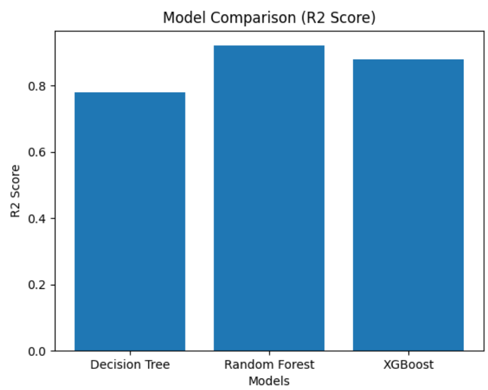
Random Forest outperformed both Linear Regression and Decision Tree models in terms of overall performance. By combining multiple decision trees, Random Forest reduced variance and improved generalization capability. The ensemble nature of the model enabled it to handle complex interactions among features more effectively, resulting in more accurate and stable predictions. The model achieved lower Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), along with a higher R^2 score compared to the other models.

The comparative analysis indicates that ensemble learning methods are more suitable for crop yield prediction tasks involving diverse and complex datasets. The results also highlight the importance of proper data preprocessing and feature selection in improving model performance.

	R2	MAE	RMSE
DecisionTree	0.959068	11.892756	191.741498
RandomForest	0.974151	11.024163	152.374225
XGBoost	0.965361	14.085093	176.386867







Overall, the findings confirm that Random Forest is the most effective algorithm among the models considered in this study. It provides a better balance between accuracy, robustness, and reliability, making it a suitable choice for real-world agricultural prediction systems.

6. CONCLUSION & FUTURE SCOPE

This study presented a comparative analysis of machine learning algorithms for crop yield prediction using agricultural and environmental data. The performance of Linear Regression, Decision Tree, and Random Forest models was evaluated using standard metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score. The results indicate that model selection plays a crucial role in achieving accurate and reliable predictions.

Among the implemented models, Random Forest demonstrated superior performance due to its ensemble learning approach, which reduces overfitting and improves generalization capability. Linear Regression, while simple and computationally efficient, was limited in capturing complex patterns within the dataset. Decision Tree showed better performance than Linear Regression but suffered from overfitting issues, affecting its consistency on unseen data. The comparative analysis confirms that ensemble methods are more suitable for handling the complexity and variability of agricultural data.

Despite the promising results, the study has certain limitations. The performance of the models depends heavily on the quality and size of the dataset. Limited features and static datasets may restrict the accuracy and adaptability of the system. Additionally, external factors such as sudden climatic changes and unpredictable environmental conditions are not fully captured in the current model.

Future work can focus on improving prediction accuracy by incorporating larger and more diverse datasets, including real-time data from sensors and satellite imagery. Advanced techniques such as deep learning models and hybrid approaches can be explored to capture complex patterns more effectively. Furthermore, the development of user-friendly applications or decision support systems can enhance the practical usability of the model for farmers and agricultural planners.

In conclusion, machine learning-based crop yield prediction offers a reliable and scalable solution for modern agriculture, enabling data-driven decision-making and improved productivity.

REFERENCES

1. J. You, X. Li, M. Low, D. Lobell, and S. Ermon, “Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data,” in Proc. AAAI Conf., 2017.
2. S. Jeong, J. Ko, and J. Yeom, “Machine Learning Approaches for Crop Yield Prediction Using Climate Data,” Computers and Electronics in Agriculture, vol. 152, pp. 292–300, 2018.
3. Prashanth Kumar, P., & Jadhav, P. P. (2023). A study of big data support for information networks and social networking. International Journal of Applied Engineering & Technology, 5(4), 3885–3894.
4. Prashanth Kumar, P., & Jadhav, P. P. (2023). Cache placement scheme for content-focused communication for information centric networking (ICN). European Chemical Bulletin, 3(1), 3138–3150.
5. Jaya Rama Krishna, V. V., Srinivasa Rao, B., Veeraiah, D., Subba Raju, S., Al Answari, M. S., & Kaur, C. (2024, February). Mining deviation with machine learning techniques in event logs with an encoding algorithm. Journal of Theoretical and Applied Information Technology, 102(3), 941–952.
6. Krishna, V., Murali Mohan, K. V., Banala, R., & Srinivas, B. S. (2023). An effective hierarchical image coding approach with Hilbert scanning. International Journal of System Assurance Engineering and Management.
7. Vadivelan, N., Bhargavi, K., & Kodati, S. (2022). Detection of cyber attacks using machine learning. AIP Conference Proceedings, 2405, 030003.
8. Suman, B., & Jadhav, P. P. (2023). Advancements in routing algorithm techniques for wireless sensor networks. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3).
9. Suman, B., & Jadhav, P. P. (2023). Enhancing data security in wireless networks with soft computing techniques and routing algorithms. International Journal of Applied Engineering & Technology, 5(4).
10. Ananthajothi, K., Balamurugan, K., Divya, D., & Latchoumi, T. P. (2026). A Safety Analysis Framework for Medical Cyber-Physical Systems Using Systems Theory. Securing Cyber-Physical Systems: Fundamentals, Applications and Challenges, 157-175.
11. Latchoumi, T. P., Parthiban, L., Balamurugan, K., Raja, K., Vijayaraj, J., & Parthiban, R. (2023). A framework for low energy application devices using blockchain-enabled IoT in WSNs. In Integrating Blockchain and Artificial Intelligence for Industry 4.0 Innovations (pp. 121-132). Cham: Springer International Publishing
12. Balamurugan, K., Deepthi, T., Subramanian, A. K., Banerjee, A., Agarwal, D., Biswas, A., & Sinha, A. (2023). A study on the mechanical properties of rare earth-based aluminium composite. Journal of The Institution of Engineers (India): Series D, 104(1), 15-25
13. Arunkarthikeyan, K., & Balamurugan, K. (2020). Studies on the effects of deep cryogenic treated WC–Co insert on turning of Al6063 using multi-objective optimization. SN applied Sciences, 2(12), 2103.
14. Pavan, M. V., Balamurgan, K., & Balamurgan, P. (2021). Wear experiments on PLA-Cu composite filament printed in different FDM conditions. Turkish Journal of Computer and Mathematics Education, 12(9), 2245-2251
15. Sneha, P., Balamurugan, K., & Kalusuraman, G. (2021). Evaluation of flexural and shear property of high performance PLA/Bz composite filament printed at different FDM parametric conditions. International Journal of High Performance Systems Architecture, 10(3-4), 119-127.
16. samples printed using fused filament extrusion by response surface method. Progress in Additive Manufacturing, 7(5), 957-969.Balamurugan, K., Pavan, M. V., & Balamurugan, P. (2022). Wear parametric analysis on PLA/Cu filament
17. Sneha, N., & Balamurugan, K. (2022, October). Micro-drilling optimization study using RSM on PLA-bronze composite filament printed using FDM. In 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon) (pp. 1-5). IEEE.
18. Balamurugan, K., Sudhakar, G., Xavier, K. F., Bharathiraja, N., & Kaur, G. (2025). Human-machine interaction in mechanical systems through sensor enabled wearable augmented reality interfaces. Measurement: Sensors, 39, 101880
19. Abshalomu, Y., Jyothi, Y., Balamurugan, K., & Selvaraj, R. (2023). Effect of varied cashew nut ash reinforcement in aluminum matrix composite. Advances in Materials Science and Engineering, 2023(1), 3383777
20. N, Bharathiraja, Minu, M. S., Vijay, R., Rajalakshmi, M., Vidyullatha, P., & Balamurugan, K. (2025). Development of Hybrid Explainable Artificial Intelligence With Swin Vision Transformer Intrusion Detection for Securing VANETs From Attacks. Transactions On Emerging Telecommunications Technologies, 36(10).
21. Latchoumi, T. P., Parthiban, L., Raja, K., Balamurugan, K., & Parthiban, R. (2023). Secured smart manufacturing systems using blockchain technology for industry 4.0. In Integrating Blockchain and Artificial Intelligence for Industry 4.0 Innovations (pp. 281-294). Cham: Springer International Publishing
22. Krishna, V., Raju, Y. D. S., Raghavendran, C. V., Naresh, P., & Rajesh, A. (2022). Identification of nutritional deficiencies in crops using machine learning and image processing techniques. In 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM). IEEE.
23. Krishna, V., Sumalatha, C., Raju, Y. D. S., & Mohan, K. V. M. (2022). Analysis of heart disease prediction using machine learning classification algorithms. Journal of Optoelectronics Laser.
24. Krishna, V., Raghavendran, C. V., & Faruk, S. K. U. (2024). Novel computer vision and color image

segmentation for agriculture application. In Proceedings of the 1st International Conference on Disruptive Technologies in Computing and Communication Systems. CRC Press.

25. Muthu, M. A. (n.d.). A hybrid deep CNN model for brain tumor image multi-classification. International Journal of Engineering Research and Science & Technology (IJERST).
26. Muthu, M. A. (n.d.). Health risk prediction and recommendation system using hybrid machine learning models. International Journal of Engineering Research and Science & Technology (IJERST).
27. Muthu, M. A. (2016). Performance analysis of cloud computing centers using M/G/m/m+r queuing systems. International Journal of Research in Engineering, Science and Technologies.
28. Muthu, M. A. (n.d.). Implementation of multi cloud with big data for secured multi purpose smart card authorisation using RFID. International Journal.